Università degli Studi di Salerno

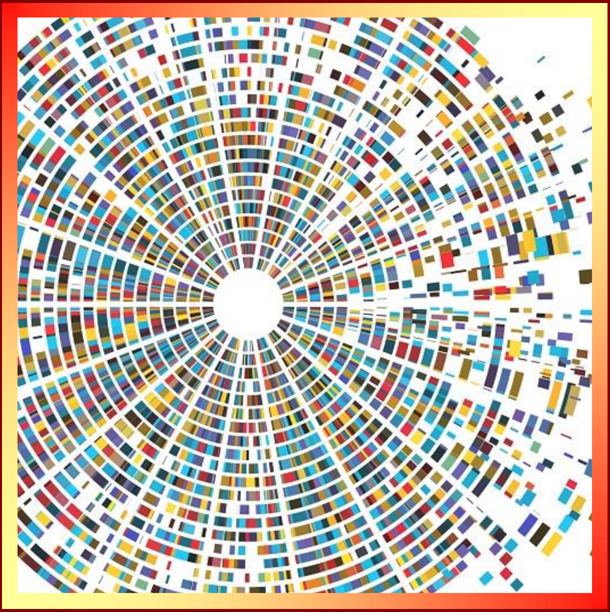
Dipartimento di Informatica

Dottorato di Ricerca in Informatica – XXXIV Ciclo

Tesi di Dottorato/Ph.D. Thesis

Quality and Privacy-aware (Linked) Open Data Exploitation

Maria Angela Pellegrino



Supervisor: **Prof. Vittorio Scarano**

Ph.D. Program Director: **Prof. Andrea De Lucia**

AA 2020/2021

Curriculum Computer Science and Information Technology

Università degli Studi di Salerno

Dipartimento di Informatica

Dottorato di Ricerca in Informatica
XXXIV Ciclo

Tesi di Dottorato / Ph.D. Thesis

Quality and Privacy-aware (Linked) Open Data Exploitation

Maria Angela PELLEGRINO

Supervisor:	**Prof. Vittorio SCARANO**
PhD Program Director:	**Prof. Andrea DE LUCIA**

A.A 2020/2021

Dedicated to my grandparents, my *past*,
for their trust in my abilities,

to my families, the one given by nature and
the one I chose for the life, my *present*,
for their constant and relentless support,

to you, my *future*, hoping you will be proud to your mum.

If you only do what you can do
you will never be more than you are now.

ACKNOWLEDGMENTS

I do not like tear-jerking thanks, and, in naming someone, I am always afraid of forgetting someone else, equally important. Surely, I desire to thank those who believed in me and gave me the strength to get up after each defeat, but also who made me fall, allowing me to get up stronger than before.

One day, Professor Scarano, after a refusal of an article, looked at me in the eyes and told me *"don't stop believing it"*. This is what I will remember about this journey: the confidence in my *perseverance*, which has always given me the strength to react.

Each is the sum of the five people who frequent in life assiduously. For sure, the people who have shown me closeness are dozens more. Thanks to my *families*, the natural one and by promise, for their constant and incessant presence. Thanks to my *friends* who listened to my sclera and shared my successes. Thanks to my laboratory, *ISISLab*, where I have always felt at home.

ABSTRACT

Data are the new oil and it is widely recognised the role of publishing them as *Open Data* to let data consumers freely access and exploit them. Data providers are not only encouraged to publish data but to ensure that available datasets are *fit-for-use*, meaning that data users can directly exploit them without investing effort, time, and money in performing data cleansing. The situation becomes even more complex when data publishers deal with data concerning individuals. Data in their raw form may contain personal and sensitive information about people and publishing them as are violate individual privacy. Hence, data publishers need to apply privacy-preserving data publishing procedures by publishing (sensitive) data without violating individual privacy.

Thus, data publishers before publishing data or data consumers before exploiting them require privacy-aware data cleansing approaches. Data publishers mainly opt for publishing data in tabular format. Hence, data cleansing approaches should be compatible with this format. As assessing and improving data quality cleansing is time-consuming and expensive, the proposed approaches should simplify as much as possible the procedures to guarantee high-quality data by proposing (semi-)automatic procedures. Moreover, data cleansing approaches usually require specific expertise that limits the applicability of the proposed mechanism. To ameliorate competencies requirements, novel proposals should limit the required skills to favorite wider exploitation of data and their cleaning methodologies.

In this context, the first pillar of my research is placed: proposing (semi-)automatic **privacy-aware data cleansing approaches** dealing with tabular data to make data users able to improve Open Data quality while preserving individual privacy. It resulted in a series of approaches and prototypes, mainly integrated into a Social Platform for Open Data (SPOD) used by Public Administrations, such as the Campania Region, associations, such as Hetor, and citizens, such as students joining activities to familiarise themselves with the Open Data directive.

While data providers mainly publish tabular data, data consumers might be interested in semantic reach data format, such as graph-like structures, as they can be easily navigated and explored thanks to their interlinked properties. However, directly querying Knowledge Graphs requires expertise in query languages and awareness in the conceptualised data, which are considered too challenging for lay users.

Hence, data consumers require Knowledge Graph exploitation means being able to mask underlying technical challenges. Moreover, data users may require to consume data according to their expertise, background, application contexts, needs, interests, capabilities. It requires designing data exploitation approaches that deal with specific requirements according to the targeted stakeholders. This dissertation mainly focuses on *people with data table manipulation and visualisations experiences*, to guide them to move from tabular data to Knowledge Graph exploitation means, *education* to guide pupils in implicitly exploiting Knowledge Graphs in knowledge management and information retrieval tasks, and the *cultural heritage community*, for their wide interest in publishing their data according to the Semantic Web technologies.

It results in the second pillar of this dissertation, the effort in designing and implementing **Knowledge Graph exploitation means**. As a general approach, users are guided in querying Knowledge Graphs by (controlled) natural language interfaces and organising results as data tables, manually or automatically perform data manipulation, and exploit results in dynamic artifacts. According to target-oriented requirements, experts in data table manipulation are provided with a mechanism to author dynamic and exportable data visualisation components; pupils are guided to navigate word clouds while implicitly consuming Knowledge Graphs; cultural heritage lovers are guided to author virtual reality-based virtual exhibitions or ready-to-use virtual assistant extensions behaving as virtual guides. The generated artifacts demonstrate our interest in letting data consumers play the role of an active user of available data and exploiting them in concrete, dynamic, reusable and shareable artifacts taking advantage of (Linked) Open Data.

In the last years, researchers and businesses are increasingly investing in Machine Learning approaches that have the potentiality to automate data analysis, such as quality improvement and privacy-preserving tasks, and make data-informed predictions in real-time without any human intervention. Machine Learning tasks expect numerical values as input, while Knowledge Graphs are graph-shape by nature. In the last decade, several different graph embedding approaches have been proposed to make them interoperable by representing Knowledge Graph nodes (and edges) as numerical vectors. Hence, the third and latter pillar at the basis of this dissertation explores how to make **Knowledge Graphs interoperable with Machine Leaning** tasks. In particular, it focuses on the definition of a fair mechanism to easily compare different graph embedding approaches, topic that is still scarcely addressed in the literature. In this context, we proposed a configurable and extensible community-shared software framework to evaluate and compare graph embedding techniques on Machine Learning and semantic tasks.

All in all, this dissertation reports different approaches and prototypes to support data producers and consumers in the entire data management process, from data cleansing to guided data exploitation mechanisms. It is based on three main pillars: 1) *Open Data quality and privacy assessment and improvement*, where different data cleansing and privacy-preserving approaches have been proposed to support data publishers in providing privacy-aware fit-for-use tabular data; 2) *Knowledge Graph exploitation*, further specialised in approaches proposed to support experts in tabular data manipulation to reuse their expertise in the Semantic Web context, pupils in exploiting Knowledge Graphs in digital storytelling and knowledge management, and the cultural heritage community in leaving a passive position and play the role of active content creation; 3) *Knowledge Graphs and Machine Learning* to fairly compare graph embedding techniques in Machine Learning and semantic tasks. Each contribution resulted in a community-shared working prototype, tested on real users belonging to the targeted stakeholder group. Each pillar is supported by peer-reviewed publications.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AC	Agglomerative Clustering
AI	Artificial Intelligence
ASQ	After Scenario Questionnaire
ATU	Attitude Toward Using
BGP	Basic Graph Pattern
BI	Behavioural Intention
CRC	Campania Regional Council
CH	Cultural Heritage
CI	Confidence Interval
CSW	Catalogue Service for the Web
DB	Database

EOU	Easy of Use
ETL	extract, transform, load
FN	False Negative
FP	False Positive
FSI	Faceted-search interface
GaaP	Government as a Platform
GIS	Geographic Information System
GLAM	Galleries, Libraries, Archives, and Museums
G2C	Government to Citizens
G2G	Government to Government
HCI	Human-Computer Interaction
KG	Knowledge Graph
KGQA	Knowledge Graph Question Answering
KM	Knowledge management
ICT	Information and communication technologies
ID	Identifier
IR	Information retrieval
LD	Linked Data
LOD	Linked Open Data
ML	Machine Learning
NL	Natural Language
OD	Open Data
OLD	Open Linked Data
PA	Public Administration
PP	Periceived Playfulness
PPDP	Privacy Preserving Data Publishing
PU	Perceived Usefulness
QA	Question-answering
QID	Quasi-identifier
RPA	Regional Public Administration

RQ	Research Question
SD	Standard Deviation
SUS	Standard Usability Survey
TAM	Technology acceptance model
TN	True Negative
TP	True Positive
VA	Virtual Assistant
VR	Virtual Reality

Part I

INTRODUCTION AND BACKGROUND

1 INTRODUCTION

Data publication is the act of releasing data for being used by others. It is a practice consisting in preparing data for public use to make them available to everyone to use as they wish. This practice is an integral part of the open science movement and let data be distributed as *Open Data*. In publishing data, governments, businesses, and entrepreneurs harness the power of data for economic, social, and scientific gains [351] and encourage data reuse. There is a large and multidisciplinary consensus on the benefits resulting from this practice [302]. Researchers are interested in publishing academic data as data represent a fundamental source for data analysis, statistics, and informed knowledge-based decision making [225]. Public Administrations achieve transparency by publishing Open Data while guaranteeing participation and collaboration via active involvement of heterogeneous stakeholders [237]. Data curators can easily spread data, to make them re-usable and easily accessible by data consumers. These examples point out the interest and the advantages in publishing data as Open Data. However, enabling reuse is strictly related to provide *fit-for-use* data, corresponding to pay attention to qualitative data properties [50].

Data quality plays an essential role in successfully learning and discovering knowledge from data [225]. It avoids misleading conclusions caused by dirty data, inaccurate data analytics results, and wrong business decisions [30]. If users aim to exploit data in Machine Learning tasks, missing values, class imbalance, and inaccurate values can heavily affect model performance and its outcome [317]. Data quality affects credibility of data providers, as it happens in health institutions [315, 319]. Moreover, it can simplify data discoverability and support lookup mechanisms. For instance, the economy field can benefit from easier access to information, content, and knowledge to contribute to innovative services and new business models [331]. Fit-for-use data guarantee the possibility to take advantage of available data and transform them into knowledge. For instance, public agencies and data providers are spur in guaranteeing high-quality data [243] to enable the value gain out of data [247]. These examples clarify that data quality assessment and improvement are mandatory steps before data exploitation [21] and they are of interest of both data producers and consumers [309] in any domain, from government to healthcare, from academia to industry.

Nevertheless the interest in making data easily accessible, data quality behaves as a barrier in data exploitation [26]. Assessing and improving data quality is challenging, error-prone, time-consuming and expensive. In fact, up to 80% of data scientist's effort is spent in adapting data to user applications and make data reliable enough to lead to trustful results [284]. Moreover, the Harvard business review reported that inaccurate data costs the companies about $3 trillion per year [268]. Ad-hoc manual approaches are widely used in real applications, and they require heavy manual labour and human expert judgments. All these aspects justify the need of (semi)-automatic approaches to guide data publishers and consumes to detect and correct quality issues.

The situation becomes even more complex when published data concern individuals. Data in their raw and original form could contain personal and sensitive information about individuals. Publishing such data violate individual privacy [102]. Rather than opting for keeping data closed, it is worth applying methods and tools for publishing useful data while *preserving individual privacy*. This practice is known as Privacy-preserving data publishing [53] and it represents the art of publishing sensitive personal data without violating individuals' privacy.

To guarantee high-quality and privacy preserving data requires providing data users with quality and privacy assurance approaches to care about data quality and privacy during the publication phase or before data exploitations means to make them easily compatible with data analysis or data exploitation means.

Data publishers usually opt for publishing data in tabular format as it requires minimum effort. It only requires organising data in data tables, caring about the selection of meaningful columns and data table structures. On the opposite side, realising them as Linked Data or as Knowledge Graphs helps data users to easily access them and supports data exploitation. The central idea of Linked Data is that they simplify exploratory applications and data integration by complying with a set of best practices in the areas of linking, vocabulary usage, and metadata provision. An increasing interest is manifested over Knowledge Graphs publication: the Linked Open Data Cloud [207] (a Knowledge Graph that collects most of the published Knowledge Graphs by academia and industry) counted 12 datasets in 2007 and currently contains 1,239 datasets. Despite the quantitative reason to exploit Knowledge Graphs, also the provenance is an important aspect: several virtuous institutions invested or are investing in publishing data in the linked format, such as, Europeana[1], Eurostat[2], ISTAT, Beni Culturali[3], the British Museum[4]. Furthermore, it is highly recommended to interlink Knowledge Graphs [28] and it implies the possibility to navigate from a Knowledge Graph to another. Because of the extensive range of heterogeneous information stored in Knowledge Graphs, for their easy navigation, thanks to their quantitative and qualitative properties, they could behave as a critical resource for information retrieval and knowledge management.

The difference in *data format* used by data curators and the one desired by data consumers requires transforming data from tabular data to a graph-like structure where concepts are modelled as nodes and relations are modelled as edges connecting interested nodes. To perform this transformation, data curators can either opt for materialising data in a Knowledge Graph by defying the underlying ontology and representing data according to Semantic Web standards or for exposing tabular data as virtual Knowledge Graphs without explicitly converting data. Besides the adopted method, there is a plethora of tools and standard approaches to perform this transformation and they result in the exposure of data as Knowledge Graphs.

By focusing on end-users in general, we can make assumption on their technical skills. Directly *querying Knowledge Graphs* is

[1] Europeana: https://pro.europeana.eu/page/linked-open-data
[2] Eurostat: https://ec.europa.eu/eurostat/web/nuts/linked-open-data
[3] Beni Culturali: http://dati.culturaitalia.it/
[4] British Museum: https://.../british-museum-collection

mainly affected by required technical competences in query languages, such as SPARQL, and in understanding the semantics of the supported operators, too challenging for lay users [75, 106], and conceptualisation issues to understand how data are modelled. Hence, there is an interest in developing tools to implicitly compose queries by hiding the underlying complexity to open Knowledge Graphs also to lay users [74]. To meliorate these challenges, in the last decade researchers and companies developed tools and interfaces to support users in interacting with Knowledge Graphs by implicitly composing queries while hiding the underlying complexity.

Among data consumers, developers, researchers and businesses should not be left behind. According to their technical skills and the role plaid as infomediary, these communities usually perform data analysis on raw data. In the last years, researchers and businesses are increasingly investing in Machine Learning approaches that have the potentiality to automate data analysis, such as quality improvement and privacy-preserving tasks, and make data-informed predictions in real-time without any human intervention. Machine Learning tasks expect numerical values as input, while Knowledge Graphs are graph-shape by nature. In the last decade, several different graph embedding approaches have been proposed to make them interoperable by representing Knowledge Graph nodes (and edges) as numerical vectors. But, the definition of a fair mechanism to easily compare different graph embedding approaches is still scarcely addressed in the literature. Hence, technical users require approaches, tools and mechanisms to efficiently and effectively identify the best graph embedding technique according to the task(s) they plan to perform.

1.1 MOTIVATIONS

All the contributions at the basis of this dissertation have a motivational *fil rouge* that justifies design and implementation choices. Besides the specific contributions discussed in the following, all the proposals have some underlying assumptions and motivations that justify adopted data format, elicited stakeholders, proposed methodologies and aspects considered during the design and implementation. This section summarises the key motivations at the basis of any proposal discussed in this dissertation.

- **Open by default.** When data publishers experience data privacy issues or expensive data cleansing approaches usually opt for preventing the publication of data at all. On the opposite, we encourage data providers to open data by default, supporting them in solving quality and privacy concerns by publishing useful data, in terms of data quality, without compromising individual privacy.

- **Data quality and privacy checks on tabular data.** Data publishers are required to guarantee fit-for-use data and they usually publish data in tabular format. Hence, to effectively support them in the data cleansing process, data providers should be supported in performing detecting and correcting actions directly on the tabular format to guarantee high-quality data and preserve individual privacy.

- **(Semi-)automatic data cleansing approaches.** Due to required expertise in performing data cleansing and technical challenges in publishing high-quality Open Data, data providers should be supported by (semi)-automatic approaches by exploring artificial intelligence and machine learning driven mechanisms. They might minimise human efforts by carefully checking for the accuracy and performance of the proposed approaches.

- **Knowledge Graph exploitation mechanisms while masking technical issues.** Data consumers might be interested in accessing and exploiting Knowledge Graphs thanks to their quantitative and qualitative reasons. However, technical challenges posed by query language obstacle data exploitation. Hence, users require mechanisms and interfaces to easily accessing data stored in Knowledge Graphs without requiring technical competences in query languages and minimising their awareness in data modelling.

- **Proposal of target-oriented Knowledge Graph exploitation approaches.** Knowledge Graph exploitation can be interesting for many different stakeholders. However, they are different for background, technical competences, needs, application context. Hence, data exploitation mechanisms should take into requirements posed by target users in designing and proposing mechanisms that effectively address end-users needs.

1.2 CONTRIBUTIONS

The purpose of the contributions presented in this dissertation is to support users in the entire process of data publication and exploitation. My contributions mainly focus on the following pillars: i) *Open Data quality and privacy assessment and improvement* aiming to propose approaches and toolkit to support data curators in assessing and improving data quality and privacy concerns by focusing of Open Data released in a tabular format; ii) *Knowledge Graph exploitation* aiming to propose approaches and prototype tools to express users' needs or explore available data by a Natural Language interface to guide end-users with different interests, types of background, age, and needs to query Knowledge Graphs and take advantage of them without requiring technical skills in query languages. This pillar can be further split according to the targeted stakeholders. In particular, we focus on requirements posed by experts in Open Data manipulation to bring them closer to Knowledge Graphs by exploiting their expertise in data table manipulation and visualisation by exploring the *Knowledge Graphs and data visualisation*. We explore the *Knowledge Graphs and storytelling* in the educational context. In the context of *Knowledge Graphs and the Cultural Heritage community*, we detect the best data exploitation means of interest of this community, such as virtual guides and virtual exhibitions, and we explore how Knowledge Graphs can let Cultural Heritage lovers to play the role of museum curators. Finally, iii) the analysis and evaluation of the compatibility between *Knowledge Graphs and Machine Learning* where Knowledge Graphs are graph shaped by nature and Machine Learning algorithms expect data in a vector form. Making Knowledge Graphs compliant with Machine Learning algorithms requires the definition and a systematic evaluation and comparison of graph embedding techniques. For each pillar, this section summarises contributions and matches them to peer-reviewed scientific publications supporting statements and evidences described in this dissertation. Table 1.1 schematically reports investigated pillars and the related scientific articles.

Table 1.1: Peer-reviewed publications supporting my dissertation

Topic	Year	Pub. Type	Title	Ref.
Open Data Quality and Privacy Assessment and Improvement	2019	Conference *Dg.O.*	A Non-prescriptive Environment to Scaffold High Quality and Privacy-aware Production of Open Data with AI	[108]
	2019	Conference *EGOV*	Orchestrated Co-creation of High-Quality Open Data Within Large Groups	[109]
	2020	Conference *EGOV*	Detecting and Generalizing Quasi-Identifiers by Affecting Singletons	[249]
	2021	Conference *CODS-COMAD*	Detecting Data Accuracy Issues in Textual Geographical Data by a Clustering-based Approach	[255]
	2021	Journal *Transforming Government: People Process and Policy*	Government as a Platform in a Regional Public Administration	Under evaluation
Knowledge Graphs and data visualisation: QueDI	2019	Conference *CSCWD*	Linked Data Queries by a Trialogical Learning Approach	[93]
	2020	Conference *Semantics*	QueDI: from Knowledge Graph Querying to Data Visualization	[94]
	2020	Workshop *TEL4FC@MIS4TEL*	Education Meets Knowledge Graphs for the Knowledge Management	[81]
Knowledge Graphs and storytelling: Novelette	2020	Conference *IV*	Visual Storytelling by Novelette	[2]
	2021	Workshop *TEL4FC@MIS4TEL*	Engaging Children in Digital Storytelling	[4]
	2021	Conference *I-CITIES*	Engaging Children in Smart Thing Ideation via Storytelling	[254]
	2021	Journal *IEEE Education Society*	Novelette, a Usable Visual Storytelling Digital Learning Environment	[3]
Knowledge Graphs and the Cultural Heritage community	2020	Journal *Cultural Heritage*	The role of Linked Open Data in Authoring Virtual Exhibitions	[216]
	2021	Conference *ESWC*	Automatic Skill Generation for Knowledge Graph Question Answering	[259]
	2021	Journal *Semantic Web*	Move Cultural Heritage Knowledge Graphs in Everyone's Pocket	Under evaluation
Knowledge Graphs and Machine Learning: GEval	2019	Conference *ESWC*	A Configurable Evaluation Framework for Node Embedding Techniques	[253]
	2020	Conference *ESWC*	GEval: A Modular and Extensible Evaluation Framework for Graph Embedding Techniques	[252]

1.2.1 *Open Data Quality and Privacy Assessment and Improvement*

Data quality and privacy concerns gain the interest of both data producers and consumers in any application domain. In this dissertation, we mainly focus on a Regional Public Administration as data cleansing target user, the Campania Region, without any loss of generality as the proposed approaches and toolkits are general enough to be used by any data curator and user.

The Campania Region performed a strategic flagship project in 2018 and 2019 to completely transform both its organisation and technological support and break down its data siloed structure by revising the data production process. By focusing on the technical and technological changes, the Campania Region administration transformed the internal workflow to produce and publish Open Data. In this context, we actively collaborate with Campania Region administration delegates to collect requirements and needs. This partnership resulted in the adaptation of a Social Platform for Open Data (SPOD) as an internal platform by the Campania Region administration. SPOD represents the main outcome from an H2020 project, ROUTE-TO-PA, coordinated by my research group, while its modification to be compliant with the Campania Region administration requests is referred to as *Campania Crea*.

The main contributions of the partnership with the Campania Region Public Administration are:

- the technical support of our Regional Public Administration to perform a strategic modification to the Open Data publishing mechanisms to favourite a more collaborative setting and break down data silos;
- design and modification of a social platform to support data curators in collaboratively publishing high-quality Open Data, referred to as `Campania Crea`;
- the extensive evaluation of the acceptance rate of `Campania Crea` by our Regional Public Administration members to verify in practice the effectiveness of the proposed mechanism in a real context. Results are rather positive, highlighting that the Campania Region succeeded in involving its

members in this revolutionary plan and members positively accept a social platform to collaboratively create Open Data by enabling multi-disciplinary and multi-departmental collaboration. It is a step forward in breaking data siloes.

The political and technological changes applied in our Regional Public Administration to change the internal mechanism to publish Open Data, break down data silos, favouring a more collaborative data publication mechanism, the design, implementation, and evaluation of *Campania Crea* has been submitted as

> Salvatore Avella, Angela Cocchiarella, Dario Fonzo, Giuseppina Palmieri, Maria Angela Pellegrino, and Vittorio Scarano. *"Government as a Platform in a Regional Public Administration"*. *Submitted* to Transforming Government: People, Process and Policy in July 2021.

Authors are placed in alphabetical order. Salvatore Avella, Angela Cocchiarella, and Dario Fonzo, as members of the Campania Region administration, curated the documentation of the flagship project from a political and organisational point of view. Giuseppina Palmieri, Maria Angela Pellegrino, and Vittorio Scarano curated the technological and technical details and the performed evaluation. Maria Angela Pellegrino curated the results analysis and the article writing under the supervision of all the co-authors.

DATA QUALITY AND PRIVACY ASSESSMENT. The data publication mechanism may involve heterogeneous stakeholders different in competencies, interest, application context, reliability. Users involved in a collaborative data curation process require to agree on data format, data structure, data quality standards, dataset content. They might benefit from a scaffolding set of toolkits and rules to simplify and coordinate their collaboration, minimise communication efforts, and maximise the quality of the resulting datasets. The main contributions of the *data quality and privacy assessment* pillar are:

- the design and definition of data quality assessment and privacy-preserving mechanisms compliant with tabular data;

- the design and implementation of a unified approach to assess data quality and privacy issues based on an automatic type inference approach;
- the evaluation of the performance of the proposed mechanism on real datasets;
- the introduction of the proposed mechanism in `SPOD`.

The proposed quality and privacy assessment approach has been presented as the following conference article:

> Giuseppe Ferretti, Delfina Malandrino, Maria Angela Pellegrino, Donato Pirozzi, Gianluigi Renzi, and Vittorio Scarano. *"A Non-prescriptive Environment to Scaffold High Quality and Privacy-aware Production of Open Data with AI."* In 20th Annual International Conference on Digital Government Research (2019).

Authors are placed in alphabetical order. Giuseppe Ferretti and Gianluigi Renzi, belonging to the Campania Region administration, dealt with requirements and motivational use case. Delfina Malandrino and Vittorio Scarano curated the coordination between the Campania Region representatives and the technical team and supervised the technological contributions. Maria Angela Pellegrino and Donato Pirozzi developed the technological solution, while Maria Angela Pellegrino tested it. The article was written by Maria Angela Pellegrino and Donato Pirozzi, under the supervision of Delfina Malandrino and Vittorio Scarano.

DATA QUALITY IMPROVEMENT. Open datasets, such as the ones published by the Campania Region, contain highly inaccurate data where the accuracy is compromised by abbreviations, inconsistent representations, misspellings. Hence, once identified quality issues, data curators should perform data cleansing.

The main contributions of the *data quality improvement* pillar are:

- the design and definition of data quality improvement mechanisms able to directly work on tabular data;
- the design and proposal of a role-based proactive data quality assurance mechanism to orchestrate data publishing;

- the design and implementation of a semi-automatic reactive data cleansing approach to detect and improve inaccurate values in data tables by a clustering-based implementation;
- the evaluation of the performance of the proposed mechanisms on real open datasets.

The role-based orchestration approach is guided by a motivational use case defined by the Campania Region administration. The proposed approach ensures data quality while dealing with large groups where roles guarantee syntactic and semantic data quality and it has been presented in

> Giuseppe Ferretti, Delfina Malandrino, Maria Angela Pellegrino, Andrea Petta, Gianluigi Renzi, Vittorio Scarano, Luigi Serra. *"Orchestrated Co-creation of High-Quality Open Data Within Large Groups"*. In the Proceedings of Electronic Government (EGOV) 2019.

Authors are placed in alphabetical order. Giuseppe Ferretti and Gianluigi Renzi, belonging to the Campania Region administration, dealt with requirements and the motivational use case. Delfina Malandrino and Vittorio Scarano curated the coordination between the Campania Region representatives and the technical team and supervised the technological contributions. Andrea Petta and Luigi Serra developed the technological solution, while Maria Angela Pellegrino documented the performed process and wrote the article under the supervision of Vittorio Scarano.

The proposed clustering-based data cleansing approach detects inaccurate values, such as typos, abbreviations, misspellings, or any other syntactical errors, on textual geographical data, such as provinces and municipalities, while the improvement phase proposes a correction for any detected error. It combines an approximate string matching and the well-known Levenshtein. By evaluating it on real open datasets, we proved that this combination improves (or not worsens) the results of using Levenshtein in isolation. Since the clustering-based approach returns datasets containing fewer errors if compared with a pair-wise comparison algorithm between each value in input with a dictionary of correct values to detect the corresponding entity, clustering helps in detecting and correcting inaccurate values in textual geographical data. Finally, our proposal obtains extremely more accurate results than OpenRefine, a popular tool used by data curators to refine data quality, thanks to the exploitation of a dictionary of correct values. The proposed approach and the related evaluation have been presented in

> Maria Angela Pellegrino, Luca Postiglione, and Vittorio Scarano. *"Detecting Data Accuracy Issues in Textual Geographical Data by a Clustering-based Approach."* In the Proceedings of the 8th ACM IKDD CODS and 26th COMAD (2021)

DATA PRIVACY-PRESERVING. Concerning the privacy assessment and improvement approach, data owners are spur in opening up their data to enable informed decision making, ensure transparency, audience engagement, and release social and commercial value. Unfortunately, data in their raw and original form could contain personal and sensitive information about individuals. Thus, data curators should perform Privacy-preserving data publishing to provide useful data without violating individuals' privacy. It requires detecting identifiers and quasi-identifiers and applying corrective actions. The proposed approach is based on a privacy issues detection step followed by an anonymity approach based on generalisation and suppression.

The main contributions of the *data privacy preserving* pillar are:

- the assessment of data privacy issues in real open datasets;
- the design and definition of data privacy-preserving mechanisms able to directly work on tabular data;
- the design and implementation of an anonymisation mechanism to deal with well-known quasi-identifiers;
- the evaluation of the performance of the proposed mechanism on real open datasets.

The proposed anonymisation approach can be defined as a modified version of k-Anonymity where k is at least equal to 2; suppression is discouraged in favour of generalisation, and changes are applied to work locally. Tests on real datasets concerning driver licenses released by the Italian Ministry of Infrastructure and Transport demonstrate that our proposal achieves the same results of widely adopted k-anonymity in terms of privacy-preserving, while obtaining better data quality thanks to a local recording. The proposed anonymisation algorithm and its evaluation has been presented in

> Matteo Pastore, Maria Angela Pellegrino, Vittorio Scarano. *"Detecting and Generalizing Quasi-Identifiers by Affecting Singletons"* In the Proceedings of EGOV-CeDEM-ePart-* (2020)

Authors are placed in alphabetical order. Maria Angela Pellegrino and Vittorio Scarano defined the work objectives and proposed the methodology. Matteo Pastore and Maria Angela Pellegrino developed and tested the technological contribution. Maria Angela Pellegrino wrote the article with the support of Matteo Pastore and under the supervision of Vittorio Scarano.

1.2.2 *Knowledge Graphs Exploitation*

A traditional knowledge management process requires i) retrieving data, ii) refining them, and iii) performing data exploitation. According to the requirement to mask syntactical challenges and minimise the required awareness in the underlying data structure, we mainly focus on data retrieval mechanisms based on query builders enhanced with (controlled) Natural Language interfaces to guide users in naturally posing questions by simulating, as much as possible, human interactions. If users have a clear objective, they can directly type or pronounce their requests. Vice versa, in exploratory search, users are guided in iteratively creating and refining questions. As a query builder, Natural Language queries can be translated to SPARQL to be run over a SPARQL endpoint that is a standard way to expose Knowledge Graph content. Among SPARQL constructs, SELECT query results can be naturally represented as tabular data. Thus, retrieved data are modelled as tables, which can be manually or automatically refined, and finally, used in data exploitation mechanisms. It may result in textual replies or concrete artifacts, perhaps customizable and exportable, such as charts, data visualizations, data stories, or virtual reality-based data representations.

Different stakeholders may be interested in accessing, querying, and exploiting Knowledge Graphs. In the projects at the basis of my dissertation, we mainly focused on Open Data experts and Public Administration, education, and the Cultural Heritage community as target groups. In the context of Open Data experts, we give for granted their capabilities in manipulating data and easily visualise them. Hence, we explore the exploitation of Knowledge Graphs in data visualization. Concerning education, we mainly explore the possibility to exploit Knowledge Graphs to support pupils in digital storytelling. Finally, according to the application context of the Cultural Heritage community, we explore the exploitation of Knowledge Graphs by Virtual Assistants in the direction of generating virtual guides and Virtual Reality in authoring virtual exhibitions. The proposed approach and the resulting prototypes have been summarised in

> Maria Angela Pellegrino. *"Knowledge Graphs within everyone's means"*. Ph.D. consortium in CHItaly 2021.

KNOWLEDGE GRAPHS AND DATA VISUALISATION. Open Data experts are usually aware of data refinement and exploitation approaches, considered their comfort zone, while they might be unaware of Knowledge Graph query languages. They might be provided with a guided approach to get closer to Knowledge Graphs without requiring any expertise on Semantic Web technologies. The main contributions at the basis of the *Knowledge Graphs and data visualisation* pillars are:

- the proposal of a *transitional approach* where Open Data experts are guided from Knowledge Graphs querying to their comfort zone;
- the prototyping of this transitional approach which resulted in `QueDI` composed of `ELODIE`, an implicit SPARQL query builder enhanced with a controlled Natural Language interface, and a data visualisation mechanism;
- the introduction of `QueDI` in `SPOD` to let the Open Data community built around the `SPOD` platform to exploit not only Open Data but also their linked version;

- the design and implementation of a *trialogical learning approach* within SPOD to let data users the possibility to exploit Knowledge Graphs either in isolation or collaboratively;
- the extensive evaluation of the performance of QueDI in terms of accuracy, scalability, and usability.

QueDI allows users to build queries step-by-step with an auto-complete mechanism and to exploit retrieved results by exportable and dynamic visualizations. First, it scaffolds users in creating a tabular representation of the dataset of interest by ELODIE, which realises an exploratory search by organising available data in facets and supporting users in automatically retrieving both user query results and data to go on with the query formulation by querying a configured SPARQL endpoint. Second, QueDI supports a manual dataset manipulation phase where users can exploit their skills in data refinement by aggregating, sorting, filtering, and cleaning data by interacting with a form-based interface that behaves as a SQL builder. Finally, it enables the creation of exportable visualisation. QueDI interaction model, its interface and evaluation have been presented in

> Renato De Donato, Martina Garofalo, Delfina Malandrino, Maria Angela Pellegrino, Andrea Petta, Vittorio Scarano. *"QueDI: From Knowledge Graph Querying to Data Visualization"* In the Proceedings of SEMANTiCS 2020.

The trialogical learning approach and the usability evaluation of ELODIE according to Open Data experts have been presented in

> Renato De Donato, Martina Garofalo, Delfina Malandrino, Maria Angela Pellegrino, Andrea Petta, Vittorio Scarano. *"Linked Data Queries by a Trialogical Learning Approach"*. In the Proceedings of Computer Supported Cooperative Work in Design (CSCWD) 2019.

All the groups considered ELODIE of a *good* usability level demonstrating that it results in a non-intimidating approach to familiarise and exploit Knowledge Graphs without requiring any technical skill in query languages.

Concerning the education context, we proposed QueDI as a Knowledge Management tool in the educational context to support future citizens in going beyond the passive inspection of results returned by a search engine, and in actively searching for the data that best answer their questions. The potentialities of using `QueDI` at school has been presented in

> Renato De Donato, Martina Garofalo, Delfina Malandrino, Maria Angela Pellegrino, Andrea Petta. *"Education Meets Knowledge Graphs for the Knowledge Management"*. In the Proceedings of International Conference in Methodologies and Intelligent Systems for Technology Enhanced Learning (MIS4TEL) 2020.

In all the articles listed above, authors are placed in alphabetical order. Martina Garofalo and Maria Angela Pellegrino designed, developed, and tested the technological contribution. Renato De Donato and Andrea Petta curated the integration of the described steps. Delfina Malandrino and Vittorio Scarano supervised the work. Maria Angela Pellegrino wrote the article under the supervision of Vittorio Scarano.

KNOWLEDGE GRAPHS AND STORYTELLING. Digital storytelling is considered an opportunity to *think without a bannister*, meaning that pupils are free to imagine, invent and tell any story of interest, without a pre-defined correct answer. It is strictly related to creativity and divergent thinking, as it is perceived as a powerful approach to spur imagination and freely invent any story of interest. In this context, we investigated the possibility to exploit Knowledge Graphs in storytelling to support pupils in overcoming the blank page syndrome. The contributions of the *Knowledge Graphs and storytelling* pillar are:

- the proposal of a synonym lookup mechanisms which exploits Knowledge Graphs to retrieve and navigate inspiring words starting from a user-defined input;

- the design and implementation of this approach in `Novelette`, a digital learning environment to support pupils in performing storytelling;

- the extensive evaluation of `Novelette` usability according to educators and pupils in a controlled and a real setting environment, respectively;

- the evaluation of the pupils' engagement in inventing and authoring stories via `Novelette`, in the presence and remotely.

The applicability of `Novelette` to heterogeneous contexts has been shown by authoring tales, data stories, and media stories, as demonstrated in the poster paper presented in

> Agnese Addone, Renato De Donato, Giuseppina Palmieri, Maria Angela Pellegrino, Andrea Petta, Vittorio Scarano, Luigi Serra. *"Visual Storytelling by Novelette"*. In the Proceedings of Information Visualisation (IV) 2020.

Authors are placed in alphabetical order. Agnese Addone, Giuseppina Palmieri and Vittorio Scarano designed `Novelette` and curated the networking with educators. Renato De Donato, Andrea Petta, and Luigi Serra developed the technological contribution. Maria Angela curated use cases, `Novelette` evaluation with educators and learners, and results analysis. The article was mainly written by Maria Angela Pellegrino, proofread by Agnese Addone and Giuseppina Palmieri, and revised by Vittorio Scarano.

If users experience writer's block, `Novelette` implements a suggestion provision mechanism. Users can type the word of interest and `Novelette` automatically retrieves synonyms by querying BabelNet and organises retrieved results in (navigable) word clouds. It represents a keyword-based interface to implicitly explore Knowledge Graphs by navigating synonyms. By testing `Novelette` and the suggestion provision mechanism in a real context at school, `Novelette` engages children in inventing and authoring stories, as reported in

> Agnese Addone, Giuseppina Palmieri, Maria Angela Pellegrino. *"Engaging Children in Digital Storytelling"* In the Proceedings of Methodologies and Intelligent Systems for Technology Enhanced Learning (MIS4TEL) 2021.

Authors are placed in alphabetical order. Agnese Addone and Giuseppina Palmieri designed `Novelette` and curated the networking with educators. Maria Angela curated use cases, `Novelette` evaluation with educators and learners, and results analysis. The article was mainly written by Maria Angela Pellegrino, and proofread by co-authors.

Usability has been assessed both according to educators and pupils. Children stated that `Novelette` is *super-adapt* for them in terms of usability as discussed in the article

> Agnese Addone, Renato De Donato, Giuseppina Palmieri, Maria Angela Pellegrino, Andrea Petta, Vittorio Scarano, Luigi Serra. *"Novelette, a Usable Visual Storytelling Digital Learning Environment." Accepted* in the IEEE Education Society. *In press.*

Authors are placed in alphabetical order. Agnese Addone, Giuseppina Palmieri and Vittorio Scarano designed `Novelette` and curated the networking with educators. Renato De Donato, Andrea Petta, and Luigi Serra developed the technological contribution. Maria Angela curated use cases, `Novelette` evaluation with educators and learners, and results analysis. The article was mainly written by Maria Angela Pellegrino, proofread by Agnese Addone and Giuseppina Palmieri, and revised by Vittorio Scarano.

In the context of letting children familiarise themselves with smart city concepts and design their smart object [76, 256, 258, 275], `Novelette` results in a non-intimidating environment to clarify terminology and exploit storytelling to model smart objects interactions, as reported in the long abstract presented in

> Maria Angela Pellegrino, Mauro D'Angelo. *"Engaging Children in Smart Thing Ideation via Storytelling."* In I-CITIES 2021.

KNOWLEDGE GRAPHS AND THE CULTURAL HERITAGE COMMUNITY. Concerning the Cultural Heritage community, in the last year, due to COVID-19 restrictions and worldwide lockdowns, virtual experiences have been widely adopted to enhance physical tours, but Cultural Heritage lovers still behave as visitors. In this context, we explored the possibility to bring the Cultural Heritage community closer to the Semantic Web technologies. The main contributions at the basis of the *Knowledge Graphs and Cultural Heritage community* are:

- the proposal of mechanisms to let the Cultural Heritage community play the position of active museum curators;

- the detection of data exploitation means compliant with the Cultural Heritage community needs;
- the design and implementation of an interface to author virtual reality-based virtual exhibitions;
- the design of a generator of virtual assistant extensions to automatically query Knowledge Graphs by implicitly composing queries;
- the prototyping of a generator of Alexa skills to automatically query Knowledge Graphs;
- the evaluation of the performance of the automatically generated skills with standard benchmarks in Knowledge Graph Question Answering;
- the assessment of the expected impact and potential utility of the proposed solution according to Cultural Heritage lovers and data curators.

The proposal to take advantage of Cultural Heritage Knowledge Graphs in an authoring platform for Virtual Reality-based virtual exhibitions by combining `ELODIE` and an automatic mechanism to create Virtual Reality-based solutions has been accepted in the Journal of Cultural Heritage and it is currently in press as

> Daniele Monaco, Maria Angela Pellegrino, Vittorio Scarano, Luca Vicidomini. *"Linked Open Data in Authoring Virtual Exhibitions." Journal of Cultural Heritage, 2022*

Authors are placed in alphabetical order. Vittorio Scarano designed the proposed approach. Daniele Monaco, Maria Angela Pellegrino, and Luca Vicidomini developed the technological contribution. The article was written by Maria Angela Pellegrino and revised by Vittorio Scarano.

The prototype of the generator has been presented as

> Maria Angela Pellegrino, Mario Santoro, Vittorio Scarano, Carmine Spagnuolo. *"Automatic VA extension Generation for Knowledge Graph Question Answering"*. In the Proceedings of Extended Semantic Web Conference (ESWC) (Satellite Events) 2021.

An extended version of this article containing a quantitative and qualitative analysis of Cultural Heritage Knowledge Graphs and an extensive overview of the proposed community shared software framework has been submitted as the following contribution and it is actually under the second round of evaluation:

> Maria Angela Pellegrino, Vittorio Scarano, Carmine Spagnuolo. *"Move Cultural Heritage Knowledge Graphs in Everyone's Pocket." Submitted* to the Semantic Web Journal in March 2021.

1.2.3 *Knowledge Graphs and Machine Learning*

While Knowledge Graphs are graph shaped by nature, most traditional Machine Learning algorithms expect data in a vector form. To transform graph elements to vectors, several graph embedding approaches have been proposed. Systematic comparative evaluations of different approaches are scarce; approaches are rather evaluated on a handful of often project-specific datasets. Hence, there is the need for a mechanism to simplify the evaluation and the comparison of graph embedding techniques. The main contributions of the *Knowledge Graphs and Machine Learning* pillar are:

- the design of a systematic comparison evaluation approach to fairly evaluate graph embedding techniques on both Machine Learning and tasks borrowed by the semantic field, such as document or entity similarity;
- the implementation of the proposed approach in a community shared software framework, `GEval`;
- the evaluation of the performance of `GEval` on well-known graph embedding techniques;
- the demonstration of the working mechanism of `GEval` and its comparison in parameter tuning.

A preliminary version of `GEval` and the evaluation of the sequential and parallel performances have been presented as

> Maria Angela Pellegrino, Michael Cochez, Martina Garofalo, Petar Ristoski. *"A Configurable Evaluation Framework for Node Embedding Techniques."* In the Proceedings of Extended Semantic Web Conference (ESWC) (Satellite Events) 2019.

An ameliorate version of the framework with an extensive overview of the inner mechanism and a guided use case demonstrating its working mechanism in parameter tuning has been presented as

> Maria Angela Pellegrino, Abdulrahman Altabba, Martina Garofalo, Petar Ristoski, Michael Cochez. *"GEval: A Modular and Extensible Evaluation Framework for Graph Embedding Techniques."* In the Proceedings of Extended Semantic Web Conference (ESWC) 2020.

1.3 DOCUMENT STRUCTURE

This dissertation is split into five main parts organised as follows:

PART I starts with this introductory chapter, provides the reader with all the background and the terminology to get in the topics discussed in this thesis, and introduces the contributions fully presented in the following parts. Each part concerns a specific pillar and contains a chapter for each contribution. Further detail follows.

PART II presents the *Open Data quality and privacy assessment and improvement* pillar. Each chapter is dedicated to a different contribution. In particular, Chapter 3 presents the role-based orchestration approach to guarantee a proactive data quality assurance mechanism. Chapter 4 discusses the unified approach to assess quality issues and privacy concerns contextually based on an automatic type inference approach. Chapter 5 reports the semi-automatic approach based on Machine Learning to automatically detect and correct quality issues in textual geographical data. Chapter 6 presents the anonymisation approach to preserve individual privacy while publishing Open Data. Finally, Chapter 7 discusses the flagship project to collaboratively create Open Data supported by `Campania Crea`.

PART III reports the *Knowledge Graph exploitation* pillar where each chapter is dedicated to a specialised sub-pillar. In particular, Chapter 8 overviews the unified approach to let lay users query Knowledge Graphs without explicitly dealing with syntactical challenges posed by SPARQL and being aware of the underlying data structure. Chapter 9 reports the transitional approach to let experts in data table manipulation explore Knowledge Graphs and exploit retrieved results in reusable data visualisations supported by `QueDI`. Chapter 10 presents the `Novelette` project and the effort in exploiting Knowledge Graphs in storytelling by supporting pupils in automatically navigating Knowledge Graphs in overcoming the blank page syndrome. Chapter 11 reports the proposal to let Cultural Heritage lovers play the museum curator role by authoring virtual exhibitions starting from data stored in Knowledge Graphs. Chapter 12 explores the effort in bringing the Semantic Web technologies closed to the Virtual Assistant world letting data curators or people interested in (Cultural Heritage) data automatically create Virtual Assistant extensions to query Knowledge Graphs by vocal commands.

PART IV discusses the *Knowledge Graphs and Machine Learning* pillar by detailing in Chapter 13 `GEval` and its mechanism to fairly evaluate and compare graph embedding techniques on Machine Learning and semantic tasks. This contribution is designed in the direction to support end-users with technical skills in performing task-oriented evaluation and comparison of graph embedding techniques. Hence, `GEval` can be exploited and extended to estimate the effectiveness of candidate graph embedding techniques in data analysis, such as quality improvement and privacy preserving tasks.

PART V concludes the dissertation by remarking key contributions, achieved outcomes, and the potential impact of the research at the basis of this thesis. Finally, it reports future directions.

2 BACKGROUND

2.1 TERMINOLOGY AND DEFINITIONS

The W3C Foundation has created a basic model for Open Data (OD) with regard to quality: the **5-Star OD model**.

★ The data are available on the Web, whatever format, under an open license.

★★ The data are structured and machine-readable, (e.g. Excel instead of image scan of a table).

★★★ The data do not use a proprietary format (e.g. CSV instead of Excel).

★★★★ The data use only open standards from W3C (RDF, SPARQL) to identify things, so that people can point at your stuff.

★★★★★ The data are linked to that of other data providers.

THE 1-STAR STAGE: PUBLISHING DATA. This stage can be achieved by publishing data in various ways, via download, bulk download, or APIs.

THE 2-STAR STAGE: MAKING THEM AVAILABLE AS STRUCTURED DATA. The power of OD lies in its re-usability and stimulates interoperability of systems and services. Data formats can be clustered into structured and unstructured data. Structured data are developed to be processed by machines and is thus different from digitally accessible information. They are machine-readable and more interoperable.

THE 3-STAR STAGE: USING NON-PROPRIETARY FORMATS. Non-proprietary implies not being bound to specific software or a specific vendor. For instance, an Excel file (.xls) might seem very open, but it is not. It is bound to Microsoft Excel. This means that every one that is not in the possession of Microsoft Office is unable to open this file. It is widely promoted to convert proprietary and non-machine readable files into open and machine-readable formats to get a high-quality open dataset.

THE 4-STAR STAGE: USE URIS TO DENOTE THINGS. If data providers publish data as 4-Star OD, they require using URIs to denote things and create the first step towards Linked Data (LD). In practice, this means they should convert datasets to RDF format and enrich metadata with URIs. This is the first step towards LD.

THE 5-STAR STAGE: LINK DATA TO OTHER DATA TO PROVIDE CONTEXT. This is a very advanced stage of OD. In this stage, data are linked to other data to provide context. This will lead to very interoperable and easily discoverable data.

ADVANTAGES AND CHALLENGES FOR DATA PUBLISHERS AND CONSUMERS. As a data publisher, data up to three stars, corresponding to tabular OD, are rather simple to publish, while data consumers lead to data manipulation and basic data exploitation. By moving to the last position of the 5-star scale, data publishers are required to further invest time and effort in linking data and making them referable by URIs. However, data consumers are provided with easy manipulation and exploitation mechanisms, leading to link data from heterogeneous sources.

While this chapter overviews key concepts related to OD, LOD, KGs, Quality dimensions, and Privacy concerns, the following ones discuss advantages and challenges posed by OD and KGs.

2.1.1 *Open Data*

According to the Open Knowledge Foundation and its Open Definition, "*OD are data that can be freely used, re-used and redistributed by anyone*", subject at most to measures that preserve provenance and openness. The Open Definition gives precise details on its interpretation that can be summarised as follow:

- **availability and access**: data should be available as a whole and, if possible, for free. However, data publishers can ask no more than a reasonable reproduction cost. Data must be available in a modifiable and convenient form, preferably over the Internet;
- **re-use and redistribution**: data must be published under terms that permit re-use and redistribution including the possibility to be **interoperable** with other datasets;
- **universal participation**: everyone must be able to use, re-use, and redistribute without discrimination against fields of endeavour, groups, or people. For example, 'non-commercial' restrictions that would prevent commercial use, or restrictions of use for certain purposes, e.g. only in education, are not allowed.

There are two dimensions of data openness:

- data must be **legally** open, which means they must be placed in the public domain or under liberal terms of use with minimal restrictions;
- data must be **technically** open, which means they must be published in electronic formats that are machine-readable and non-proprietary so that anyone can access and use the data using common, freely available software tools. Data must also be publicly available and accessible on a public server, without password or firewall restrictions.

BENEFITS OF OPEN DATA. Making data open has economic and social benefits. For instance, when government data are made accessible and re-usable, they enable individuals, organizations, and even governments themselves to innovate and collaborate in new ways. OD builds connections between government, private, and research sectors, stimulating business activity, and developing knowledge that everyone can exploit. OD benefits include:

- **transparency**: OD supports public oversight of governments and helps reduce corruption by enabling greater transparency. For instance, OD makes it easier to monitor government activities, such as tracking public budget expenditures and impacts. It also encourages greater citizen participation in government affairs and supports democratic societies by providing information about voting procedures, locations, and ballot issues.

- **public service improvement**: OD gives citizens the raw materials they need to engage their governments and contribute to the improvement of public services. For instance, citizens can use OD to contribute to public planning or provide feedback to government ministries on service quality.

- **innovation and economic value**: OD provides new opportunities for governments to collaborate with citizens and evaluate public services by giving citizens access to data about those services. Businesses and entrepreneurs exploit OD to better understand potential markets and build new data-driven products.

- **efficiency**: OD make it easier and less costly for government ministries to discover and access their data or data from other ministries, which reduces acquisition costs, redundancy, and overhead. OD can also empower citizens with the ability to alert governments to gaps in public datasets and to provide more accurate information.

OD benefits may differ according to the type of stakeholder involvement. These stakeholders can be divided into 3 main groups: governmental organisations, citizens, and re-users.

Governments are one of the main re-users of the data they collect themselves. Publishing OD enables the sharing of information within governments in machine-readable interoperable formats, which results in reducing costs of information exchange and data integration, error reduction by having one copy instead of multiple ones. This results in improved data management, in terms of both quality and efficiency, as well as an overall reduction in administrative costs. Breaking down the silos that exist between the various departments, bodies, and layers of government and allowing a fluid data flow can have substantial efficiency gains.

Citizens are better informed and can actively participate and cooperate with the (local) government. Besides the creation of social value, OD opens up possibilities for entrepreneurs. OD creates value for both citizens and private businesses after the release of a specific application. Social value for the public sector can generate commercial value for the private sector. Data have an extraordinary commercial value. Since governments typically hold large amounts of information stored in all kinds of systems, opening up this data would lead to freeing up this potential.

OD stimulate re-users to create new innovative products and services, guarantee long-term commercial success, and stimulate economic growth. In addition, OD re-use can stimulate processes improvements, such as planning, quality, and digitalisation. For some businesses, this means an in-depth transformation of business models and therefore internal innovation can be achieved.

OPEN DATA STANDARDS. Open standards are reusable agreements that make it easier for people and organisations to publish, access, share and use better-quality data. OD standards can be broadly split in shared vocabularies and a common lexicon to agree on the adopted terminology, exchange data formats and shared rules, and guidance for creating high-quality data.

A **shared vocabulary** helps people and organisations communicate the concepts, people, places, events, or things that are important to meet their needs or solve their problems. A good shared vocabulary focuses on a specific area and uses clear, unambiguous definitions of the words and concepts it contains. Shared vocabularies range from simple lists of words and their meaning to more complex products. The complexity of a vocabulary depends on the complexity of the problem being solved. Typical vocabulary formats include registers, taxonomies, collections of defined words, and ontologies that describe concepts and relationships. A shared vocabulary can standardise

- **concepts** that represent important information,
- **words** used in the application context of interest,
- **attributes** that are properties of concepts,

- **relationships** between concepts,
- **identifiers**,
- **unit of measurement**,
- **models** that describe involved stakeholders and information flows.

Standards for exchanging data specify common formats and shared rules that lead to consistent data. A good standard for data exchange leads to the identification of tools to check that data has been properly structured. Typical data exchange standards define a common format for data that describes how data should be serialised or structured for sharing. In this context, data curators can standardize:

- **data format** that describe how data is structured for sharing or storage, for example, files and data formats like CSV, JSON, and XML. In particular, CSV is the plain text format for structuring data files using rows and columns.
- **datatypes** describing how attributes are expressed, for examples people's name is text, while their age is a whole number,
- **data transfers**, such as APIs, to find data and exchange information,
- **rules** that define what data should be shared, their schema, format, and the shared vocabularies to use.

Standards for guidance provides support to people and organisations in understanding and documenting information flows and data models. In this context, data curators can standardise:

- **unit and measures** to collect data,
- **processes** that describe protocols or methods for measuring, capturing or sharing data consistently,
- **codes of practice** that guarantee data consistency, for example, best practices and recommendations.

TECHNOLOGY OPTIONS. This section guides the selection and implementation of various technologies used to develop OD platforms. It is intended to support IT specialists playing the leader or coordinator role in managing the technical infrastructure of an OD initiative.

- A **data catalog** is a list of datasets available in an OD initiative. Essential elements of a data catalog include searching, metadata, clear license information, and access to the datasets themselves. Typically, a data catalog is the online centerpiece of an OD initiative.
- A **platform** provides an online front door for users to access all resources available under an OD initiative. A platform includes the data catalog along with other information and services that are part of the OD ecosystem. These typically include an online forum for questions, technical support, and feedback; a knowledge base of background and training materials; and a blog for communications and outreach. The services within a platform are often implemented with a suite of technologies, not a single one.

OD catalogs should fulfil the following requirements:

- **easy access**. OD catalogs make it very easy for users to access data quickly, freely, and intuitively. Access to OD catalogues requires no registration or login since such requirements would discourage exploration and use.
- **search**. OD catalogues make data easy to find. Most data catalogues sort data by subject, organization or type, and support full-text searching of catalogue contents. Many OD catalogues implement search engine optimization to expose data to conventional search engines.
- **machine-readable data access**. Data are available for download in machine-readable, non-proprietary electronic formats. To the extent possible, the preference is to have all data in a dataset available as a single download file.
- **metadata**. Key metadata, such as publication date and attribution, are prominently displayed for each dataset.

- **clear data license**. Data licenses are clearly and prominently displayed for each dataset.

- **data preview and visualisation**. Many OD catalogues include some facility to preview data before downloading or visualising data using built-in graphical tools.

- **standards compliance.** Most OD catalogues have built-in support for various standards, such as data formats (e.g., CSV, XML, JSON) and metadata (i.e., Dublin Core). OD catalogues typically make each dataset available as a unique and permanent URL, which makes it possible to cite and link to the data directly.

- Application Programming Interface (**API**). APIs allow software developers to access the OD catalogue – and often the data itself – through software. APIs facilitate data discovery, analysis, catalogue integration, harvesting metadata from external sites, and a host of applications.

- **security.** OD catalogues implement security measures to protect data and metadata from being changed by unauthorized users.

Commonly used OD catalogues are:

- **CKAN** is an open-source data catalogue formally supported by the Open Knowledge Foundation. CKAN is designed for publishing and managing data either through a user interface or an API.

- **DKAN** is designed to be "feature compatible" with CKAN. This means that its underlying API is identical, so systems designed to be compatible with CKAN's API should work equally well with DKAN. DKAN is also open-source, but it is based on Drupal, a popular content management system written in PHP instead of Python.

- **OpenDataSoft** is a cloud-based SaaS platform that offers a comprehensive suite of OD and visualization tools. The front end is fully open source. The platform supports common OD formats such as CSV, JSON, and XML, along with geospatial formats such as KML, OSM, and SHP. Search functionality is easy to use and the platform is available in multiple languages.

- **Semantic MediaWiki** is an extension of MediaWiki – the wiki application best known for powering Wikipedia. While traditional wikis contain only text, Semantic MediaWiki adds semantic annotations that allow a wiki to function as a collaborative database and data catalog. Semantic MediaWiki is an RDF implementation, meaning that both data and metadata are stored as LD and are accessible via LD interfaces such as SPARQL.
- **Socrata** is a cloud-based SaaS OD catalog that provides API to access data and data manipulation tools. One distinguishing feature of Socrata is that it allows users to create views and visualisations of published data, and save them for future usage. Additionally, Socrata offers an open-source version of its API, intended to facilitate transitions for customers that decide to migrate away from the SaaS model.

EXAMPLES AND USE OF OPEN DATA. There are no constraints on data curators or application contexts. OD initiatives can be organized at a country-level, such as at the national level, and below according to the city and subnational initiatives. Individual agencies or sectors may have their data with a specific thematic focus. Other sources may contain specific kinds of data, such as statistical indicators, geospatial data, or microdata, such as business and household surveys. Data repositories contain transport, geospatial, healthcare, environmental, demographic, and real-time emergency data.

OD can be exploited by the user interface to simplify data access, can enable research, support evidence-based decision making, develop a business plan for creating goods or services, or simply improve knowledge and understanding of social, economic, and environmental trends.

OD can help make governments more transparent. For instance, they can demonstrate that public money is being well spent and policies are being implemented. As an example, according to leading open government activist David Eaves, OD allowed citizens in Canada to save the government $3.2bn in fraudulent charitable donations in 2010.

OD is opening up new opportunities for business. As an example, English transport agencies have released OD about transports that developers have used to build over 800 apps.

OD can support decision-making and data-driven research. For instance, by analysing data about weather, researchers might detect an early warning system for environmental disasters. As an alternative scenario, OD might also support decisions, such as helping consumers to understand their impacts on the environment and letting them take steps to improve it.

2.1.2 *Linked Open Data and Knowledge Graphs*

The concept of Linked Data (LD) increases the interoperability and discoverability of datasets. LD is not the same as OD. Whereas OD concerns the openness of the data itself, LD is a way of publishing OD as LD or enriching datasets with metadata. Tim Berners-Lee defined LD as "a set of design principles for sharing machine-readable data on the web to be used by public administrations, business and citizens". LD are pieces of information that are linked through a graph connection. Opposed to other relational descriptions of data, in LD, a machine can walk through the graph and understand the content. This is seen as a revolution in the area of data storage and sharing: a computer can, to some extent, qualitatively interpret data. This is possible because the data is enriched with uniform descriptors. Using these descriptors, the data is no longer a set of static content, but is described and can therefore be interpreted, regardless of any distinguishing factor such as language or file type.

SEMANTIC WEB TECHNOLOGIES. The Semantic Web term refers to W3C's vision of the Web of LD. Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data. LD are empowered by technologies such as RDF, SPARQL, OWL, and SKOS.

Which is the name rule or the convention to identify data?

To avoid ambiguity and name clashes, names of resources and properties must conform to the format for Web resource names, that is Uniform Resource Identifiers (or **URIs**). The **RDF** Framework (Resource Description Framework) is the basic principle of LD. It is the new general syntax for representing data on the web. Every piece of information expressed in RDF is represented as a **triple**:

- **subject** - a resource identified with a URI,
- **predicate** - a relationship identified with a URI,
- **object** - a resource or a literal related to the subject.

Each object has a unique identifier, i.e, a URI, and is usually assigned to ontological classes, such as 'city' or 'person', and an arbitrary number of properties that define links between the objects. One of the advantages of using URIs is that they can be dereferenced using the HTTP protocol. According to the so-called Linked Open Data (LOD) principles, such a dereferenced URI should result in a document that offers further data about the given URI.

How can I create a knowledge base?

The knowledge base is a triple collection. **XML** provides an elemental syntax for content structure within documents, yet associates no semantics with the meaning of the content contained within. XML Schema is a language for providing and restricting the structure and content of elements contained within XML documents.

XML is not at present a necessary component of Semantic Web technologies in most cases, as alternative syntaxes exist, such as **Turtle**. Usually, this is the used notation, in which statements are formed simply by listing the elements of the triple on a line, in the order subject-predicate-object, followed by a full stop, with URIs possibly shortened through the use of namespace abbreviations defined by “prefix" and “base" statements. The drawback of the Turtle notation is that it is a de facto standard, but has not been through a formal standardisation process.

How can I describe information?

RDF is a simple language for expressing data models, which refer to objects (“web resources") and their relationships. An RDF-based model can be represented in a variety of syntaxes, e.g., RDF/XML, N3, Turtle, and RDFa. RDF is a fundamental standard of the Semantic Web.

What has a name in RDF?

RDF decomposes the information into small chunks with the help of simple rules about the meanings of these chunks. The objective is to provide a simple and flexible way to express facts, but, at the same time, they should be so well structured that computers can process the expressed knowledge. In this framework, formal names can be assigned to what is called **resources** which would include Titanic, its director, all the actors of its cast. Names can also be assigned to types (or **classes**) of resources (such as, movies, movie directors, actors) and to **properties** that link resources (e.g., the “directed-by" relationship between movies and the corresponding movie director). By reasoning over facts encoded in this way, a query system can confirm that an actor plays a role in a movie directed by an English director and can retrieve the answer to users’ requests.

More formally, all the information in RDF is articulated with the help of a triple pattern also known as a statement. A triple or a statement consists of a subject, a predicate, and an object in the form of Subject-Predicate-Object, and this form never changes. The subject and object are names of two things and the predicate is the relationship between them.

How can I model information?

On the Semantic Web, **vocabularies** define the concepts and relationships (also referred to as terms) used to describe and represent an area of concern. Vocabularies are used to classify the terms that can be used in a particular application, characterise possible relationships, and define possible constraints on using those terms.

There is no clear division between what is referred to as vocabularies and ontologies. The trend is to use the word

ontology for a more complex, and possibly quite formal collection of terms, whereas vocabulary is used when such strict formalism is not necessarily used or only in a very loose sense. Vocabularies are the basic building blocks for inference techniques on the Semantic Web. Data curators can use **OWL** or RDFS to build vocabularies or ontologies, and SKOS for designing knowledge organization systems.

RDFS stands for RDF Schema and it is a language used to create a vocabulary, which is most of the time domain-specific. It is an extension of RDF which allows resources to be classified explicitly as classes or properties; also supports some further statements that depend on this classification, such as class-subclass relationships (resources *rdfs:Class* and *rdfs:subClassOf*), and domain and range of a property (*rdfs:range* and *rdfs:domain*). RDFS also contains some predicates for linking a resource to information useful in presentation or navigation, but not for inference, such as *rdfs:label*.

An ontology is a common term used to describe a domain. It contains concepts and relationships among concepts. Thanks to ontology technology, data curators can encode general facts about classes, for example, they can write *Every book has a writer*. Data curators can store tens and tens of book titles, for each of them the query application knows that there is at least a writer, exactly as a person could expect. To allow automatic inference, ontologies may be encoded in some version of mathematical logic. There are many formal logics, which vary in expressiveness (the meanings that can be expressed) and tractability (the speed with which inferences can be drawn). To be useful in practical applications it is necessary to trade expressiveness for tractability. Description logic, which is implemented in the Web Ontology Language **OWL**, does precisely this. We can conclude that OWL consists of RDFS and new constructs to deal with complex situations. OWL adds more vocabulary for describing properties and classes: among others, relations between classes (e.g. disjointness), cardinality (e.g. “exactly one"), equality, richer typing of properties, characteristics of properties (e.g. symmetry), and enumerated classes.

How can I query the data?

The first process towards achieving re-usability is data access. LOD is usually accessible on data portals or catalogues through a **SPARQL endpoint**, that is a conformant SPARQL protocol service that enables users (human or other) to query a knowledge base via the SPARQL language. Results are typically returned in one or more machine-processable formats. A SPARQL endpoint is mostly conceived as a machine-friendly interface towards a knowledge base. Both the formulation of the queries and the human-readable presentation of the results should typically be implemented by dedicated software, not manually.

SPARQL is a language for formulating queries over RDF data. It is the Semantic Web counterpart to SQL. Exactly as SQL retrieves information from data organized into tables, SPARQL retrieves information from sets of triples. SPARQL makes it possible to send queries and receive results, e.g., through HTTP or SOAP. SPARQL queries are based on (triple) patterns. RDF can be seen as a set of relationships among resources (i.e., RDF triples); SPARQL queries provide one or more patterns against such relationships. These triple patterns are similar to RDF triples, except that one or more of the constituent resource references are variables. A SPARQL engine would return the resources for all triples that match these patterns. SPARQL consumers can extract possibly complex information (i.e., existing resource references and their relationships) which are returned, for example, in a table format. This table can be incorporated into another Web page. Hence, SPARQL provides a powerful tool to build, for example, complex mash-up sites or search engines that include data stemming from the Semantic Web.

The definition of the Knowledge Graph (KG) term remains contentious [34, 98], where several (sometimes conflicting) definitions have emerged, varying from specific technical proposals to more inclusive general ones. This dissertation adopts the widely referred definition according to a KG is *a semantic knowledge base (or, put alternatively, a knowledge base organized as a graph) where relationships between facts are formally described by an ontology. The graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent potentially different relations between these entities*.

While LD technologies are generally considered fundamental to building a KG, LD is not in itself a KG. Put another way, LD is a necessary but not sufficient condition for KG construction.

2.1.3 *Data quality dimensions*

Data have been referred to as *the new oil* because while both data and oil have intrinsic value, they both must be **refined** or otherwise transformed to realise their full potential.

Quality is often defined as *fit-for-use* meaning that data should be published by making data consumers free to reuse data in their application contexts without wasting time in cleaning data.

Eurostat defined six quality dimensions originally applied to statistical data, but can also be extended to other types of data:

- **relevance** in terms of degree to which data meet current and potential users' needs. It corresponds to the *fit-for-use* data quality definition.

- **accuracy and reliability**, i.e., the degree to which data are free of errors,

- **timeliness and punctuality**,

- **accessibility and clarity** that is the ease with which users can access the data and the degree to which they are explained through metadata;

- **comparability** that is the degree to which data can be compared across time, regions or other domains,

- **coherence** with definitions and data production methodologies.

Usefulness can be determined by data quality. The quality of OD, next to its discoverability, is one of the largest influencers of the success of OD. Some of the widest popular quality dimensions follow:

COMPLETENESS refers to the comprehensiveness or wholeness of the data. There should be no gaps or missing information for data to be truly complete.

CLEANNESS concerns lack of empty fields, management of default values, and if they correctly model the truth, the presence of wrong values or duplicates. Moreover, it also concerns privacy-sensitive information.

ACCURACY mainly concerns a direct usage in potential purposes, the specification of interval and data values reported, if data aggregation or disaggregation is required.

TIMELINESS Data changes over time. Historical data will remain stable, but recent data will be updated over time. Therefore, it is important to check data concerning its timeliness regularly. For consistency purposes, it is wise to create an update process that keeps the data up-to-date. Be sure that the data contains a notion of its timeliness. This topic is closely related to the maintenance of datasets.

CONSISTENCY Supposing re-users desire to merge data from various sources, but all datasets differ in accuracy, use of terms, and timeframe. As an example, if data users change the field names of the data collected for managing waste each year, the data cannot be compiled from one year to the next. This makes it difficult to use datasets. It will require a large effort of manipulation. Therefore, data curators should make sure to use standards and be consistent in publishing datasets of equal quality.

2.1.4 *Privacy concerns*

While data curators are invited in publishing OD, data in their raw and original form could contain personal and sensitive information about individuals. When opening up data, the focus is on non-personal data, i.e., data which do not contain information about specific individuals.

To prevent privacy leakages, data curators mainly opt for close data. This is a missed opportunity. According to the principles of OD, they should be *open by default*. However, the data publication process requires performing a Privacy Preserving Data Publishing (PPDP) process to spread useful data without violating individuals' privacy. According to the PPDP principles, data publishers have tables containing *Identifiers, Quasi-Identifiers, Sensitive Attributes, Non-Sensitive Attributes* where

- Identifier (ID) is a set of attributes that identifies record owners;
- Quasi-identifier (QID) is a set of attributes that could potentially identify record owners;
- Sensitive attributes are person-specific information, such as, diseases, salary, religion, political party [102];
- Non-sensitive attributes correspond to all the remaining attributes.

Among the PPDP approaches, **anonymization** is the process of obscuring or removing information from a dataset that could be used to identify individuals, households or businesses so that their anonymity is preserved and protected. Anonymization and the imperative to protect confidentiality are especially important for governments releasing public data. Equally important is the need for organizations to clearly articulate their privacy policies concerning data management, both to individuals that provide data and individuals that use that data. That said, many, many types of government data do not entail confidential information, and thus have little or no need for anonymization techniques.

2.2 BACKGROUND LITERATURE

This section introduces an overview of background literature, focusing on key aspects and examples of prototypes proposed in data quality assessment and improvement, privacy issues detection and anonymization, and LOD exploitation. This section guides inexperienced readers in become familiar with the topics covered in this dissertation and the related state-of-the-art, which will be further explained in detail in the corresponding chapters.

DATA QUALITY ASSESSMENT AND IMPROVEMENT Data quality assessment identifies errors contained in available data and estimates their impact on data-driven processes [203]. Researches and companies heavily invested in supporting the data quality assurance process, as can be observed by contributions proposed in the literature by IBM [10] or in the Messytable project [189].

Detected data quality problems concern structure, format, or value inconsistencies, data completeness, spelling errors, free-text fields [10], and metadata completeness [92, 189].

The proposed approaches include data profiling techniques to collect statistics and insights about data [229], as in the Messytables project [189] or in the project presented by Döhmen et al. [92]; Machine Learning (ML)-based approaches to detect duplicates [304], patterns and data correlations [144, 166, 315], assess data completeness [104, 146, 226, 352] or data accuracy [298, 353], outlier and anomaly detection [82, 100, 164, 184, 185, 231].

PRIVACY BREACHES IDENTIFICATION AND ANONYMIZATION APPROACHES An *anonymization* approach prevent linking attacks by applying generalisation, suppression, anatomisation, permutation, and perturbation [105]. The most famous privacy models include k-Anonymity [57] and its multiple versions such as (X- Y)-anonymity [336], MultiRelational k-Anonymity [232], l-diversity [191], (α, k)-anonymity [343], t-closeness [182].

Due to the power of modern re-identification algorithms [227], removing personally identifying information does not guarantee that the remaining data protect individuals [228, 312]. Hence, it is crucial to perform the privacy breaches identification step on any set of columns, both inspecting isolated fields and unstructured textual description [127].

A typical content-based data leakage prevention system identifies sensitive data by using regular expressions, data fingerprinting, and statistical analysis, as can be observed in DgSECUREE [78]. In the last years, there is an increasing interest in using ML and Artificial Intelligence (AI)-based approaches [198, 316] to reduce privacy leakage effectively and efficiently.

LINKED OPEN DATA EXPLOITATION Several tools and interfaces have been designed to support users in interacting with KGs by implicitly composing queries while hiding the underlying complexity. Users are provided with graph-like query builders (such as FedViz [350], RDF Explorer [328]), visual query builders (e.g., OptiqueVQS [307]), facets based interfaces (e.g., SemFacet [13]) also enhanced by keyword search interfaces (such as SPARKLIS [107] and Tabulator [29]), query completion mechanisms (such as YASGUI [270]), summarization approaches (such as Sgvizler [300]), or a combination of them.

Then, collected data can be used to generate data visualization, as in SPARKLIS [107] or YASGUI [270]. Alternatively, they can be used to obtain more advanced data exploitation, such as virtual exhibitions [130] to engage visitors in virtual guides, VA-based Question-answering (QA), natively offered in Google Assistant and Alexa and explored by researchers and businesses [11, 72, 129, 171, 187, 192, 262].

2.3 CHALLENGES IN DATA MANAGEMENT

Even if this thesis mainly focus on technical and technological challenges, it is worth clarifying that OD diffusion and exploitation require overcoming organisational, economical, cultural, political, and legal challenges. In the following, we will focus on technical and technological aspects.

MULTIPLE STAKEHOLDER COORDINATION The production and consumption of OD require the coordination of multiple and heterogeneous stakeholders, different for background, skills, interests, and goals. The government, public sector, private organisations, citizens, developers, and researchers can play the stakeholder role. The OD management process should also deal with groups of an arbitrary number, asking for coordinating large groups of stakeholders.

LOW DATA QUALITY Low data quality is a recognised as a critical barrier in the OD consumption due to the use of abbreviations, inconsistent representation, misspellings, mixed styles.

COSTS IN DATA QUALITY IMPROVEMENT Improve data quality requires time, human effort, and data literacy. It may require introduce a (semi-)automatic data cleansing phase based on a proactive or reactive approach to support data producers in proactively improving OD quality before publishing them or data consumers in reactively improving OD quality before using them.

SENSITIVE INFORMATION MANAGEMENT Data in their raw and original form could contain personal and sensitive information about individuals. Publishing such data violate individual privacy [102]. Hence, it is crucial to care about data quality preservation approaches and avoid individual privacy leakage during the OD publication process. It requires data providers performing PPDP [53] to make useful data available without violating individuals' privacy.

CHALLENGES IN KNOWLEDGE GRAPH QUERYING The KG exploitation is mainly affected by i) required technical skills in query languages (e.g., SPARQL) and in understanding the semantics of the supported operators [329], too challenging for lay users, and ii) conceptualization issues to understand how data are modelled [24, 329]. Moreover, KG querying is threatened by data heterogeneity, heterogeneity in access point (mainly SPARQL endpoint and APIs), instability of the SPARQL endpoint status, different interfaces for humans and machines, low level of interoperability.

PROLIFERATION OF DATA EXPLOITATION EXPECTATIONS Any stakeholder have specific interests, goals and common data exploitation means. Children should be scaffolded by simplified and user-friendly interfaces to implicitly exploit data in learning, performing researches, getting inspiration for storytelling. Journalists are mainly interested in processing data to obtain infographics, data visualization, maps that can be used in articles and data stories. Researches might be interested in performing data analysis on (linked) open data. In the case they are interested in performing ML tasks on LOD, they should be guided in evaluating and comparing graph embedding techniques that convert KG nodes (and edges) into numerical vector to make KGs compatible with ML model requirements. Hence, it is crucial to provide communities with target-oriented exploitation means, taking into account their skills and objectives.

Part II

OPEN DATA QUALITY AND PRIVACY ASSESSMENT AND IMPROVEMENT

3 OPEN DATA QUALITY ASSURANCE BY ORCHESTRATION

Alone we can do so little; together we can do so much.
– Helen Keller

In the last years, e-government manifests great interest in OD as they satisfy the Open Government principles [154, 202]. In fact, OD have the potential to guarantee transparency, drive innovation [51, 154], and empowering citizens [154].

Simply providing OD does not automatically result in value for society: the potential benefits of OD will not be realised unless data are actually used [154, 323]. Even if many datasets are available, often repositories contain OD that users do not need and, on the contrary, datasets that citizens need are not available or not published by Public Administrations (PAs) [356].

Involving citizens in the OD publication process results in the possibility to rely on heterogeneous skills and the opportunity to reply to real needs. On the opposite, it requires coordinating many stakeholders, trading off interests and goals, and guaranteeing data quality preventing duplication and mixed styles.

The research question that guided this proposal is:

> *how to support PAs and citizens in publishing high-quality OD without posing any limit on the group size?*

This question replies to real needs emerged by our Regional Public Administration (RPA), the Campania region council, during the collection of structured data about the "*Land of fires*", a phenomenon that takes place in the Campania region, and reported in this chapter as motivating use case. The requirement to collect information by involving heterogeneous stakeholders motivates the research at the basis on this chapter. While orchestration let several heterogeneous stakeholders working together, a proactive data quality assurance approach is introduced to guarantee data free from syntactical and semantic errors.

The research presented in this chapter has been published in the following contribution and it is the result of strict cooperation between the Campania region council and researchers in technical and technological solutions:

> Giuseppe Ferretti, Delfina Malandrino, Maria Angela Pellegrino, Andrea Petta, Gianluigi Renzi, Vittorio Scarano, Luigi Serra: *Orchestrated Co-creation of High-Quality Open Data Within Large Groups*. In the Proceedings of Electronic Government (EGOV) 2019.

3.1 MOTIVATING USE CASE

The need to distinguish among several roles and the identification of the proposed profiles are motivated by a concrete use case claimed by a PA, the Council of the Campania Region in Italy. Our RPA has established a Special Regional Committee since 2015, named *Land of Fires*, which takes care of precise monitoring of the uncontrolled phenomenon of the occurrence of garbage scattered over a vast territory, 90 municipalities between the province of Naples and Caserta. This monitoring is performed by involving heterogeneous and qualified stakeholders dedicated to the collection of both structured and non-structured data. Data concern the structural characteristics of the territory and its municipal resources dedicated to the problem of the rubbish fires. Moreover, data are used to assess the usefulness of the legislative disposal that qualifies and encourages the municipalities in this zone, also with money. First, this commission must verify the effectiveness of the application of the Regional Law (n. 20/2013) by direct co-operation with the involved municipalities. Moreover, the information and communications technology department of the RPA has been involved to streamline the process via automated tools. Several experiences were matured with questionnaires reported by using European tools (EUsurvey[1]). From the analysis of recovered data, the legislator will obtain valid tools to identify and implement more precise and timely intervention rules for the elimination of this dangerous phenomenon.

[1] https://ec.europa.eu/eusurvey/home/welcome

In this context, our RPA is interested in investing in a direct and effective process of collecting and sharing data as OD by actively involving citizens. It is crucial to deal with a wide set of contributors and avoid low-quality data. Usually, data production tools only allow a simple collect-and-send data process. It is insufficient to support a complex mechanism of data collection, joint analysis and discussion, and publication as OD. For instance, previous experiences of required data collection were further elaborated through several successive meetings with other competent bodies and institutions in the field, with national government authorities, a list of chosen delegates from the major municipalities involved, and a series of public hearings held at our RPA site, where the results were presented and discussed.

The Special Regional Committee is evaluating the possibility to extend the data collection activities to a much wider audience of municipalities, involving all the towns of the Campania region. According to the plan to cover more than 500 towns, traditional collect-and-send data collection approaches are no longer sustainable. It requires the design of a supportive environment that guides heterogeneous stakeholders in collecting, assembling, evaluating data for publication. Furthermore, automated tools for quality checking are needed. Because of the large number of participants, the orchestration is more suitable than peer working to separate roles and tasks. There should be a supervision role to define data constraints and a semantic guarantee by a manual inspection of the dataset content. There is also the necessity to involve a big number of stakeholders who play the role of filler.

3.2 AN ORCHESTRATION-BASED APPROACH

Small groups of $7-8$ persons can exploit peer-to-peer methodologies without losing the overall picture of the rest of the group. When the group size increases and there is a consistent diversity of partners and contributors, an orchestrator might ensure valuable inputs and mitigate concerns raised by involved actors [85].

Orchestration is a well-known strategy to deal with large groups [117]. It ensures the creation and extraction of value, without the introduction of hard hierarchical authority [117]. Roles are clearly defined by specifying tasks and responsibilities, distributed in an agile way by taking into account individual competencies and overall needs. Agile approaches and motivated stakeholders are key factors of a successful co-creation process [206]. However, it is crucial to guarantee the introduction of data quality preservation mechanisms as it is easy to duplicate data or leave them incomplete when several people are involved in the data publication process.

We propose to scaffold PAs and citizens in working together by orchestrating the OD co-creation process through roles definition. Roles keep responsibilities and tasks divided and they should be distributed according to stakeholders' competencies and global needs to make data fit-for-use. By collaborating with the Campania region administration and according to their use case, the proposed orchestration-based approach is based on the following roles and tasks:

CREATOR This role corresponds to an expert in the field who can opportunely model the dataset under the definition. It is in charge of defining the structure of the dataset, specifying columns data types and any useful constraint. The dataset structure is also referred to as *form* as all the requirements and constraints will be modelled as a form in the dataset population phase. The creator can start from an empty or a partially filled dataset, at least exposing a column header. During the form creation, the creator is guided in specifying for each column or a subset of them any desired constraint. The form can specify the column datatype among basic data types, such as text, number, date, or more specific datatypes, such as geo-coordinates, images, and documents, drop-down lists (also referred to as select options). A select option can either be manually populated by the creator or correspond to built-in lists, such as the list of all Italian regions, provinces, or municipalities. Based on the datatype, the form will guide the creator in specifying extra parameters, if necessary. For example, by opting for a numeric value, the creator can also bind minimum and maximum values. The form can also specify constraints on values to automatically validate inputs inserted by the filler. For instance, by selecting email as datatype, the form will prevent the insertion of syntactically wrong emails. The creator can also specify extra information, such as placeholders or tooltips, labels, or descriptions, ask for mandatory fields or define a default value, which will help fillers in interpreting more easily which information should be inserted into the dataset and in which format.

FILLER This role is in charge of filling in the dataset under definition by accessing the form defined by the creator. The form will prevent syntactically wrong inputs and trivial errors which could compromise the overall dataset quality. The filler role is further specialised in *advanced* and *plain* filler. This role distinction is only on the visibility of the whole dataset under construction. While the advanced filler has the privilege of having an overall vision of the whole dataset, the plain one cannot inspect rows authored by others. This distinction is due to security requirements: based on the situation there could be the need to involve a huge number of fillers. Since also unreliable people might be involved by accident, there is the need to avoid the suggestion of rows that can deliberately change the overall statistics of the dataset. Therefore, the proposed mechanism offers the opportunity to give access to the dataset in reading mode only to reliable people by the advanced filler role, and preventing access to others by the plain filler role.

VALIDATOR By correctly and deeply defining the form it is possible to minimise syntactical errors in the dataset filling. It does not prevent semantic errors. For example, by restricting the data type of a column to dates, fillers will not be able to specify incorrect dates but there is no validation about its correctness. For also guaranteeing a semantic check, the validator role is required. This role represents a super parties inspector who is in charge of checking the dataset content and discarding all rows conceivably semantically incorrect. The validator has legal responsibility for the published dataset. Therefore, this role has the power to decide which rows should be included in the final version of the dataset and it takes the responsibility for all the information which are included and (also!) of those discarded.

Data suggested by fillers are not automatically added to the final dataset but are left in a grey zone until the validator explicitly accepts them. The validator has to manually inspect proposed data to filter out the wrong ones. Approved data will be moved from the grey zone to the actual dataset under construction. By the validation step, also semantic errors are reduced, by ensuring high-quality OD.

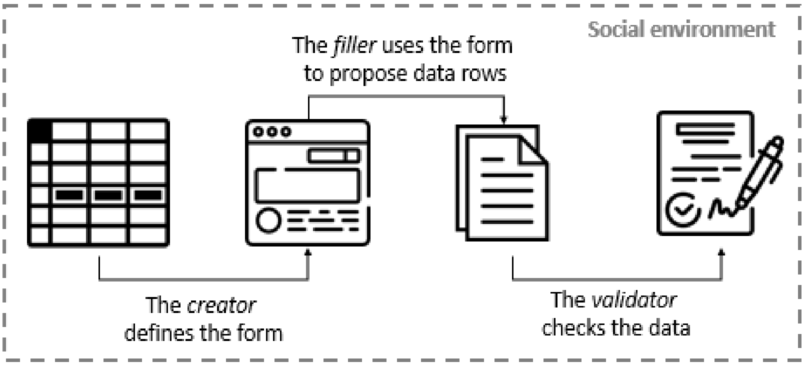

Social environment
The *filler* uses the form to propose data rows
The *creator* defines the form
The *validator* checks the data

Figure 3.1: Tabular data quality preservation approach in SPOD

Figure 3.1 summarises the entire workflow: the creator role defines the form starting from an empty or already partially filled dataset; the filler proposes new candidate rows by filling in the form; the validator inspects the candidates' rows and takes responsibility for the data effectively added to the dataset under the definition. Users are not forced to follow these steps in this particular order. The creator can create the form at any time. The filler can start producing rows also before the form definition. The filler is not locked by the validator verification step. Therefore, users can choose the best operative approach according to their needs.

3.3 TOOLKIT: DATA QUALITY PRESERVATION IN spod

The proactive data quality assurance approach described in the previous section has been introduced in SPOD, literally Social Platform for Open Data, resulting from an H2020 project, ROUTE-TO-PA. SPOD is freely accessible at http://spod.routetopa.eu/

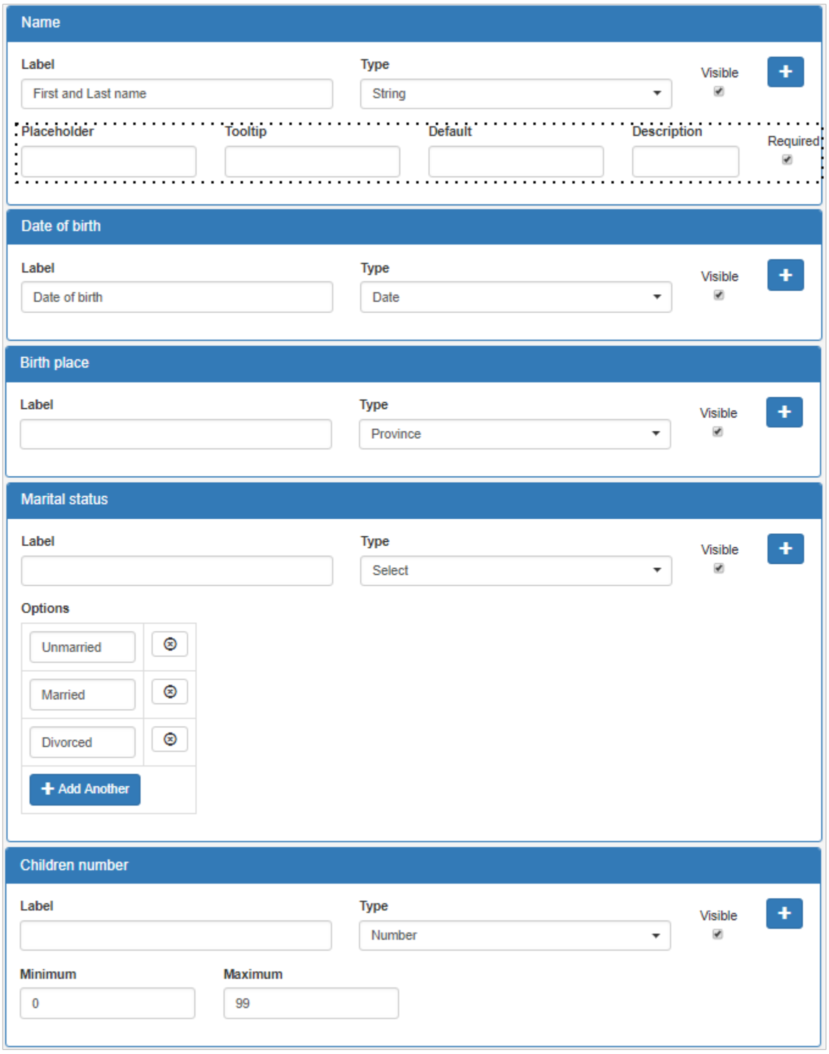
and it is released as an open-source project on GitHub at https://github.com/routetopa/spod. As a social platform, SPOD supports users in discussing and exchanging information and opinions by news-feed, groups, chats similarly to a traditional social network. However, the term choice stresses the interest in keeping conversations and sharing focused on OD. SPOD supports data-driven discussions where citizens can discuss any topic of interest by exploiting OD as evidence. Thanks to its full interoperability with existing OD portals, users can i) access open datasets, ii) create reusable visualizations, and iii) share use them as evidence in discussions in a seamless way.

In the direction of producing OD, SPOD supports co-creation of datasets by creating dedicated rooms accessible by any interested stakeholder, such as PAs and citizens, that is interested in working with other partners and creating shared datasets.

Thanks to the orchestration roles, users who play the creator role can attach a form to the dataset under definition and guarantee syntactically correct data. Fig. 3.2 represents an example of form from the creator side. When the creator opens the form template, a box is created for each column. In the example reported in Fig. 3.2, the dataset represents citizen profiles where name, birthplace and birth date, marital status, and children number are specified as column headers. The creator decides to model names as strings. By clicking on the plus icon on the right, the section delimited by a dotted line is opened. These options will provide extra information to fillers during the dataset populating step. Dates of birth are modelled as dates while the birthplaces are modelled as provinces. The latter represents an example of auto-filled select: the creator has to simply decide the type of the column as a province, while the filler will have access to all the available provinces. The marital status column is a select filled by the creator who specified the list of valid options. Children number is modelled as a number and we can specify the minimum and maximum value.

3.4 FINAL REMARKS

SPOD has been proposed as a social platform to support citizens, PAs and any interested stakeholders in co-creating OD and ex-



Name

Label	Type	Visible
First and Last name	String	☑

Placeholder	Tooltip	Default	Description	Required
				☑

Date of birth

Label	Type	Visible
Date of birth	Date	☑

Birth place

Label	Type	Visible
	Province	☑

Marital status

Label	Type	Visible
	Select	☑

Options

- Unmarried
- Married
- Divorced

Children number

Label	Type	Visible
	Number	☑

Minimum	Maximum
0	99

Figure 3.2: Tabular data quality preservation layout in SPOD

ploit them in practice, in debates, to disseminate information supported by data-driven evidence, to visually detect patterns, and perform data analysis. According to the motivating use case presented by the Council of the Campania Region, the co-creation phase might require coordinating a huge number of users, different in competencies, requirements, ambitions. Orchestration represents a promising approach to clearly define and assign roles and tasks. Roles should be defined to guarantee data quality assurance dealing with the proposal of a uniform and consistent data modelling, data format, and preventing syntactical and semantic data errors. The identified roles deal with the `creator` of a syntactic check attached to co-created datasets, the `filler` role to encourage large participation of citizens and interested stakeholders, dealing with security requirements, and the `validator` role that prevents semantic errors in the published dataset. The proposed approach is implemented in the `SPOD` platform and it has been used by the Special Regional Committee of the Campania Region to collect and publish data concerning the Land of Fires phenomenon.

The research described in this chapter has been conducted in strict cooperation with our RPA officials and their information and communications technology department by discussing, designing, and implementing an orchestration-based approach to co-ordinate large groups of data curators while guaranteeing data quality.

4 OPEN DATA QUALITY AND PRIVACY ASSESSMENT BY DECISION TREES

> *We should treat personal electronic data with the same care and respect as weapons-grade plutonium – it is dangerous, long-lasting and once it has leaked there is no getting it back.*
>
> – Cory Doctorow

Data owners are spur in opening up their data to enable informed decision making, ensure transparency, audience engagement, and release social and commercial value [242]. Unfortunately, data in their raw and original form could contain personal and sensitive information about individuals. Publishing such data violate individual privacy [102]. Hence, it is crucial to care about data quality preservation approaches and avoid individual privacy leakage during the OD publication process. It requires data providers performing PPDP [53] to make useful data available without violating individuals' privacy.

While a *proactive* data quality improvement program identifies data quality issues in an early stage of the dataset design, a *reactive* approach responds to data quality issues either during the data population phase or, as an extreme case, after data publication, before data exploitation. In the context of the data publishing workflow, proposing a proactive data quality and privacy assurance approach before data are materialized, at the definition step. On the contrary, a reactive approach can intervene when data are (partially) collected.

This chapter proposes a reactive approach to recognise quality problems and potential privacy risks exposed by a dataset. The proposed approach assesses quality and privacy issues at once in an early stage of the dataset population. As an assessment approach, it only reports the risk of a quality or privacy issue, without proposing any corrective action.

The research presented in this chapter has been published in the following contribution and it is the result of strict cooperation between the Campania region council and researchers in technical and technological solutions:

> Giuseppe Ferretti, Delfina Malandrino, Maria Angela Pellegrino, Donato Pirozzi, Gianluigi Renzi, Vittorio Scarano: *A Non-prescriptive Environment to Scaffold High Quality and Privacy-aware Production of OD with AI*. In the Proceedings of Digital Government (DG.O) 2019.

4.1 RELATED WORK

This section reports solutions proposed by academies and companies to deal with data quality and privacy breaches concerns, by analysing well-known and commonly adopted solutions and the role plaid by AI.

4.1.1 *Data quality assessment*

Data quality assessment identifies errors in available data and estimates their impact on data-driven processes [203]. Errors can be identified by data profiling techniques to collect statistics and insights about data [229]. Data profiling can work on single or multiple columns. Single column approaches include the identification of i) the amount of null values column, ii) column data type(s) and iii) the occurrence of additional patterns [229]. The research discussed in this chapter exploits traditional data type inference methods on single columns to discover syntactic heterogeneity [229] and assess completeness.

Researches and companies struggle to support the data quality assurance process. The IBM Knowledge Centre developed *IBM WebSphere Information Analyser* [10] to help users in easily identifying data quality problems concerning structure, format, or value inconsistencies, data completeness, spelling errors, and free-text fields. The IBM approach relies on a type inference approach by identifying a datatype that matches all the values in each column. The approach proposed in this chapter works similarly, but, rather than generalising datatypes, it reports values mismatches in the same column as a quality concern.

Type inference is usually performed by employing the cast operator, such as in the Messytables project [189]. However, there are values for which the casting fails but using regular expressions datatypes can be correctly identified. For instance, a numerical string with the currency symbol, very common in OD, can not be cast, but it can be recognised easily by a proper regular expression. Regular expressions in the type inference step are gaining an increasing interest in the literature, as in the project presented by Döhmen et al. [92] that iteratively parse all the column values by identifying a unique representative datatype. If the process fails, columns are attached to the string value. Differently, the approach presented in this chapter does not force all values to be compliant with a single data type but attaches to each column the most common datatype inferred by its values and reports all the mismatches as errors.

4.1.2 *Privacy breaches identification*

Privacy leakages may occur both in isolated fields and unstructured textual description, as observed by Green et al. [127] that look for privacy breaches in IDs, addresses, and unstructured text, such as comments and descriptions.

Due to the power of modern re-identification algorithms [227], removing some personally identifying information - such as social security numbers, names, surnames - does not guarantee that the remaining data does not identify individuals. Indeed, while some bits of information may not be uniquely identifying individuals on their own, in other cases individuals can be identified by combining several and different attributes [228, 312]. These attributes are referred to as QIDs or pseudo-identifiers [102]. Hence, it is crucial to perform the privacy breaches identification step on any set of columns.

A typical content-based data leakage prevention system identifies sensitive data by using regular expressions, data fingerprinting, and statistical analysis. Regular expressions are normally used under a certain rule such as detecting social security numbers and credit card numbers. For instance, Dataguise, a leader in data privacy protection and compliance, proposes DgSECUREE [78] identify any privacy breach that occurs in enterprise cloud repositories through a sophisticated regular expression pattern builder. It combines structured, semi-structured, and unstructured content and it finds sensitive data - such as credit card numbers, social security numbers, names, email addresses, medical IDs, bank account numbers, and financial codes.

Data leakage prevention systems using regular expressions mainly suffer the risk to offer limited data protection and having high false-positive rates [294] if you are looking for a specific word or name instead of a pattern. In contrast, there are a series of studies [198, 316] which focus on finding more efficient methods to reduce leakage with a growing interest in the usage of machine learning and AI. In this direction, the approach discussed in this chapter deal with evaluating the effectiveness of exploiting AI approaches on OD.

4.2 A DECISION TREE-BASED APPROACH

The proposed methodology assumes that the dataset is in tabular format which is one of the formats widely available on the OD portals around Europe. Typical tabular formats are CSV and TSV files and with no semantics attached to the table nor any additional information at the schema level. For instance, it does not specify columns datatypes, something that is typical in the context of relational databases. This lack of details is the main reason for the poor quality of many datasets found over the governmental portals.

The proposed approach performs both quality and privacy checks based on a type inference step able to automatically infer datatypes by dataset values. As it combines quality and privacy checks, it would be referred to as a *qualicy* approach by trivially combining the two objective names. Further details of the proposed procedure follow:

TYPE INFERENCE - The process infers the datatype for each column. This phase attaches a datatype to each value and, transitively, to each column. Besides *basic datatypes* recognising dates, numbers, and strings, it is also infer the semantic content of each column. Semantic datatypes, referred to as *meta datatypes*, model surnames, names, addresses, SSNs, telephone numbers. By default, each value is categorised a *string*. Then, the type inference module tries to refine the datatype by choosing the most suitable basic datatype (and a meta datatype, if applicable).

QUALITY CHECK - Based on the inferred data types, the dataset *accuracy* and *completeness* are assessed. If a column contains more than one datatype, the quality check underlines that it lacks accuracy, interpreted as the presence of heterogeneous datatypes. If a column contains empty values, it is classified as incomplete. For each value not classified by a meta datatype, the quality check module verifies if a typo occurs. In particular, it verifies if by applying a single transformation to the current word it can easily be traced back to a recognised meta datatype.

CONTENT PRIVACY BREACHES - This phase verifies if a privacy breach occurred in an unstructured text. For instance, a text can accidentally contain an SSN, a telephone number, or other sensitive information. These leakages must be prevented. Thus, for all the string values not attached to a meta datatype, the process verifies if a content privacy breach occurred, i.e., if the text contains structured personal data. For example, if an SSN or an IBAN occur in a description, it will be not recognised by the meta datatype inference step since the personal information is not isolated. In this phase, the process looks for structured information within the text and reports it as a content privacy breach. It represents the presence of personal data in a textual content.

STRUCTURAL PRIVACY BREACHES - At this stage, each column is labelled by a basic and/or a meta datatype. By considering only the meta datatypes, this phase checks if there are columns labelled by meta datatype which corresponds to ID or QID, i.e., if a single column or a combination of them exposes personal information. Emails, SSN, IBANs, mobile phone numbers are considered single-column personal data, while any combination of user identifiers, such as SSN or name and surname, and information about users' location, birthplace, or birthdate, religious preferences are multi-column personal data.

The order in which the steps are presented reflects the pipeline effectively run.

4.3 TOOLKIT: A DECISION TREE-BASED IMPLEMENTATION

Technically speaking, the proposed type inference approach is based on the pattern matching technique, enhanced by AI to guarantee efficiency and accuracy.

The described process has been included into SPOD and, specifically, in the co-creation process. Once a data curator has created a co-creation room and imported a dataset, any user can perform for the *qualicy* check. This request triggers the guided process described so far composed of type inference, quality check, content and structural privacy breaches to collect an immediate feedback of quality and privacy problems of the dataset under definition.

4.3.1 *Type inference phase*

Basic datatypes are inferred by running a set of syntactical heuristics based on regular expressions to determine column types.

The implementation of the proposed approach extended the agile methodology described by De Donato et al. [80] which defines a quality-aware OD publication process by introducing the a contextual quality and privacy assessment approach. The technique considers each column in the dataset independently. The objective is to probabilistically infer the type of each column. In this case, basic datatypes are introduced. Each datatype corresponds to a set of regular expressions. The type inference approach tests each value against these regular expressions and infers the value datatype. Then, it derives the column datatype picking the one corresponding to the majority of its value datatypes. The order in which the regular expressions are matched is crucial as data types are organised in the hierarchy shown in Fig. 4.1. A depth-first tree visit generates an array of types that implies the order in which regular expressions must be evaluated. Hence, the more specific type is evaluated before the more general type. This means that the default type - text type - is the last one and corresponds to the default datatype attached to a value.

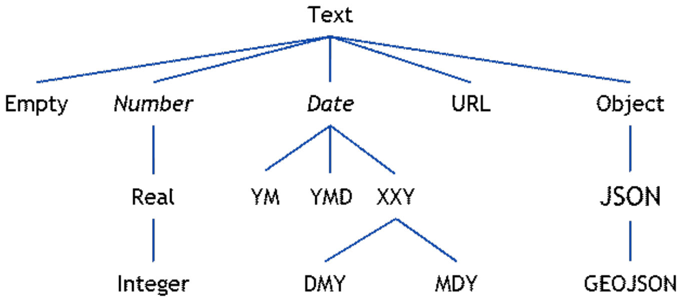


Figure 4.1: Basic datatypes recognised by the decision tree-based approach.

The meta datatype inference step annotates each column of the dataset with additional information concerning its semantic content. For instance, the meta datatype infers if a column models surnames, names, addresses, phone numbers. This information is instrumental to improve the dataset correctness and to prevent the publication of datasets containing personal.

It is based on a combination of regular expressions and dictionaries. To improve the efficiency of the type inference phase, we proposed a decision tree-based approach to ignore all the pattern matching rules not applicable to the current value simply looking at the value format. For instance, if a value contains digits and letters, it is useless testing patters to infer datatypes concerning numbers. A decision tree learning approach uses a decision tree as a predictive model to go from observations about an item represented in the branches to conclusions about the item target value represented in the leaves. It is one of the predictive modelling approaches used in statistics, data mining, and machine learning. Tree models where the target variable takes as input a discrete set of values are called classification trees where leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Instead of testing all the meta datatypes in sequence, the proposed approach uses a combination of pattern matching and decision trees.

The meta datatype inference phase relies on the output of the basic datatype inference step by refining numbers, dates, and text. The meta datatypes that can be inferred by the text basic datatype are detailed as follows:

- *text without letters*: it contains only numbers and special characters, such as %, comma, $, and so on;
- *text without numbers*: it contains only letters (and perhaps special characters);
- *alphanumeric text*: it contains both numbers and letters (and perhaps special characters).

By both considering the basic datatype and the occurrence of specific characters, the proposed approach detects all candidate meta datatypes that must be tested, as represented by the decision tree visible in Fig. 4.2: leaves identify a group of candidate meta datatypes and the conjunctions filter out all the not applicable pattern rules. Each branch called *Run tests on...* hides another decision tree.

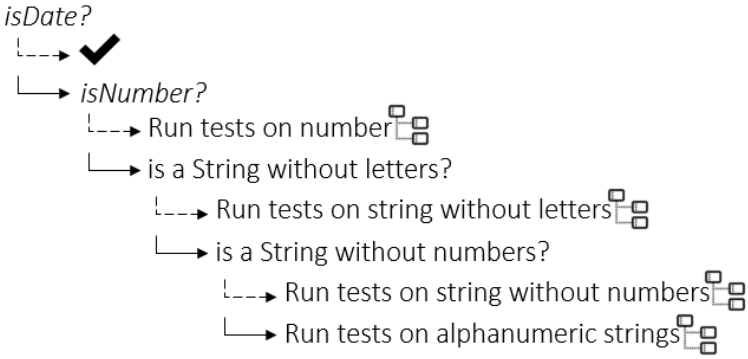

isDate?
✓
isNumber?
Run tests on number
is a String without letters?
Run tests on string without letters
is a String without numbers?
Run tests on string without numbers
Run tests on alphanumeric strings

Figure 4.2: Decision-tree based data quality and privacy assessment approach

The following reports the sub-tree related to strings without letters. However, considerations can be generalised to reconstruct also remaining subtrees. All the values classified as text which contains only numbers and special characters belong to this category. Fig. 4.3 reports the schema of all the tests executed on texts without letters. To read the decision trees, the reader has to keep in mind that they contain

- *test labels*: the labels in italic are the tests done to filter the meta datatypes to verify; in uppercase are reported all the actual tests to verify if the value matches a specific meta datatype;

- *plug icons* representing extension points and modelling all the patterns that must be customised according to the dataset language;
- *lines* modelling questions replies. Dashed lines represent positive answers to the posed question, while solid lines represent negative answers.
- *icons*: the check icon represents a recognised meta datatype, while the cross icon represents the possibility to provide an evocative message to explain to the user the leaf reached in the parsed decision tree.

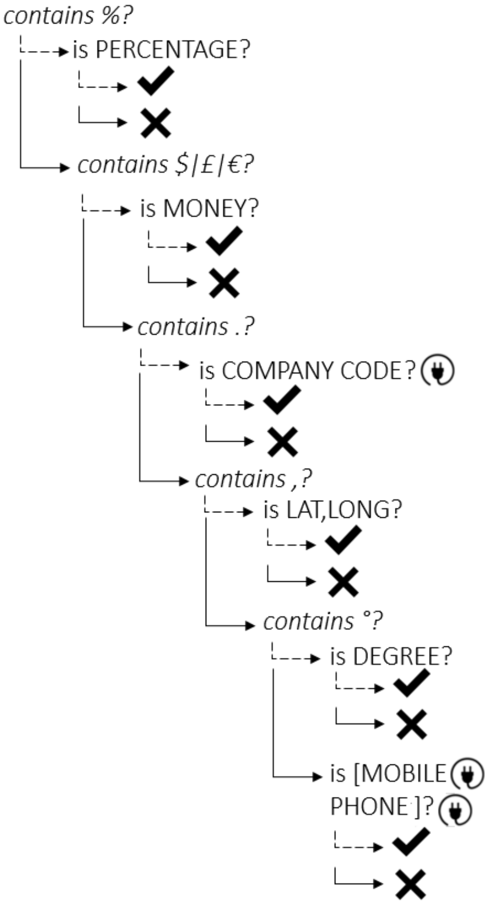


Figure 4.3: Decision-tree to infer datatypes of words containing numbers and special characters

The tested meta datatypes in this branch are:

PERCENTAGE , i.e., a number followed by the special character %;

MONEY , i.e., a number followed by a special character representing a currency, such as $, €, £;

COMPANY CODE , i.e., a sequence of numbers split by dots;

LAT,LONG , i.e., a combination of latitude and longitude comma separated;

DEGREE , i.e., a number followed by the special character °;

PHONE NUMBER , i.e., a sequence of digits, preceded by a symbol to introduce the national prefix (for instance, +). The symbol can not be omitted, otherwise the phone number would have been classified as a sequence of digits and not as a string without letters.

As a result of the type inference phase, each column will be attached to a basic datatype and - if possible - to a meta datatype.

4.3.2 *Quality check*

Each column is parsed to compute the accuracy and the completeness scores. The accuracy is interpreted as datatype homogeneity. Consequently, it is compromised when a column contains heterogeneous datatypes. The accuracy score reports the percentage of values attached to the inferred column meta datatype. The completeness score reports the percentage of null values in a column.

Concerning the typo verification, the module verifies which values have no meta datatype attached to them. For all of them, it computes all the possible combinations of words obtained by inserting, removing or substituting a letter. Then, the module verifies if at least one of these word alterations correspond to a valid meta datatype.

4.3.3 *Privacy breach detection*

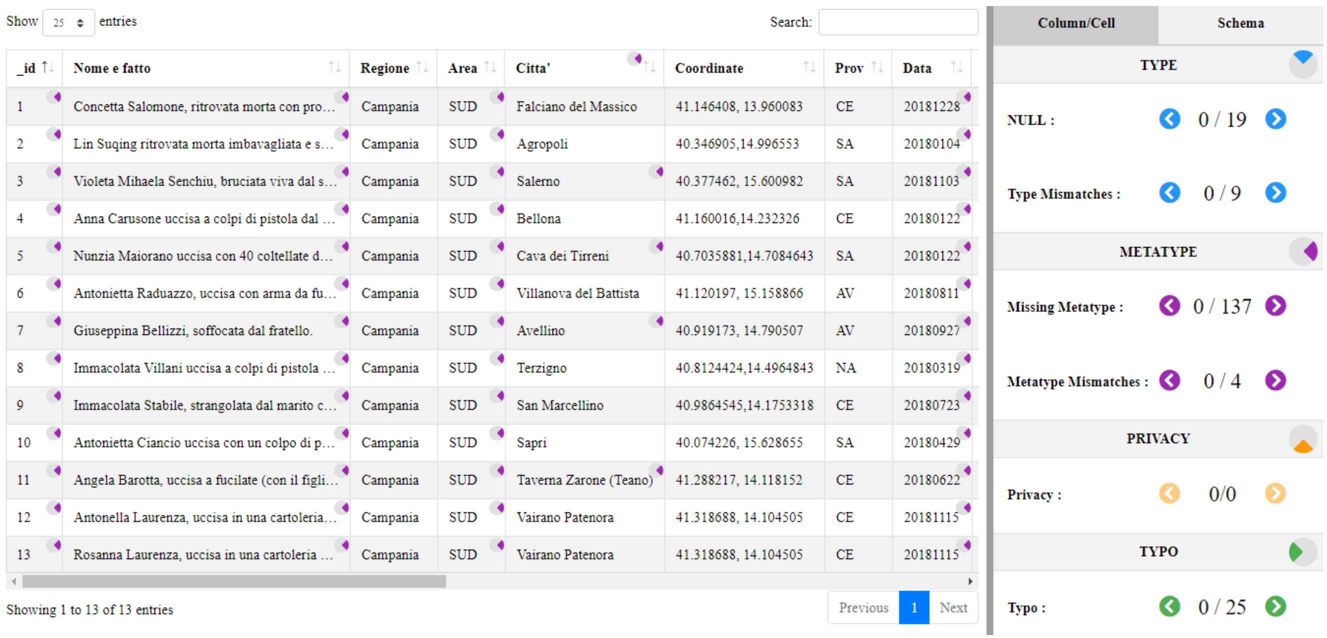
This module verifies if a textual description contains any structured sensitive information, such as SSNs or telephone numbers. Thus, all the values recognised as strings without any refined meta datatype are parsed by regular expressions to detect the presence of privacy leakage. It is referred to as a *content privacy breach* at it is related to a specific value.

Finally, this modules verifies if any combination of columns correspond to well-known ID or QID. It is referred to as *structural privacy breach* and it is related to the entire dataset.

4.4 USE CASE

Our methodology and implementation have been employed in a real use case. In 2018, Dr. Rosaria Bruno, the President of the Observatory on the Phenomenon of Violence against Women, a body of the Campania Regional Council (CRC), requested support for monitoring data on the femicide phenomenon. The CRC used SPOD to exploit the potential of a guided workflow to create datasets and their graphical representations. The dataset content has been extrapolated by Dr. Gianluigi Renzi from the media and press through a continuous process of journalistic sources monitoring. The dataset contains the information about victims, such as their nationality, age, and photos; details about their families, such as the number of children; details about the femicide, such as the date and the place; data about the perpetrator, e.g., the relationship between the murderous and the victim.

Figure 4.4 reports the result of the qualicy assessment on the dataset about femicides. On the left, the dataset is visible. An overview of the detected problems is reported on the right of the interface. Each cell affected by one or more defeats is decorated by the corresponding colour. TYPE and METATYPE group the statistics about the basic and meta data types. The PRIVACY box is related to content privacy breaches. The structural privacy breaches are visible at the schema level. By clicking on each cell, its types, and the typo correction candidate - if any - are reported. By clicking on a column, its stats (e.g., the completeness and accuracy scores) are reported.



Show 25 entries

Search:

_id	Nome e fatto	Regione	Area	Citta'	Coordinate	Prov	Data
1	Concetta Salomone, ritrovata morta con pro…	Campania	SUD	Falciano del Massico	41.146408, 13.960083	CE	20181228
2	Lin Suqing ritrovata morta imbavagliata e s…	Campania	SUD	Agropoli	40.346905,14.996553	SA	20180104
3	Violeta Mihaela Senchiu, bruciata viva dal s…	Campania	SUD	Salerno	40.377462, 15.600982	SA	20181103
4	Anna Carusone uccisa a colpi di pistola dal …	Campania	SUD	Bellona	41.160016,14.232326	CE	20180122
5	Nunzia Maiorano uccisa con 40 coltellate d…	Campania	SUD	Cava dei Tirreni	40.7035881,14.7084643	SA	20180122
6	Antonietta Raduazzo, uccisa con arma da fu…	Campania	SUD	Villanova del Battista	41.120197, 15.158866	AV	20180811
7	Giuseppina Bellizzi, soffocata dal fratello.	Campania	SUD	Avellino	40.919173, 14.790507	AV	20180927
8	Immacolata Villani uccisa a colpi di pistola …	Campania	SUD	Terzigno	40.8124424,14.4964843	NA	20180319
9	Immacolata Stabile, strangolata dal marito c…	Campania	SUD	San Marcellino	40.9864545,14.1753318	CE	20180723
10	Antonietta Ciancio uccisa con un colpo di p…	Campania	SUD	Sapri	40.074226, 15.628655	SA	20180429
11	Angela Barotta, uccisa a fucilate (con il figli…	Campania	SUD	Taverna Zarone (Teano)	41.288217, 14.118152	CE	20180622
12	Antonella Laurenza, uccisa in una cartoleria…	Campania	SUD	Vairano Patenora	41.318688, 14.104505	CE	20181115
13	Rosanna Laurenza, uccisa in una cartoleria …	Campania	SUD	Vairano Patenora	41.318688, 14.104505	CE	20181115

Showing 1 to 13 of 13 entries

TYPE

NULL : 0 / 19

Type Mismatches : 0 / 9

METATYPE

Missing Metatype : 0 / 137

Metatype Mismatches : 0 / 4

PRIVACY

Privacy : 0/0

TYPO

Typo : 0 / 25

Figure 4.4: Decision tree-based quality assessment interface in SPOD.

4.5 Evaluation: Correctness Assessment

This evaluation focuses on the assessment of the type inference step as it represents the core of the entire process. It evaluates the correctness of the inferred datatypes to verify if and to what extent AI guarantees correctness if compared with a traditional approach.

4.5.1 *Evaluation design*

The following paragraphs detail the Research Question (RQ) based on the performed evaluation, the used datasets, the performed protocol, and the collected metrics.

methodology. The RQ that guides this evaluation is: *"Which is the correctness achieved by the proposed decision tree-based approach?"*. *Correctness* is interpreted in terms of correctly recognised meta datatypes during the type inference step.

datasets. The evaluation is performed on three Italian datasets already published as OD. All of them are related to public services. Hence, they do not compromise individual privacy. Details are schematically reported in Table 4.1. The paediatricians and the libraries datasets used in the evaluation are subsets of the original datasets available in the Campania region data portal.

Table 4.1: Datasets details

Dataset	Columns	Rows	Cells	Link
Associations	14	57	798	Link
Paediatricians	8	150	1,200	Link
Libraries	21	99	2,079	Link

protocol. Each dataset has been manually inspected to identify the number of expected values for each meta datatype. The non-automatic checking justifies the reduced size of the datasets taken into account. Then, each dataset is parsed by the proposed decision tree-based approach and the number of inferred meta datatypes is returned as output. By comparing the actual and the expected numbers of meta datatype, the evaluation reports the metrics discussed in the data-gathering section.

DATA GATHERING. For each dataset and each meta datatype contained in the dataset, the evaluation reports

- the number of correctly recognised instances by the type inference module, reported as True Positive (TP);
- the number of wrongly recognised instances by the type inference module, reported as False Positive (FP);
- the number of not-recognised instances by the type inference module, reported as False Negative (FN);
- the number of correctly not-recognised instances by the type inference module, reported as True Negative (TN).

Moreover, the evaluation reports the Accuracy (A), Precision (P), and Recall (R).

4.5.2 *Results*

Table 4.2 contains results concerning the national association dataset. All the provinces and municipalities have been rightly recognised since the initialisation of the vocabularies for these two meta datatypes has been exhaustive - all the national provinces and the municipalities have been taken into account. Only one surname has been wrongly recognised as Municipality (the false positive) since one person has the surname exactly as a municipality. All the addresses, ZIP CODEs, mobile phone numbers, emails, and URLs have been rightly recognised. Even if the dataset contains a column dedicated to phone numbers, all of them are wrongly parsed as numbers and, for this reason, they are not recognised. Surnames and names vocabularies are initialised considering only the 200 most common Italian names (100 female and 100 male) and the 20 most common surnames for each region. Since the initialisation is not exhaustive, it is easy to forecast a big number of FN. There are also some FP since there are people which have as surname common names and

Table 4.2: Type inference results on the national association dataset

	TP	FP	FN	TN
Province	70			728
Municipality	40	1		757
Address	56			742
ZIP CODE	57			741
Mobile phone	45			753
Email	146			652
URL	29			769
Name	37	2	20	739
Surname	6	1	51	730
Region		1		788

vice versa. The system recognises also a region since one person has as a surname a word that refers to a region, e.g., *Piemonte*.

Table 4.3 contains results concerning the subset of the paediatricians dataset. Only a few addresses - CORSO ITALIA N251 and VIA 24 MAGGIO 3 - are not recognised as wrongly formatted. All ZIP CODEs have been rightly recognised. All the provinces have been rightly recognised since the initialisation of the vocabulary to recognised the provinces has been exhaustive. Only a few municipalities are not recognised. About surnames and names, it could be repeated as in the previous test the same consideration about the not exhaustive initialisation. Despite this consideration, all the values recognised as a surname are correctly classified; instead few values recognised as names are wrongly classified. However, the wrongly classified values are usually used as Names instead of as Surnames. These ambiguities increase the number of wrongly recognised values.

Table 4.4 contains results concerning the subset of the libraries dataset. All the addresses, emails, and URLs have been rightly recognised. Also, all the provinces, municipalities, and regions have been rightly recognised. Even if the dataset contains a column dedicated to phone numbers, all but one are wrongly parsed and, for this reason, they are not recognised. The same can be repeated for FAX and mobile phone number columns. There

Table 4.3: Type inference results on the paediatricians dataset

	TP	FP	FN	TN
Address	80		2	1,120
ZIP CODE	128			1,072
Province	105			1,091
Municipality	87		4	1,105
Surname	22		128	1,050
Name	79	3	121	997

are also longitudes and latitudes in the dataset, but the format is not correct and, for this reason, they are not recognised. All the ZIP CODEs are correctly recognised but the system wrongly classifies them as ZIP CODE also numeric codes, since both of them comply with the same pattern. The system recognises one surname since there is one municipality - for instance Sala - which is usually a surname.

Table 4.4: Type inference results on the libraries dataset

	TP	FP	FN	TN
Address	94			1,985
Email	41			2,038
URL	31			2,048
Municipality	89			1,990
Province	109			1,970
Region	99			1980
Phone number	1			2,078
ZIP CODE	99	99		1,881
Surname		1		2,078

Table 4.5 reports the summary statistics related to the described datatests. The proposed approach recognises the considered meta datatypes with excellent accuracy in all the tests. The lower values of the recall are caused by the surnames and names. Name and surname recognition can be improved i) by considering a

bigger number of names and surnames in the initialisation phase, ii) by implementing a strategy to learn by its errors, such as by updating the vocabularies storing all the unrecognised names, iii) working on composite names that are names which combine common names.

Table 4.5: Overall type inference results

Dataset	TP	FP	FN	TN	A	P	R
Associations	486	5	71	7,399	0.99	0.99	0.87
Paediatricians	501	3	255	6,435	0.96	0.99	0.66
Libraries	563	100	0	18,048	0.99	0.85	1.00

4.6 evaluation: performance assessment

This evaluation focuses on the assessment of the type inference step as it represents the core of the entire process. It evaluates the performance of the approach to verify if and to what extent the use of AI gains better performance if compared with a traditional approach.

4.6.1 *Evaluation design*

The following paragraphs detail the RQ at the basis of the performed evaluation, the used datasets, the performed protocol, and the collected metrics .

methodology. The RQ that guides this evaluation is: *"Which are the performance of the proposed decision tree-based approach, where performance is interpreted as execution time?"*.

dataset. This evaluation tests a file of incremental size by repeatedly the rows of a national dataset published as OD until the reported size is achieved. The number of columns of the dataset (21 columns) has been left unchanged in all the tests.

protocol. This evaluation compares the execution time required by the proposed decision-tree-based approach which is referred to as experiment with AI and a traditional approach where all the pattern rules are tested in order, referred to as experiment without AI.

DATA GATHERING. This evaluation assesses the execution time required to complete the entire type inference phase.

4.6.2 *Results*

The average execution times are reported in Table 4.6. The test without AI and with a data set of 100,000 rows fires memory error. The evaluation underlines that i) the results are accurate; ii) the decision tree approach improves the efficiency without losing in correctness; iii) the implemented library is able to manage data set of considerable size in reasonable time.

Table 4.6: Execution times of decision tree based quality and privacy assessment approach

rows	Time with AI	Time without AI
1,000	382.3 ms	621.0 ms
10,000	2,231.9 ms	35,284.7 ms
100,000	15,077.9 ms	-

4.7 FINAL REMARKS

This chapter is grounded in the context of performing the data publication process while caring about data quality and preserving individual privacy. While data providers are encouraged to publish data, they are usually refrained by privacy concerns and they often opt for closing data. On the contrary, we propose a decision tree-based approach to assess the presence of quality and privacy issues. The proposed approach works on tabular datasets without any semantic attached to values and columns. Hence, it first performs an automatic type inference approach to categorise values by their syntactic and semantic content and, then, verifies the presence of quality and privacy issues. To the best of our knowledge, they are rarely managed together, representing the main novelty of our approach.

The proposed approach is based on AI as it exploits decision trees to speed up the type inference phase. The performed evaluation shows that the decision tree-based approach is more accurate, fast, and scalable if compared with traditional approaches based on brutal force.

The presented approach does not assume the dataset content format and size. Hence, it can be run on any CSV to verify quality issues and privacy breaches.

As with any data assessment approach, it only reports the risk of a quality or privacy issue, without proposing any corrective action. Data curators should verify the reported issues and perform correction, manually or supported by alternative tools. Hence, this approach represents only the first step to support data curators in performing a conscious data publication process caring about quality dimensions and avoiding privacy leakages.

5 OPEN DATA QUALITY ASSESSMENT AND IMPROVEMENT BY A CLUSTERING-BASED APPROACH

I personally think that humans and Artificial Intelligence need to handle together the global decision-making process.
– Zoltan Andrejkovics

Open datasets may contain highly inaccurate data where the accuracy is compromised by the use of abbreviations, inconsistent representations, misspellings. For instance, it can be observed in OD published by the Campania region.

Ad-hoc manual approaches are widely used in real applications, and they require heavy manual labour and human expert judgements. However, because of the enormous amount of available data, data consumers may require (semi-)automatic approaches to clean data by limiting human effort as much as possible. Hence, multiple attempts are being undertaken to develop automatic corrective proactive or reactive solutions. For instance, data providers could proactively attach constraints to tabular data to avoid syntactical errors [109], while ML techniques are becoming the core part of reactive data quality monitoring systems.

ML-based solutions bring several advantages as they may provide fast and reliable data quality assurance while, in the meanwhile, reducing the human resources requirements [82], discovering relevant patterns, and gain crucial knowledge from data [30]. On the other side, automation could affect the accuracy of the result. Anomalies caused by detector malfunctioning or sub-optimal data processing are difficult to enumerate apriori and rarely occur, making it difficult to use supervised classification [82]. Semi-automatic approaches can help users in detecting quality issues to speed up checks and assess quality dimensions, but there is the need to verify that the accuracy obtained by automated cleansing approaches is comparable with the one achieved by human effort [337].

This chapter proposes a (semi-)automatic approach to assess and improve data accuracy, by focusing on textual geographical data. The research presented in this chapter has been published in the following contribution:

> Maria Angela Pellegrino, Luca Postiglione, Vittorio Scarano: *Detecting Data Accuracy Issues in Textual Geographical Data by a Clustering-based Approach*. In the Proceedings of CODS-COMAD 2021.

5.1 RELATED WORK

ML and clustering-based approaches are gaining an increasing interest in duplication detection [304], in detecting patterns and data correlations [144, 166, 315], in assessing data completeness [104, 146, 226, 352], and it is widely used in outlier and anomaly detection [82, 100, 164, 184, 185, 231].

From a technical point of view, the most recurrent clustering approaches are k-Means in performing anomaly detection and completeness assessment [352]; the local outlier factor in outlier detection and accuracy assessment [347], k-nearest neighbours (k-NN) in the correction process [183]. Exploited character-based similarity metrics are Levensthein [178], Jaro [157] or Jaro-Winkler distance [341].

Several different solutions have been proposed by researchers and companies. Jingling Zhou et al. [353] propose a semi-automatic approach that requires human-in-the-loop for performing the accuracy assessment manually. Hence, they exploited a limited set of datasets to due the required human effort. The ML component of the proposed approach relies on Bayesian Networks that takes as input both manually annotated and original datasets. By computing, Euclidean and Jaccard distances between those datasets are calculated to determine the overall accuracy of each dataset. This approach requires heavy human intervention to start the process. By limiting the human effort, Sessions et al. [298] propose an automatic accuracy assessment algorithm based on probability theory without relying on any prior knowledge.

OpenRefine[1] is widely used in data profiling and quality improvement by journalists and data consumers [173]. It is based on a well-established approach in the context of textual facets: based on the Levenshtein similarity metric and the k-NN clustering approach, OpenRefine splits columns of the tested dataset into bags of words, a.k.a. clusters, and users have the opportunity to correct detected errors if any. In particular, users can merge clusters and are guided in making the content of a cluster uniform. In this way, users can easily correct typos and misspelling errors. OpenRefine can be considered a referring point during the evaluation phase due to its wide usage and the professionalism of the proposed solutions.

Usually, the accuracy assessment relies on an external dictionary of correct values [225]. For instance, the approach defined by Batini and Scannapieco [21] tests the occurrence of values to correct against the options listed in the dictionary. The same can be observed in the approach proposed by Robert Crone [71] whose model requires a source of correct data to compare with.

About the best way to provide the information related to the accuracy level of the tested dataset, the accuracy could be reported as a simple ratio or by reporting errors and proposing corrections [263]. The simple ratio reports the information related to the proportion of detected errors concerning the total number of sampled data.

5.2 APPROACH AND METHODOLOGY

This chapter proposes a clustering-based approach for assessing and improving the quality of textual geographical data corresponding to provinces and municipalities, at the instance level. While the assessment phase detects inaccurate values, such as typos, abbreviations, misspellings, the improvement phase proposes a correction for any detected error.

The clustering-based approach receives as input the collection of values and split them into clusters modelling the same geographical entity, such as the same province, by the Agglomerative Clustering (AC) approach. To compute the word similarity it re-

[1] http://openrefine.org/

lies both on the well-known Levenshtein similarity metric and the approximate string matching. Moreover, it also exploits a vocabulary of correct values in computing clusters and proposing corrections.

AC [274] is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. It works in a bottom-up manner. The algorithm starts by treating each item as a singleton cluster corresponding to a single-element cluster as a leaf. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects. At each step of the algorithm, the two most similar clusters are combined into a new bigger cluster, and it is represented as a new node posed as the father of the merged clusters. The result is a tree-based representation of the objects, named dendrogram.

To perform agglomerative hierarchical clustering, first, the dissimilarity between every pair of objects in the dataset must be computed; second, based on the similarity metrics, objects or clusters that are more similar are linked together by the linkage method; third, there is the need to determine the cut of the hierarchical tree to create the data partitions.

There are many methods to calculate the similarity information. Euclidean and Manhattan distances are widely used to work with numbers and digits. Since we aim to work with strings, we required string similarity metrics, such as Levenshtein.

Levenshtein [178] is a string metric for measuring the difference between two words as the minimum number of single-character edits (i.e., insertion, deletion, substitution) to change a word into the other. We also consider the two chars swap as an atomic edit.

Since datasets are usually manually fed by users' input, the proposed approach also requires a mechanism to measure the semantic similarity of words in the human language. In other words, it requires approaches to find strings that match a pattern approximately rather than exactly. These techniques, such as approximate string matching, can find matches even when users misspell words or enter only partial words, e.g., use abbreviations. The closeness of a word match is measured in terms of single-character edits necessary to convert the string into an exact match, like the Levenshtein metric. However, it can work not only with entire words but also on its sub-strings. It is exploited in search engines, such as Google or Expedia.

The proposed algorithm to detect and correct inaccurate values in collections of string focuses on columns of values that contain textual geographical information, such as provinces and municipalities. The input of the proposed approach is a single column containing strings related to geographical information. These values are compared with a set of valid options for that field, stored in a dictionary. The approach relies on a complete set of valid options to check if any syntactical error occurs. A complete list of provinces and municipalities is released by every national administration in common data formats. We exploit the list provided by the Italian administration as a CSV file. As output, this approach detects all inaccurate values in the tested list of words and proposes a correction by detecting the closest word in the dictionary.

5.3 TOOLKIT: A CLUSTERING-BASED IMPLEMENTATION

The clustering-based approach is based on the following steps:

SIMILARITY COMPUTATION computed by combining Levenshtein and Fuzzy matching similarity metrics to both consider similarities at the word and substring levels;

CLUSTERING phase performed by AC to split columns into clusters, which are iteratively refined to obtain coherent groups;

CORRECTIONS PROPOSAL where for each inaccurate value a correction is proposed according to the dictionary of correct values and the obtained clusters.

The following analyses each step separately. Its implementation is release as open-source on GitHub at this link.

5.3.1 *Similarity computation*

The similarity score between each pair of words in the tested column is computed once. In this step, the approach exploits the dictionary of valid options to calculate the similarity among words opportunely. If the two words are similar, but both valid, the similarity score will consider them different. If (at most) one value matches a valid option, the tool will calculate their *similarity score*. It is computed by combining the *Levenshtein distance* and the *Fuzzy string searching*. The *input* of this step is a column of words and the dictionary of valid options. The *output* is the similarity matrix storing the similarity for each pair of column values. The used algorithm is the following:

```
# distance computation
def wombo_combo(words, dictionary):
   HIGH_AVG_FUZZY = 95
   LOW_AVG_FUZZY = 85
   HIGH_LEV = 20
   LEV_TOLLERANCE = 1
   LOW_LEV = 5
   for w1 in words:
      for w2 in words:
         w1 = w1.lower()
         w2 = w2.lower()
         # pure Levensthein computation
         lev = levensthein_distance(w1, w2)
         # if both words are valid,
         # it forces high different values
         if dictionary.get(w1) is not None
            and dictionary.get(w2) is not None
            and w1 != w2:
               dist[w1][w2] = lev + HIGH_LEV
               # if they are very similar,
               # it forces very low distance
               if lev <= LEV_TOLLERANCE:
                  dist[w1][w2] = 0
                  # fuzzy matching computation
                  fuz1 = fw.ratio(w1, w2)
                  fuz2 = fw.partial_ratio(w1, w2)
                  fuz3 = fw.token_set_ratio(w1, w2)
                  fuz = (fuz1 + (fuz2*1.2) + fuz3)/3
                  # if they are very similar,
                  # it forces very low distance
                  if fuz >= HIGH_AVG_FUZZY:
                     dist[w1][w2] = 0
                  if fuz < LOW_AVG_FUZZY:
                     dist[w1][w2] = lev + LOW_LEV
```

As reported in comments, if the two words are both valid, the algorithm forces very different values. If they are similar according to the pure Levenshtein metric or the Fuzzy Matching score, they are forced to shallow distance values. Otherwise, the algorithm returns the Levenshtein similarity and guarantee that the minimum distance is not lower than a threshold (LOW_LEV).

Technically speaking, `fw` is the fuzzywuzzy[2] and Levensthein is computed by the pyxdameraulevenshtein[3] Python libraries.

5.3.2 *Clustering phase*

This phase relies on the AC by exploiting the pre-computed similarity scores. AC is a bottom-up hierarchical approach: starting from all singleton clusters, it tries to merge clusters by minimising the distance within the cluster and maximising the distance intra-clusters. The *input* of this step is the pre-computed similarity matrix. The *output* is a partition (decomposition without overlapping) of column values into clusters. The used algorithm is the following:

```
# word list in clusters
def clustering_phase():
   # range of clusters
   min, max = cluster_range(words, dict)
   current_num = int((min_cluster + max_cluster)/2)
   # AC run
   model, clusters = AC(matrix, current_num, words)
   # refinement step
   change_num = check_clusters(clusters, dicty)
   while change_cluster_number != 0 and
      current_cluster_num>=min and
      current_cluster_num<=max:
         if current_num==min and current_num<0:
            break
         if current_num==max and current_num>0:
            break
         current_num = current_num + change_num
         model, clusters=AC(matrix,current_num,words)
         change_num = check_clusters(clusters, dict)
```

[2] https://github.com/seatgeek/fuzzywuzzy

[3] https://pypi.org/project/pyxDamerauLevenshtein/

The minimum number of clusters is computed by counting the number of distinct correct values in the word collections against the dictionary, while the maximum number of clusters is the number of unique values in the word collection. At each step, the algorithm verifies if the clusters must be refined, i.e., if it has to collapse clustering by reducing the number of required clusters or by splitting clusters by increasing the number of clusters. The `check_clusters` function returns -1 if clusters must be collapsed, 1 if clusters must be split, 0 otherwise. The choice to either increase or decrease the number of clusters depends on the homogeneity of entities in the same cluster. In other words, if each cluster contains entities that represent the same geographical information, clusters should not be further refined. If all clusters contain at most one single value, which occurs in the vocabulary of correct values, and all the values model the same information, clusters satisfy the *well-formed* requirement. If clusters represent heterogeneous entities, clusters require subsequent refinement steps. If two valid options are in the same cluster, clusters must be split. If two entities representing the same geographical information are in two clusters, clusters must be merged. For example, the cluster formed by [New York, Nwe York, New Yor] is considered a well-formed cluster since only New York is a valid value, but all of the other ones are misspelled versions of the same value. On the other side, the cluster formed by [New York, Washington, New Yor] is not well-formed since there are two valid values and it merges two different states.

Technically speaking, the proposed clustering-based approach is implemented in Python by using the sklearn implementation of AC[4]. For the linkage method and the cut, we exploit default values of the library, while the similarity is pre-computed as described before. Linkage is set to *ward*, and it minimises the variance of the clusters being merged to spur uniform clusters. *compute_full_tree* and *distance_threshold* control the cut. By default, the compute_full_tree parameter is false, and it decreases the computation time.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

5.3.3 *Correction proposal*

For each cluster, the dictionary sample closer to cluster elements (where the similarity is computed as before) behaves as the correction for the entire cluster. This phase exploits the dictionary of valid values to elect corrections. Thanks to the refinement step, the clustering-based approach can rely on the assumption that all the cluster elements represent the same geographical entity. The correction proposal phase still requires to detect which is the most promising entity that might behave as the correction. Supposing that both *Budapest* and *Bucharest* are right word in our dictionary. We have a cluster containing [Budarest, Budrest]. Budarest has distance 1 to Budapest (r in p) and distance 2 to Bucharest (d in c and add h); while Budrest has distance one to Budarest (add a), distance 2 from Budapest (add and r in p), and distance 3 from Bucharest (d in c, add h and add a). The best candidate correction is Budapest since there is one cluster element, Budarest, with distance 1 from Budapest. The other elements, such as Budrest, will be transitively corrected as Budapest.

5.4 evaluation: performance assessment

The evaluation reported in this section assesses different aspects of the proposed approach concerning the role plaid by the clustering and the fuzzy matching while compared with traditional and popular approaches.

5.4.1 *Evaluation design*

The following paragraphs detail the RQ based on the performed evaluation, the used datasets, the performed protocol, and the collected metrics.

methodology. TheRQs at the basis of the performed evaluation are:

RQ1 - does approximate string matching (often referred to as fuzzy string searching) give a contribution to detect and correct errors in string similarity?

RQ2 - does the clustering help in detecting and correcting inaccuracies in textual geographical information?

DATASETS. The evaluation is performed on datasets released as OD by the Campania region containing at least one column containing textual geographical data, such as provinces or municipalities[5]. It results in more than 60 datasets, belonging to 7 different categories, from culture to health, from transport to the environment. The used datasets are the original ones, without introducing any extra error. Table 5.1 reports the information related to each category of tested datasets. Categories are defined by the data provider, and they are not mutually exclusive. We manually refined categories by forcing a single category (the most representative according to the dataset content) for each dataset. Table 5.1 also reports the number of datasets belonging to each category, statistics about the size of the datasets in terms of the number of rows (minimum and maximum size, average size, and its standard deviation), and the statistics of the errors (minimum and maximum number, average number and its standard deviation). Each geographical textual value that does not match any valid city or municipality is counted as an error.

Table 5.1: Datasets used during the evaluation of the clustering-based approach

		Size statistics				Errors statistics			
Category	**Size***	**Min**	**Max**	**Mean**	**St. dev**	**Min**	**Max**	**Mean**	**St. dev**
culture	9	4	1062	255.44	451.02	0	5	1.67	1.94
economy	1	265	265	265.00	0.00	4	4	4.00	0.00
environment	7	43	299	128.14	83.33	0	6	1.43	2.07
farming	6	81	285	147.00	72.44	0	14	5.50	5.09
health	5	45	4083	1456.40	1588.60	0	68	17.20	28.53
society	25	7	16184	1715.48	3411.48	0	116	13.88	23.63
transport	11	79	159	108.09	29.04	0	3	1.18	1.17

Errors can be classified in the following groups:

[5] https://dati.regione.campania.it/catalogo/

- *abbreviations*: for example *Saint* is simply reported as *S.*;
- *typos*, such as extra characters, misspelled words, single-character missing;
- *improper splitting*, e.g., missing or improper spaces,
- *wrong names*;
- *improper formatting*, such as missing apostrophes;
- *improper usage of accent*, often replaced by apostrophes or completely missing;
- *incomplete names*, e.g., truncated values;
- *not in the domain*, such as `not specified value`.

PROTOCOL. Each dataset is parsed by i) the proposed clustering-based approach relying on both Levenshtein and Fuzzy Matching during the similarity computation, ii) a clustering-based approach that performs the similarity computation by only relying on Levenshtein, and iii) a pair-wise approach that compares each value in the dataset under assessment with the dictionary of correct values, both considering Levenshtein in isolation and by combining Levenshtein and Fuzzy Matching. The corrections returned by any of the performed algorithms are manually checked. Comparing results achieved by the clustering-based approaches and the pair-wise comparison lets assessing the role plaid by ML is assessing and correcting inaccurate values in textual geographical data. By comparing the results achieved by Levenshtein in isolation and the ones achieved by combining Levenshtein and Fuzzy Matching, the best similarity metric can be detected empirically.

DATA GATHERING. Each algorithm run returns the number of corrected values. The manual refinement verifies all the proposals that correspond to valid corrections and the ones corresponding to introduced errors.

5.4.2 *Results*

THE ROLE OF FUZZY MATCHING IN THE SIMILARITY COMPUTATION PHASE. This section reports the role plaid by fuzzy matching when combined with Levenshtein in performing the similarity computation phase. The evaluation is based on the hypothesis that by combining both approaches we can detect more error instead of the well-known Levenshtein metric used in isolation as works at the sub-string level, while Levensthein works at the string level.

Table 5.2 reports results for each dataset according to all the compared approaches. The dataset is referred to by an ID obtained by the English dataset name and, mostly, the publication year. Empty cells mean 0. Values in italic represent identical results obtained by both similarity approaches, while in bold are reported best results. All datasets without errors (17 datasets) have been omitted as both approaches correctly recognise the absence of errors. Moreover, also the datasets in which the results with both metrics return the same results in both the algorithms (19 datasets) are not reported.

Concerning the similarity metrics, it can be observed that the combination of Levenshtein and Fuzzy matching can detect and correct the greatest number of errors in the lookup algorithm as in all cases but two ones (i.e., RS_2015 and RS_2017) it achieves the best results. In the clustering algorithm, both similarity approaches return almost the same results, but in a few cases. Since, in general, the exploitation of Fuzzy Matching improves (or not worsens) the results of using Levenshtein in isolation, it can contribute to detect and correct inaccurate values in textual geographical data.

THE ROLE OF CLUSTERING IN ASSESSING AND CORRECTING INACCURATE VALUES. This section reports the empirical verification that a clustering-based approach is more accurate than a pair-wise comparison between values to validate and a dictionary of correct values to detect the most similar correct word. This algorithm will be referred to as *dictionary lookup* or simply *lookup*.

It is based on the hypothesis that a clustering-based approach can correct errors by exploiting word similarity transitivity. In

Table 5.2: Comparison between string comparison approaches

Dataset ID	Errors	Lookup		Lookup		Original			Clustering		
Dataset ID	Errors	Lev & FM		Lev		approach			and Lev		
		TP	FP	TP	FP	TP	FP	FN	TP	FP	FN
PS_2014	23	**20**	3	18	5	*22*		1	*22*		1
PS_2015	17	**14**	3	13	4	*15*		2	*15*		2
RS_2015	7	4	3	**5**	2	*6*		1	*6*		1
AB_2015	7	**5**	2	4	3	*5*		2	*5*		2
AB_2016	4	**3**	1	2	2	*2*		2	*2*		2
RS_2017	11	8	3	**9**	2	*11*			*11*		
AB_2017	7	**6**	1	5	2	*5*		2	*5*		2
AB_2018	12	**10**	2	7	5	*9*		3	*9*		3
Fh_Na_2018	7	**6**	1	4	3	*4*	1	4	*4*	1	4
Coop_2019	42	**39**	3	36	6	*36*		6	*36*		6
Reg_2019	4	**2**	2		4	**2**	1	3	1	2	5
CarSa_2018	3	**3**		2	1	2		1	2		1
AirQual_2019	6	**6**		2	4	*1*		5	*1*		5
MinCom_2018	18	**18**		13	5	*18*			*18*		
Ph_2018	6	**4**	2	2	4	3		3	**4**		2
ParaPh_2018	5	**3**	2	1	4	*3*		2	*3*		2
Show&Cine	25	**19**	6	12	13	*16*	2	11	*16*	2	11
EduFarm2018	14	**11**	3	8	6	*9*	1	6	*9*	1	6
LibAdd_2019	5	**3**	2	2	3	*3*		2	*3*		2
HighIns_2019	3	**3**		2	1	*3*			*3*		
GenPra_2019	68	**55**	13	44	24	**43**	4	29	37	6	37
Moto_Sa_2018	3	**3**		2	1	2		1	2		1
UnavAssets	1	**1**			1	*1*			*1*		
Ped_2019	7	**5**	2	3	4	*4*		3	*4*		3
SocProm_1_2019	2	**2**		1	1	*1*		1	*1*		1
Volunteers	116	**95**	21	68	48	**60**	6	62	58	5	63
G_Sa_2017	1	**1**			1	*1*			*1*		
Lib_2019	4	**3**	1	2	2	*3*		1	*3*		1

other words, clustering partitions records into disjoint clusters of items, where each cluster corresponds to one real-world entity, and records in a cluster are different representations of the same entity. The proposed approach first finds groups of similar words during the clustering phase, and, then, proposes corrections according to the elements that co-occur in the same cluster. It avoids forcing the correction of words that are extremely far from any correct word by isolating them in dedicated clusters. Moreover, it exploits the transitive property in proposing corrections. In fact, during the clustering phase, the approach ensures that each cluster represents a uniform concept by the refinement step. Then, during the correction phase, it detects the centroid in each cluster, corresponding to the element that exactly matches a correct word or is the closest to one of them, and all the others are transitively corrected.

Table 5.3 reports results for each dataset. as in the previous evaluation, each dataset is referred to by an ID obtained by the English dataset name and, mostly, the publication year. Empty cells mean 0. Values in italic represent equal results obtained by both similarity approaches, while in bold are reported best results. Datasets without errors and the ones in which both algorithms return the same results (17 datasets) have been omitted. Table 5.3 reports the number of corrected errors (TP) to quantify the number of correctly identified errors, while the column `Err.` is dedicated to the errors that each dataset contains at the end of the algorithm run.

By manually inspecting errors at the end of the clustering-based approach, it results that errors, in the end, match errors in the original dataset that are not correctly identified. The lookup approach introduces extra errors by forcing the correction even if words are extremely far away from each other. For instance, some datasets containing *not specified* (in the Italian language) and the lookup algorithm forced the correction to a valid municipality name similar for pronunciation, while the original algorithm simply ignores it. Moreover, while final errors in the original approach are easier to detect since they are replaced by empty values, the lookup approach forces the correction and requires the human-in-the-loop to identify the incorrect values manually.

By only considering the corrected errors (TP columns), it seems that the lookup algorithm performs better than the original approach while considering the final errors, it introduces more errors than the clustering-based approach. It implies that the ML based approach is more effective than traditional approaches. Since the clustering-based approach returns datasets containing fewer errors if compared with the lookup algorithm, clustering helps in detecting and correcting inaccurate values in textual

Table 5.3: Comparison between the clustering-based and a lookup approach

Dataset ID	Init.	Clustering			End	Lookup		End
	Err.	TP	FP	FN	Err.	TP	FP	Err.
PS_2014	23	22		1	**1**	20	3	3
RS_2014	9	8		1	**1**	6	3	3
AB_2014	5	4		1	**1**	4	1	3
PS_2015	17	15		2	**2**	14	3	3
RS_2015	7	6		1	**1**	4	3	3
AB_2015	7	5		2	2	5	2	2
RS_2016	22	21		1	**1**	18	4	4
AB_2016	4	2		2	2	3	1	**1**
RS_2017	11	11			**0**	8	3	3
AB_2017	7	5		2	2	6	1	**1**
PS_2018	14	13		1	**1**	11	3	3
AB_2018	12	9		3	3	10	2	**2**
Fh_Bn_2018	7	6		1	1	7		**0**
Fh_Na_2018	7	4	1	4	5	6	1	**1**
Coop_2019	42	36		6	6	39	3	**3**
Reg_2019	4	2	1	3	4	2	2	**2**
YouAssoc_2019	3	1		2	2	1	2	2
CarSa_2018	3	2		1	1	3		**0**
AirQual_2019	6	1		5	5	6		**0**
Ph_2018	6	3		3	3	4	2	**2**
ParaPh_2018	5	3		2	2	3	2	2
Show&Cine	25	16	2	11	13	19	6	**6**
EduFarm2018	14	9	1	6	7	11	3	**3**
State-ownProp	2	1		1	2	1	1	**1**
LidAdd_2019	5	3		2	2	3	2	2
GenPra_2019	68	43	4	29	33	55	13	**13**
Moto_Sa_2018	3	2		1	1	3		**0**
Ped_2019	7	4		3	3	5	2	**2**
SocProm_1_2019	2	1		1	1	2		**0**
Volunteers	116	60	6	62	68	95	21	**21**
Lib_2019	4	3		1	*1*	3	1	*1*

geographical data. Moreover, since both algorithms do not return datasets free from errors, it underlines that cleansing approaches can support data producers and consumers, but a human check is still required.

5.5 EVALUATION: TIME PERFORMANCE ASSESSMENT

5.5.1 *Evaluation design*

METHODOLOGY. The RQ at the basis of the proposed approach is *"What impact the overall performance of the clustering-based approach?"*.

DATASETS. The performance evaluation is evaluated on all the datasets exploited in the previous evaluation.

PROTOCOL. For each dataset, the clustering-based approach returns the total execution time and the time needed by each phase, i.e., the similarity computation, the whole clustering time (and the number of runs of the clustering algorithm), and the correction time. Experiments are performed on a system with an Intel(R) Core(TM) i7-8700T CPU at 2.40GHz and 16 GB RAM.

DATA GATHERING. The performance is computed according to the execution time, in seconds.

5.5.2 *Results*

While the average time is 19 seconds, its standard deviation is 0.0007. Up to 1,500 lines, it requires less than 30 seconds. When the size increase (up to 4,000 lines), it requires 1 minute to 3 minutes and a half. With 6,000 lines, PS_2014 requires 6 minutes. Therefore, the time execution seems to have quadratic complexity. However, with 16,000 rows, RS_2016 requires precisely 1 minute. Moreover, the time complexity seems not to be related to the number of errors. While PS_2014 contains 23 errors, and it requires 6 minutes, Volunteers with the maximum number of detected errors (116) require 95 seconds. Table 5.4 reports the correlation between each algorithm phase and both the size and the number of occurred errors in the considered datasets. While the dataset size affects the similarity computation and the total time, the number of mistakes affects both the clustering and, above all, the correction time.

Table 5.4: Relation between time of each phase of the clustering-based approach, dataset size and contained errors.

	Total time	Similarity time	Clustering time	Correction time
Size	**0.56**	**0.56**	0.43	0.34
Error	0.47	0.39	**0.75**	**0.91**

5.6 COMPARISON WITH THE STATE-OF-THE-ART

5.6.1 *Evaluation design*

The following paragraphs detail the RQ based on the performed evaluation, the used datasets, the performed protocol, and the collected metrics.

METHODOLOGY. The RQ at the basis of the performed evaluation is *"Is the proposed clustering-based approach competitive with widely adopted tools to perform quality assessment and improvement?"*. The reference tool is OpenRefine.

DATASETS. This evaluation is performed on a subset of the datasets used in the performance evaluation. For each dataset category, the comparison evaluation considers the biggest dataset and the one containing the greatest number of errors. If both metrics elect the same dataset, only one dataset is considered for the comparison.

PROTOCOL. For each considered dataset, the evaluation compares the clustering-based approach with results achieved by OpenRefine[6]. This evaluation verifies if the proposed approach is competitive with widely adopted tools in the state-of-the-art and it also verifies the role plaid by a dictionary of correct val-

[6] OpenRefine: https://openrefine.org/

ues in detecting inaccurate values in textual geographical data as OpenRefine does not exploit external sources containing the correct values. For each dataset, a dedicated project is opened in OpenRefine, and a text facet for the column containing textual geographical data is generated. Then, a clustering phase is performed where OpenRefine exploits the k-NN algorithm to cluster data and the Levenshtein metric to compute word similarity.

DATA GATHERING. Each algorithm run returns the number of corrected values. The manual refinement verifies all the proposals that correspond to valid corrections and the ones corresponding to introduced errors.

5.6.2 *Results*

Table 5.5 reports the considered datasets, their size and number of manually detected errors, the number of errors detected by the proposed approach, and the error detected by OpenRefine.

Table 5.5: Comparison between the clustering-based approach and OpenRefine

Dataset ID	Dataset size	# errors	# errors Original Approach	# errors OpenRefine
Lib_2019	1,062	4	3	0
LibAdd_2019	1,037	5	3	0
Reg_2019	265	4	2	0
BO_2018	299	1	1	0
AirQual_2019	43	6	1	0
EduFarm2018	285	14	9	1
GenPra_2019	4,083	68	43	9
RS_2016	16,184	22	21	0
Volunteers	2,471	116	60	5
CarSa_2018	159	3	2	0
Moto_Sa_2018	159	3	2	0

OpenRefine correctly recognises alternative forms to write the same word (lowercase and uppercase), such as in Reg_2019. It proposes to correct words that are both valid options but very similar (such as `Laviano` and `Saviano` in Lib_2019 and LibAdd_2019, or `Fisciano` and `Visciano` in LibAdd_2019 or BO_2018).

OpenRefine proposes the right correction only if at least one time each value is correctly written, for instance, in EduFarm2018 `Massa Lubrense` is written two times in this correct form and a single time it is misspelled by omitting the space `Massalubrense`. OpenRefine recognises that they represent the same concept and propose the right correction since the right form occurs more often than the inaccurate value. Otherwise, when the occurrence of write and wrong values corresponds, OpenRefine chooses one version without any guarantee on the choice of the right value. For instance, in GenPra_2019, `Colla Sannita` and `Colle Sannita` are correctly merged in the same cluster, but the wrong value `Colla Sannita` is proposed as a correction. In the same dataset, `atena lucana` and `atenalucana` are correctly merged in the same cluster and the right value (`atena lucana`) is proposed as correction. OpenRefine probably opts for the longest version in these cases, and not always it represents the right choice. It can not propose corrections for words that are never correctly written in the dataset, as it can be observed in Lib_2019.

OpenRefine proposes clusters and corrections, but users have to accept or regret all the suggestions manually. It implies that OpenRefine requires expert users in distinguishing right suggestions from wrong ones.

5.7 FINAL REMARKS

By analysing geographical textual information in real open datasets released by the Campania region, we detected many inaccurate values, including improper splitting and formatting, truncated values, improper usage of accents, and typos. Hence, this chapter proposes a clustering-based approach where the detection and correction of quality issues are based on the AC, the combination of the Levenshtein metric and the approximate string matching to compute string similarity, and the exploitation of a dictionary of correct values.

According to the discussed results, the combination of Levenshtein and Fuzzy string matching improves the similarity recognition in textual geographical data if compared with Levenshtein used in isolation, positively replying to RQ1. Moreover, the reported results show that the clustering-based approach introduces less error than the direct comparison of each word against the dictionary of valid values. Hence, it positively replies to RQ2. Finally, the exploitation of a dictionary containing correct values improves the correction of inaccurate values, according to the comparison with the widely used OpenRefine.

Future direction might assess the generalisation of the proposed approach by investigating its applicability in other contexts where values must be compliant with a pre-defined set of valid options, such as fields modelling diseases, religions, genders, school levels.

6

OPEN DATA PRIVACY ASSESSMENT AND IMPROVEMENT BY AN ANONYMISATION APPROACH

> *Arguing that you do not care about the right to privacy because you have nothing to hide is no different than saying you do not care about free speech because you have nothing to say.*
>
> – Edward Snowden

Data owners, such as PAs, health care, and financial institutions, may release the data they collect by de-identifying them, i.e., by masking, generalising, or deleting IDs. However, even anonymised, public information may be re-identified by exploiting other pieces of available data. A 2002 study found that 87% of the U.S. population can be identified using gender, birth date, and ZIP CODEs as QID by matching anonymised hospital visit records and voting lists [313]. These data are not problematic if isolated but lead to the re-identification of individuals by exploiting additional information.

Individuals whose data are re-identified at risk of having their private information, such as their finances, health or preferences, and their identity, sold to organisations without user consent [264] or disclosed to undesired end-users [264], or even it can cause the refusal of an insurance provision [313].

When a privacy leakage is reported, PAs usually react by closing data or publishing poorly informative datasets. As an example, the datasets analysed in the evaluation of this approach have been substituted with a version with significantly lower informativeness, as only the province of residence, driving license category, and release date are provided. These datasets are reduced to *pointless OD*. Instead of making data useless, data curators should be invited to invest in further sanitation actions.

This chapter proposes an approach to support data publishers in guaranteeing datasets significantly more informative than the one currently available on their websites while preserving citizens' privacy. The research at the basis of this chapter has been published in the following contribution:

> Matteo Pastore, Maria Angela Pellegrino, Vittorio Scarano: *Detecting and Generalizing Quasi-Identifiers by Affecting Singletons*. In the Proceedings of EGOV-CeDEM-ePart-* 2020.

6.1 RELATED WORK

The choice of the QID is an open question [115] since it depends on attributes that attackers can exploit to link actual data to any external source. Braghin et al. [41] define an approach to detect QIDs by considering both single columns and their collections and counting the unique occurrence of values. Motwani and Xu [220] exploit the separation and the distinct ratio to quantitatively describe the ability of attributes to distinguish an individual from another. While the distinct ratio measures the percentage of distinct values, the separation ratio measures the proportion of tuple pairs that can be uniquely distinguished.

Concerning *anonymization* approaches, they explicitly remove IDs and hide the sensitive information assuming that the latter should not be used in data mining algorithms. ID removal might not be enough: it is still possible to recognise individuals by QIDs. To prevent linking attacks, datasets must be sanitised [285] by applying anonymization operations such as generalisation, suppression, anatomisation, permutation, and perturbation [105].

Among the most famous privacy models, k-Anonymity [57] is based on the fundamental concept that if a record has a particular value for a QID, at least other k-1 records will have the same value for that QID. Multiple versions of k-Anonymity have been proposed to overcome some of its limitations. For instance, (X- Y)-anonymity [336] face the case in which multiple rows of the dataset are related to the same individual; MultiRelational k-Anonymity [232] focuses on the anonymization of multiple relational tables; l-diversity [191] guarantees that each equivalence class has at least *l* well-represented values for each sensitive attribute overcoming the risk of the homogeneity attack; (α, k)-anonymity [343] experiences local recording by reducing data distortion; t-closeness [182] requires that the distance between the distribution of a sensitive attribute and the distribution of the attribute in the overall table should be no more than t.

6.2 APPROACH AND METHODOLOGY

The proposed approach is based on a privacy issues detection step followed by an anonymity approach based on generalisation and suppression. The workflow starts from the human provision of the dataset to test, and it automatically returns the best QID. If the best QID matches (year, municipality, gender), it also provides the corresponding generalisations.

This approach interprets the detection of privacy issues as the occurrence of unique values by considering a single column or a combination of them. The *best* QID detector identifies the number of uniquely occurrences of a combination of values in a dataset. These unique occurrences will be referred to as *singletons*. The best QID is the *minimum* number of columns/attributes that leads to the disclosure of the *highest* number of singletons. By recalling that PPDP is interested in minimising both the information loss and the privacy leakage [115], a privacy leakage corresponds to the number of occurred singletons, and the information loss is estimated as the number of suppressed and modified rows, decreasing thereby the overall dataset quality.

Concerning the anonymisation phase, the proposed approach can be defined as a modified version of k-Anonymity where:

- k is at least equals to 2;
- we discourage suppression in favor of generalization;
- while k-Anonymity operates at a global level, we can also locally work.

6.3 TOOLKIT: `qid` IDENTIFIER AND ANONYMISATION APPROACH

This section presents the implemented approach to identify the best QID and perform anonymisation. It implementation is freely available on GitHub[1].

[1] https://github.com/isislab-unisa/qid_identifier_and_anonymizer

qid IDENTIFIER. The implemented pseudo-code to identify IDs and QIDs follows:

```
def detect_ID_and_QID(dataset):
   identifiers = []
   stats = {}
   for size in range(1, num_columns):
      # all the dataset column subsets of "subset_size"
      subsets = get_subsets(columns_to_check, size)
      # IDs: columns containing all distinct values
      temp_IDs, still_to_check = get_IDs(dataset, subsets)
      IDs += temp_IDs
      # for each subset, it stores the singletons number
      stats.update(get_stats(dataset, still_to_check))
      columns_to_check = list(set(still_to_check))
   # best QID: the smallest subset of columns exposing
   # the greatest number of singletons
   best_QID = get_best_QID(stats)
```

The best QID election procedure first takes into account the number of singletons detected by each set of attributes. If more than one set of columns shares the same number of singletons, it elects as *best* QID the minimum set of columns.

ANONYMITY BASED ON GENERALISATION AND SUPPRESSION. To ameliorate the presence of privacy issues, the proposed approach anonymise the dataset by performing generalisation and suppression. While incomplete rows are suppressed at the beginning of our technique; then, only the generalisation is permitted. The generalisation is preferred to the suppression as it is better to publish *incomplete* information rather than preventing the publication at all.

The proposed approach focused on the well-known QID composed of (date_of_birth, ZIP, sex). In particular, it focuses on a slightly simplified version of this QID, where it only has access to the year_of_birth. It is simplification without loss of generality since it can be easily generalised to the entire date. However, years can be straightforwardly generalised by the mean value of the years' interval. Moreover, in Italy, there is a two-way correspondence between ZIP codes and Municipalities. Therefore, they can be used interchangeably. The proposed approach detects which column or combination of columns is *worth* to generalise to achieve the minimum number of singletons by modifying the minimum amount of rows.

Numerical attributes (i.e., year_of_birth) are generalised by the standard approach of substituting values by intervals. Therefore, rows are first sorted by year_of_birth and then split into groups of at least k values. If two rows containing the same year are separated into two consecutive groups, rows are iteratively merged in the same group until the cut splits rows containing different years. Finally, each year is substituted with the interval [*min_year*, *max_year*) of the corresponding group. The proposed approach also applies a second strategy where the average value of the interval replaces each interval. This practice is based on the hypothesises that if current years mainly correspond to the mean value of the range, fewer rows will be modified while still reducing the number of singletons.

Concerning the sex column, the proposed approach replaces *male* and *female* values by *any gender*. In this case, the generalisation plays the same role as cell suppression. About the municipality column, the hierarchy induced by the Italian national administrative levels is exploited: municipalities are generalised by provinces, provinces by regions, regions by states. The reported experiments only consider the first level of this hierarchy by generalising municipalities by provinces. Categorical attributes (i.e., sex and municipality) can be generalised by global or local recording. While the global recording affects the entire dataset, the local one only modifies rows related to the singletons disclosed by the best QID. We hypothesise that the local recording can introduce a sufficient level of privacy protection while affecting a minimal number of rows interpreted as a slight decrease of the overall dataset quality. The implemented approach can be resumed as follows:

1. the rows containing empty values are dropped out, and the removed rows alter the counter of affected rows. The *full* version of the dataset (i.e., only rows without any empty cell) is considered in the following steps;

2. the following generalisations are performed:
   - ***all the municipalities*** are generalised by the corresponding province;

- only the ***municipalities*** of the rows corresponding to the ***singletons*** detected by the best QID are generalised by the corresponding province;
- ***all*** the values of the ***sex*** column are generalised by *any gender*;
- only the values of the ***sex*** columns corresponding to the ***singletons*** detected by the best QID are generalised by *any gender*;
- ***all the birth_years*** are generalised by the corresponding ***intervals***;
- ***all the birth_years*** are generalised by the ***mean value*** of the intervals generated by the previous step;
- all the attributes are generalised by combining every pair of the generalisations described so far and by generalising all fields at once;

3. for each performed generalisation, we compute the *number of singletons*, the *percentage of singletons*, the number of *distinct values*, the *number of modified and removed rows*;
4. the best generalisation is elected by considering the one achieving the minimum number of singletons while affecting the minimum number of rows.

6.4 EVALUATION: PERFORMANCE ASSESSMENT

This section reports the evaluation of the proposed approach in terms of the empirical verification of the set of columns that is worth generalising to minimise the privacy leakage while maximising the dataset quality.

6.4.1 *Evaluation design*

METHODOLOGY. The RQ at the basis of the performed evaluation is "*Which is the set of columns in the well-known QID (date_of_birth, ZIP, sex) that is worth generalising to achieve the minimum privacy leakage while altering the dataset the least possible?*

DATASETS. The proposed approach has been evaluated on real datasets released by the Italian Ministry of Infrastructure and Transport. These (anonymised) datasets contain information related to the driver's licenses of all the Italian regions. The datasets used in the evaluation have been downloaded in October 2019 from the official site. However, in January 2020 the Ministry updated the online version by significantly reducing the (already minimal) available content. The tested datasets (in their original form) are available on GitHub[2].

PROTOCOL. The evaluation only considers columns related to personal information (i.e., the municipality and the province reported as the driver residence, the year of birth, and the sex), while it ignores all the Non-sensitive information, such as the driver's license details, and numerical IDs.

For each dataset, the best QID is identified and, if it corresponds to the well-known QID (date_of_birth, ZIP, sex), also the anonymisation step is executed.

DATA GATHERING. For each returned anonymised dataset, the evaluation considers the anonymisation approach that minimises the privacy leakage interpreted in terms of singletons and maximises the dataset quality interpreted as the maximum coherence with the original dataset.

6.4.2 *Results*

The results of the proposed privacy issue detecting approach are available on GitHub[3]. The QID identifier module reports [Birth_year, ZIP, Sex] as *best* QID in all the regions. Even if datasets are anonymised, our approach highlights the possibility of distinguishing up to 2% (1.93%) of singletons uniquely. If 2% seems to be a negligible amount of disclosed identities, it is worth noting that the maximum number of disclosed singletons is more than 25K. It implies that removing IDs is not enough, and further anonymization actions must be performed to publish sanitised datasets.

[2] https://github.com/isislab-unisa/driver-license-datasets

[3] https://github.com/isislab-unisa/qid_identifier_and_anonymizer

We evaluate the impact of the anonymisation approach on three datasets used in the previous phase, heterogeneous in the disclosed percentage of singletons. Results are reported in Tables 6.1 and 6.2. The algorithm is linearly correlated to the dataset size, and it takes 0.0152 seconds for processing datasets with 6M rows.

YEAR RANGE VS YEAR MEAN VALUE. By comparing the year generalisation by intervals (row **Y_ran** of Tables 6.1 and 6.2) and by mean values (row **Y_avg** of Tables 6.1 and 6.2), the mean value gains 95% fewer singletons while affecting 7% rows less than the range approach.

GLOBAL VS LOCAL RECORDING. By comparing the global generalisation of the Municipality (**M** row of Tables 6.1 and 6.2) and its local recording (**M_loc** row of Tables 6.1 and 6.2), both obtain nearly the same number of singletons, but the local recording affects only the 2% of the rows while the global one modifies the entire dataset. By considering the number of singletons, the local recording achieves results close to 0, while the global recording succeeds in completely avoiding the disclosure of singletons. But, the global recording affects the entire dataset, significantly decreasing the dataset quality. We consider a good thread-off between privacy-preserving and data quality the generalisation of the municipality (and the sex) only of singletons disclosed by the QID (year_of_birth, sex, ZIP).

6.5 COMPARISON WITH THE STATE-OF-THE-ART

This section empirically compares the performance of the proposed approach with the well-known k-anonymity, considered the most similar approach proposed in the literature and a widely adopted anonymisation algorithm in the state-of-the-art.

6.5.1 *Evaluation design*

METHODOLOGY. The RQ at the basis of the performed evaluation is *"Is the proposed approach competitive with state of the art?"*

Table 6.1: Results of the anonymisation approach - 1

	Molise			
	#S	**%S**	**Size**	**DV**
	869	1.30	597,243	13,391
Cols	**#S**	**%S**	**MR**	**DV**
S_loc	745	0.12	869	13,329
M_loc	7	∼0	860	12,539
S,M_loc	32	∼0	868	12,680
S	295	0.05	597,243	7,101
M	6	∼0	414,584	338
Y_ran	127	∼0	597,243	3,641
Y_avg	1	∼0	556,875	736
S,M	1	∼0	597,243	171
S,Y_ran	48	∼0	597,243	1895
S,Y_avg	0	0	597,243	368
M,Y_ran	0	0	597,243	44
M,Y_avg	0	0	584794	8
S,M,Y_ran	0	0	597,243	44
S,M,Y_avg	0	0	597,243	8
	Umbria			
	#S	**%S**	**Size**	**DV**
	2,569	0.15	198,312	16,628
Cols	**#S**	**%S**	**MR**	**DV**
S_loc	2,129	1.07	2,569	16,408
M_loc	7	∼0	2,556	14,078
S,M_loc	3	∼0	2,569	14,446
S	909	0.46	198,312	9,490
M	7	∼0	198,312	324
Y_ran	325	0.001	198,312	4,806
Y_avg	3	∼0	184,111	1,089
S,M	7	∼0	198,312	324
S,Y_ran	116	∼0	198,312	2,606
S,Y_avg	1	∼0	198,312	545
M,Y_ran	0	0	198,312	84
M,Y_avg	0	0	194,932	16
S,M,Y_ran	0	0	198,312	43
S,M,Y_avg	0	0	198,312	8

Table 6.2: Results of the anonymisation approach - 2

	Valle d'Aosta			
	#S	**%S**	**Size**	**DV**
	1,684	1.93	87,464	9,174
Cols	**#S**	**%S**	**MR**	**DV**
S_loc	1,264	1.45	1,684	8,964
M_loc	4	∼0	1,679	7,501
S,M_loc	1	∼0	1,684	7,785
S	621	0.71	87,464	5,166
M	4	∼0	87,464	167
Y_ran	198	0.002	87,464	2,739
Y_avg	9	0.001	81,476	607
S,M	1	∼0	87,464	85
S,Y_ran	70	∼0	87,464	1,442
S,Y_avg	5	∼0	87,464	308
M,Y_ran	0	0	87,464	43
M,Y_avg	0	0	85,948	8,817
S,M,Y_ran	0	0	87,464	22
S,M,Y_avg	0	0	87,464	4

DATASETS. The proposed approach has been evaluated on a real dataset released by the Italian Ministry of Infrastructure and Transport. The (anonymised) dataset contains information related to the driver licenses of the Valle d'Aosta Italian region. The dataset used in the evaluation has been downloaded in October 2019 from the official site[4]. However, in January 2020 the Ministry updated the online version by significantly reducing the (already minimal) available content. The tested dataset (in their original form) is available on GitHub[5].

PROTOCOL. The evaluation only consider columns related to personal information (i.e., the municipality and the province reported as the driver residence, the year of birth, and the sex), while it ignores all the Non-sensitive information, such as the driver's license details, and numerical IDs.

[4] http://dati.mit.gov.it/catalog/dataset/patenti

[5] https://github.com/isislab-unisa/driver-license-datasets

For each dataset, the best QID is identified and, if it corresponds to the well-known QID (date_of_birth, ZIP, sex), also the anonymisation step is executed.

The best QID generalisation elicited by the proposed approach, i.e., the one minimising the number of singletons and minimising the affected rows is compared with the anonymised dataset returned by k-anonymity[6]. k-Anonymity has been run by generalising Municipalities by their Provinces, Sex by *any gender*, and the Year by intervals of width 4. It performed the generalisation of any set of attributes. By setting the algorithm run, the evaluation allowed suppression equals 0.01 in all cases, but in the sex, generalisation is set to 0.05.

DATA GATHERING. For each returned anonymised dataset, the evaluation considers the anonymisation approach that minimises the privacy leakage interpreted in terms of singletons and maximises the dataset quality interpreted as the maximum coherence with the original dataset.

6.5.2 *Results*

At the global level, both approaches modify almost the entire dataset. While the proposed approach obtains a generalised version of the dataset information, k-Anonymity removes many rows. When a small number of singletons occurs, k-Anonymity drops the corresponding rows. In all the other cases, it drops at least 200 rows and modifies all the other ones. Thanks to the local recording, the proposed approach obtains a minimum number of singletons (near to 0) while affecting a small portion of the dataset (up to 2%). Concluding, the proposed approach achieves the same results of k-anonymity in terms of privacy-preserving, while it obtains better data quality thanks to the local recording.

[6] K-anonymity Python implementation, released by the DZone community: https://dzone.com/articles/an-easy-way-to-privacy-protect-a-dataset-using-pyt

6.6 FINAL REMARKS

While data publishers are encouraged to publish OD, they have to pay attention to avoid breaking individual privacy. This section proposes an anonymisation approach to detect QIDs by counting *singletons* in a dataset. An empirical evaluation performed on real OD published by the Italian Ministry of Infrastructure and Transport demonstrates that the well-known QID (date_of_birth, sex, ZIP) discloses up to 2% (and up to 25K) of singletons in already anonymized datasets.

When a privacy leakage is reported, data publishers usually react by closing data or publishing poorly informative datasets. As an example, the datasets exploited in the evaluation have been substituted with a version with significantly lower informativeness, as only the province of residence, driving license category, and release date are provided. These datasets, in our opinion, are reduced to *pointless OD*. Instead of making data useless, we suggest investing in further sanitation actions. This section empirically proves that the proposed approach achieves the minimum number of modified rows (up to 2% of affected rows) while obtaining the number of singletons close to 0 thanks to a local recording.

In the future, further well-known QIDs can be explored to verify the effect of a local recording while performing anonymisation. Moreover, despite proving the effectiveness of the proposed approach, data curators should be provided with working tools to easily perform sanitation actions. Consequently, the most promising approach should be wrapped in a framework to support data publishers, such as PAs, in guaranteeing datasets significantly more informative than the one currently available on their websites while preserving citizens' privacy.

7 QUALITY AWARE DATA PUBLICATION IN SPOD

It is the long history of humankind that those who learned to collaborate and improvise most effectively have prevailed.
– Charles Darwin

The public sector and, mainly, PAs are often organised in silos [68], i.e., poorly coordinated bureaucratic structures. Silos have been created as a way to organise and manage processes keeping tasks and responsibilities separated. But, today, they are perceived as a limitation as they impede the development of a more collaborative, multi-disciplinary approach to manage resources [40]. Moreover, they are considered highly inefficient both for citizens and employees [19]. Thus, siloed structured organisations are slowly breaking down vertical silos by enabling horizontal interoperability. It implies moving from closed, structured, and hierarchical into open, flat, and unstructured organisations by performing structural changes [68].

Information and communication technologies (ICTs) behave as significant enablers of public sector transformations to serve more and better public services [68]. The recent trend is referred to as Government as a Platform (GaaP) and it sees PAs empowering citizens to create public value via platforms. Some of the most outstanding advantages achieved by adopting platforms follow:

- bring data, services, technologies, and people together to respond to mutable and heterogeneous societal needs [45, 68, 110, 153, 155, 245];
- increase PAs efficiency as they enable external actors' participation in co-producing public services, helping organisations to deliver more value with fewer investments;
- easily coordinate actors involved in the service production and dissemination;
- guarantee transparent service accessibility, creation, and modification [245].

GaaP enables the Government to Citizens (G2C) collaboration where PAs represent the Government while any external actor, such as non-profit organisations or the private sector, behaves as citizens [237]. Furthermore, it encourages the Government to Government (G2G) collaboration by enabling wider cooperation among departments [208].

The GaaP concept is based on several premises [245]. It requires, first, the adoption of open-source software platforms and, second, the OD philosophy that states that data should be freely accessed, used, and modified to be shared for any purpose, accompanied by a license that ensures free re-use [101]. While open-source software enables developers to contribute to the platform ecosystem, the adoption of OD enables the possibility to autonomously create public value by freely combining them unexpectedly and creatively [245]. But GaaP also relies on the participatory design approach by emphasising standardisation, modularity, component reuse, and agile development that can reduce data management and data exploitation barriers.

This section focuses on a regional level and retraces changes performed by our RPA, the Campania Region, to realise the GaaP concept. In particular, it explores how PA members can collaboratively create, refine, and exploit OD by a multi-disciplinary, multi-departmental, and cross-domain collaboration. As a result, it discusses the adaptation of SPOD [67] as a back-office platform, i.e., as an internal platform [68], for the Campania Region. While SPOD is already used by the third sector to co-create, publish and exploit OD, this article details how SPOD features have been adapted and enhanced to satisfy RPA requirements by a co-design approach actively involving RPA delegates. The remaining section refers to the customised version of SPOD to satisfy the Regione Campania needs by naming it Campania Crea, while it names SPOD only when it refers to the original platform resulting from the ROUTE-TO-PA project.

It is worth noting that the work at the basis of this paper relies on strict coordination, cooperation, and collaboration among the ICT field experts and Campania Region delegates. As a result, the Campania Region revised the OD production and publication workflow by introducing Campania Crea to enable the co-creation and exploitation of high-quality OD.

The research at the basis of this chapter has been submitted as the following academic contribution and it is actually under evaluation:

> Salvatore Avella, Angela Cocchiarella, Dario Fonzo, Giuseppina Palmieri, Maria Angela Pellegrino, and Vittorio Scarano. *"Government as a Platform in a Regional Public Administration"*. *Submitted* to Transforming Government: People, Process and Policy in July 2021.

7.1 THE FLAGSHIP PROJECT OF THE CAMPANIA REGION

The Campania Region, one of the 20 Italian regional authorities, is a large and complex institution with over 5,000 employees. In 2016, it was classified as the last one in a national ranking concerning OD due to a low level of data management maturity, poor integration between information systems, lack of data culture, and the adopted data siloed model. Thus, the Campania Region experienced national political pressure in breaking silos, performing digitisation and automation processes, and guaranteeing more comprehensive interoperability and data exchange within RPAs as part of a national digital transformation program.

Starting from 2017, the Campania Region promoted a technological and organisational intervention to acquire ICT services and adopt production methods to satisfy national constraints concerning mandatory datasets that any RPA must publish as OD. It resulted in the proposal of a strategic flagship project, named *OD Campania*, that aims to realise OD to spur and encourage public sector information reuse.

According to RPA delegates, digital transformation projects mainly concern people and processes. Consequently, the Campania Region performed consistent organisational transformations by revising the governance model and setting up a cross-department work group, i.e., the *OD Team*, based on multidisciplinary competencies in ICT, data analysis, communication, and content management skills. These organisational transformations overcame limitations and constraints posed by a data siloed structure enabling G2G collaboration. The staff has been involved through a bottom-up strategy aimed to engage RPA members by increasing their awareness about data value and encouraging their participation in the whole process of opening up public data. This strategy is based on RPA members' skill development by periodic activities and training courses. Hopefully, in the future, further efforts might be invested in hiring new personnel with data management skills as the OD team is aware of the benefits of introducing other skilled people to their team.

Concerning technological changes, the most outstanding project achievement consists in the revision of the OD production process and adoption of the `Campania Crea` infrastructure to support stable data production and publishing processes. `Campania Crea` has been integrated as a part of this technological infrastructure to enable the collaborative creation of datasets among the RPA departments. Adopting `Campania Crea`, the Campania Region aims to promote and sustain an increment in the internal data production quality as a crucial step towards ensuring new data-driven services and a significant impact on citizens and the community.

To fully take advantage of `Campania Crea`, it is required to involve both RPA members and citizens and stimulate their imagination and interest, demonstrating them the opportunities to adopt `Campania Crea` to create, modify, discover and exploit open datasets. For this reason, the Campania Region organised the *Open Data Academia Campania* program, i.e., activities and training sessions with strategic groups to share objectives, perform actions to break data silos in our RPA and support G2C collaboration. It mainly focused on RPA members and categories interested in OD, such as learners and journalists as separate sessions.
First, it encouraged RPA members to get aware of the fundamentals and crucial aspects of OD and their management through meetings, workshops, and seminars on data value, data creation, data visualisation. Participants were gently introduced to OD concepts by discussing how data can be defined and reused according to users' background and skills. Then, they were guided to get aware of the OD creation process by focusing on data modelling, licenses, and data exploitation by visualisations to enhance the public information assets. Finally, they learnt how to enhance and guarantee OD sustainability. The evaluation of this chapter reports and discusses results related to a training session involving 54 RPA members.

In March and April 2019, the Campania Region organised two meetings with 300 high school learners from 18 regional institutes involved in a project concerning open cohesion, which uses OD to perform civic monitoring and communication actions.

In April 2019, the Campania Region organised a seminar ending with a hands-on session with 100 professional journalists to demonstrate to them how to use data exploitation tools offered with `Campania Crea` in data journalism activities.

From May to June 2019, the Campania Region promoted a contest named "Represent your dataset!" to encourage learners, citizens, and data enthusiasts to choose one of the regional datasets and represent it through the data visualisation mechanism offered within the regional CKAN, which is the same data exploitation mechanism implemented in `Campania Crea`. This activity resulted in a large engagement of citizens in exploiting and taking advantage of data.

In October 2018, the Campania Region presented the "Campania Open Data" project in Brussels, at the region headquarters, during a meeting scheduled in the European Week of Regions and Cities European Commission and attending European regions. Moreover, it was also presented in May 2019 during the event "Europe in my region" dedicated to best practices within the European Regional Development Fund. The aforementioned project was inserted in the Department of PA's catalogue of experiences within the Steering Committee to coordinate interventions for strengthening the administrative capacity of a PA (Thematic Objective 11 - OT11) and implementation of the Digital Agenda (Thematic Objective 2 - OT2). Thus, it represents an inspiring experience recognised at the national level as a model for implementing an intervention to strengthen administrative capacities and realise RPA digitisation. It figured among the finalists of Open Government Champion 2019 (in the Transparency and OD category) by radically changing the Campania Region ranking at a national level for OD management processes maturity.

7.2 THE MODIFIED OPEN DATA PRODUCTION PROCESS

The representation of a PA as a platform is linked to the existence of a participatory ecosystem that enables third parties to co-produce public services [43, 245]. PA as a platform is based on a stable centralised core, such as platforms [18] integrated with various ecosystems modelling different domains existing in the PA scope. The centralised platform contains all the needed regulations, policies, services, security, quality, and privacy aspects, to favour productive ecosystems where PA members and external actors can co-produce public services [139].

The Campania Region OD infrastructure, graphically represented in Fig. 7.1, is delivered in a hybrid cloud. It represents the data production infrastructure which has been modified to introduce `Campania Crea`. Some components are deployed in the RPA datacenter and others in the environment of a Cloud Service Provider. The infrastructure is designed with a modular architecture satisfying decomposition and modularity requirements [68].

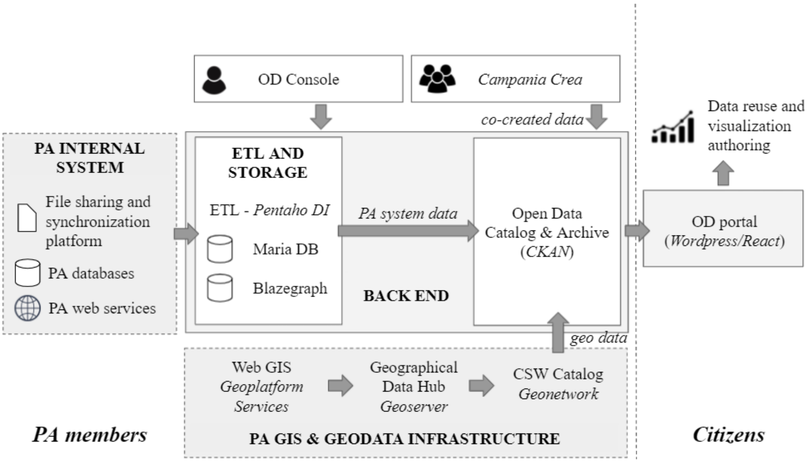

OD Console
Campania Crea
co-created data
Data reuse and visualization authoring
PA INTERNAL SYSTEM
File sharing and synchronization platform
PA databases
PA web services
ETL AND STORAGE
ETL - Pentaho DI
Maria DB
Blazegraph
PA system data
BACK END
Open Data Catalog & Archive (CKAN)
OD portal (Wordpress/React)
geo data
Web GIS Geoplatform Services
Geographical Data Hub Geoserver
CSW Catalog Geonetwork
PA GIS & GEODATA INFRASTRUCTURE
PA members
Citizens

Figure 7.1: OD management architecture within the Campania Region.

The *PA Internal system* corresponds to file sharing and synchronisation platforms, databases, web service end-points and provides access to PA internal data.

The *OD Console* represents a single access point from which the authorized high-level structure employees and RPA office members can upload datasets or modify existing ones, manage related metadata, such as data sources, load new sources or define REST services to expose data.

Both PA internal data and datasets produced by OD console behave as input for the extract, transform, load (ETL) *and storage* component. First, users can manually clean data by improving their syntactic and semantic data and, then, data are automatically transformed and stored as OD thanks to an ETL server, the Pentaho Data Integration. The ETL workflow leads to meeting the scheduled publication frequency. While relational Database (DB)s are stored in Maria DB, LOD are stored in Blazegraph DB, which supports RDF/SPARQL APIs.

Datasets can also be generated by `Campania Crea` that offers co-creation features and OD exploitation and reuse. RPA employees can create thematic online communities and socially interact in co-creation rooms as an opportunity to collect and produce data collaboratively. The implemented OD co-production mechanism enables real-time collaboration as multiple actors can concurrently modify data under the definition. Moreover, `Campania Crea` also offers quality-checks mechanisms within the platform to deal with the lack of quality that obstacles to data exploitation [113]. Finally, co-created datasets are published as OD. The same publication process is followed both for data returned by the ETL and storage component and geo data resulting from PA Geographic Information System (GIS) and geodata infrastructure.

The *PA GIS and geodata infrastructure* component is based on the Free and Open-Source (FOSS) GeoNetwork application, which is a cataloguing application for spatially referenced resources. The geographical data, hosted on a geo-server, are harvested and stored on a catalog, i.e., the Catalogue Service for the Web (CSW), dedicated to the management of resources with spatial references.

Finally, the OD Catalogue and Archive (CKAN) harvests data. CKAN is an Open Source platform for managing, publishing, and researching OD-based archiving components. Datasets harvested or produced by heterogeneous workflows, e.g., manual co-production, CSW catalog harvesting, ETL processes, are published as OD respecting the national DCAT-AP metadata guidelines.

The CKAN of Campania Region behaves as a source for the OD portal accessible by citizens. The OD portal corresponds to the front-end layer, and it is a WordPress component that enables content management concerning events, news, latest publications, published datasets overview, and their details. Moreover, datasets behave as input for the data reuse and visualisation authoring mechanism to create reusable and dynamic data visualisations.

Therefore, users are guided in exploiting available datasets by a data visualisation workflow to move from raw data to data visualisations. The user interface is developed using React to create the catalogue front-end layout using the API services exposed by CKAN.

7.3 TOOLKIT: CAMPANIA CREA

This section describes requirements defined by our RPA to create or customise a platform able to support RPA members in collaboratively creating, publishing, and exploiting high-quality OD supporting both G2G and G2C collaboration.

To satisfy the Campania Region requirements, we propose to adapt SPOD, a Social Platform to create and publish high-quality OD collaboratively, as a back-office platform. SPOD is already adopted by the third sector to co-create and exploit OD as a supply chain platform [68]. In this article, we propose SPOD as an internal platform [68] for the Campania Region internally referred to as Campania Crea. To make evident how Campania Crea satisfies RPA requirements, we make a parallel between RPA requests and how Campania Crea satisfies them.

A demo of the original SPOD is online available for free via registration at http://spod.routetopa.eu/, accessible also by a mobile application. All the source code, as well as documentation, is published on GitHub at https://github.com/routetopa/spod.

A USER-CENTRIC PLATFORM. According to the Open Government requirements [237], PAs achieve transparency by publishing OD while guaranteeing participation and collaboration via active involvement of heterogeneous stakeholders. Information belonging to PAs should be shared, interpreted, and organised to clarify and discuss its meaning and achieve transparency. To encourage participation and collaboration, PAs require a user-centric platform where data curators can easily create and exploit data.

SPOD is a social platform based on OD that encourages collaboration and naturally enables discussions thanks to its social environment. It is worth noticing that we refer to SPOD by calling it a social platform and not a social network, as it cannot be considered a general-purpose social network, but it is a social platform founded on OD where data curators can co-create data, discuss and exploit them. To support co-creation, SPOD enables the definition of co-creation rooms where interested and authorised users can access and contribute to the dataset population. About data consumption, SPOD supports data-driven discussions in public rooms, named agoras, where users can exploit datasets or their visualisations as evidence in discussions [67, 95].

ROLES AND RESPONSIBILITIES. While guaranteeing transparency, security, interpreted as unauthorised access, should not be compromised. Hence, PAs require the possibility to provide access only to trustful users. Consequently, our RPA requires distinguishing roles and responsibilities, as detailed in a previously published motivating use case by the same research group [109]). For this reason, `Campania Crea` implements an agile orchestration by distributing roles to users [109]. To each role, tasks and responsibilities are attached, as reported in the Chapter 3. The implemented roles are:

- *creator*, the expert in the field, is in charge of defining constraints to the dataset under the definition. In a nutshell, besides defining the dataset structure, the creator can also attach data type constraints to each column or their subset. Based on the chosen data type, users are guided in specifying additional constraints.

- *filler*, the dataset populator that insert data within the dataset. Once the creator has defined the data constraints as described so far, the filler is guided in specifying syntactically correct content. The filler role can be further distinguished in *advanced* and *plain* filler for security reasons.

- *validator*, the legal representative of the dataset who is in charge of accepting or discarding rows proposed by fillers but not confirmed yet. By assigning the validator role to experts in the field, this manual check avoids, or at least limits, semantic mistakes.

PA members can play the creator and validator roles, while they can rely on the participation of several different stakeholders by asking them to play the filler role.

DATA QUALITY AND PRIVACY REQUIREMENTS. Limited data quality is a common barrier in data exploitation, mainly, as it takes time to clean data and obstacles to data integration. Therefore, it is crucial to care about data quality during the data creation or publication phase to guarantee the shared data's immediate exploitation Thus, `Campania Crea` offers both a proactive and a reactive quality mechanism.

As a proactive mechanism, the dataset can be attached to data types constraints to simplify the dataset population. More in detail, starting from a dataset, users who play the creator's role can define a form, as described in the Chapter 3. Once confirmed the form, all the users who play the filler's role are guided in filling in the dataset by an advised template. The form prevents the insertion of syntactically wrong data. Therefore, it represents a proactive quality assurance approach. Moreover, as stated before, the validator can also avoid semantic mistakes.

`Campania Crea` also offers a reactive mechanism to assess both quality and privacy concerns by parsing datasets, even if partially completed, as described in the Chapter 4.

Regarding security concerns, our RPA received a dedicated instance of `SPOD`, completely independent by any other instance distributed to other PAs and organisations. In this way, stakeholders and partners have a security and privacy guarantee. Moreover, it guarantees effectiveness since a dedicated platform avoids misleading topics and focuses on specific discussions.

DATA AND METADATA PUBLICATION. Data publishers, in general, and PAs, in particular, require to be interoperable with standard data portal solutions to simplify data sharing, to be compliant with data quality requirements, to guarantee data discoverability, and, consequently, maximise data exploitation. Among data portal solutions, `SPOD` is completely interoperable with CKAN. Once registered to CKAN, users can update and refine datasets by completing them with metadata, e.g., title, description, revision history, licence, tags. CKAN provides many features for end-users, such as multi-lingual support, full-text search. However, data stored on the OD portals are also valuable for PAs because they have one central repository for their data accessible to everyone within the administration. It leads to breaking silos and avoiding centralising data by enabling data sharing within the PA and with any interested user.

DATA EXPLOITATION MECHANISMS. Besides producing and publishing data, OD should be enriched with further information or presented in a fascinating way to catch citizens' and public agencies' interest, encouraging data enthusiastic about using them for analysis, inquiry, economic, civic, monitoring purposes, or any other application context. Behind each dataset, there can be one or more stories to tell, stories about the strategy of a public institution, performed actions, achieved objectives and still open gaps, social phenomena, strengths, and resources of a territory.

In the OD context, data visualisation has the potentiality to be a powerful tool to understand, interpret and get insight into datasets' content. Data tell more things when represented as visual communication and have a higher impact and effectiveness than a textual one as "*Data displayed in a chart rather than a table are easier to understand and trends or patterns are easier to identify*" [314]. End-users can be interested in accessing OD for various motivations. For example, businesses in road maintenance could be interested in accessing road data or statistics in subcontracting, start-ups can reuse data to propose novel services, journalists may exploit data visualisations as evidence in their articles.

Visualising tabular datasets may require technical skills in data exploitation. Thus, SPOD offers a guided workflow to support non-experts users, such as citizens, and expert ones, such as data journalists, in visualising datasets by the most appropriate visualisation options without requiring any technical skill [194]. Users are guided in selecting columns of interest, performing data manipulation or transformation [147] and, finally, visualising them starting from a dataset. During the data manipulation phase, users can filter and aggregate rows by a form-based interface. For instance, by drop-down menus, users are guided in selecting a column to group by the table content, choosing the aggregation function, and filtering only rows compliant with the desired pattern. Technically speaking, this interface masks the creation and execution of SQL queries on the dataset. Once the dataset is ready to be visualised, users are guided in choosing charts and visualisations options. They can choose a bar, column, pie chart, or a histogram to compare values and attributes; tree-maps to render hierarchies and relationships; timelines to explore temporal information or maps to represent geographical details; audio and media sliders to render media content; tables to list results. SPOD assists users in selecting visualisation options that are compatible with the given dataset. The visualisation process is a cyclic task where users can iteratively modify fields of interest, visualisation, and parameters until they are satisfied.

The same guided workflow is directly accessible from our RPA CKAN. Consequently, both data curators and end-users can access the same exploitation mechanism to access datasets of interest and visualise their content.

The achieved visualisation can be downloaded as a traditional image or embedded in any web page, e.g., blogs, institutional websites, forums, as a dynamic, interactive, and real-time visualisation. As an interactive visualisation, users can interact with chart items by zooming in and out in the data or asking for additional information. As a real-time representation, it always visualises up-to-date data. In other words, when the visualisation is created, SPOD does not store the output of the visualisation process, but it keeps track of the dataset source and all the needed information to rebuild the desired visualisation. If the dataset is updated, the visualisation will be automatically updated. Furthermore, the proposed data representation mechanism guarantees data provenance enabling interested users in verifying the trustfulness of information and data source.

As a use case to underline the expected advantage in supporting data visualisation mechanism, we recall the contest “Represent your dataset!" where citizens and data enthusiasts have been encouraged to exploit open datasets to create engaging data visualisations. The winners exploited the dataset concerning air quality to discuss damages induced by air pollution. In particular, they studied the adopted mechanism to monitor air quality by considering data collected via air quality control units and their geographical position. As a result, they created a map representing the geographical distribution of control units in the Campania region and identified municipalities corresponding to the highest monthly mean value. Data also enable comparison. For instance, participants compared air quality data concerning different Italian regions to identify the areas most at risk and any correlations between air pollution and respiratory diseases.

7.4 EVALUATION: ACCEPTANCE ASSESSMENT

This section describes the performed evaluation to assess the technology acceptance level of `Campania Crea` within the Campania Region, reports, and discusses collected results concerning a training session that took place in April 2019 and involved 54 RPA members.

7.4.1 *Evaluation design*

METHODOLOGY. The performed evaluation replies to the Research Question (RQ) *"to what extent our RPA members accept* ***`Campania Crea`*** *for creating, refining, and exploiting OD?"*

This RQ can be split into two sub-questions:

RQ1 What is the acceptation level of `Campania Crea` in creating and exploiting OD?

RQ2 How do RPA members perceive the `Campania Crea` exploitation to perform basic OD creation and visualisation tasks?

PARTICIPANTS. 54 RPA members belonging to the Campania Region voluntarily joined the training session without being compensated for taking part. They were informed that all the collected information remains confidential and is stored in an anonymous form.

Table 7.1 reports participants' demographic details. Most of the sample is male (67%), and all of them are over 40. Participants belong to 3 different PA fields, i.e., technical, administrative, and communication sectors. This implies that there is interest in a multidisciplinary and multi-departmental collaboration which behaves as a step forward to break data silos. Moreover, it evidences the interest in investing in OD in different crucial sectors related to the internal organisation, i.e., administration, in the technical and technological field, and in the communication area, curating the communication with external partners and citizens.

Table 7.1: Participants' demographics, ICT and OD experiences.

	Number	Percentage (%)
Total Participants	54	
Gender		
Male	36	67
Female	18	33
Age		
41-45	11	20
46-50	13	24
>50	30	56
PA field		
Technical	23	43
Administrative	21	39
Communication	10	18
ICT experience		
Inexperienced	3	5
Beginner	22	41
Competent	14	26
Advanced	9	17
Expert	6	11
OD expertise		
Inexperienced	6	11
Beginner	31	57
Competent	8	15
Advanced	8	15
Expert	1	2
Previous OD querying experience		
Yes	21	39
No	31	57
Don't know	2	4
Previous OD visualisation experience		
Yes	20	37
No	33	61
Don't know	1	2

Fig. 7.2 graphically represents the comparison between the auto-assessed OD expertise and the RPA field. Administration members considered their expertise at most as advanced by mainly considering themselves at a beginning level. The same pattern can be observed in the communication field. As probably it may be suspected, users who defined their knowledge at an expert level are technicians. Interesting to notice that also among technicians, PA members consider themselves as beginners. It underlines that no field can be left behind in training sessions concerning OD production and exploitation.

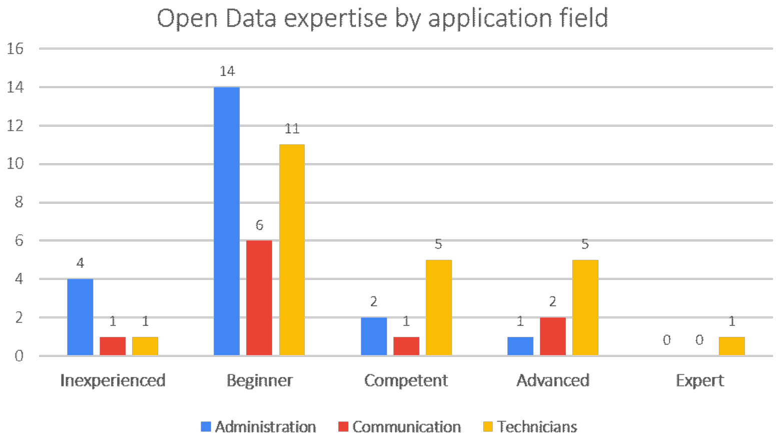


Figure 7.2: Comparison between OD expertise and the PA field

We compared the OD expertise level and previous experiences using OD querying and visualisation tools (Figure 7.3). Inexperienced users never used OD tools. Most of the participants considered their OD experience at a beginning level, and only one out of three beginners used a tool for querying or visualising OD. It can be justified by the consideration that the OD expertise auto-assessment is interpreted as general knowledge on the OD creation and exploitation concepts. One out of three participants considered at least "*competent*" in creating and exploiting OD, even if some of them has never used OD querying or visualisation tools.

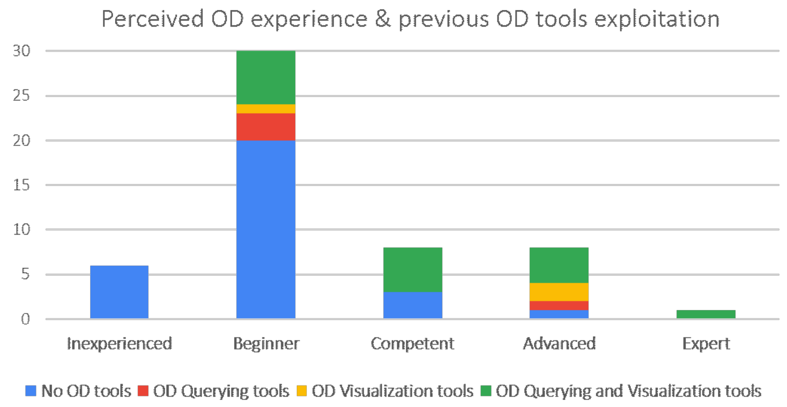


Figure 7.3: Comparison between OD expertise and previous experiences with OD querying and visualisation tools

PROTOCOL. The training session and the related evaluation have been conducted within the Campania Region, spanning over three days, 4 consecutive hours per day. Each session introduced a `Campania Crea` feature, describing the functionality, showing how to perform it within the platform, and describing use cases in which PA members may take advantage of that feature. Then, ICT experts behaving as moderators moved to the hands-on sessions focusing on a task concerning the overviewed feature, and each participant completed it autonomously. However, participants were allowed to cooperate and ask for ICT experts' help during the entire hands-on session. Introduced features concerned:

- *OD creation and modification*. The moderator overviewed how to generate a co-creation room, import an incomplete dataset in CSV format, and modify it in the co-creation room.

- *Quality check.* The moderator described how to perform quality checks to detect typos and inconsistencies in the dataset under definition automatically.

- *Data publication.* The moderator showed how to attach metadata to the complete dataset and publish it internally.

- *Data visualisation.* The moderator demonstrated how to visualise datasets both as a data exploitation mechanism and as an approach to visually detect errors and inconsistencies.

DATA GATHERING. As made evident both by the RQ in its original formulation and its reformulation in RQ1 and RQ2, the evaluation concerns the assessment of the acceptance level of `Campania Crea` to create and exploit OD in the Campania region. As part of the strategic flagship project performed by our RPA, ICT experts representing the `SPOD` developers were invited to organise and conduct a training session for RPA members who voluntarily joined. The evaluation is based on the technology acceptance model (TAM) [79], a theoretical construct widely used to assess users' behavioural acceptance and intentions when accessing a new technology or system. TAM includes perceived usefulness and perceived ease of use, where perceived usefulness refers to *"the degree to which a person believes that using a particular system would enhance his or her job performance*", while perceived ease of use refers to *"the degree to which a person believes that using a particular system would be free of effort*". A system that is perceived to be easy to use is also likely to be accepted by users [79]. TAM model has been extended by introducing playfulness and attitudes to use over time [217]. We take into consideration the Perceived Usefulness (PU), the Easy of Use (EOU), the Periceived Playfulness (PP), and the Attitude Toward Using (ATU).

At the beginning of the first day and the end of the last day of the training session, participants were invited to fill in a questionnaire composed of two main parts. The first part concerned 1) demographic details, i.e., gender and age, 2) ICT expertise, 3) OD expertise, and previous experiences in OD querying and visualisation tools. The second part was based on TAM. The remaining sections only consider results collected at the end of the training session as most participants completed it.

During the hands-on sessions, the moderator asked participants to complete tasks (by detailing all the required subtasks) as reported in Table 7.2, which tested the acquired experience in all the introduced features.

At the end of each task, participants were invited to reply to a structured questionnaire, referred to as ASQ [179] concerning 1) the degree of the perceived difficulty of the task by performing it through Campania Crea, 2) if the time to complete the task is reasonable, 3) if the provided knowledge in the training phase is sufficient to complete the task. Finally, they also reported the required time to complete the task.

Table 7.2: Tasks provided during the evaluation phase

Tasks	**Task description**
Task 1	***Creation of High-Quality OD***
	Create a co-creation room.
	Upload a given CSV file in the co-creation room.
	Perform quality checks to detect and correct errors.
Task 2	***Geo-localized data visualisation***
	In a co-creation room, create a map.
	Create a map by filtering data related to Naples.
Task 3	***Data manipulation and visualisation***
	By filtering and visualisation options, returns how many protected areas are in our region.
	By filtering and visualisation options, returns how many protected areas are in each province of our region.
	Create a regional map to represent the number of protected areas for each province in our region.
	Create a word cloud to visualize different types of protected areas.

7.4.2 *Results*

This section reports results related to the TAM and ASQ questionnaires. While results related to TAM replies to RQ1, results related to ASQ replies to RQ2.

7.4.2.1 *TAM Results*

We checked the internal consistency of multi-item scales using Cronbach's alpha [70], and we reached an alpha value greater than 0.9 in all the metrics (see Column Cronbach's alpha in Table 7.3), proving the consistency of the presented results. Table 7.3 also reports TAM questionnaire results in general and for each involved RPA field. We computed the minimum, the mean value and its standard deviation, and the maximum level for each group. While the minimum and maximum levels for each TAM section are not relevant at a global level as they cover the entire range, it is interesting to notice that the minimum level changes in some PA fields. For instance, the communication group per-

Table 7.3: TAM questionnaire results

TAM	**Cronbach's alpha**	**Metrics**	**All**	**Admin.**	**Comm.**	**Techn.**
range	0-1		1-7	1-7	1-7	1-7
PU	0.96	**Min**	1.0	1.0	4.0	2.0
		Mean	5.6	5.5	5.6	5.7
		St.Dev.	1.3	1.3	1.3	0.9
		Max	7.0	7.0	7.0	7.0
EOU	0.97	**Min**	1.0	1.0	1.0	2.0
		Mean	5.1	5.3	5.4	5.0
		St.Dev.	1.3	1.3	1.0	1.0
		Max	7.0	7.0	7.0	7.0
PP	0.96	**Min**	1.0	1.0	3.0	1.0
		Mean	5.4	5.4	5.7	5.6
		St.Dev.	1.5	1.5	1.3	1.1
		Max	7.0	7.0	7.0	7.0
ATU	0.93	**Min**	1.0	4.0	1.0	2.0
		Mean	6.0	5.9	6.1	6.1
		St.Dev.	1.1	0.9	0.9	1.0
		Max	7.0	7.0	7.0	7.0

ceived Campania Crea extremely useful as the minimum value of PU for communication is 4. Moreover, the same group underlines that Campania Crea is perceived as a playful approach, probably for the emphasis posed to OD exploitation during the training. Administration demonstrated a remarkable intention to use Campania Crea as the minimum score for ATU is 4 for the administrative group. About the maximum level, there is at least a user in each group that appreciates Campania Crea according to each reflection lens, i.e., in each TAM section. In fact, in each group and for each metric, the maximum value is 7, corresponding to the TAM questionnaire's maximum allowed score. In all the considered metrics, the mean value lies between 5 as 6, achieving the highest scores in ATU, demonstrating the interest in adopting Campania Crea in their daily activities. As the mean value is at least 5 and the maximum level is always 7 in all the metrics, RPA members accept Campania Crea to deal with OD (RQ1).

7.4.2.2 *ASQ Results*

This section reports on the results of the ASQ question items compared with the auto-assessed OD skill level, the previous experience with querying and/or visualisation tools, and the PA field. While ASQ results related to OD auto-assessed level and previous experience in OD tools in reported in table 7.4, ASQ results related to the PA field are reported in table 7.5. As described in table 7.2, task 1 (T1) concerns High-quality OD creation, task 2 (T2) concerns basic visualisation, while task 3 (T3) concerns data manipulation and visualisation. Higher values correspond to higher confidence in completing tasks.

In T1, there are no significant differences among perceived difficulties, required time, and required skills or training grouping results by auto-assessed OD skill level. Differently, in T2 and T3, the higher is the auto-assessed OD skill level better are ASQ results in all the metrics. Concerning previous OD tool exploitation, already in T1, it is evident that participants who already experienced OD querying or visualisation tools are more confident in completing the proposed tasks.

By grouping ASQ metrics by the RPA field, it is interesting to notice that administration members and technicians keep a constant

Table 7.4: ASQ results by participants OD expertise

Task	ASQ	OD skill level			Prev. OD tool use		
		I + B	C	A + E	None	Q || V	Q & V
T1	Q1	5.0 (1.5)	5.1 (1.5)	5.4 (1.6)	4.9 (1.6)	5.4 (1.0)	5.8 (0.6)
	Q2	5.9 (0.7)	5.8 (0.7)	6.0 (1.2)	4.9 (1.5)	5.9 (0.7)	5.9 (0.8)
	Q3	5.6 (0.7)	5.8 (0.4)	6.2 (0.7)	5.2 (1.6)	6.1 (1.1)	6.1 (1.1)
T2	Q1	5.1 (1.4)	6.1 (0.8)	6.1 (0.8)	5.0 (1.6)	5.4 (0.5)	6.1 (0.9)
	Q2	5.1 (1.6)	5.8 (0.8)	6.3 (0.7)	5.0 (1.7)	5.8 (0.8)	6.1 (0.9)
	Q3	5.3 (1.6)	6.0 (1.1)	6.2 (0.4)	5.1 (1.7)	6.1 (0.7)	6.1 (1.0)
T3	Q1	5.0 (1.5)	5.5 (1.0)	5.8 (0.8)	5.0 (1.6)	5.1 (0.7)	5.6 (1.0)
	Q2	4.9 (1.7)	5.8 (0.9)	6.1 (0.8)	4.9 (1.8)	5.6 (0.5)	5.8 (1.0)
	Q3	5.2 (1.6)	6.0 (0.9)	6.2 (0.7)	5.1 (1.7)	6.1 (0.7)	6.0 (1.1)

Table 7.5: ASQ results by RPA field

Task	ASQ	Field		
		Administration	Communication	Technicians
T1	Q1	5.3 (1.3)	5.0 (1.7)	5.3 (1.2)
	Q2	5.4 (1.3)	5.0 (1.7)	5.4 (1.2)
	Q3	5.4 (1.4)	5.5 (1.8)	5.8 (1.4)
T2	Q1	5.4 (1.3)	4.9 (1.7)	5.7 (1.2)
	Q2	5.5 (1.3)	5.0 (1.8)	5.5 (1.4)
	Q3	5.4 (1.4)	5.3 (2.1)	5.7 (1.3)
T3	Q1	5.5 (1.3)	4.0 (1.7)	5.5 (1.0)
	Q2	5.6 (1.3)	4.0 (2.1)	5.5 (1.2)
	Q3	5.6 (1.4)	4.8 (2.1)	5.8 (1.2)

confidence level while increasing complexity tasks. In contrast, the confidence level for communication members decreases while increasing the task complexity. Thus, the communication field members registered more difficulties in completing assigned tasks even if they were the more motivated users. In all groups, the provided limited training and the required skills (Q3) are perceived sufficient to achieve the highest scores in all the tasks.

As the mean value is almost greater than 5 - besides few outliers equals to 4 - in all the metrics, it demonstrates that RPA members invested a considerable effort in learning how to exploit `Campania Crea` in accomplishing their daily tasks concerning OD creation and exploitation (RQ2).

7.5 FINAL REMARKS

The public sector, such as Regional or National PAs, usually has a data siloed structure, which impedes collaboration and multi-disciplinary approaches. Recently, the exploitation of platforms gains an increasing interest to perform public sector transformations. Thus, this chapter explores the GaaP paradigm in a RPA, the Campania Region. Thanks to its strategic flagship project, the Campania Region invested in 2018 and 2019 to completely transform both its organisation and technological support. By focusing on the technical and technological changes, our RPA transformed the internal workflow to produce and publish OD and adopted `Campania Crea` to support RPA members is (co-)creating, refining, publishing and exploiting OD. In particular, this article overviews how `Campania Crea` satisfies RPA requirements and analyses its acceptance rate by the TAM questionnaire. Results are rather positive, highlighting that the Campania Region succeeded in involving RPA members in this revolutionary plan and they positively accept a social platform to collaboratively create OD by enabling multi-disciplinary and multi-departmental G2G collaboration. It represents a step forward in breaking data siloes.

Besides the technical and technological aspects analysed in this chapter, members of Campania Region also noticed a contribution concerning cultural factors related to dissemination of the data culture in the institution, creation of a network of contacts, involvement of students, journalists, and other stakeholders in training activities, the empowerment achieved by the adopted communication means. By focusing on the network of contacts, the community of the Campania Region members has been instrumental in satisfying OD requests, also during the pandemics. During the COVID-19 several stakeholders, such as businesses, professionals, and citizens, asked for datasets concerning production activities, job or training opportunities, pressing, and cultural sites, to cite a few examples. Moreover, everyone was particularly interested in OD concerning public health related to the spread of the COVID-19 pandemic. Consequently, the Campania Region satisfied these requests due to the presence of the internal network of contacts interested in caring OD as a result of the organisational changes. Based on the experience and the developed skills concerning OD creation, manipulation, and refinement, our Region succeeded in easily identifying which member(s) might satisfy each received request. It is worth noting that the Campania Region OD experience has been cited as a good practice in the catalogue of national experiences in an initiative edited by the Presidency of the Council of Ministers and related to the "Strengthening the Administrative Capacity and Digitisation of the PA" by demonstrating the positive and promising effect of the performed strategic flagship project. During this nomination, some representatives of the Campania Region attended a seminar where several local authorities in our Region joined due to their interest in replicating the regional OD project in their administrations. The reuse of the catalogue experiences has no territorial limits and can be activated by all national PAs. During this initiative, the Directorate General's regional official and contact person for social and socio-health policies overviewed organisational and operational phases, which led to the identification, analysis, management, and publication of a series of datasets and their visualisation. It represents a real application context of `Campania Crea` as a collaborative tool to participate in taking care of open datasets and as an exploitation tool in creating data representation. `Campania Crea` results in an effective and decisive means to spread data and information.

Part III

LINKED OPEN DATA EXPLOITATION

8 LINKED OPEN DATA EXPLOITATION

You can have data without information, but you cannot have information without data.

– Daniel Keys Moran

Over the past decades, hundreds of datasets have been published using the Semantic Web standards covering any topical domain [290]. The LOD Cloud [207] (a KG that collects most of the published KGs) counted 12 datasets in 2007 and currently contains 1,239 datasets. Some of these KGs are proprietary, maintained internally by companies such as Google, Microsoft, Apple; while others, like DBpedia and Wikidata, are openly available and maintained by dedicated communities. The central idea of LD is that data publishers support applications in discovering and integrating data by complying with a set of best practices in the areas of linking, vocabulary usage, and metadata provision [290]. Because of the extensive range of heterogeneous information stored in KGs, for their easy navigation, thanks to their quantitative and qualitative properties, they could behave as a critical resource for Information retrieval (IR) and KM.

The KG exploitation is mainly affected by i) required technical skills in query languages (e.g., SPARQL) and in understanding the semantics of the supported operators [329], too challenging for lay users, and ii) conceptualization issues to understand how data are modelled [24, 329].

These drawbacks have led to the development of tools and interfaces to support users in interacting with KGs by implicitly composing queries while hiding the underlying complexity.

My research is situated in this context and proposes approaches and prototyped tools to express users' needs or explore available data by a Natural Language (NL) interface to guide end-users with different interests, types of background, age, and needs to query KGs and take advantage of them without requiring technical skills in query languages. Instead of proposing a unified tool to address the heterogeneity of the target audience, we opt for proposing a unified approach to guide KG exploitation and instantiate it in different interfaces to fulfil specific requests of each target group.

The research presented in this chapter has been published as a Ph.D. consortium in the following contribution:

> Maria Angela Pellegrino: *Knowledge graphs within everyone's means*. PhD consortium in CHItaly 2021.

8.1 A UNIFIED APPROACH

The KM process includes 1) data retrieval, 2) data refinement, and 3) data exploitation (Fig. 8.1). It requires the contamination of IR, Information Visualization, and Human-Computer Interaction.

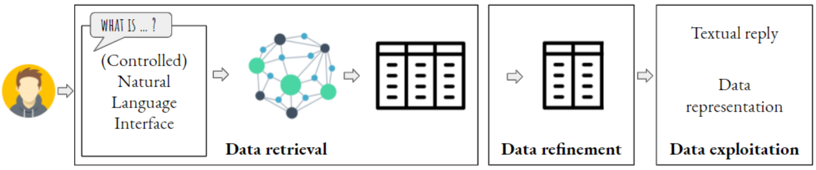


Figure 8.1: Unified approach for KGs exploitation

Data retrieval requires search activities that can be classified in lookup and exploratory search [200]. Lookup is a search task where users know what they are looking for and can formulate it as a direct question, as in QA applications, while an exploratory search task is an open-ended search that usually starts with vague information needs and requires iterative query formulation [138], facets or taxonomies [322], keyword search paradigm which includes auto-suggestions, instant results, partial matches [219, 279] to explore data.

As a data retrieval interface, we propose query builders enhanced with (controlled) NL interfaces to guide users in naturally posing questions by simulating, as much as possible, human interactions. If users have a clear objective, they can directly type or pronounce their requests. Vice versa, in exploratory search, users are guided in iteratively creating and refining questions.

As a query builder, NL queries can be translated to SPARQL to be run over a SPARQL endpoint (i.e., a way to publicly expose

KG content). Among SPARQL constructs, SELECT query results can be naturally represented as tabular data. Thus, retrieved data are modelled as tables, which can be manually or automatically refined, and finally, used in data exploitation mechanisms.

It may result in textual replies or concrete artifacts, perhaps customizable and exportable, such as charts, data visualisations, data stories, or Virtual Reality (VR)-based data representations.

To instantiate the general approach represented in Fig. 8.1 in concrete interfaces to fulfil end-users needs, we considered target audience who might be interested in accessing KGs, and consequently, we designed, prototyped, and evaluated tools to satisfy users' requirements. We focus on OD experts and PAs, education, and the CH community as target groups. It resulted in the prototypes summarised in Fig. 8.2 which will be fully discussed in the following chapters.

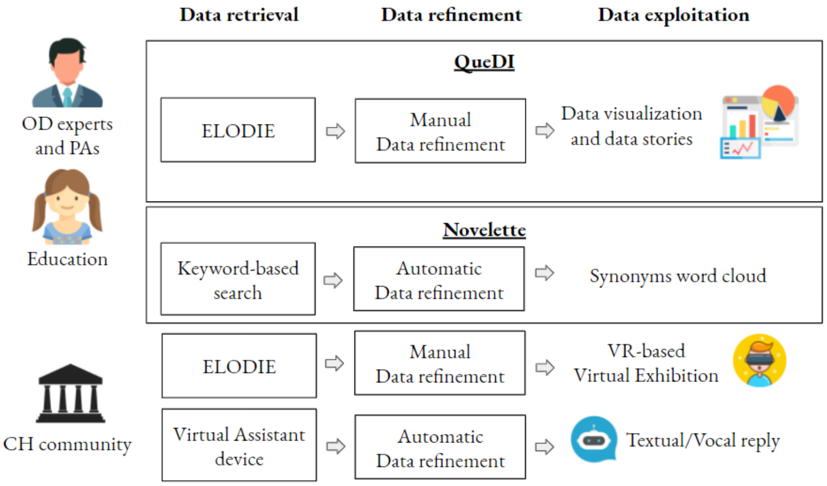


Figure 8.2: KG exploitation prototypes proposed during my Ph.D.

OPEN DATA EXPERTS AND DATA VISUALISATION. Since 2018, my research lab, ISISLab, led an H2020 project, ROUTE-TO-PA, to support citizens and public institutions in publishing high-quality OD and effectively exploiting them. The main outcome of the project is SPOD a Social Platform for OD [67] to co-create and exploit OD. SPOD is used by data producers represented by PAs, associations, communities, or data enthusiastic, such as citizens and learners, which developed OD management skills, such as, table manipulation and chart creation.

The problem we aimed to solve is how to make this target group able to access LOD and KGs without requiring explicit usage of SPARQL with the possibility to rely on their expertise in OD management. We proposed a *transitional approach* where OD experts are guided from LOD querying to table manipulation and chart creation, considered their comfort zone. It resulted in `QueDI` (Query Data of Interest) which allows lay users to build step-by-step queries thanks to an auto-completion mechanism and to exploit collected results by exportable and dynamic visualizations. Data visualisation can also be combined into coherent data stories and can be used as evidence to support discussions, write journalistic articles, spread and disseminate information supported by graphical representations.

EDUCATION AND KNOWLEDGE MANAGEMENT TOOLS. `QueDI` implements a *trialogical learning approach* where end-users can query KGs working individually and learn how to exploit results in reusable data visualisations according to the *monological learning approach*. As `QueDI` has been embedded in a social platform, end-users can easily share realised artefacts with the community, achieving the *dialogical learning*. Finally, users can exploit shared artefacts as a starting point for their research, enabling the possibility to learn by examples, reuse available artefacts and transform them into new knowledge. The latter realises the *trialogical learning approach* where artefacts are collaboratively created thanks to the layering of individual efforts. It represents a powerful KM tool in the educational context to support future citizens in going beyond the passive inspection of results returned by a search engine, and in actively searching for the data that best answer their questions.

EDUCATION AND DIGITAL STORYTELLING. Due to the difficulties in directly poising queries on KGs for lay users, such as (young) learners, we investigated the implicit exploitation of KGs in retrieving synonyms lookup in `Novelette`, a digital storytelling environment where storytellers can create stories to graphically represent tales, data or media stories. If users experience writer's block, Novelette has a suggestion provision mechanism that automatically retrieve definitions, synonyms, and analogies by querying available KGs and organising retrieved results in (navigable) word clouds. It represents a keyword-based interface to implicitly explore KGs by navigating synonyms and analogies.

CULTURAL HERITAGE COMMUNITY AND VIRTUAL EXHIBITIONS. In the last year, virtual exhibitions have been widely adopted to enhance physical tours, but CH lovers still behave as visitors. The CH community requires mechanisms and tools to be actively involved in CH data exploitation, for instance, by playing the role of exhibition curators. As the CH community represents one of the biggest contributors of the Semantic Web, I propose to take advantage of CH KGs in an authoring platform for VR-based virtual exhibitions by combining ELODIE and an automatic mechanism to create VR-based solutions. As a result, the proposed approach lets the CH community to author their virtual exhibitions to disseminate data of interest; they implicitly exploit CH data modelled as KGs without being explicitly aware of the Semantic Web technologies, and virtual exhibitions have the potential to enhance (without substituting) real tours.

CULTURAL HERITAGE COMMUNITY AND VIRTUAL ASSISTANTS. While KGs require mechanisms to pose questions without asking for technical skills, VAs are widely adopted as a natural approach to pose questions by simulating conversations with humans. We investigated how to make VAs compatible with KGs. It resulted in a community shared software framework (a.k.a. generator) that enables lay-users to create ready-to-use custom extensions for performing QA over KGs. This proposal represents a step forward in enabling direct search and lookup over KGs. While commercial VAs are already configured to automatically query KGs based on human requests, our proposal represents a step forward in letting end-users, such as the CH community, in designing and developing VA extensions able to query their data of interest without asking for technical skills in VA extension design and query languages. It may empower local and minor communities to disseminate and make their data easily accessible to others, librarians to implement VA extensions to simplify consultation of archives and book lookup, museum curators in developing customised virtual guides, teachers to develop innovative mechanisms to look up for domain-specific terms, tour curators to propose original virtual guides able to reply visitors' questions instead of reproducing traditional and packaged tracks.

8.2 LINKED OPEN DATA HETEROGENEITY, AVAILABILITY AND LIMITATIONS

This section points out advantages and challenges in using LOD by automatic query building mechanisms. The performed analysis focuses on CH KGs, but can be easily extended to other fields.

It analyses the CH community effort in publishing CH data as KGs, in making them accessible by either SPARQL endpoints or APIs, in maintaining working SPARQL endpoints in most of the cases, in attaching human-readable labels to resources to making them accessible by NL interfaces. The performed analysis makes evident the potentialities of proposing exploitation tools in this application domain due to the vast amount of available data. In particular, this survey quantifies the amount of available CH KGs behaving as a source for the proposed generator and it estimates some of the aspects that are crucial for making data accessible by any data exploitation tool, such as, accessibility by a working SPARQL endpoint, and by NL interfaces, such as VA providers, that require the use of labels attached to resources.

The first advantage is LOD heterogeneity. Second, the wide availability of CH KGs demonstrates why technicians should enable the CH community to easily exploit the vast amount of published data. Concerning challenges, we verify the easiness of querying SPARQL endpoints by `ELODIE`, representing a zero-configuration and general approach to query KGs by exploiting their public SPARQL endpoints.

The performed analysis has been described in the following scientific article:

> Daniele Monaco, Maria Angela Pellegrino, Vittorio Scarano, Luca Vicidomini. *Linked Open Data in Authoring Virtual Exhibitions*. Journal of Cultural Heritage. 2022.

8.2.1 *Selection approach*

It is worth clarifying that we do not aim to provide a complete overview of all published KGs in the CH context, but the described selection process seeks to point out the absence of bias in the selected KGs and, consequently, the impartiality of the considerations reported in the performed analysis.

We performed the KG selection as a non-technical user, by looking at available aggregators of published KGs and querying their user interfaces. We exploit LOD cloud [207] (updated in May 2020) as it is one of the biggest aggregators of published KGs, and a combination of datasets and articles search engines. In particular, we explored datasets aggregators not specifically related to the Semantic Web, such as DataHub [244]. Finally, we considered recent publications available in Scopus to identify also KGs published recently. The variety of queried sources aims to demonstrate the lack of bias in the performed analysis.

Our selection process results in `55` KGs covering `20` countries.

1. We exploited the search mechanism provided in the LOD cloud [207] to retrieve KGs containing *museum*, *library*, *archive*, *cultur**, *heritage* or *bibliotec**. It is worth noticing that the search engine requires that the dataset title includes English terms, but it does not pose any constraint on the provider's country.

2. We retrieved datasets registered in the datahub [244] with format equals to `api/sparql`. We manually inspected the `710` returned datasets by looking for *museum*, *library*, *archive*, *culture*, *heritage*, *bibliography*, and similar terms in dataset title and description. DataHub also returns the SPARQL endpoint attached to retrieved datasets. When the specified endpoint is no more available, we searched the dataset name attached to "SPARQL endpoint" on the Google search engine to determine if any URL migration took place.

3. We inspected the articles indexed by Scopus and matched the article title, abstract, and keyword filter `("cultural heritage" and ("semantic web" OR "linked data" OR "knowledge graph"))` from 2020 to 2018. It resulted in `150` articles. We manually checked them to verify the presence

of a KG publication and if so, we further checked if authors expose a public SPARQL endpoint.

8.2.2 *Heterogeneity considerations*

CH are characterised by a noticeable heterogeneity. In particular, according to the taxonomy of the CH term, it includes *tangible* CH, which can be further specialised in i) movable (such as paintings, sculptures, coins, manuscripts); ii) immovable (such as monuments, archaeological sites), and iii) underwater (shipwrecks, underwater ruins), and *intangible* CH, such as oral traditions, folklore, performing arts, rituals. Moreover, it also encompasses *natural heritage*, i.e., culturally significant landscapes, biodiversity, and geo-diversity. In retrieving CH KGs, we categorised them according to the provided content following this definition. Consequently, we classified KGs as tangible (refined as museums and libraries representing immovable goods and movable ones), intangible CH and natural heritage. Moreover, for each CH KG we identified if it can be accessed by APIs or by a SPARQL endpoint.

Results are reported in table 8.1. For each KG, we report the *original name* and the *category* and *sub-category* data belong to, the *country of the provider* and the *service* that enables data exploitation (SPARQL endpoint or API). For each KG, we also generate a *short name* (mainly combining country and category) to easily refer them in the following analysis. The online version of the same table[1] also provides access to the SPARQL endpoint, if any, and useful links to retrieve KG details. Main observations follow.

1. There is a substantial interest not only in materializing data but also in defining models (mainly tailored to libraries, archives, and museums [91]) and precise terminology by thesaurus (22%). For instance, the `CIDOC-CRM` [55] is a theoretical model for information integration in the field of CH. It can help researchers and interested people in modelling CH collections and documents.

2. Most of the KGs are related to tangible goods, while the effort in intangible ones is limited to a few datasets related

[1] CH KG collection details: `https://bit.ly/ch-kg-collection`

Table 8.1: CH KG overview

Short Name	Name	Category	Sub-category	Country	Service
DBTune_classical	DBTune Western Classical Music	Intangible CH		GB	SPARQL
EventMedia	EventMedia	Intangible CH		FR	SPARQl
FI_folklore	Semantic Folklore	Intangible CH		FI	SPARQL
Munnin	First World War	Intangible CH		CA	SPARQL
MusicKG	MusicKG	Intangible CH		FR	SPARQL
WarSampo	WarSampo	Intangible CH		FI	SPARQL
NaturalFeatures	Natural Features	Natural Heritage		GB	SPARQL & API
ARTIUM	Library and Museum of ARTIUM	Tangible CH	Library & Museum	ES	SPARQL
B3Kat	Bavaria, Berlin and Brandenburg Library	Tangible CH	Library	DE	SPARQL
Bibliopolis	Dutch National Library	Tangible CH	Library	NL	SPARQL
BNB	British National Bibliography	Tangible CH	Library	GB	SPARQL
Cervantes_lib	"Miguel Cervantes" Library	Tangible CH	Library	ES	SPARQL
CL_library	Chilean Library	Tangible CH	Library	CL	SPARQL
DE_library	German National Biographies	Tangible CH	Library	DE	SPARQL
DE_uni_library	Mannheim University Library	Tangible CH	Library	DE	SPARQL
DigitalNZ	DigitalNZ	Tangible CH	Library	NZ	API
DPLA	Digital Public Library of America	Tangible CH	Library	USA	API
ES_library	Spanish National Library	Tangible CH	Library	ES	SPARQL
FI_library	BookSampo (Kirjasampo)	Tangible CH	Library	FI	SPARQL
FR_library	French National Library	Tangible CH	Library	FR	SPARQL
GR_library	National Library of Greece Authority Records	Tangible CH	Library	GR	SPARQL
HEBIS	HEBIS – service for libraries	Tangible CH	Library	DE	SPARQL
Hedatuz	Hedatuz	Tangible CH	Library	ES	SPARQL
JP_library	Japanese National Library	Tangible CH	Library	JP	SPARQL
KR_library	Korean National Library	Tangible CH	Library	KR	SPARQL
LIBRIS	LIBRIS: Swedish National Library	Tangible CH	Library	SE	SPARQL
NL_library	National Dutch Library	Tangible CH	Library	NL	SPARQL
PLV	Public Library of Veroia	Tangible CH	Library	GR	SPARQL
British museum	British museum	Tangible CH	Museum	GB	SPARQL
Europeana	Europeana	Tangible CH	Museum	EU	SPARQL & API
FI_museum	Finish Museum	Tangible CH	Museum	FI	SPARQL
HU_museum	Museum of Fine Arts Budapest	Tangible CH	Museum	HU	SPARQL
IT_museum	Italian museums	Tangible CH	Museum	IT	SPARQL
NZ_museum	Auckland Museum	Tangible CH	Museum	NZ	API
Rijksmuseum	Rijksmuseum	Tangible CH	Museum	NL	SPARQL
RU_museum	Russian Museum	Tangible CH	Museum	RU	SPARQL
Yale_GB_museum	Yale center of British Art	Tangible CH	Museum	GB	SPARQL
USA_museum	Smithsonian Art Museum	Tangible CH	Museum	USA	SPARQL
ARCO	ARCO	Tangible CH	Movable	IT	SPARQL
FondazioneZeri	Zeri Foundation	Tangible CH	Movable	IT	SPARQL
HU_archive	National Hungarian Digital Data Archive	Tangible CH	Movable	HU	SPARQL
Logainm	Placenames Database	Tangible CH	Movable	IE	SPARQL
NL_maritime	Dutch Ships and Sailors	Tangible CH	Movable	NL	SPARQL
Nomisma	Nomisma	Tangible CH	Movable	EU	SPARQL
CIDOC-CRM	CIDOC Conceptual Reference Model	Vocabulary	Model	-	SPARQL
MMM	Mapping Manuscript Migrations	Vocabulary	Model	FI	SPARQL
MONDIS	MONDIS	Vocabulary	Model	CZ	API
AAT	Art & Architecture Thesaurus	Vocabulary	Thesaurus	-	SPARQL
ADL	Alexandria Digital Library Gazetteer	Vocabulary	Thesaurus	USA	SPARQL
ES_thesaurus	Public Libraries Headings	Vocabulary	Thesaurus	ES	SPARQL
FR_archive	Thesaurus for Local Archives	Vocabulary	Thesaurus	FR	SPARQL
GB_thesaurus	English Heritage Thesaurus	Vocabulary	Thesaurus	GB	SPARQL
Loanword	World Loanword Database	Vocabulary	Thesaurus	DE	SPARQL
Thesaurus BNCF	Florentine National Library Thesaurus	Vocabulary	Thesaurus	IT	SPARQL
UNESCO	The UNESCO thesaurus	Vocabulary	Thesaurus	EU	SPARQL

to music, heritage in the event of armed conflicts, folklore, or event (12%), e.g., MusicKG [103].

3. Few KGs provide APIs (8%), while most opt for SPARQL endpoints. Europeana [133] invested in both.

4. European datasets mainly opt for enriching aggregators, e.g., Europeana [133] or well-known KGs. Aggregators host multiple collections, creating an integrated point of access to artifacts of different institutions. As example, the BNB [83] and the PLV [170] are independent by any aggregator. On the opposite side, Rijksmuseum [89] and British Museum[2] opt for being wrapped in Europeana.

This analysis underlines a consistent heterogeneity in terms of i) available content published as KGs, ii) countries behaving as providers, iii) accessing mode. Moreover, it is worth noticing that thanks to their linked nature, KGs behave as a unique point of access to heterogeneous information (images, textual description, multiple language support) that can be exploited to create tailored datasets. Independently from the queried KGs, users can aggregate artifacts by author, date, location, topic among other filtering options. With the following examples, we aim to underline the opportunity that CH lovers have in querying LOD and model any exhibition of interest.

8.2.3 *Technical challenges*

While designing KG exploitation tools, developers may experience several technical challenges. If they aim to query the SPARQL endpoint, they have to check its status, the (timing) performances to avoid time-out errors, and the language support to provide users with an international tool besides users' speaking language.

For each KGs provided with a SPARQL endpoint, we verify the SPARQL endpoint status, the CORS-enable option, the easiness in retrieving classes and predicates, and the English support in this order. As soon as a test fails, we stop the exploration chain. Some considerations follow. A detailed description of the considered parameters follows.

[2] British museum: http://collection.britishmuseum.org

- *The SPARQL endpoint status.* Any KG exploitation approach that relies on SPARQL endpoints as accessing mechanism requires the assessment of the SPARQL endpoint status.

- *CORS-enabled* SPARQL endpoint. CORS is a specification enabling truly open access across domain boundaries [333]. Client-side scripts are prevented from accessing much of the Web of LOD due to same-origin restrictions implemented in all major Web browsers. While enabling such access is important for all data, it is crucial for LOD, and related services [333]. Without this, data are not open to all clients. An option to deal with the CORS-enable option requires the exploitation of a proxy, as in OD exploitation tools [67].

- *Possibility to retrieve classes and predicates.* Generic and standard queries posed against the configured SPARQL endpoint by HTTP GET requests can be used to retrieve both classes and predicates. In verifying the easiness in retrieving available data, we queried used classes (i.e., classes behaving as objects in triple patterns of the form `?instance a ?class`) and used predicates (i.e., predicates `p` in triple patterns of the form `?s ?p ?o`) even if they are not explicitly defined in the ontology.

- *English support.* Data curators that expose also English labels attached to the published resources demonstrate the interest in being interoperable with the international community. In particular, we verified if at least a subject of any KG triple is attached to an English label.

Table 8.2 reports the assessed technical challenges that any KG exploitation tool has to face in automatically querying (CH) KGs.

As underlined by the SPARQL endpoint status in Table 8.2, there is a discontinuous effort in maintaining SPARQL endpoints or the lack of attention in updating the dataset search engines when a SPARQL endpoint URL migration took place. Fig. 8.3 represents the available SPARQL endpoints for each country and makes evident their status: 6 countries (e.g., Finland) manifest continuous maintenance of SPARQL endpoints, while 4 countries (e.g., the USA) stop maintaining their SPARQL endpoints.

Table 8.2: CH KG challenges

	SPARQL endpoint	CORS-enabled	Retrieved	English support
			classes & predicates	
AAT	✓	✓	*empty results*	✓
ADL	✘			
ARCO	✓	✓	✓	✓
ARTIUM	✘			
B3Kat	✓	✘		
Bibliopolis	✘			
BNB	✓	✓	*empty results*	✓
British museum	✘			
Cervantes_lib	✓	✘		
CIDOC-CRM	✘			
CL_library	✓	✓	✓	✘
DBTune_classical	✘			
DE_library	✘			
DE_uni_library	✘			
ES_library	✓	✘		
ES_thesaurus	✓	✘		
Europeana	✓	✓	✓	✓
EventMedia	✓	✓	*predicates fail*	✓
FI_folklore	✓	✓	*restricted access*	✓
FI_library	✓	✓	*restricted access*	✘
FI_museum	✓	✓	*restricted access*	✓
FondazioneZeri	✓	✘		
FR_archive	✘			
FR_library	✓	✘		
GB_thesaurus	✓	✘		
GR_library	✘			
HEBIS	✘			
Hedatuz	✘			
HU_archive	✓	✓	*predicates fail*	✓
HU_museum	✘			
IT_museum	✘			
JP_library	✓	✘		
KR_library	✓	✘		
LIBRIS	✘			
Loanword	✘			
Logainm	✓	✘		
MMM	✓	✓	*restricted access*	✓
Munnin	✓	✓	✓	✓
MusicKG	✘			
NaturalFeatures	✓	✓	✓	✓
NL_library	✓	✓	✓	✓
NL_maritime	✓	✘		
Nomisma	✓	✓	*unavailable service*	
PLV	✘			
Rijksmuseum	✘			
RU_museum	✘			
Thesaurus BNCF	✓	✘		
UNESCO	✓	✓	✓	✓
USA_museum	✘			
WarSampo	✓	✓	*restricted access*	✓
Yale_GB_museum	✘			

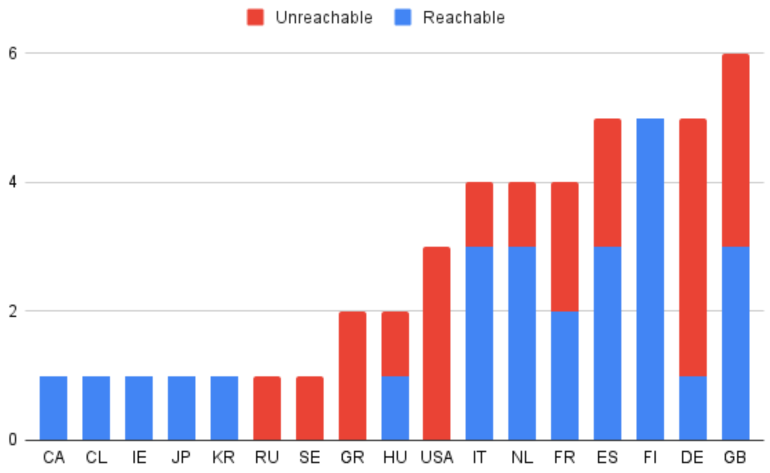


Figure 8.3: Distribution and status of SPARQL endpoints over countries

This discontinuous effort obstacles KG exploitation by standard mechanisms and required KG exploitation tool developers to check the status of queried SPARQL endpoints periodically.

Concerning the English support, the `CL_library` successfully passed all the tests, but it does not provide any resource with an English translation. It is not an isolated case. The `Cervantes_lib` only provides access to Spanish labels, `FL_library` only to Finnish results, `Thesaurus_BNCF` only supports Italian labels.

8.2.4 *Availability considerations*

In this section, we focus on KGs exposing a working SPARQL endpoint correspond to a ✓ in the SPARQL endpoint status column in the Table 8.2. We query them to quantify the amount of published and accessible data. Table 8.3 quantifies their size in terms of declared classes, relations, triples, and class instances to estimate the availability of CH data that may be queried by KG exploitation tools, such as `ELODIE`. To retrieve the available classes, we queried all the classes attached to any subject by the `rdf:type` property. While the returned classes are counted as classes, all the distinct retrieved subjects are counted as class instances. To count properties, we collect all the distinct properties used in at least one KG triple. Queries are executed directly on the SPARQL endpoint. If the SPARQL endpoint does not support the COUNT operator, results are downloaded and manually counted.

Even if all the listed SPARQL endpoints are online, some of them (completely or partially) fail in returning query results due to timeout (empty values in Table 8.3).

By looking at the amount of available data, the used classes are of the order of hundreds. Besides a single outlier represented by `Munnin`, also the used predicates are of the order of hundreds. Concerning available triples, independent CH KGs (without considering aggregators, such as `Europeana`) provide access to roughly $400M$ triples that can potentially be queried by any data exploitation tool (`ELODIE`, among others). Some SPARQL endpoints, i.e., the ones corresponding to values in italic in Tab. 8.3, pose a results limit requiring multiple queries to retrieve all data.

It is also interesting to notice the gap between the minimum and maximum value of each column underlying the differences in size any KG exploitation tools should be able to deal with.

As a conclusive consideration, the CH community paid a substantial effort in publishing their content in LOD format, as demonstrated by the amount of retrieved KGs (see Table 8.2) provided with a SPARQL endpoint and the amount of available data (see Table 8.3). Thus, it is crucial to support this community by providing them with KG exploitation tools. However, developers have to pay attention to technical issues that may arise in proposing a SPARQL query builder to unlock the potentialities of (CH) KGs.

Table 8.3: Availability of CH KG data.

Short Name	class	relation	triple	instance
AAT	74	29	77,556,215	**35,056,063**
ARCO	396	803	216,433,045	
B3Kat	31			
Cervantes_lib	22	128	16,060,445	18,212,48
CL_library	11	32	12,310,613	674,773
ES_cultura	31	354	867,535	51,334
ES_library	27	308	289,777,525	20,916,789
ES_thesaurus	2	14	1,757,437	175,787
Europeana	11		**2,836,270,332**	
EventMedia	39	188	6,377,133	240,999
FI_folklore	17	43	306,549	28,824
FI_library	61	114	4,363,198	486,689
FI_museum	18	60	210,986	13,485
FondazioneZeri	105	124	11,827,416	
FR_library	93		378,356,947	
GB_library	46	174	197,418,446	14,723,083
GB_thesaurus	123	52	*500*	*500*
HU_archive	**469**	295	48,378,455	4,211,734
JP_library	4	36	*100*	*100*
KR_library				
Logainm	114	170	1,344,903	212,901
MMM	58	152	22,472,633	3,274,463
Munnin	181	**18,136**	*2,000*	*2,000*
NaturalFeatures	322	135	918,664,981	
NL_library	30	128	167,056,240	22,237,016
NL_maritime	92			137,7634
Nomisma				
Thesaurus BNCF	7	38	728,272	65,499
UNESCO	8	40	84,911	4,250
WarSampo	90	310	14,322,426	1,797,027

9 KNOWLEDGE GRAPHS AND DATA VISUALISATION

> *A picture is worth a thousand words.*
>
> – Henrik Ibsen

During the past years, several different approaches have been proposed to hide the complexity of SPARQL and enable query building. Users are provided with graph-like query builders (such as FedViz [350], RDF Explorer [328]), visual query builders (e.g., OptiqueVQS [307]), facets based interfaces (e.g., SemFacet [13]) also enhanced by keyword search interfaces (such as SPARKLIS [107] and Tabulator [29]), query completion mechanisms (such as YASGUI [270]), summarization approaches (such as Sgvizler [300]), or a combination of them. The expressiveness of the query method can be affected by the interaction model, required usability, efficiency.

Once retrieved results of interest, users should be guided also in exploiting them according to their needs. This section focuses on the exploitation of collected data by visualising them. In particular, this section presents the design and implementation of a guided workflow to move from KG querying to data visualisation. It results in `QueDI` Query Data of Interest, whose design mainly relies on abilities in data table manipulation and chart creation. It is not a strong limitation since many data visualisation tools start from CSV files or, in general, data tables. We refer to this target users as *experts in data table manipulation*, and we aim to guide them in manipulating LOD through their tabular representation.

The research at the basis of this chapter has been published in the following contributions:

- Renato De Donato, Martina Garofalo, Delfina Malandrino, Maria Angela Pellegrino, Andrea Petta, Vittorio Scarano: *QueDI: From Knowledge Graph Querying to Data Visualization*. In the Proceedings of SEMANTiCS 2020.

- Renato De Donato, Martina Garofalo, Delfina Malandrino, Maria Angela Pellegrino, Andrea Petta, Vittorio Scarano: *Linked Data Queries by a Trialogical Learning Approach*. In the Proceedings of Computer Supported Cooperative Work in Design (CSCWD) 2019.

- Renato De Donato, Martina Garofalo, Delfina Malandrino, Maria Angela Pellegrino, Andrea Petta: *Education Meets Knowledge Graphs for the Knowledge Management*. In the Proceedings of International Conference in Methodologies and Intelligent Systems for Technology Enhanced Learning (MIS4TEL) 2020.

9.1 RELATED WORK

Table 9.1 overviews tools to query LOD and visualise retrieved results by schematically comparing the query building mode, reached expressiveness, and the need for SPARQL awareness by users, supported visualisation modes, and customisation and export options, if available. The reported tools propose heterogeneous exploitation modes, spanning from facets and forms to graph, keyword-based to NL interface when SPARQL complexity is masked, while they implement text-based interfaces when users have to explicitly formulate SPARQL queries. Among data visualisation options, timelines and maps are widely used to graphically organise temporal and geographical information.

Table 9.1: LOD querying and visualisation tools comparison

Tool	**Year**	**Query Builder**			**Visualisation**		
		Mode	**Expressiveness**	**SPARQL masking**	**Mode**	**Custom**	**Export**
TABULATOR	2006	facet	Path Traversal	✓	time, map		
NITELIGHT	2008	graph	SPARQL 1.0−	~	time, map		
VISINAV	2010	facet+ keywords	Basic Graph Pattern (BGP)s	✓	time, map		✓
Sgvizler	2012	text	SPARQL	×	Google Charts	✓	✓
VISU	2013	text	SPARQL	×	Google Charts	✓	✓
Visualbox	2013	text	SPARQL	×	chart, map, time	✓	✓
rdf:SynopsViz	2014	form	BGPs−	✓	chart, treemap, time		✓
YASGUI	2017	text	SPARQL 1.1	×	Google Charts	✓	✓
SPARKLIS	2018	facet+NL	SPARQL−	✓	Google Charts + map, image	✓	✓
WQS	2018	form	BGPs	✓	chart, map, time, image, graph	✓	✓
QueDI	2020	facet+NL	BGPs+	✓	chart, time, image, map	✓	✓

Tabulator [29] leads to query and modify KGs without being aware of SPARQL. Users interact with a Faceted-search interface (FSI) where pairs of predicates and objects pairs are retrieved and organised for each element behaving as subject. Users performs path traversal recursively choosing element by element and follow the available paths. Besides the tabular representation of retrieved results, Tabulator offers basic visualizations: if results contain temporal or geographical information, users can create timelines or maps. Authors do not mention if the authored visualisation can be either customised or exported.

NITELIGHT [280] support SPARQL query building by a graphical interface. However, authors explicitly state that it is intended for users that already have a SPARQL background since the complexity is not masked during the query definition. Queries are formulated by a keyword-based interface that supports classes and properties lookup. All the collected results can be visualised either on a map or on a timeline. It seems that the resulting visualisation can neither be customised nor exported.

VISINAV [132] implements a keyword-based search completely masking the underlying data modelling mechanism. User-defined keywords are literally searched into the KG, without extending it with synonyms and related terms. Users can iteratively refine results by performing path transversal and selecting facets to manipulate and extend the result set. Once users are satisfied with the retrieved results, VISINAV supports basic temporal and spatial visualisations. While the tool supports the export, it is not clarified if the customisation can be performed.

VISU [8] and Sgvizler [300] let users interact with a single or multiple SPARQL endpoints without masking the underlying complexity as users have to explicitly pose SPARQL queries. Retrieved results are returned as data tables and users can manipulate the resulting data representation, create customizable and exportable visualisation by Google Charts. While Sgvizler is a general-purpose tool, VISU is bounded to academic data concerning universities.

Visualbox [126] let users query KGs by SPARQL and visualise results by a set of visualization templates which are referred to as filters. Authored visualisations can be downloaded and embedded in any hyper-textual documents, blogs or wikis.

rdf:SynopsViz [31] implements a FSI to query and filter classes and properties. It computes statistics and infers hierarchies from data without requiring any further user interaction. Once users are satisfied with the retrieved results, data can be visualised by charts, such as treemaps and timelines according to available data and specific data exploitation requirements. Visualisation can be exported, but not customised by users.

YASGUI [270] guides users in querying KGs by explicitly using SPARQL and visualising data through Google Charts. YASGUI is enhanced by auto-completion, while Google Charts offers exportable and customisable visualisations.

SPARKLIS [107] is a query builder based on a FSI enhanced by a controlled NL interface. SPARKLIS offers basic visualisation options, such as maps and image viewers. Furthermore, it is integrated with YASGUI and, hence, it inherits its visualisation approach, including the export and the customisation options. SPARKLIS successfully masks the SPARQL complexity, without losing its expressiveness.

Wikidata Query Service (WQS) [196] is specifically designed for working on Wikidata. It implements a form-based query builder and offers several visualisation modes, such as charts, maps, timelines, image viewers, and graphs.

QueDI is a guided workflow from KG querying to data visualization. Users can query LOD by FSI enhanced by a controlled NL interface, by masking an automatic and on-the-fly generation of SPARQL queries. While the SPARQL query generation phase covers Basic Graph Patterns (BGPs), such as path traversal, union, filters, negation, and optional patterns, the dataset manipulation phase covers aggregation and sorting. Thus, by considering the total expressiveness of QueDI it supports more than BGP. Finally, users can author customizable and exportable visualization. Users can export the visualization as an image or as a dynamic and live component that can be embedded in any hyper-textual page, such as HTML pages, WordPress blogs and Wikis.

9.2 APPROACH AND METHODOLOGY

QueDI implements a *transitional approach* where users are guided from LOD querying to their comfort zone, that is a tabular representation of data, table manipulation, and chart generation. QueDI let users build queries step-by-step with an auto-complete mechanism and exploiting retrieved results by exportable and dynamic visualisations. To guide users in the entire workflow, the querying and visualisation process is split into three steps, graphically represented in Fig. 9.1.

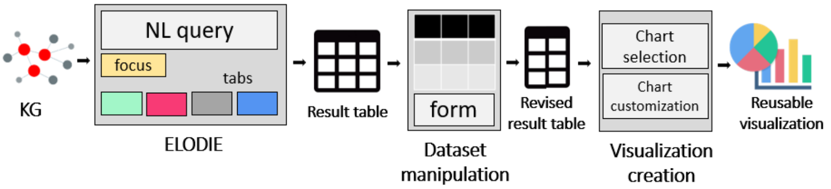


Figure 9.1: Transitional approach in QueDI

DATASET CREATION. Users starts choosing the KG of interest provided with a working SPARQL endpoint. This step focuses on the creation of a tabular representation of data of interest representing results collected by the user-query. The dataset creation phase is implemented by ELODIE, Extractor of Linked Open Data of IntErest, pronounced elədē, which is a FSI enhanced by a controlled NL interface to query the configured KG and organise results replying to users' requests as a data table. ELODIE completely masks the SPARQL complexity as user queries are automatically translated into SPARQL queries without requiring any user interaction. Furthermore, ELODIE automatically retrieves classes and relations to go on in the query formulation according to the current user query letting users exploring data by inspection. Hence, ELODIE does not require any previous knowledge of queried data. By representing retrieved results as a data table, QueDI moves LOD to target users' conform zone, implementing the transitional approach from LOD to data table representation.

DATASET MANIPULATION. When users are satisfied with retrieved results, they can start the dataset manipulation phase to refine collected information and to make it compliant with the desired visualisation. In this stage, QueDI relies on data manipulation skills of the target group: users can refine results, aggregate values, and sort columns. This step let clean data tables returned by ELODIE and make them compliant with the visualisation requirements.

VISUALISATION CREATION. This phase creates exportable and reusable visualizations that can be customised according to users' needs and preferences. It realises the immediate gratification for information consumers of seeing the results of their effort in a concrete artifact.

The main *novelties* of QueDI are:

- the provision of a query mechanism articulated in two steps: first, a SPARQL query building phase to automatically retrieve results from KGs without requiring skills in Semantic Web query languages, and, second, a SQL building phase to manipulate retrieved results. Query builders have to trade off usability of the proposed mechanism and its expressiveness. While ELODIE covers BGPs, such as path traversal, union, filters, negation, and optional patterns, users can perform aggregations, filtering, sorting during the table manipulation phase. Hence, by combining the expressiveness of ELODIE and table manipulation, QueDI covers SELECT queries which results can always be represented as a table, BGPs directly in SPARQL, sorting, GROUP BY, aggregation operators and filtering by table manipulation;
- a guided workflow from data querying to knowledge representation instead of the juxtaposition of visualization mechanisms to query builders.

9.3 TOOLKIT: QUEDI

This section provides the reader with further details on the inner mechanism of QueDI, the proposed interface and technical aspects.

QueDI is released open-source on GitHub[1], it is freely available online[2] and quick tutorials[3] are available on YouTube.

DATASET CREATION. The dataset creation phase is implemented by ELODIE, a SPARQL query builder provided with an FSI and enhanced by a controlled NL interface. The dataset creation phase starts from the user selection of the SPARQL endpoint of interest by choosing an option from the list of suggestion which includes, among others, general purpose KGs, such as DBpedia[4] or Live DBpedia version[5], CH KGs, such as the Italian endpoint "*Beni Culturali*"[6] or the National library of Chile[7], and academic KGs, such as the French endpoint Persée[8]. By default, ELODIE queries DBpedia. Then, users move to the proper querying phase.

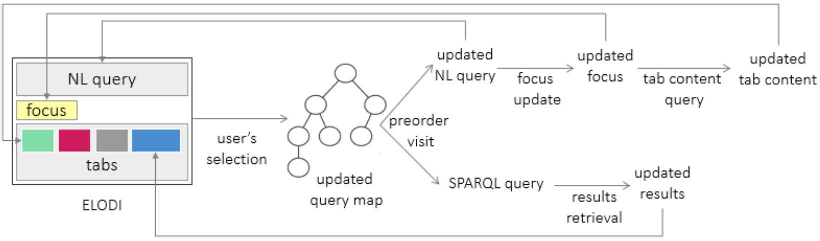


Figure 9.2: Operating mechanism of ELODIE

The operating mechanism of ELODIE is graphically represented in Fig. 9.2. Users formulate their queries by iterative interactions by choosing one of the available options which are organised in facets. The user query is automatically verbalised as a controlled NL query. Hence, herein NL query and user query will be used as synonyms. While the user query represents the query under construction, the focus represents the insertion position for ap-

[1] QueDI on GitHub: https://github.com/routetopa/deep2-components/tree/master/controllets/splod-controllet
[2] Online demo of QueDI: http://deep.routetopa.eu/deep2t/COMPONENTS/controllets/splod-visualization-controllet/demo.html
[3] YouTube tutorials of QueDI: https://youtu.be/e_o32GP-l1c
[4] DBpedia SPARQL endpoint: https://dbpedia.org/sparql
[5] Live DBpedia SPARQL endpoint: http://live.dbpedia.org/sparql
[6] Beni Culturali SPARQL endpoint: http://dati.culturaitalia.it/sparql
[7] National Library of Chile SPARQL endpoint: https://datos.bcn.cl/sparql
[8] Persée SPARQL endpoint: http://data.persee.fr/sparql

plying query transformation. User query and focus determine the state of the system. According to the focus, concepts that are classes, predicates that are relations, and resources are automatically retrieved by ELODIE by posing queries on the SPARQL endpoint and resulting data are organised in facets, also referred to as tabs. More in detail, all the sub-classes that can refine the focus are listed in the classes tab; all the predicates that have the focus as subject, which are referred to as direct predicates, or as the object, which are referred to as reverse predicates, are listed in the predicate tab. As ELODIE autonomously retrieves data to suggest how to go on in the query formulation, it solves the conceptualisation issue as users are not required to explicitly look for data of interest and the source is explored by inspection.

Suggestions are retrieved by path traversal queries, generic enough to be used to retrieve data from any endpoint, by solving the portability issue. At each query refinement, the inner data structure, referred to as map, that models user interactions is updated by modifying the focus neighbourhood. Then, ELODIE performs a pre-order visit of the map to auto-generate both the NL and the SPARQL queries. While the NL query verbalises users' interactions, the SPARQL query is posed against the SPARQL endpoint to retrieve user query's results. All the retrieved results are organised as a data table. The last selected element behaves as the new focus, and, according to it, all the facets are automatically and consistently updated by querying the SPARQL endpoint. This process is repeated to each user interaction.

Users are guided step-by-step in the query formulation thanks to the FSI. At each step, ELODIE offers a set of suggestions, organised in the concepts, predicates, operator, and results tabs, to go on in the query formulation by preventing empty results. A clarification is required: empty results are a desirable result in complete KGs interpreted as a close world. However, as common KGs suffer from incompleteness, empty results can be interpreted either has a real desired result or as missing information. As data exploitation tools can not automatically distinguish these scenarios, ELODIE prevent empty results by providing all the navigable edges outgoing from the focus as a suggestion and the actual user query. In other words, suggestions are focus-dependent. This exploratory search provides an intuitive guide in query formulation. Once a suggestion has been selected, it will be incorporated and verbalised into the user query.

As ELODIE automatically translates user questions in SPARQL queries, it makes ELODIE a query builder, without asking for the SPARQL knowledge. SPARQL is completely masked to the final users by solving the problem of technical complexities. However, ELODIE also provides the possibility to inspect the SPARQL query under construction if explicitly requested by the user. ELODIE enables users to compare the NL and the SPARQL query side by side and in increasing their expertise in formulating SPARQL queries. Users can interact with NL queries words and ELODIE automatically highlights the corresponding SPARQL constructor.

The query, suggestions, and results are verbalised in NL to solve the readability issue. Instead of returning URIs to users, ELODIE retrieves labels obtained by looking for `rdfs:label` predicate attached to the retrieved data and by asking for labels in the user language. If labels are missing, ELODIE looks for the English ones. If also this attempt fails and resources are not attached to `rdfs:label`, the URL local names are exploited as labels. Suggestion labels are contextualised by phrases. For instance, instead of showing the author as a predicate, the predicate label is wrapped into a meaningful phrase, such as `that has an author`. The user query always represents a complete and meaningful phrase. Therefore, ELODIE implements a NL interface. However, it is worth noting that users cannot freely input queries, but ELODIE implements a controlled NL interface used to verbalise the iteratively created user query. It makes query formulation less spontaneous and slower instead of directly writing the query in NL, but it provides intermediate results and suggestions at each step, prevents empty results, and avoids ambiguities issues of free-input NL query and out-of-scope questions. ELODIE translates queries and suggestions in English, Italian, and French, and new supported languages with the same syntax, such as Spanish, can be easily incorporated.

Only a limited number of results and suggestions are retrieved to address scalability issues. However, this limit can be freely changed by users. The main drawback of limited suggestions is that they can prevent the formulation of some queries. Therefore, ELODIE implements an intelligent auto-completion mechanism at the top of each suggestion list. At each user keystroke, it filters the corresponding suggestion list for immediate feedback. If the lists get empty, the list of suggestions is re-computed by querying other suggestions that also include the user filter.

ELODIE is implemented as a client-side web application, without any server-side computation. To promote portability, ELODIE is entirely based on Web standards: the entire application is written in Javascript and the interface with HTML/CSS, with zero configuration. It only requires CORS to enable SPARQL endpoint URL, i.e., a specification that enables truly open access across domain boundaries.

DATASET MANIPULATION. This phase implements a SQL query builder provided with a form-based interface. Users can select columns of interest, perform aggregation, filtering, sorting, and complete the selected operation by any required parameter. For instance, users can remove empty cells from a column, drop out all values but numbers, filter a column by number or string operations, group tables by column values, aggregate values by counting or summing them, compute the average, or detect the minimum or maximum value. Sorting is intuitively enabled on the top of each column. These patterns enhance the BGPs of ELODIE. Thanks to the aggregation feature, users can perform `group by` and compute statistics of retrieved data, such as `count`, `average`, `sum`. Thanks to the filtering feature, users can remove empty cells or perform textual and numeric filters, such as keeping only cells containing a user-defined sub-string or numerical values lower than a user-defined number. The SQL query builder implicitly transforms user interactions in a SQL query and automatically updates the result table.

VISUALISATION CREATION. This step implements the exploitation phase, where users are guided in representing the acquired knowledge by exportable and reusable visualisation. Besides proposing the realisation of mere images, users are guided in authoring dynamic artifacts that can be downloaded and embedded in any blog, web page as an HTML5 component. It is a dynamic component as it embeds the query to retrieve and refine the dataset by always ensuring up-to-date results. Therefore, if data in the queried endpoint change, their visual representation will change as well.

According to the guidance principle, users are provided with a vast pool of charts, such as timelines, maps, media players, histograms, pie charts, bar charts, word clouds, treemaps. Only charts compliant with the retrieved data are enabled. According to the chosen visualisation mode, users can customise both the chart content and its layout. Then, the realised chart can be downloaded as an image or as a dynamic component.

9.4 NAVIGATION SCENARIO

This section details a navigation scenario querying DBpedia through `QueDI`. Table 9.2 contains iterative queries as verbalised by `QueDI` of a navigation scenario that retrieves the *geographical distribution of the Italian architectural structures*. At each step, the bold part represents the last suggestion selected by the user and the underlined part represents the query focus. Suggestions can be classes, such as `city`, direct and inverse properties, respectively, `has a thumbnail` and `is the location`, operators, such as, `that is equals to`, and resources, such as, `dbr:Italy`[9].

Fig. 9.3 is a collage of screenshots of the different steps of the workflow implemented in `QueDI`. The top of Fig. 9.3, Fig. 9.3.1, shows the user query at the end of its formulation by `ELODIE`. When users are satisfied with the retrieved results, they can move to the second step that is the dataset manipulation. During the dataset manipulation phase, users can group data by city and count the architectural structures in each group, as shown in Fig. 9.3.2. In other words, they perform data aggregation. Moreover, they can also sort data by the number of structures. Finally, they are ready to visualise the retrieved results and represent the achieved knowledge by an exportable visual representation. Fig. 9.3.3 represents the geolocalised distribution of architecture structures on an Italian map.

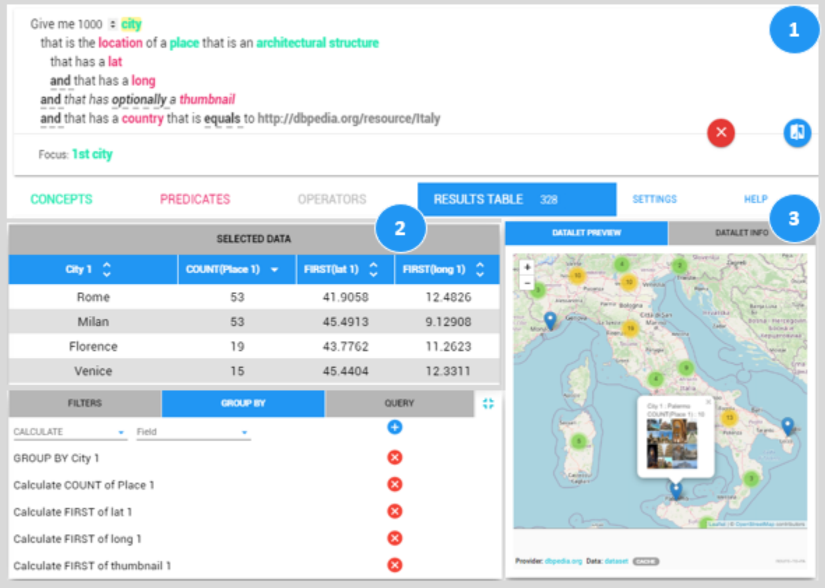
[9] `dbr` is the prefix corresponding to `http://dbpedia.org/resource/`

Table 9.2: Navigation scenario using QueDI on DBpedia

Step	Query
1	Give me something
2	Give me a **city**
3	Give me a city **that is the location of something**
4	Give me a city that is the location of a **place**
5	Give me a city that is the location of a place **that is an architectural structure**
6	Give me a city that is the location of a place that is an architectural structure **that has a lat**
7	Give me a city that is the location of a place that is an architectural structure that has a lat and **that has a long**
8	Give me a city that is the location of a place that is an architectural structure that has a lat and that has a long
9	Give me a city that is the location of a place that is an architectural structure that has a lat and that has a long and **that has a thumbnail**
10	Give me a city that is the location of a place that is an architectural structure that has a lat and that has a long and that has **optionally** a thumbnail
11	Give me a city that is the location of a place that is an architectural structure that has a lat and that has a long and that has optionally a thumbnail
12	Give me a city that is the location of a place that is an architectural structure that has a lat and that has a long and that has optionally a thumbnail and **that has a country**
13	Give me a city that is the location of a place that is an architectural structure that has a lat and that has a long and that has optionally a thumbnail and that has a country **that is equals to http://dbpedia.org/resource/Italy**

9.5 Evaluation: accuracy, expressiveness and scalability assessment

This section reports the accuracy, expressiveness, and scalability evaluation of `QueDI`. As `QueDI` implements the query formulation into two phases, the SPARQL query generation to retrieve results of interest and a SQL query generation to aggregate and sort results, the reported evaluation assesses if and in which cases the accuracy is compromised. Moreover, the performed analysis assesses the expressiveness level by testing `QueDI` on standard benchmark for question answering and its scalability when tested against real KGs, such as DBpedia.



Give me 1000 city
that is the location of a place that is an architectural structure
that has a lat
and that has a long
and that has optionally a thumbnail
and that has a country that is equals to http://dbpedia.org/resource/Italy

Focus: 1st city

SELECTED DATA

City 1	COUNT(Place 1)	FIRST(lat 1)	FIRST(long 1)
Rome	53	41.9058	12.4826
Milan	53	45.4913	9.12908
Florence	19	43.7762	11.2623
Venice	15	45.4404	12.3311

GROUP BY City 1
Calculate COUNT of Place 1
Calculate FIRST of lat 1
Calculate FIRST of long 1
Calculate FIRST of thumbnail 1

Figure 9.3: QueDI interface in the navigation scenario on DBpedia

9.5.1 *Evaluation design*

methodology. The RQ that guides this evaluation is: *Does the proposed approach lose in accuracy?* This evaluation compares the two-phase query generation implemented in QueDI with the formulation of SPARQL queries only exploiting SPARQL query building. The performed evaluation is based on the hypothesis that the accuracy is affected only when the complete set of query results is so huge that the queried endpoint does not return all the results or QueDI can not manage them.

dataset. The evaluation is performed by testing ELODIE and the data manipulation phase on the QALD-9 [326] challenge dataset. This dataset behaves as a benchmark in comparing KGQA tools. The QALD-9 DBpedia multilingual test set[10] contains 150 questions which can be replied to over DBpedia and, for each question, it reports the question verbalisation in English, the related SPARQL query, and the result collection.

[10] QALD-9 dataset https://github.com/ag-sc/QALD/blob/master/9/data/qald-9-test-multilingual.json

PROTOCOL. For each question of the QALD-9 dataset, a trained and focused user reproduced it by combining features offered by ELODIE and the dataset manipulation interface. Finally, at the end of the evaluation, results are computed by exploiting the system used to evaluate the QA systems joining the QALD challenge, GERBIL. The performed evaluation reports the minimum number of interactions and the related needed time starting from the empty query, that is "*Give me something...*". Since this evaluation assesses the accuracy of the proposed two-step querying approach, the expressiveness of QueDI, and the scalability on real datasets and *not* the usability, it minimises the exploration and thinking time required by users to conceptualise queries. Thus, the evaluation both considers the English NL formulation of the query and the related SPARQL query while performing them on QueDI. The measured time represents the best interaction time for a trained and focused user in performing questions on QueDI. In real use, interaction time increases according to unfamiliarity with QueDI and the queried dataset and lack of focus in exploratory search.

DATA GATHERING. The evaluated metrics are accuracy, precision and F-measure, both for each question (*macro-measure*) and for the entire dataset (*micro-measure*).

9.5.2 *Results*

This section compares results achieved by QueDI[11] with the results achieved by the systems that joined the QALD-9 challenge [326] (Table 9.3).

EXPRESSIVENESS. QueDI can reply to 143/150 questions. Not supported patterns cause the failures, i.e., make computation by SPARQL operator (3/7 cases), field correlation and not exist), and, in 2/7 cases, too many results.

[11] QueDI replies to QALD-9: https://github.com/mariaangelapellegrino/QueDI_evaluation,
QueDI evaluation over QALD-9: http://gerbil-qa.aksw.org/gerbil/experiment?id=202110210025

Table 9.3: Evaluation results of QueDI over QALD-9

Tool	**Micro results**			**Macro results**		
	P	**R**	**F1**	**P**	**R**	**F1**
ELON	0.095	0.002	0.003	0.049	0.053	0.050
QASystem	0.039	0.021	0.027	0.097	0.116	0.098
TeBaQA	0.163	0.011	0.020	0.129	0.134	0.130
wdaqua	0.033	0.026	0.029	0.261	0.267	0.250
gAnswer	0.095	0.056	0.070	0.293	0.327	0.298
QueDI	**0.762**	**0.948**	**0.845**	**0.951**	**0.956**	**0.950**

ACCURACY. In 20/143 cases, we both exploited ELODIE expressiveness and data manipulation features. By considering the queries that required further refinement, sorting, or aggregates, we observed that: in 8/20 cases we performed a `group by` to remove duplicates; in 4/20 cases we performed `group by`, `count` as aggregation and `sort`; in 8/20 cases only sorting has been required. It is worth noting that ELODIE returns the count of table tuples without requiring any further interaction. Only one failure is caused by a too wide pool of results (*all books and their numbers of pages*) that QueDI can not manage. In conclusion, we can assess that by splitting the querying phase into two steps, we only lose accuracy when the required results are too much to be first collected and then refined.

SCALABILITY. By considering the interaction time for the 143 successful questions, we observe that: more than half of the questions (75/143) can be answered in less than 40 seconds (with 30 s as median and average time); 115 of them can be replied in less than 60 s (with 0,4 as average time and 0,37 as median time); only 6 of them requires a time that lies between 2 minutes and 3 minutes and a half (median time 40 s and average time 60 s).

9.6 EVALUATION: USABILITY ASSESSMENT

This section estimates the usability and the execution time in real use by providing a list of tasks to i) users familiar with OD management, ii) computer scientists, and iii) lay users and by collecting results of standard questionnaires to assess usability and by comparing the needed time of lay users with the execution time of focused expert in accomplishing the same tasks.

9.6.1 *Evaluation design*

METHODOLOGY. The RQs that guided this evaluation are:

- Are ELODIE operating mechanism and its interface considered *usable*? By whom?
- Can users without skills in querying languages *quickly* learn how to exploit ELODIE in retrieving data of interest?

PARTICIPANTS AND SETTINGS. The group of *users familiar with OD* corresponds to 22 Italian learners of a High School institute of Caserta, 16-20 years, familiar with OD management with SPOD, but completely unaware of LOD, KGs and SPARQL. Unfortunately, some of the tests were incomplete and, we only could consider 11 complete result sets. All the users could either work individually or exchange opinions in groups. The group of *computer scientists* includes 11 students still studying and already graduated. The group of *lay users* includes 12 users without any technical skill in query language and heterogeneous background.

While the users familiar with OD tested QueDI in the context of a partnership between schools and universities, the other groups tested it in an informal setting. All the participants were completely unaware of QueDI and had no expertise in KGs and SPARQL. All the activities have been performed in person. All the participants voluntarily joined the evaluation and agree that all the collected data in an anonymous form would be used for academic research.

PROTOCOL. The performed protocol follows:

- **training phase** to demonstrate how to exploit QueDI by guided examples of incremental complexity;
- **testing phase**: six tasks are submitted in the Italian language. The tasks are sorted by increasing complexity. The

group of users familiar with OD tested the Italian endpoint *Beni culturali*[12] to be coherent with other activities performed in the context of CH preservation and performed the tasks listed in Table 9.4. The other two groups tested DBpedia and performed the tasks listed in Table 9.5.

Table 9.4: Tasks assigned during the usability assessment of `QueDI` by users familiar with OD.

Task #	**Query**
Task 1	The disciplines of the archaeological parks.
Task 2	The archaeological parks that provide assistance for disabled as service.
Task 3	The events taking place until 2018.
Task 4	The libraries whose site has address in Campania region.
Task 5	The events hosted by the Royal Palace of Caserta and take place until 2018.
Task 6	The museums of Caserta that either have discipline art or provide the Book store as service.

Table 9.5: Tasks assigned during the usability assessment of `QueDI` by lay users

Task #	**Query**
Task 1	The Italian museums.
Task 2	The games with at least 2 players.
Task 3	The presenters who are the presenter of a TV Show.
Task 4	The female scientists born and dead in Germany.
Task 5	The athletes which are not dead.
Task 6	The artists born in the same place of an athlete.

DATA GATHERING. For each task, users filled an ASQ [180] to evaluate 1) the degree of difficulty of the task, 2) if the time to

[12] Beni Culturali SPARQL endpoint: `http://dati.beniculturali.it`

complete the task is reasonable, 3) if the provided knowledge in the training phase is sufficient to complete the task. While the degree of the satisfaction gives an insight on the engagement and the desire to reuse the tool, the required time gives information about the perceived usability in terms of complexity in performing the task, the satisfaction in the support models the easiness in learning how to use the proposed interface.

At the end of the evaluation, all the participants are invited to fulfil a final questionnaire to evaluate i) users' satisfaction based on a Standard Usability Survey (SUS) [181] and ii) their interest in using and proposing the tool by a Behavioural Intention (BI) survey. The questions of the BI survey are i) *"I will use `QueDI` regularly in the future"*; ii) *"I will strongly recommend others to use `QueDI`."* and users can use a 7-point scale to reply.

9.6.2 *Results*

This paragraph reports and comments results achieved by the ASQ, the SUS questionnaire, considerations concerning the execution time, and the BI for each participating group.

AFTER SCENARIO QUESTIONNAIRE. The responses to the ASQ, reported in Table 9.6, show that in general everyone has been quite satisfied with the ease of completing every task and with the amount of time that took to complete them. The highest values are related to the satisfaction with the knowledge given in the training phase, reaching an average value higher than 6 on a 7-point scale (6.11).

Table 9.6: ASQ in `QueDI` usability evaluation by users familiar with OD.

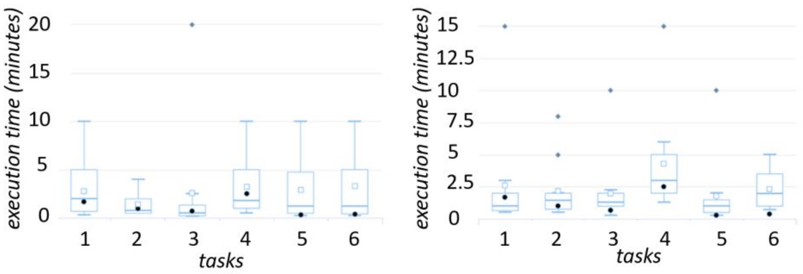
	Question 1	**Question 2**	**Question 3**
Task 1	6.36 (0.88)	6.27 (0.86)	6.45 (0.78)
Task 2	5.72 (1.05)	6.00 (0.85)	6.18 (0.83)
Task 3	5.00 (1.13)	5.55 (1.23)	5.73 (1.29)
Task 4	5.55 (1.30)	5.36 (1.43)	6.27 (1.13)
Task 5	4.82 (1.40)	5.45 (1.44)	5.91 (1.44)
Task 6	5.55 (1.07)	5.82 (1.03)	6.09 (1.08)

EXECUTION TIME. This section compares the execution time of each task in each group with the time required to one expert of the field and familiar with QueDI to accomplish the same task. This value is considered the *optimal* one.
Table 9.7 reports the average time and, among brackets, the standard deviation of participants familiar with OD related to the time needed by an expert of the field to execute the same query. Execution times are taken in minutes and include all the steps to perform the task, from the track reading to the sharing of the results. In general, participants familiar with OD needed at least double time compared to a focused user, expert of QueDI. However, it is interesting noting that in the last case, execution time is almost the same. It is worth recalling that tasks are sorted by complexity. It implies that, despite the fact the last task is the most complex one, most of the participants succeeded in executing it being indistinguishable by an expert user in terms of execution time. It creates the conditions to assess that in a short time users familiar with OD can learn how to query KGs by QueDI. The results related to computer scientists are reported in Fig. 9.4.a, while the results related to the lay user group are reported in Fig. 9.4.b.

Table 9.7: Execution time during the usability assessment of QueDI by users familiar with OD.

	Mean time of users familiar with OD	**Time of an expert**
Task 1	3.91 (2.11)	1.38
Task 2	5.91 (3.15)	1.50
Task 3	7.45 (3.94)	2.30
Task 4	6.18 (2.25)	2.35
Task 5	8.18 (2.91)	2.46
Task 6	8.64 (4.27)	7.12

In all the queries but the last query for the second group, the minimum time needed by the participants either matches the optimal one or is even better. It is a surprising result, and it means that there are users (at least one in each group) able to get familiar with QueDI and learn how to use it in a short



a) Execution times of the *computer science* group.

b) Execution times of the *lay users*.

Figure 9.4: Execution time during the usability assessment of `QueDI` by computer scientists and lay users

time. About the outliers, in the open questions, it is evident that the main difficulties are in *"finding the exact way to refer to an asked predicate or concept"* (reported by 6 out of 23 users). The participants suggest to *"insert inline help, tool-tips to help the users during the usage, examples of usage"* (reported by 9 out of 23 users). The start is considered a small obstacle to face: *"After a bit of experience, the system is pretty easy to use"* (by 6 out of 23 users).

usability assessment. The SUS score is 71 for the group of users familiar with OD, 70 for computer scientists and 68 for lay users. According to the SUS score interpretation, all the values at least equal to 68 classify systems usability above the average. That means that `QueDI` reached a good level of usability according to all the considered groups. The perceived usability seems to be proportional to the awareness in OD.

Both the statement *"I found the various functions in this system were well integrated"* and *"I thought there was too much inconsistency in this system"* have an excellent average score (respectively 3.82 and 1.82), that, about the low values of the standard deviation (respectively (0.94) and (1.03)), underline a good design of the tool, consistency of the features and predictability of consequences to each action. The statement *"I found the system unnecessarily complex"* with an average score of 1.82 (0.83) is proof for the accomplishment of hiding the underlying complexity of semantic technologies. Moreover, as *"I felt very confident using the system"* obtains the score 4.36 (0.64), `QueDI` offers an intuitive and not intimidating interface to become familiar with and visualise KGs.

The statement *"I think that I would need the support of a technical person to be able to use the system"* has an average score of 1.90 (0.51) in the group of users familiar with OD underlying that they considered a preliminary training phase to successfully exploit the system useful, but not strictly required. In contrast, lay users asses and their modelling.

BEHAVIOURAL INTENTION. The BI survey requires to express if i) participants are interested in using `QueDI` in the future and ii) if they would be prone to propose `QueDI` to others. As both questions obtained a mean score above 5, there is an overall intention to reuse and propose `QueDI` to others.

9.7 KNOWLEDGE GRAPHS FOR KNOWLEDGE MANAGEMENT

Usually, users passively accept the list of results returned by a search engine without further investigating the ones ranked as less important. Ironically, Elon Musk says that "*the safest place to hide a dead body is the second page of Google search results*" since most people stop on its first page. Thus, users should be encouraged and provided with approaches and tools to actively search on the Web, to critically choose results of interest, and to actively create their knowledge. While it is generally important for any users, it assumes a crucial role for learners to make them leave the position of indifferent spectators of web search, going beyond the passive inspection of results returned by a search engine, actively searching for the data that best answers their questions, learning how to read, work with, analyse, and argue with data [197], taking responsibilities, and contributing to the process of building and manage their knowledge [246].

The knowledge concept is widely studied from several different perspectives, and there is no consensus on its definition. In this context, the *knowledge* term refers to *information in action* [240]. It represents one of the phases of the data pyramid [277] composed of data, information, and knowledge. Data represent observations or facts; information is inferred from data by the process of answering questions and retrieving data useful for actions or decisions; the knowledge is achieved by processing, organising, or structuring information. The process to reach the knowledge starting from data is referred to as KM.

This section focuses on the exploitation of KGs for supporting the KM process in the *educational context*. In the last decades, many KG exploitation tools have been proposed, as overviewed in Section 9.1, but they mainly rely on a *monological* learning approach, that is a learning process performed individually [248].

When information needs become complex, it might be useful to collaboratively explore the information space and participate in shared learning [339]. This practice is well-known in different fields, from healthcare [131] to tourism [215], from bioinformatics [116] to data visualisation applications [167]. Collaboration is also a key aspect to ensure the creation of value from OD [357], and it is easy to extend this consideration also to LOD. Users may opt for cooperating with like-minded peers, by communicating and discussing their insights, building social networks and establishing a sense of community [48], participating in social interactions and sharing their knowledge with other members according to the willingness to freely reveal their expertise and openly work together [172]. This dynamic would enable collaborative learning, also referred to as *dialogical* learning [248].

According to the metaphors of the *trialogical approach* [248], while the monological learning corresponds to acquiring knowledge individually, the dialogical learning is enabled by participation, the trialogical approach enables the knowledge creation [248]. These metaphors are based on the creation of useful artifacts alternating individual and collaborative activities [287].

9.7.1 *The trialogical learning approach*

We propose to place the trialogical learning metaphors in the context of KG exploitation for KM to enable users, and mainly learners, in exploiting KGs in the knowledge acquisition and management process [81, 93]. Users can either formulate queries individually or can borrow a query created by others, modify and publish it. It results in the collaborative creation of artifacts by layering individuals' efforts. Queries express information needs and have the potential to behave as a starting point to guide other searchers with similar information needs to perform query reformulation. Thus, starting from a shared query, users can replay parts of the search process using different parameters, modifying artifacts authored by the community, and spreading new knowledge. Consequently, query reformulation is interpreted as collaborative querying [114] which refers to a family of techniques that assist users in formulating queries to meet their information needs by harnessing other users' expert knowledge or search experience.

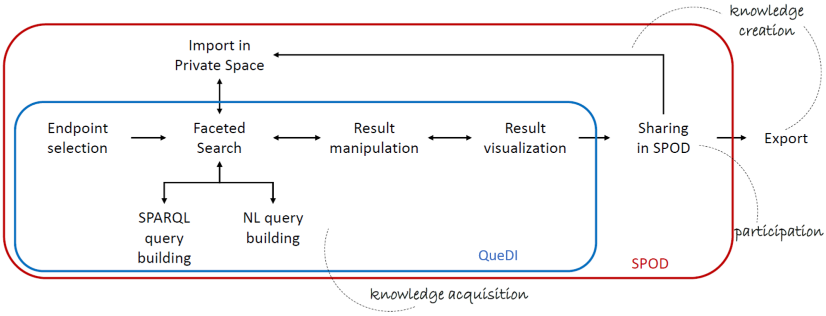


Figure 9.5: Trialogical learning approach to query KGs

The proposed approach is graphically represented in Fig. 9.5 and it is based on 1) the *knowledge acquisition* scaffolded by QueDI 2) within SPOD to gain *participation*, 3) providing the opportunity of collaboratively developing mediating artifacts by the refinement of intermediate releases and export the new gained knowledge in- and out-side the community to achieve *knowledge creation*. This approach encompasses different types of knowledge conversion: individuals learn tacit knowledge, which represents knowledge embedded in individual experience, by participating in an expert community during the socialisation phase; learners transform the tacit knowledge into the explicit one during the externalisation step, sharing evidence of the previous learning steps with the group; the expert knowledge is synthesised during the combination phase and, then, internalised thanks to the spreading of new knowledge with others [235]. Creating a parallel with the proposed mechanism to query KGs and perform KM, individuals learn how to query KGs by borrowing already defined queries, modifying and re-share them, combining their own and community effort to provide new knowledge within and outside the community.

knowledge acquisition and management by quedi. The general KM process, graphically represented in Fig. 9.6 (a), starts from data that models facts, reaches information by querying data and collecting results, and realises knowledge by analysing and structuring information in shareable artifacts. In a KG-based KM process, graphically represented in Fig. 9.6 (b), queried data are modelled as KGs. To move from data to information, a querying mechanism is required. As explained in the challenges posed by KGs, the proposed querying mechanism should hide the underlying syntactical complexities while raising questions.

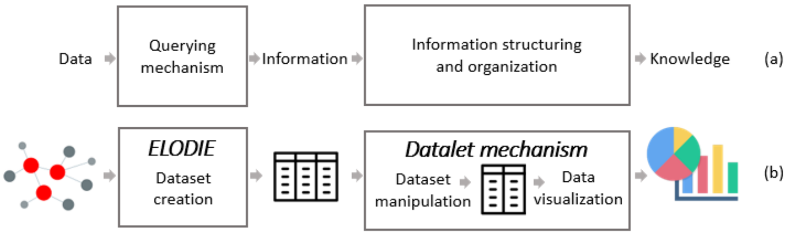


Figure 9.6: General and KG-based KM process

Visual query systems are a popular approach where lay users formulate queries according to their needs without involving technical experts. Most of the time users are only provided with a graphical interface where they navigate options, select attributes, apply constraints, and, then, finalise their request. This request is converted into an actual query behind the scenes to fetch the data. Thus, we proposed a KG querying mechanism where SELECT queries are posed in NL and retrieved results are organised in data tables. This phase is referred to as dataset creation and it is performed by ELODIE.

To move from information to knowledge, users need to process and organise retrieved results. Starting from the first year of their educational plan, learners get used to reading tables, creating simple charts, performing basic manipulation, and interpreting results to reply to questions [283]. Thus, QueDI models the gained knowledge as shareable artifacts, such as charts, that can support discussions and justify decisions. To achieve knowledge, results can be manipulated to be compliant with visualisation modes and, then, charts can be realised by the datalet mechanism.

participation - community-based approach empowered by spod. Learners can work individually building their queries until they are not ready to share acquired knowledge with the community of SPOD, as made evident in Fig. 9.5. Since QueDI has been integrated into SPOD, users can perform all the typical actions of a social platform, such as sharing queries and artifacts, commenting and re-posting visualisation accessible by the news feed, joining evidence-based discussions with colleagues and friends by exploiting query results in debates.

knowledge creation - collaborative artifacts developing. Besides inspecting the results of the performed query execution, users can even replay the search process performed by others using different parameters. This implies that not only the visualisation of the query result is shared, but implicitly also the process to obtain and elaborate data. Thus, users can look for the query closest to their needs among all the shared ones in the social platform, import and work on them, using QueDI until new knowledge has been reached. Finally, achieved results can be published and, cyclically, it could be borrowed by other less-skilled users to face a slightly different challenge. This process of borrow-modify-share can be repeated iteratively producing intermediate artifacts, and, thanks to the individual and community effort, a final artifact can be developed.

Besides the sharing within SPOD, users can export either the dataset as a table or the obtained visualisation and use it in blogs, websites, articles as evidence in discussions. It is worth recalling that artifacts can be downloaded not only as a mere representation but can be exported as dynamic representations of the performed query which reflect in real-time the queried source. In other words, if data curators update the queried KG, the exported visualisation automatically reflects available data without requiring any additional interaction.

9.7.2 *Discussion*

potentialities of knowledge graphs in educational settings. General-purpose KGs, such as DBpedia, can behave as useful sources in educational contexts thanks to the heterogeneity of covered topics, from sport to art, from science to geography, and covering many different educational subjects. KGs can be queried to reply to domain-specific questions or curiosities and general knowledge. KGs can be queried to look for natural places, countries and capitals, mountains and rivers to learn about geography; celestial bodies, chemical substances or natural species of animals and plants to deep learners' knowledge in science; military conflicts and wars for learning about history; sports facilities, athletes, sports seasons and teams in the context of sports learning; museums, monuments, artists and artworks to master art knowledge or musical artists and musical work concerning the music subjects; writers and written work for a deeper knowledge of the literature; devices and programming languages for mastering technology and computer science.

REQUIRED SKILLS FOR KNOWLEDGE MANAGEMENT BY KNOWLEDGE GRAPHS. The KM process requires the ability to actively extract information out of data and represent the acquired knowledge by a shareable artifact. It implies that future citizens must learn how to locate and query data of interest, critically evaluate retrieved information, learn how to synthesise it to reply to the initial query, and decide how to represent and share the knowledge. To address data literacy, learners must become familiar with chart interpretation to properly use them as evidence in discussions while arguing with data.

In working with KGs, future citizens have to try to imagine how data of interest can be abstracted and, then, verify how data are modelled. For instance, if we are interested in museums and their paintings, we have to realise that the concepts we need to retrieve are museums and paintings. Then, we have to think about how this information can be linked: the museum exposes a painting or a painting is located in a museum. Thus, the modelling ability plays a crucial role.

9.8 FINAL REMARKS

While data publishers are spur in providing data as LOD to boost innovation and knowledge creation, the complexity of RDF querying languages, such as SPARQL, threatens their exploitation. Hence, lay users require support in easily querying and successfully exploiting LOD.

In particular, this chapter focuses on experts in table manipulation, such as OD experts, and takes advantage of their expertise in table manipulation and chart creation. It proposes QueDI, a question-answering and visualisation tool that implements a scaffold transitional approach to query LOD without being aware of SPARQL and representing results by data tables; let users perform table manipulation and visually representing data by exportable and dynamic visualizations by relying on their expertise in OD management. The main novelty of the proposed approach is the split of the querying phase in SPARQL query building and data table manipulation.

The discussed evaluation empirically shows that the query approach in two phases loses in accuracy only when results are too much to be first retrieved and then filtered. Concerning usability, the 70 score according to the SUS questionnaire reports that QueDI is considered usable by users (with and without table manipulation skills) without expertise in the Semantic Web technologies and KG query languages. The needed time by users with a computer science background to interact with ELODIE is almost indistinguishable by the execution time of focused users, experts in QueDI features.

The described evaluation is a preliminary experiment to assess QueDI performance and still requires a structured comparison with the state of the art. As a technical issue, it requires enriching the supported endpoints by also considering the integration of a proxy to overcome the issue of not CORS-enabled endpoints. Moreover, we aim to further simplify the exploratory search in retrieving suggestions by also considering synonyms and alternative forms of the queried keywords.

10 KNOWLEDGE GRAPHS AND DIGITAL STORYTELLING

Thinking without a banister.

– Hannah Arendt

Storytelling is defined as the skill of inventing and authoring stories, behaving as a simple but powerful method to share experiences and convey knowledge [320]. Storytelling is also metaphorically defined as *thinking without a banister* [14] by stressing the inter-connection between storytelling and creativity. It has the potentiality to let storytellers think out-of-the-box, imagine, reason in a different manner, developing critical and divergent thinking that is the thought approach used to develop creative ideas or not conventional solutions by augmenting problem-solving skills. It empowers learning practices, enables knowledge sharing, communicative competences, and skills development [218]. In fact, storytelling enables learners to develop literacy skills [65, 218, 273, 349, 354], creativity [241], digital and technological skills, the ability to communicate through images [272] clearly and effectively [282] to ensure the story memorability [20], experience computational thinking and problem-solving skills, including synthesising, analysing, evaluating, and presenting information [272] by exploiting narratives [301].

Narratives are resources that scaffold storytellers in inventing and authoring stories [301]. They concern mechanisms to guide authors in telling a story and include artifices and the art of visually analysing and presenting any topic of interest [320]. In this direction, the Italian writer Gianni Rodari (1920-1980), considered worldwide a master in children's literature, defines techniques, methods, literary and narrative artifices to let children invent and author stories. His books are an extraordinary exercise of creativity, an attempt to nonsense plots, and a sophisticated exploration of the many possibilities of the Italian language. He encourages storytellers to use imagination to create imaginary scenarios and trigger metaphorical consequences. His masterpiece entitled "Grammar of Fantasy"[273] is a theoretical essay on the art of telling stories by providing educators and children with artifices, techniques, and approaches to narrate stories. Among others, he describes the art of playing on words (referred to as the *suggestion mechanism* in the remaining dissertation) and continues someone else's story (referred to as the *incipit mechanism* in the following).

The suggestion mechanism represents any technique to get inspired and overcome the blank page syndrome. Rodari introduces this mechanism by using the metaphor of "*throwing a stone in the pond*": when the stone touches the water surface, many concentric waves are generated. Similarly, when a word comes to mind, a human brain recalls mental images, associations, metaphors, personal experiences, and feelings. These memories represent a fundamental source to invent memorable stories. As an exercise, Rodari suggests thinking about a small set of words, a single word in the simplest case, and exploring synonyms and analogies by playing on terms. The process generates a word chain that leads narrators to recall images that may inspire stories.

The incipit mechanism represents the possibility to continue someone else's story. As an exercise, Rodari encourages narrators to think about "*what happened to Pinocchio when became a child?*". While the Pinocchio tale ends with the transformation of Pinocchio into a real child, Rodari uses it as a starting point of another story. In this way, basic narrative components, such as characters, time, and place, may take unexpected paths and lead to a surprising development. According to these literary artifices, pupils may be provided with an incomplete story to continue by their creativity taking surprising and unexpected developments.

Digital storytelling is the storytelling practice where users are supported by technological solutions [52], such as authoring interfaces or digital learning environments. As a result, besides authoring stories by traditional means, e.g., paper and pencils, users can experience visual storytelling by relying on digital media, including images and text [52]. It has become increasingly popular in many fields, such as human health [38] also as narrative therapy [193], media [296], entertainment [289], journalism, data visualisation [320], tourism [59] and education [291].

Educational digital storytelling has a robust tradition that relies on Bruner's researches concerning the role of narratives as an opportunity for learners to share knowledge [46]. It is perceived as a powerful technology-enhanced learning approach [344], widely adopted at each education level, from primary to secondary [344]. It has also been investigated in a broad range of educational domains, from religion [141] to healthcare [128], from anti-bullying [17] to first aid intervention [201], from social issues [214] to computational thinking [212], from science learning [276] to social science [205], economy and cultural heritage [266], applicable to any topic of interest.

All the contributions mentioned above prove that narratives and principles for empowering storytelling skills are widely explored in education. However, according to a survey conducted in 2020, educators feel that there are not enough opportunities for developing creativity at school [311], and the situation seems not changed much in 10 years, as already in 2010 teachers claimed they need more support and training to help pupils fully develop their creative potentiality [49]. Creativity potentially lies in everyone, but it is crucial to provide learners with stimuli and a favourable context for letting these skills emerge. Thus, educators and children require being supported by digital solutions in improving and developing their creativity.

To mitigate the lack of tools to support learners in improving their creativity by digital storytelling, my research group proposed `Novelette` [2, 4], a free and open-source digital learning environment to invent and author visual stories. As a visual storytelling authoring interface, `Novelette` supports pupils in creating, refining, and rendering stories containing textual and visual components. As its main novelty, `Novelette` embeds narrative artifices proposed by Gianni Rodari to scaffold pupils in inventing their stories. In particular, `Novelette` embeds the suggestion mechanism to guide pupils in playing on words by iteratively exploring synonyms and analogies and enables pupils to continue someone else's story through the incipit mechanism. As a digital learning environment, `Novelette` scaffolds educators in managing groups or classes.

To identify learners' needs and take care of learning requirements, `Novelette` results from a collaborative design approach where educators have been actively involved in proposing features and revising the resulting prototype. As a result, `Novelette` has been designed not only *for* education, but also *with* educators.

Within the context of collaborative design and implementation of `Novelette`, my responsibility was the identification, design, and implementation of the suggestion provision mechanism. Educators report the requirement of supporting pupils in developing literacy skills and in being supported by approaches to explore word meaning, analogies, use synonyms properly, be aware that each word may be interpreted differently according to senses. We explored the possibility to implement the suggestion mechanism by also exploring available KGs. Since users are mainly children, unaware of the KG modelling approach and mechanisms to query them, it required the definition of an implicit mechanism to make KG accessible. The research presented in this chapter has been published in the following contributions:

- Agnese Addone, Giuseppina Palmieri, Maria Angela Pellegrino: *Engaging Children in Digital Storytelling*. In the Proceedings of Methodologies and Intelligent Systems for Technology Enhanced Learning (MIS4TEL) 2021.
- Agnese Addone, Renato De Donato, Giuseppina Palmieri, Maria Angela Pellegrino, Andrea Petta, Vittorio Scarano, Luigi Serra: *Visual Storytelling by Novelette*. In the Proceedings of Information Visualisation (IV) 2020.
- Maria Angela Pellegrino, Mauro D'Angelo. *Engaging Children in Smart Thing Ideation via Storytelling*. As long abstract presented at I-CITIES 2021.
- Agnese Addone, Renato De Donato, Giuseppina Palmieri, Maria Angela Pellegrino, Andrea Petta, Vittorio Scarano, Luigi Serra: *Novelette, a Usable Visual Storytelling Digital Learning Environment*. *Submitted* to IEEE Education Society in September 2021.

10.1 RELATED WORK

Several digital storytelling platforms have been proposed to create media or data stories, such as Tableau-Stories [6], iStory [22], and Gravity [238]. Although journalists and media curators widely adopt these tools, they cannot be easily adopted as educational digital storytelling platforms as they lack features considered crucial in the learning settings, such as support at the class level, groups management, and literary support.

Table 10.1 provides an overview of the related work supporting educational digital storytelling, implementing group management features, and/or literary support as defined by Rodari to support the story invention phase. Tools are classified as *digital storytelling editors* if they are provided with story authoring and publishing mechanisms. The *Class management* feature represents the support at the class level, group management, and supervision from the educator's side. *Incipit* and *Suggestion* represent any approach to embed Rodari-style artifices in the digital storytelling platform.

Table 10.1: Digital Storytelling tools comparison.

Tool	**Year**	**Digital Storytelling editor**	**Class management**	**Incipit**	**Suggestion**
Fabula	2017	-	-	✓	✓
Communics	2020	✓	-	-	✓
Wakelet	2019	✓	✓	-	-
UTellStory	2020	✓	✓	-	-
StoryJumper	2020	✓	✓	-	-
Storyboard That	2020	✓	✓	-	-
Pixton	2020	✓	✓	-	-
Storybird	2020	✓	✓	-	-
Comic Life	2020	✓	✓	-	-
Novelette	2020	✓	✓	✓	✓

INCIPIT AND SUGGESTIONS. *Fabula*[1] scaffolds professional writers in creating stories supported by narrative suggestions, such as synonyms, rhymes, pre-defined sentences, and the incipit mechanism, here interpreted as proposed sentences pertinent to the user typed words. Fabula enables the creation of stories

[1] https://www.scribis.it/fabula/index.html unpublished tool, available since 2017 (Internet Archive WayBackMachine)

in written form without any media support. Thus, it cannot be classified as a digital storytelling editor. *Communics* [281] is a web-based digital tool designed for supporting the creation of comics. Communics provides users with a library of suggestions implemented as pre-defined sentences to overcome the blank page syndrome. Even if they are good proposals in Rodari-style support, not satisfy the class management requirement.

CLASS MANAGEMENT. By focusing on platforms that implement both the storytelling creator and the class management feature, *Wakelet* [66] and *UTellStory* [221] are online platforms which enable the possibility to export the authored stories and embed them into any other website. *Storyboard That* [334] and *Pixton* [321] enable the creation of comic strips. *Storybird* [54] and *Comic Life* [163] are commercial tools, while *StoryJumper* returns a paid book by posing a strong limitation on the use of the tool. They represent significant contributions to educational digital storytelling environments, but they do not implement any literary artifice to support pupils in inventing stories.

`Novelette` fills this gap as a digital learning environment to scaffold educational digital storytelling with Rodari-style support. In other words, it is a digital storytelling creator enhanced by the class management feature and embedding the incipit and the suggestion mechanisms to support pupils in inventing stories.

10.2 A CO-CREATION DESIGN APPROACH

The role of end-users in designing and implementing user interfaces is at the basis of many Human-Computer Interaction (HCI) methodologies [213]. Researchers argue and motivate the crucial role plaid by end-users in developing technological solutions [213]. By looking at the available literature, end-users are usually only involved in the testing phase [27, 321, 334]. However, during the evolution of HCI methodologies, the end-users role became more and more active to such extent to involve them not only in the evaluation phase but in the entire design process.

In designing technology-enhanced learning systems, involving educators in the collaborative design would lead to selecting goals and objectives that are crucial for the learning settings and able to address concrete pupils' needs [162]. Moreover, their involvement in designing tools and technology-based learning solutions leads them to experience the process of placing technology, content, and pedagogy together [169]. When an organisation and a defined group of end-users work together to design and refine a product, the co-creation takes place [286]. In the co-creation approach, there are different HCI design methodologies involving end-users that can be applied, such as user-centered design and participatory design [213]. While *user-centered design* approach [236] relies on designers focusing on users' needs without any actual involvement of real end users [213], in the *participatory design* [292], users play an active role by contributing to the design and the development processes proposing functionalities and rising design issues [213].

`Novelette` has been designed and developed combining user-centered and participatory design addressing educators' needs and actively involving them in proposing and revising supported features. At the beginning of the `Novelette` project, developers identified educators' and learners' needs regarding creativity support, expected features in a digital storytelling platform, interface requirements mainly by looking at the state-of-the-art. Consequently, the requirement elicitation phase of `Novelette` took place as a user-centered design approach. Once identified the project goals, elicited initial requirements, and prototyped the proposed interface of `Novelette`, developers actively involved interested educators in testing the platform, verifying in first person the ease of use, and suggesting either integration or modifications to supported features. `Novelette` has been iteratively refined according to educators' feedback by following the action research process, widely explored in the literature for improving educational practice [151]. This process involves a plan, action, evaluation, and reflection to gather evidence to implement change in practices [60]. Similarly, during the `Novelette` design, developers plan and implement functionalities according to the feedback collected during the previous round, end-users are inquired to evaluate the resulting prototype, and reflections are collected to revise the tool features during the following round.

REQUIREMENTS ELICITATION PHASE - *user-centered design.* Novelette supports educators and learners in inventing and authoring stories by improving their creativity and posing no limit to their imagination. Goals, requirements, features, user interface, and interaction model have been proposed by taking into account the state-of-the-art, collecting educators' needs, and analysing educational digital storytelling tools.

Digital storytelling platforms usually mainly support users in presenting their stories. On the opposite, Novelette also scaffolds learners in inventing them by embedding techniques proposed by Rodari. The introduction of narrative artifices in a digital learning environment represents the most noticeable novelty of our proposal. It provides learners with a digital tool to develop their creativity and experience divergent thinking by combining linguistic and literary artifices borrowed by educational literature.

It resulted in the proposal of the Novelette project, a digital learning environment to support educators and learners in inventing and authoring stories. The proposed approach is general enough to be adopted in a wide range of educational subjects and contexts, from the literature and classic tales to historical events and war undertakings, from theorem demonstrations to project presentations.

ACCEPTANCE PHASE - *participatory design.* In December 2019, the Novelette project was introduced to more than one hundred schools. Participants voluntarily joined the 1-hour presentation session in person. This event was part of a series of project demonstrations to schools and educators organised by an Italian educational company, BIMED[2], as an opportunity to bring together developers, technicians, research groups, and companies interested in proposing a solution for the educational field and schools as end-users. This event behaved as a user acceptance phase and as an opportunity to collect a list of interested school representatives participating in periodic meetings with our research group and proposing Novelette at their schools. Since the planned meetings had been thought to be in person, the Novelette project mainly collected participation from local partners.

[2] BIMED: https://www.bimed.net

The event concluded with a hands-on session where every interested educator was encouraged to think about a use case that might take advantage of Novelette. Participants were 17, mainly professors and a school principal. This phase resulted in a list of educators interested in participating in the Novelette project, the collection of inspiring use cases, and suggestions to move to the prototyping phase.

PROTOTYPING - *user-centered design.* During the prototyping phase, developers designed and implemented the Novelette architecture, interface, workflow, and features according to common patterns and expectations observed in the literature by analysing similar tools.

Novelette addresses privacy issues related to creating accounts for minors, the proposal of an interface provided with all the features expected by a graphical editor, such as the provision of an image library to select characters and background, the selection of colour as a background, components manipulation such as image size control or text modification. The proposed interface limits as much as possible the used text attached to supported features often replaced by icons and images to address internationalisation issues, be inclusive and also enable its exploitation by very young learners.

This phase resulted in a working prototype used to collect educators' feedback and suggestions to improve Novelette.

1ST ROUND OF THE novelette REVISION - *participatory design.* In January 2020, a hands-on session with 10 educators took place. It was performed in the ISISLab at the University of Salerno. Every educator was provided with a computer to test the platform in person.

First, a tutorial session took place to overview Novelette features already available at this stage, mainly concerning the editor interface offering basic features to manage scenes, such as add, clone and delete scenes, the image library to select characters and background, the use of media content such as images and text. Then, a hands-on session was performed by letting participants free to create their stories without any constraint or pre-defined track by playing the student role. Three observers joined this phase to take note of all the raised issues and educators' comments, such as proposed features, improvements, needed clarification. A brief questionnaire was administrated to organise feedback and suggestions to inspire new features incorporated in the following implemented version. We report the questionnaire questions and the most common collected replies:

- *Comments and suggestions on the layout.* Import images from local storage and enhance the control of the educator side by enabling learners' stories inspection and modification.
- *Suggestions to improve produced stories.* Provide a wider support for image modification, e.g., mirroring and positioning.
- *Other comments.* Continue the collaboration between the `Novelette` developers and end-users.

This phase resulted in a collection of technical improvements to make `Novelette` easier to use, both on the teacher and the student side. Moreover, educators manifested a great interest in the performed collaborative design and in continuing it to improve `Novelette` further and use it at school.

2ND ROUND OF THE `novelette` REVISION - *participatory design.* In February 2020, the 2nd round of revision took place. As in the previous round, all the interested partners joined. In particular, 7 educators succeeded in participating, fewer than in the previous round due to the short notice. The performed protocol reflected the one followed during the previous round by first, overviewing the platform and the introduced features, second, performing a hands-on session, and finally collecting suggestions and feedback by a questionnaire. The collected proposals were mainly related to the undo mechanism, the possibility to attach metadata to the authored story, such as a title and a brief description, and minor issues related to text management in the creator component. As in the previous round, this phase resulted in suggestions to improve `Novelette`.

10.3 TOOLKIT: NOVELETTE

`Novelette` is a digital learning environment that supports educators and learners in inventing and authoring linear stories by providing target users with an editor interface to perform digital storytelling and features to support learners in inventing stories by thinking out of the box and improving their creativity. The inventing step is enabled by embedding in `Novelette` the incipit and the suggestion mechanisms. `Novelette` poses no limits to users' imagination and no constraint on the story topic and layout. It is based on *student* and *teacher* roles. While learners play the storyteller role as they can invent and tell any story of interest, teachers can organise classes, supervise created stories, and publish them according to the class management features.

CLASS MANAGEMENT FEATURES. Users who play the teacher's role are provided with managerial dashboard (visible in Fig. 10.2) which enables class management features summarised in Fig. 10.1. First, teachers have to create classes to organise learners. They may create an arbitrary number of classes that can model real classes, merge two or more classes, or split learners attending a class into groups. For each class created in `Novelette`, teachers can register learners' accounts to let them access the `creator` component, which is the story editor interface. `Novelette` supports teachers in automatically generating a set of anonymous accounts to overcome privacy issues with minors.

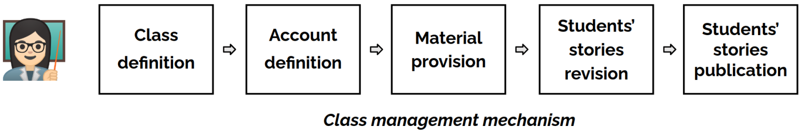
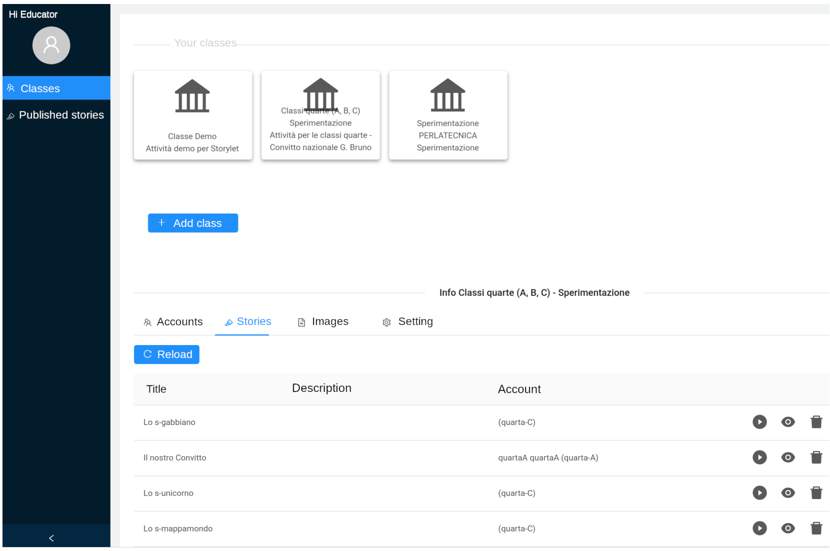

Class definition
⇨
Account definition
⇨
Material provision
⇨
Students' stories revision
⇨
Students' stories publication
Class management mechanism

Figure 10.1: Class management features workflow in `Novelette`

Educators can revise learners' stories and, finally, publish them to make stories visible also out of the class sandbox. While each learner has only access to the authored stories, teachers have visibility at the class level. Educators can perform basic story manipulation, such as rendering or removing.

Novelette also simplifies the mechanism to share material among teachers and learners. It represents an opportunity to avoid asking pupils to look for images of interest of the Web without a supervised approach. Furthermore, it supports educators in distributing material with the entire class at once via an inner sharing mechanism. For each class, educators can upload suggested images. Learners can access both images shared by educators and default images that are part of the internal picture library of Novelette via the creator interface.



Classe Demo
Attività demo per Storylet

Classi quarte (A, B, C)
Sperimentazione
Attività per le classi quarte - Convitto nazionale G. Bruno

Sperimentazione
PERLATECNICA
Sperimentazione

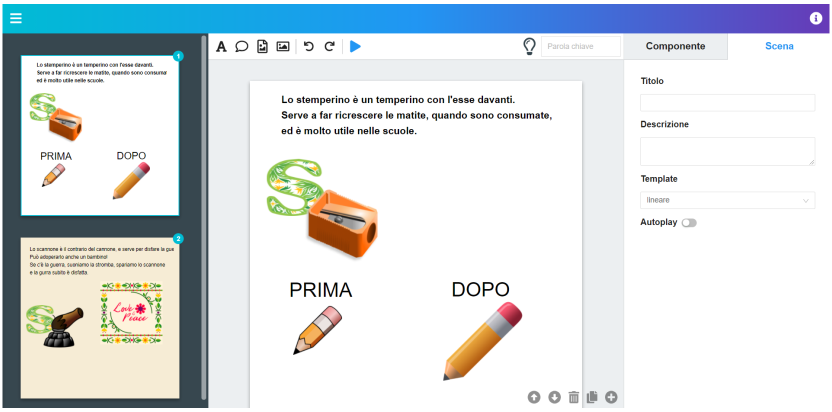
Info Classi quarte (A, B, C) - Sperimentazione

Title	Description	Account
Lo s-gabbiano		(quarta-C)
Il nostro Convitto		quartaA quartaA (quarta-A)
Lo s-unicorno		(quarta-C)
Lo s-mappamondo		(quarta-C)

Figure 10.2: Educators' dashboard in Novelette

THE INCIPIT MECHANISM. Educators may provide learners with incipits as templates. Teachers can introduce a topic by providing learners with the initiation of a story, and learners are encouraged to continue the story starting from this premise. In the alternative, teachers can bound the story initiation and its ending and ask learners to interpose their story by satisfying these constraints. For instance, it can be used to narrate different versions of the same story where there are two or more events that everyone knows, and learners have a different view of what happened in the meanwhile.

Fig. 10.3 shows the Novelette creator interface when learners start with an educator incipit. In this case, teachers have provided



Lo stemperino è un temperino con l'esse davanti.
Serve a far ricrescere le matite, quando sono consumate,
ed è molto utile nelle scuole.

Lo scannone è il contrario del cannone, e serve per disfare la gue
Può adoperarlo anche un bambino!
Se c'è la guerra, suoniamo la stromba, spariamo lo scannone
e la gurra subito è disfatta.

Titolo

Descrizione

Template: lineare

Autoplay

Figure 10.3: `Novelette` creator interface and the Incipit mechanism

their classes with a Rodari's tale "*The Land starting with S*" that narrates about an island where all the object names start with S, and this feature completely alters object properties. For instance, the *s*-harpener can stretch pencils instead of shortening them, or the cannon becomes a *s*-cannon and, instead of attacking enemies, brings peace to all. It is worth noticing that in the Italian language, the prefix *s* is usually used to define the opposite of a term. For instance, given the term *legare*, literally *to tie*, its opposite is created by putting an *s* in front of it. Consequently, *s-legare* means *untie*. Thus, the prefix *s* in Italian behaves as *in, un, im, ir, il, over* or *dis* prefixes in English. However, just saying that pupils have to tell stories about the opposite of a given word is restrictive. The *S* should be perceived as an opportunity to make ordinary objects able to behave in a completely different manner. Thus, pupils are invited to think about an ordinary object and imagine how the prefix *S* can completely change one of its features, leading to an unexpected and surprising development.

DIGITAL STORYTELLING EDITOR. Users who play the student role can author stories according to the workflow described in Fig. 10.4. Once authenticated, learners have to select one of the available templates. Some of them are default templates proposed by the platform, e.g., the empty template, while others represent the educators' defined incipits. Once chosen the starting point,

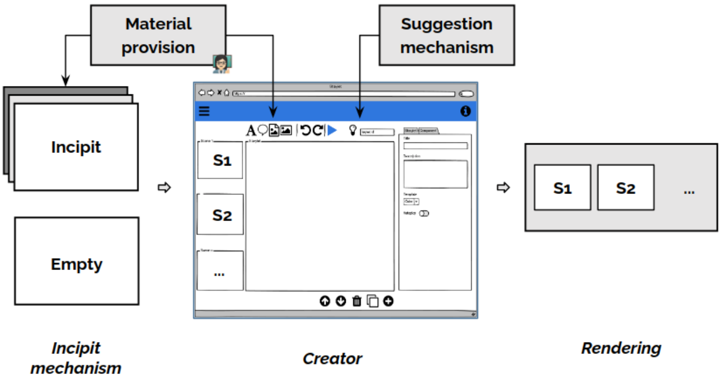


Figure 10.4: Workflow to create stories on the student side

learners access the `creator` component, i.e., the creator interface to invent and author stories visible in Fig. 10.3.

The `Novelette` creator component offers a clear and concise interface where icons and tooltips clarify all the performable actions to meet intuitiveness requirements. The massive usage of icons partially overcomes the internationalisation issue, guarantees inclusiveness, and poses no age limit to its target audience.

Stories are a sequence of scenes, and each scene can contain media content such as text, characters, and a single background. As a story editor, the `Novelette` creator provides access to a library of images, including characters and backgrounds. Educators can share images with the entire class, as described before, to enhance the default library of images. Finally, learners can freely upload images from the local storage, e.g., images drawn by themselves. In this way, `Novelette` behaves as a learning system able to integrate traditional learning mechanisms, e.g., pencil and papers artifacts, with technology-enhanced learning strategies [12].

The possibility to import images of any style, such as images downloaded by the Internet, drawn or authored by learners, reinforces the absence of constraints of the `Novelette` applicability. Learners can upload charts and create media or data stories, pupils' authored images to personalise the story style, photos, or web content to narrate tales, deeds, or any topic of interest. Once selected images, `Novelette` supports learners in performing basic manipulation, such as changing image size, rotation, orientation.

Novelette offers the forgiving mechanism, also known as the undo mechanism, to erase performed errors. The recovery operation is a crucial feature for a graphical interface, above all in learning systems, as it enables exploratory learning without fear [224] with the possibility to revisit work under definition [348].

Once learners have invented and created their stories, they can visualise them by selecting a rendering template that determines the visualisation approach in the viewer component. For instance, learners can opt for a linear layout to show the scenes in sequence. Alternatively, they can opt for a circular layout, where scenes are presented on a wheel, or a cube, where scenes are put on cube faces. Some of the rendering templates pose a constraint on the number of scenes that can be visualised, e.g., the cube can use at most six scenes. The realised artifact can be exported as an HTML component and embedded in any website.

THE SUGGESTION MECHANISM. If users experience the blank page syndrome, Novelette offers a suggestion mechanism that proposes analogies, synonyms, and rhymes based on a user-defined word. Starting from a word of interest, Novelette groups synonyms by senses. For example, given heart, it can be interpreted as love with affection as synonym, organ with cardiac organ as synonym, centre with nutshell and core as synonyms, or as card seed.

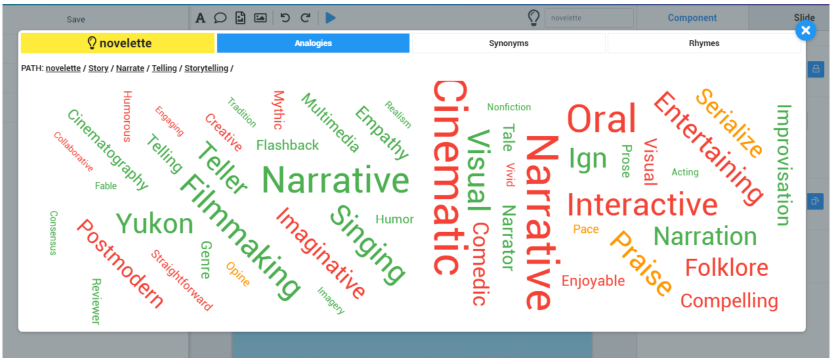


Figure 10.5: Suggestion mechanism in Novelette

Besides senses and synonyms, users can also navigate analogies and associations, i.e., words, adjectives, and verbs related to the word of interest due to the co-occurrence of in-jokes, idiomatic expressions, spoken language, or literature. For instance, a dragon is a fictional animal, usually a character in a magician tale where also wizards appear. Therefore, *animal, magician*, and *wizard* are analogies of *dragon*. Suggestions should not be merely considered a means to avoid repetitions but should inspire writers by letting them play on words and explore associations.

According to Rodari, analogies should enable narrators to create a chain of words and rapidly follow thoughts by moving from one word to another. Thus, analogies are navigable in `Novelette`. Starting from a word, users can move from one word to another until they feel inspired by the obtained chain of words.

`Novelette` visualises senses and analogies via word clouds, widely explored in the literature to facilitate storytelling [299]. As an example, Fig. 10.5 reports analogies *Storytelling* where storytelling is the last word of the path starting from `Novelette` and corresponding to `Novelette > Story > Narrate > Telling > Storytelling`. Word size reflects the association weight, while colours distinguish different parts of the speech. For generating suggestions, `Novelette` relies on external services. In particular, it exploits BabelNet to look for synonyms, WordAssociations[3] for analogies, and RhymeBrain[4] for rhymes.

BabelNet [230] is a huge multilingual semantic network that integrates lexicographic and encyclopedic knowledge from WordNet and Wikipedia. A semantic network is a knowledge base that represents semantic relations between concepts in a network. It is a directed graph consisting of vertices, which represent concepts, and edges, which represent semantic relations between concepts [306]. Semantic networks, a.k.a. knowledge graphs, are crucial for KM and IR [81]. However, their query languages are difficult for lay users, such as pupils [24, 329]. KG query mechanisms should mask the underlying complexities to allow users without technical skills in querying languages to exploit semantic networks' content easily. Thanks to navigable word clouds, learners, regardless of their age and technical skills, can implicitly query huge semantic networks and take advantage of them during learning phases without being forced to develop additional expertise in data modelling and query languages.

[3] WordAssociations: https://wordassociations.net/

[4] RhymeBrain: https://rhymebrain.com

ARCHITECTURE. Novelette, freely available online with an open-source license[5], realises the portability as it is a web-based platform composed of four independent ReactJs web applications. Each application is implemented as a modular and reusable component meeting the modularity requirement.

Novelette is based on a client-server architecture, as represented in Figure 10.6. The server-side part is implemented in PHP. It is modelled by the model-view-controller pattern. The model represents all the persistent data, such as accounts and classes, templates, scenes, and stories. Each class model is made persistent in the database by a dedicated table by avoiding the query execution. The control is represented by a set of RESTful services that give access to the available resources to populate the user interface by a profile access control. The Rule Manager is in charge of verifying the permissions according to the user role. A WordPress plugin manages data and accounts and we map WordPress roles to Novelette roles. The *view* represents the components of the interface: the editor or creator, the viewer, the administrative dashboard, and the student area. Each client-side component is an independent ReactJs web application by guaranteeing the portability and modularity requirements.

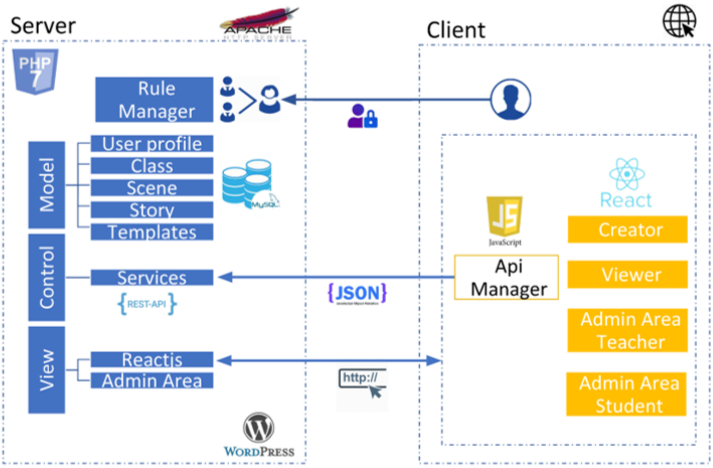


Figure 10.6: Novelette architecture

[5] Novelette homepage: http://www.isislab.it:19984/en/home-page-2. Novelette source code: https://github.com/routetopa/storylet

Internalisation is achieved through application settings and a third-party library that dynamically loads the language resources according to the system configuration, such as the language set by the user. At the moment, Novelette is available in English and Italian, but it is designed to be easily extended with other languages, thanks to its modular implementation.

10.4 EVALUATION: USABILITY ASSESSMENT

This section reports the *usability* assessment of Novelette. According to the ISO 9241 standard definition, usability is defined as *"The extent to which targeted users can use a product to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use"* [150]. Usability is recognised as a crucial factor when evaluating educational technology in schools as it can affect the teaching-learning process, such as learners' attitudes towards the software could be influenced by usability issues that were not carefully taken into account during the evaluation [39]. A poorly designed interface may make learners feel lost, confused, or frustrated, and it will hinder effective learning and information retention [12]. Including end-users in design phases and detecting usability problems at an early stage can lead to adequate solutions and effective learning systems [39] as lack of usability may obstacle pupils' learning [12, 39, 69]. As Novelette supports educators and pupils in performing educational digital storytelling, collecting opinions from both target groups is crucial. Thus, we conducted a two-round assessment. First, we tested Novelette by involving educators participating in the design approach and performing the assessment in a controlled environment. This educators' evaluation round verifies if and to what extent Novelette satisfies end-users expectations maturated during the design approach. Second, we evaluated the usability of Novelette in a real setting by involving three primary school classes that used Novelette in a real formal context, at school, without being supported by developers.

The following paragraphs analyse each round of evaluation by detailing the research questions, participants, evaluation settings, and achieved results. It describes the adopted protocol and the used questionnaires in as many details as possible to make it replicable. Table 10.2 summarises the key aspects of the performed evaluation and the achieved results.

10.4.1 *Evaluation with educators*

The main goal of the research described in this article is the `Novelette` usability assessment according to educators involved in the co-creation process.

10.4.1.1 *Evaluation design*

METHODOLOGY. By posing the evaluation objective as a RQ, *RQ1* corresponds to "To what extent educators consider `Novelette` usable?".

PARTICIPANTS AND SETTING. 11 users from the educational context were involved in the evaluation: 10 of them were educators, while a single user collaborated with schools. Participants were all female from different institutions. 9 out of 14 are primary school classes, 3 out of 14 are secondary school classes and a single High-school institute. In one case, a participant replied that she is interested in contacting several schools without quantifying the number or qualifying the school level. All participants were Italian. Thus, both the tasks and the evaluation questionnaire were in their native language. Although they were geographically distributed in Italy, there was a predominant concentration of participants from the Campania region, due to the desire to perform meetings and experiments in person, when possible.

While the meetings described in the collaborative design have been performed in person, the evaluation described in this section has been held online due to the constraints posed by the COVID-19. Most of the participants also joined previous meetings, while two users are completely new.

During the evaluation, a `Novelette` developer played the moderator role in overviewing the `Novelette` features and conducting the protocol, while 3 developers were always available to clarify doubts and solve any technical issue.

Table 10.2: Summary of usability assessment of Novelette

	Educators	Pupils
Participants	11	49
Setting	controlled environment	formal setting, at school
	PROTOCOL	
1.	Novelette features tutorial	
2.	The incipit challenge	
3.	The suggestion mechanism	
	PROTOCOL PHASES' DURATION	
1.	15 minutes	1 hour
2.	40 minutes	3 hours
3.	40 minutes	3 hours
	DATA GATHERING	
	SUS [181, 265]	
	BI	
	RESULTS	
SUS	76	75
	BI - intention to use Novelette again	
Range	1–7	1–5
Min	4	1
Mean	5.5	4.6
St.Dev.	1.2	0.7
Median	6	5
Max	7	5
	BI - intention to suggest Novelette to others	
Range	1–7	1–5
Min	4	1
Mean	6	4.6
St.Dev.	0.9	0.9
Median	6	5
Max	7	5
Favorite feature	incipit (10/26) usage simplicity (6/26) suggestions (5/26) interface (5/26)	creativity support (19/49) usage simplicity (17/49) incipit (9/49) suggestions (4/49)

PROTOCOL. The evaluation took place online remotely, and it lasted 2 hours. The moderator asked all the participants to play the student role and assess the usability of the `Novelette` creator. The moderator, first, introduced `Novelette` by focusing on its key aspects to deal with participants who have never experienced `Novelette` before and enabling all the participants to be aware of the available features. Second, participants were challenged to solve pre-defined tasks concerning literary artifices inspired by Rodari embedded in a digital learning environment. Finally, participants were invited to fill in a questionnaire to note task-based observations and assess the `Novelette` usability.

Phase 1 - Novelette features tutorial - 15 **minutes.** The moderator performed a tutorial of `Novelette` main features by pointing out the working mechanism from the student side and the novelties introduced after the second meeting of the participatory design approach. Participants have been spectators of this phase.

Phase 2 - The Incipit challenge - 40 **minutes.** Participants were provided with a template narrating the initial part of a Rodari's tale, *The Land starting with S*, visible in Fig. 10.3. As anticipated before, this tale narrates of land where object names start with S, and it completely alters an object's property. By putting it in `Novelette`, we provided participants with an incipit to continue. This task assesses the easiness of starting from a non-empty template and continuing someone else's story.

Participants were invited to start from this incipit and invent a short story composed of one or a few scenes by inventing original objects whose features are completely revised by putting an S in front of its name. Participants were invited to select one of the images proposed by the `Novelette` library, add the image corresponding to an *S* in front of the chosen object, and add a caption describing the invented object. Once the time elapsed and all the participants submitted their stories, a showcase of the realised artefacts completed the task.

Phase 3 - Suggestion provision challenge - 40 **minutes.** The second task concerned the assessment of the usability of the suggestion mechanism. Participants were spurred in freely choosing a word, navigating the resulting analogies, and creating a story based on the experienced breadcrumb. To introduce this challenge, the moderator showed a guided example (1 minute long) where, starting from `dragon`, navigated the proposed analogies by experiencing the path `dragon > breathe > breeze > cool > mint > chocolate > ice cream > kiosk > itinerant`. Finally, the moderator showed the authored story concerning a little dragon whose breath smelled of mint because it ate an ice cream bought from an itinerant kiosk.

DATA GATHERING. Educators filled in an ASQ [180] using a 7-scale questionnaire for each task to evaluate 1) the degree of the perceived difficulty of the task by performing it through `Novelette`, 2) if the time to complete the task is perceived reasonable, 3) if the provided knowledge in the training phase is sufficient to complete the task. While the first question provides an insight into the perceived usability, the second question subjectively quantifies the pleasure in using `Novelette`, while the latter estimate the facility in learning how to use `Novelette`. Participants were free to notify any experienced difficulty as an open question. At the end of the evaluation, the moderator provided participants with a questionnaire to i) evaluate the system usability by a standardised questionnaire, widely adopted in the educational field [330], the SUS [181]), and ii) verify the interest in using and proposing `Novelette` by the BI survey. Finally, participants were invited to report any comment and suggestion mainly as open questions. Among others questions to collect comments, participants rated their favourite feature(s) among the interface, the usage simplicity, the possibility to start from an incipit, and the suggestion provision. Multiple selections were enabled.

10.4.1.2 *Results*

The SUS score is 76. According to its interpretation, all the values between 68 and 80.3 classify the system as *above the average*. Hence, `Novelette` reaches a *good* usability level according to educators.

Concerning the BI results, there is an overall intention to reuse (mean score of 5.6 out of 7) and the real intention to propose the system to others (mean score of 6.2 out of 7).

More qualitative insights are enabled by the ASQ results, direct observations of moderators, and the collected comments as open questions. Participants assess that it is easy to start from an incipit (the 1st question related to the perceived complexity of beginning from an incipit within `Novelette` has a mean score, M of 6 out of 7). They consider the employed time to create a story starting from an incipit (average time of 23 minutes) reasonable (the 2nd question gained M equals 5.8 out of 7). The most promising result is related to the required explanation to create the story: they assess that the short introduction to the `Novelette` environment is enough to use the platform (M of 6.7 out of 7).

Concerning the analogy lookup phase, they assess that the difficulty in inspecting the analogies is reasonable (M of 5.1 out of 7). During this challenge, participants experienced connection problems that negatively affected the user experience and the required time perception. The question related to the observation related to the time needed in analogies lookup and exploitation gains a mean score of 4.9. In this case, the most promising result is related to the need for a concise introduction to the feature exploitation (M of 5.7 out of 7).

About favourite features, the incipit is the favourite feature (10 out of 26 votes), followed by the usage simplicity (6 out of 26 votes), while the interface and the suggestion provision gain 5 out of 26 votes. Besides aspects strictly related to the `Novelette` environment, educators appreciated the continuous involvement in designing and testing the platform and proposed several exploitation contexts. They were also interested in proposing `Novelette` to their school principals in the next years.

They suggested introducing audio support, improving text management, and providing a wider library of images and characters. These suggestions inspire features implemented in future `Novelette` releases.

10.4.2 *Evaluation with children*

This section reports about the usability evaluation of `Novelette` according to children in a real setting at school.

10.4.2.1 *Evaluation design*

METHODOLOGY. By posing this evaluation objective as an RQ, *RQ*2 "To what extent children consider Novelette usable?".

PARTICIPANTS AND SETTING. 49 children of an Italian primary school, "Convitto Nazionale Giordano Bruno", are involved. 49% females, 10 years old. Participants were divided into three groups according to the attended class. The same protocol has been followed in each class. The activity took place during April and May 2021 in a formal setting at school. It has been proposed and moderated by school educators as a curricular activity. Children's parents authorised their participation through a written consent form, signed by both parents, by demonstrating to be aware of the exploitation of Novelette and to authorise the collection of feedback and comments. The University of Salerno undertook post-workshop data processing in an anonymous form to meet data protection requirements.

PROTOCOL. The activity took place in presence during 1-hour lessons per week dedicated to computer science and technology. It spanned over 8 weeks to give every child the possibility to actively join the activity. Learners used a shared laptop during each lesson and, collaboratively, invent and tell their stories based on themes proposed by the moderator. First, the moderator introduced Novelette and its main features as none of the involved children experienced Novelette before. Among the proposed activities, learners experienced the incipit and the suggestion mechanisms, and all of the participants invented and authored a story via the Novelette creator by playing the student role. In both the invention and storytelling stages, participants were assisted by the moderator if needed.

Phase 1 - Novelette features tutorial - 1 **hour.** The moderator introduced Novelette by showing how to access the platform, choose a template, create a few story scenes, inspect available characters and backgrounds, add text, and apply basic manipulation to images (e.g., reduce size) or text (e.g., change the font, colour, background colour). It allowed participants to familiarise themselves with the Novelette interface. Participants have been spectators of this introductory phase.

Phase 2 - The Incipit challenge - 3 hours. As in the educators' evaluation, the moderator participants to test the incipit feature by continuing *The Land starting with S*. The moderator guided pupils in choosing a character among the available ones, inventing how the *S* can alter its behaviour, and creating a scene to tell the invented story. One by one, children were invited to create a story via Novelette by working independently, asking for the mentor's or peer support if needed.

This challenge was experienced as the possibility to make the imagination bloom without constraints. Pupils made ordinary objects magicians thanks to the superpower infused by the prefix *S*. They had fun in proposing objects behaving exactly the opposite of their regular behaviour. For instance, Pinocchio is a tale where the main character is a liar, instead of the *s*-Pinocchio always tells the truth and gives suggestions to pupils. Most of the pupils invented objects able to make them happy, e.g., the coffee cup that contains liquid happiness or the *s*-vacuum and the *s*-hair drier that blows dreams, the *s*-rubber that tells jokes. Some stories went in the direction of overcoming fears, e.g., the whale that does not kill but plays with children. Others preferred playing on words, e.g., the starfish behaves like a star by shining in the space instead of staying underwater.

Phase 3 - Suggestion provision challenge - 1 hour. The second task tested the suggestion provision mechanism to verify the usability of the proposed mechanism and to assess to what extent children get inspired by analogies. During this task, educators encouraged learners to start from a word assigned by the teacher, i.e., *participation* or *happiness*, and use Novelette to find out analogies that mainly surprised and inspired them. It is worth noting that pupils were unaware of the meaning of these words and tried to figure it how by exploring suggested analogies. Learners were invited to select 10 words from the suggested ones and note them on a sheet. Then, they picked a single word from the pool of 10 and were challenged to justify the relationship between the starting word and the chosen one by a graphical representation and a short textual depiction.

Starting from *participation*, pupils mainly chose i) *athlete* as athletes participate to matches and enjoy competition; ii) *football* as a team sport to play and team members participate as a group; iii) *manifestation* as an opportunity to share opinions and participate with people interested in the same field to disseminate information on crucial topics, e.g., on the environment and pollution; iv) *collaboration* as participate also means collaborate as in the participatory design; v) *share* as participants may share thought, objects, feelings, experiences and expertise.

Starting from *happiness*, pupils chose i) *serenity* as you are happy if you are at peace with everyone; ii) *joy* as a synonym of happiness; iii) *contagious* as people surrounded by happy people may become happy as well; iv) *bliss* as a fundamental feeling that every child should experience, always; v) *toast* as a way to manifest happiness by making a toast with friends.

DATA GATHERING. For each task, pupils were invited to fill in a simplified ASQ asking for i) the reached engagement level formulated as "Did you like to perform this task?", ii) the easiness in completing the task, and iii) if they required help. While the first two questions asked a rate according to a 5-Likert scale, the latter required a Yes/No reply. Moreover, they can provide feedback and notify difficulties as a free text.

At the end of the evaluation, the moderator provided participants with a final questionnaire to i) evaluate the system usability according to a simplified version of the SUS [265], and ii) the interest in using it again and proposing the tool to friends by a BI survey. In the BI, children rated if they i) desire to use `Novelette` again and if they ii) would propose `Novelette` to friends. Ratings were given with the Smiley-o-meter 5-point Likert scale, ranging from not at all (1) to very much (5). Finally, final comments were collected as open format questions inviting children to i) suggest what they would add, modify, or improve, ii) highlight what they liked more, and iii) what they liked less. Questionnaires used a language adequate to the age range of participants and adopt visual analogue scales for answers, which employed icons or images, easy to interpret by children, and concerning their feelings and opinions.

10.4.2.2 *Results*

Concerning the usability evaluation, the SUS score is 75. According to the SUS score interpretation, all the values between 68 and 80.3 classify the system as *above the average*. That means that `Novelette` is classified as *good* by children. 17 out 49 reports complimented about `Novelette` and its features stating that "*`Novelette` is super-adapt to child and it is easy to use*" and "*it is perfect as it is despite minor issues*". 2 out 49 asked for further simplifying `Novelette` by fixing technical aspects concerning the text or image manipulation.

About BI, there is an overall intention to propose `Novelette` to friends (mean score $M = 4.6$ out of 5, standard deviation Standard Deviation (SD)$=$ 0.9, 5 as median, minimum value $min = 1$ and maximum value $max = 5$). Moreover, there is an overall intention to reuse (mean score $M = 4.6$ out of 5, standard deviation $\text{SD} = 0.7$, 5 as median, minimum value $min = 2$ and maximum value $max = 5$). 9 out of 49 asked for unlimited access to the platform by their PC or mobile devices. They also suggested providing access to `Novelette` without a password to make it accessible to a wider range of users.

More qualitative insights are offered by the modified ASQ results, direct observations returned by moderators, and feedback and suggestion reported as open questions. According to the first question of the modified ASQ, its mean score is at least equals to 4.49 out of 5 (45 replies), and its value is slightly higher in the task of narrating funny stories ($M = 4.62$ out of 5, 45 replies). According to the second question, modelling the easiness in performing the invention and the story authoring reached a mean score of at least 4 out of 5 (44 replies), but increased by experience. It reaches a slightly higher mean score during the second task, 4.14 out of 5 (43 replies). Similarly, during the first task, slightly less than half of the participants (20 out of 45) stated that they required help in accomplishing the assigned task, while they decreased to 16 out of 43 during the second task.

The experienced challenges correspond to features that children suggested to improve. In particular, children suggested the introduction of a wider set of images to the default library as background and characters, simplification of the text management, and image manipulation features.

19 out of 49 participants manifested a great appreciation for `Novelette` stating that they felt supported and encouraged in improving their creativity by inventing and creating stories, stating that *"`Novelette` enables them to succeed in letting free their imagination"*. Concerning Rodari's style techniques, 9 out of 49 participants enjoyed the proposed incipit as they experienced the possibility to invent objects altered by the *S* as an opportunity to free their imagination as there were no wrong or right replies, but only novel ideas. Moreover, 4 out of 49 explicitly stated that they loved the suggestion mechanism to navigate analogies and play on words, and they felt scaffold in thinking out of the box and improving their creativity.

All of them agreed that `Novelette` could be exploited to create any story of interest. Once educators asked them what stories they would like to invent and create, they replied by citing traditional tales, stories with their favourite fictional or real characters. Moreover, they would like to narrate about their personal experience (e.g., birthdays or ceremonies), their daily life (e.g., talk about what they do at school), stories about their pets (also transformed into fictional characters with superpowers). Among the most original proposals, we can cite "My crazy year" meaning that this child would like to talk about education at a distance during the COVID-19 pandemic and how it affected their education, friendships, habits, and daily life; stories about serious topics, such as, "The land of books" to highlight the importance of reading; or stories where characters swap their properties, e.g., the story of little red riding hood where the kid is the villain, perceived as an extension of the experienced incipit.

10.4.3 *Discussion*

`Novelette` has been designed not only *for* education but *with* educators as they have been actively involved in the design approach addressing concrete needs that they face daily with pupils. As a result, Novelette is considered *usable* by educators (RQ1) and pupils (RQ2) as it achieved an SUS score of at least 75. Qualitatively speaking, involved participants successfully created stories stating that the interface is easy to be used both according to educators and children. According to involved educators, it is easy to learn how to exploit Novelette, underling the intuitiveness of the tool. Moreover, it demonstrates that users do not require technical support to use Novelette. The same pattern can be observed by working with pupils. They rarely asked for any support and, in most of the cases, peer support was enough to overcome challenges. By discussing with the moderators, children loved working collaboratively, and they asked for peer support to invent and author stories together. They think about the inventing part and the characters together to share and exchange opinions with friends. However, they rarely have concrete difficulties that required the moderators' intervention, and these difficulties were mainly concerned with formatting text that resulted in a slightly artificial functionality. Participants stated that they mainly experienced obstacles in the invention phase, while Novelette fully supported them in authoring the invented story. In fact, according to the ASQ, inventing stories is not an easy task. However, since the experienced difficulties decreased with experience, it seems that there is the possibility to improve creativity and inventiveness *by doing*.

Concerning the intention to use Novelette again or suggest it to others, both target groups seem to be particularly enthusiastic about working again with it. Educators promised to propose it to their classes in the immediate future, and classes joining the evaluation decided to include it in the set of used tools also the following year. Pupils asked for continuing working on it in the summer and requested access to Novelette also at home, by their devices. A similar pattern has been observed by related work, such as Communics [281], where primary school pupils were so excited to use visual storytelling platforms at school that they asked to access them also at home.

Concerning favourite features, educators were particularly interested in the incipit mechanism, probably as they perceived it as a way to distribute and assign tracks to all at once. Furthermore, they considered it easy to use both for them and their learners. Looking at favourite aspects with pupils' eyes, they were enthusiastic about the suggestion mechanism, perceived as the removal of any bannister, complete freedom in inventing unexpected story developments, and the possibility to invent and tell any story of interest.

Use studies concerning the assessment of the role played by digital storytelling at school are not rare. They covers heterogeneous dimensions, such as the collaborative dimension [163, 281], pupils engagement [4, 281], entertainment level [223], usability [345, 346], learning outcomes [66] or improvement in creativity [334]. All these contributions report positive results, demonstrating that digital storytelling is a promising approach to spur imagination and let pupils let in actively improve their literary skills, both individually and collaboratively.

ACCESSIBILITY EVALUATION. As classes might be composed of learners with special needs, it is crucial to assess the accessibility of the proposed tool to everyone, specifically those who have disabilities, allowing them to perceive, understand, navigate and interact with the Web. These disabilities cover all levels, including auditory, physical, speech, cognitive and neurological. Many tools have accessibility barriers that make it difficult for a person with a disability to use their interface. Automatic tools to measure accessibility assists behave as a preliminary check to sure that people with all disabilities do not face roadblocks when accessing the platform. Among others, WAVE[6] and TAW[7] help developers make their web applications more accessible to individuals with disabilities. WAVE can identify many accessibilities issues and Web Content Accessibility Guideline errors but also facilitates human evaluation of web content. They focus on issues that heavily impact end-users, facilitate human evaluation, and educate about web accessibility. We used both these tools to quantify the accessibility of `Novelette` automatically. While TAW returns a single problem concerning the difficulty in automatically retrieving the system language, WAVE states that no errors were detected. It represents a quantitative and automatic

[6] WAVE: https://wave.webaim.org

[7] TAW: https://www.tawdis.net

check that Novelette might be accessible by anyone irrespective of their characteristics.

10.5 EVALUATION: ENGAGEMENT ASSESSMENT

This section reports the engagement achieved by Novelette in remote and in-person activities. Engagement in learning settings is often correlated to learning [123, 204, 234]. Research demonstrated that engaging learners in the learning process increase their attention, motivates them to practice higher-level critical thinking skills, and promotes meaningful learning experiences.

However, engaging children at a distance, during the pandemic, can be very demanding: children sit in front of computer screens and cannot share or jointly work on the same material, as in in-presence design workshops.

To analyse and compare the engagement level achieved in performing digital storytelling by Novelette, we performed and compared an in-presence activity in a formal setting at school and an online remote experience in an informal setting. The following paragraphs analyse each round of evaluation by detailing the RQs, participants, evaluation settings, and achieved results. It describes the adopted protocol and the used questionnaires in as many details as possible to make it replicable.

10.5.1 *In-presence children engagement in formal settings*

This section reports about the engagement evaluation of Novelette according to children in an in presence activity in a formal setting at school. It is worth noting that the same sampling of participants, setting, and protocol has been performed to evaluate both the usability and the engagement of children. However, this information is repeated for completeness.

10.5.1.1 *Evaluation design*

METHODOLOGY. The RQ at the basis of the performed evaluation is "Are children engaged in inventing and authoring stories via Novelette?"

PARTICIPANTS AND SETTING. 49 children of an Italian primary school, "Convitto Nazionale Giordano Bruno", are involved. 49% females, 10 years old. Participants were divided into three groups according to the attended class. The same protocol has been followed in each class. The activity took place during April and May 2021 in a formal setting at school. It has been proposed and moderated by school educators as a curricular activity. Children's parents authorised their participation through a written consent form, signed by both parents, by demonstrating to be aware of the exploitation of `Novelette` and to authorise the collection of feedback and comments. The University of Salerno undertook post-workshop data processing in an anonymous form to meet data protection requirements.

PROTOCOL. The activity took place in presence during 1-hour lessons per week dedicated to computer science and technology. It spanned over 8 weeks to give every child the possibility to actively join the activity. Learners used a shared laptop during the lesson and, collaboratively, invent and tell their stories based on themes proposed by the moderator. First, the moderator introduced `Novelette` and its main features as none of the involved children experienced `Novelette` before. Among the proposed activities, learners experienced the incipit and the suggestion mechanisms, and all of the participants invented and authored a story via the `Novelette` creator by playing the student role. In both the invention and storytelling stages, participants were assisted by the moderator if needed.

Phase 1 - Novelette features tutorial - 1 hour. The moderator introduced `Novelette` by showing how to access the platform, choose a template, create a few story scenes, inspect available characters and backgrounds, add text, and apply basic manipulation to images (e.g., reduce size) or text (e.g., change the font, colour, background colour). It allowed participants to familiarise themselves with the `Novelette` interface. Participants have been spectators of this introductory phase.

Phase 2 - The Incipit challenge - 3 hours. As in the educators' evaluation, the moderator participants to test the incipit feature by continuing *The Land starting with S*. The moderator guided pupils in choosing a character among the available ones, inventing how the *S* can alter its behaviour, and creating a scene to tell the invented story. One by one, children were invited to create a story via `Novelette` by working independently, asking for the mentor's or peer support if needed.

Phase 3 - Suggestion provision challenge - 1 **hour.** The second task tested the suggestion provision mechanism to verify the usability of the proposed mechanism and to assess to what extent children get inspired by analogies.

DATA GATHERING. A standardised self-report questionnaire was administered to assess the children's engagement level. The questionnaire used a language adequate to the age range of participants and adopts visual analogue scales for answers, which employ icons or images, easy to interpret by children, and concerning their feelings and opinions. It had a one-close format question, asking to what extent a child liked to tell stories by `Novelette`. Ratings were given with the Smiley-o-meter 5-point Likert scale, ranging from not at all (1) to very much (5).

10.5.1.2 *Results*

Participants declared to be generally engaged in the inventing and telling stories by `Novelette`: the minimum value is $min = 2$ out of 5, the maximum one is $max = 5$ out of 5, the mean score is $M = 4.8$, with standard deviation SD $= 0.6$ and 95%-$ConfidenceInterval$(CI) equal to $[4.62, 4.98]$. Fig. 10.7 reports actual distribution of ratings over the 1 to 5 range, by splitting results by gender. Results are slightly better in the male group. More precisely, by focusing on females (49%), $min = 2$, $max = 5$, $M = 4.6$, SD $= 0.82$, 95%-CI equal to $[4.25, 4.95]$. By focusing on males (51%) , $min = 4$, $max = 5$, $M = 4.9$, SD $= 0.28$, 95%-CI equal to $[4.8, 5]$. Results are overall positive, albeit limited to being mainly descriptive. Children's ratings were rather high, with male participants showing slightly higher engagement than the female ones.

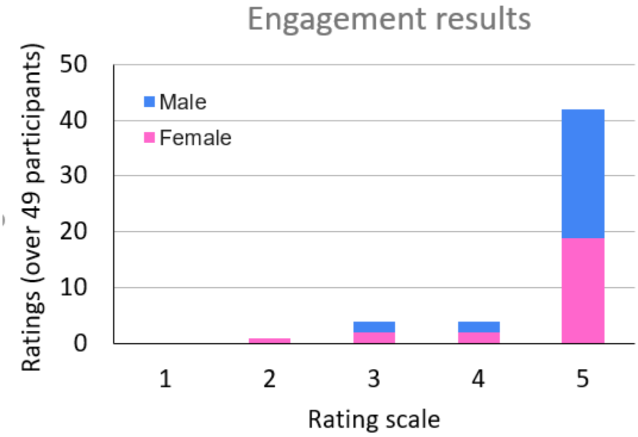


Figure 10.7: Engagement results of `Novelette`

By discussing with involved teachers as moderators to identify the reasons for such a positive engagement, they stated that children felt remarkably free in inventing strange objects with original features thanks to the *S* as initial. The absence of correct replies let participants free to imagine and spur imagination. Funny stories were used by pupils as means to overcome fears (e.g., jellyfishes that kiss instead of biting), solutions to the COVID-19 pandemic (e.g., tanks to be safe), everyday objects with a novel application (e.g., fridges transformed into ovens), approaches to make dreams come true (e.g., globe behaving as a teleporter). Pupils explicitly stated that *they felt free from any constraint and banister*. They started looking around and thinking to what extent the *S* as prefix can alter objects behaviour.

Pupils were so enthusiastic to such an extent to organise themselves in groups (of 4 or 5 members) during the homework time to invent stories together.

Teachers also reported satisfaction for `Novelette`. They appreciated having a controlled and safe environment to exchange materials between teachers and learners. They assessed that the designed interface was so clear and easy-to-use so that most of the children were able to tell their stories without being supported by mentors and, in such cases that they required help, peer-support was enough to clarify doubts and accomplish the task. Teachers had no doubts about the contribution given by the technological solution to engage participants, mainly males. They assessed that females are particularly engaged in inventing and graphically representing stories by traditional means, such as pencils and papers, while males are pretty sceptical about this task. However, by introducing the technological component, males are successfully engaged, as demonstrated by our preliminary analysis. Further exploration is required to verify if the technological component is a critical variable to engage pupils in storytelling and the role plaid by the participants' gender.

10.5.2 *Remote children engagement in informal settings*

This section reports about the engagement evaluation of `Novelette` according to children in remote activity in an informal setting.

10.5.2.1 *Evaluation design*

METHODOLOGY. The RQ at the basis of the performed evaluation is "Are children engaged in inventing and authoring stories via `Novelette`?"

PARTICIPANTS AND SETTING. A total of 9 participants voluntarily joined the workshop, 45% females. Participants are in the range from $min = 9$ to $max = 13$, with a mean value of $M = 10.45$ and standard deviation SD $= 1.24$. As they were all underage, their parents authorised their participation through a written consent form signed by both parents for each participant. A total of 3 researchers participated in the workshop, one as moderator and the others as independent observers. Finally, post-workshop data processing was undertaken by researchers of the University of Salerno in an anonymous way to meet data protection requirements. The workshop took place in June 2020, entirely online and at a distance because of COVID-19 restrictions.

PROTOCOL. First, the moderator performed the exploration phase by defining the smart city concept, detailing smart thing components, such as sensors and actuators, and challenging participants with questions concerning the smart city and smart things. Then, the moderator explored the smart city, sensor, and actuator definitions via `Novelette`Ṡecond, the participants were invited to start from the Roobo template embedded into `Novelette` and author a story concerning an invented smart thing and describe its interaction with the environment. Finally, children reflected on the invented smart things by detailing the required sensors and actuators and narrating the interaction model through the authored story. All the participants worked independently, asking for moderator support if required. The workshop took 2 hours.

DATA GATHERING. Qualitative data concerning affective engagement were collected by two independent observers that classified learners' interest, attitude, curiosity, and task absorption by the BROMP approach [239]. According to it, observers classified each child as *concentrated*, *delight*, and *surprised* modelling positive engagement, *bored*, *frustrated*, and *confused* as negative engagement. A third researcher summarised observers' evaluations in a meta-review.

10.5.2.2 *Results*

The observers explicitly stated that all the participants were highly engaged and interested in the project. They agreed that participants were engaged and delighted to invent smart things via `Novelette`. It demonstrates that storytelling and `Novelette` successfully engage children in inventing smart things.

The observers noticed that participants were concentrated until they were employed in a task. Immediately after the story completion, they became eager to receive another task and bored while waiting for slower storytellers.

Children rarely participated in the reflection phase of other participants, probably because they were still focused on their stories or due to their young age. The protocol should be revised to encourage participants to join the reflection phase of all projects actively.

Finally, observers underlined their difficulty in being sure of the perceived mood and attitude due to the remote nature of the workshop. It requires further investment in supporting observers in drawing conclusions during at-a-distance activities.

10.5.3 *Discussion*

This section reported about two different experiences to verify the engagement level achieved by `Novelette` according to pupils. The performed experiences differ from the number of participants, setting, proposed themes, and data gathering approach. However, in both cases, participants are positively engaged in inventing and authoring stories by `Novelette`. Thus, `Novelette` is perceived as an inspiring and not intimidating environment to learn and speculate about any topic of interest.

By focusing on the smart city experience, the reported activity is part of a series of workshops designed to support children in familiarising themselves with smart city and smart object concepts. While the term smart city and related concepts are frequently mentioned in the news, their meaning is unclear for people without a technical background and mainly for children. Thus, children require opportunities to familiarise themselves with smart things components and ideate their smart things [318]. On one hand, the Roobopoli workshops [76] mainly involved participants in the programming of smart cities and autonomous driving. On the other hand, the `SNaP` workshops encompassed an entire design process of novel smart things [256, 275].

The Roobopoli workshops are focused on technologies future citizens may experience during their lives, e.g., autonomous vehicles. They propose an imaginary smart city with pre-defined smart things. They encourage participants in analysing possible behaviours that these smart things may have, expressed as missions, in which participants should propose an algorithmic solution that may creatively solve a standard and well-known problem, and program it. The Roobopoli workshops are more programming-oriented and enable older teens from secondary schools to delve into advanced programming challenge.

The `SNaP` workshops, on the other hand, are focused on the active role future citizens may play in proposing and designing original and novel smart things, and reflecting on them as part of a complex ecosystem. Through their missions, participants need to use their imagination and creativity to ideate their smart things, according to their preference, and make them concrete by programming them.

The SNaP workshops are centred around SNaP, a card-based game toolkit for children and young teens for smart-thing design [118–121]. The main game elements of SNaP are cards and boards for conceptualising smart things. Cards of SNaP are divided into mission cards (i.e., goals that smart things should address), environment cards (i.e., things related to a park to be made smart), and technology cards divided into input and output cards (i.e., sensors and actuators). Technology cards are matched with the physical computing toolkit that children will later use for prototyping their ideas (e.g., [112, 175]). Boards serve to organise cards and conceptualise smart things. SNaP comes into two versions, a physical one and a digital one.

Digital SNaP[8] supports children or teens in the entire design workflow, by, firstly, enabling them to familiarise themselves with smart-thing components (e.g., sensors, actuators) and, secondly, it guides them in ideating novel smart things by conceptualising them on boards. Finally, SNaP automatically generates the programming code corresponding to the idea created, and it displays it in the MakeCode programming environment, with an infinite loop and an embedded if-else conditional [209]. For example, a smart-thing idea uses a button card as input, a card for showing an icon as output, and a basketball card for the environment. Starting from it, SNaP generates an infinite-loop with a nested conditional of the form: *if the button is pressed then show a happy icon, else show nothing*. Therefore, once the ideation phase is completed, children are gently introduced to the programming stage in the MakeCode environment to create more and more complex smart things.

The SNaP workshops encompass all the design stages, starting from the exploration stage, continuing with the ideation, and concluding with the programming and prototyping stage. During the exploration phase, children become familiar with smart things and their components and are guided to create their smart-thing ideas according to the chosen mission. During the ideation phase, children can create as many smart-thing ideas as they wish which will be automatically transformed into a programming code by SNaP. Children can inspect an authored idea, see the resulting code in the MakeCode programming en-

[8] Digital SNaP website: https://snap.inf.unibz.it

vironment, and proceed to the programming and prototyping stage. In this stage, participants explore and get familiar with basic programming constructs and MakeCode blocks with the help of a facilitator. Participants are introduced to the following programming constructs: loops (e.g., *for*, *repeat*), conditionals (e.g., *if-then-else*), comparisons (e.g., $>$, $<$, $=$), and Booleans (e.g., *and*, *or*, *not*) by introducing each concept separately via example programs. Next, participants are challenged to expand the programs of their smart things generated by SNaP by reflecting on the programming constructs they were taught and implementing them in their ideas if they wish so. Subsequently, they can move on to make a physical prototype of their smart things by using physical computing toolkits and paper-based material. Prototyping is not conducted "physically" in case workshops are held at a distance but only simulated.

The SNaP workshop took place in December 2020, in two rounds: the first round involved teen participants aged 11–16 years old, referred to as S-W1; the second involved children aged 8–10 years old, referred to as S-W2. Both were conducted digitally and at-a-distance, in the form of a teleconference. Participants were at home, after school, and they participated voluntarily. Each round lasted 6 hours in total, spanning over 3 consecutive days (2 hours per day). The workshop involved 20 participants (40% females), 10 per round.

A quantitative analysis on children's engagement during the SNaP workshops showed that all the participants were highly engaged [257]. Nevertheless, the digital SNaP may be perceived as a limitation for the oldest participants who wish to acquire more advanced technical skills. They suffered limitations from the block-based interface and asked for a similar workshop where they might be free to use also textual programming languages, such as Python and JavaScript (the ones integrated into MakeCode). Concerning the youngest participants, they had difficulties in visualising the smart thing under construction, how sensors and actuators should be placed on the thing made smart, and how it should communicate with the environment and other smart things.

To help children better experience the interaction mode between the authored smart thing and the surrounding city, we proposed to take advantage of storytelling and `Novelette`.

Storytelling is perceived as a non-intimidating approach in actively involving lay users in the smart city design process [152, 190, 303, 308] as stories represent a natural way of beginning a dialog with users [99]. This approach generalises well-established user stories in the agile development [64], i.e., an informal, general explanation of a software feature from the perspective of end-users. Thus, users, regardless of their technical skills, are encouraged to hypothesise features for the system under design and model requirements, usage modality, users' expectations.

The performed experience demonstrates that `Novelette` supported children in ideating smart things, and defining their interaction model with external actors, such as other smart things and human beings. All the participants completed their stories and detailed smart thing components. Some of the cleverest stories follow. The *observant bin* uses a camera to inspect the surroundings and checks if someone throws the rubbish on the floor and, in such a case, encourages passers-by to use the appropriate bins. The smart *lifeguard seagull* perceives the quiet level of the sea via a sound sensor and recognises the presence of dangerous fish via a camera enhanced by a classification algorithm. The smart *teapot* senses the liquid temperature by a thermometer and turns on LEDs accordingly. The smart *bed* monitors the sleep quality via a movement sensor and, if it registers a scarce sleep quality, turns on a relaxing and soft melody. The *thoughtful toaster* prepares toasts if it perceives a stomach rumbling via a sound sensor.

The described experience demonstrates that children were engaged in narrating smart thing-oriented stories and modelling interactions between objects and people. Further exploration is required to understand if this approach behaves as a valuable starting point to scaffold children in the entire design process.

10.6 FINAL REMARKS

Despite the wide awareness of the advantages of educational digital storytelling, educators still feel that schools lack opportunities, skills, and tools to guide pupils in developing their creativity.

Hence, we proposed Novelette, a digital learning environment for educational digital storytelling. It supports both educators in class management and pupils in inventing and authoring stories. While most of the available tools only behave as editor interfaces, Novelette embeds well-known approaches defined by Rodari to inspire storytellers and guide them in the invention phase.

Novelette has not only to be designed *for* educators, but it resulted by a collaborative design approach to propose a tool that actively involved educators in the entire design and development phases. Hence, Novelette has been actively designed *with* educators. As a result, this chapter also retraces the entire design process by reporting the performed steps and the educators' role, describing the resulting tool, and demonstrating how Novelette can be exploited in a concrete use case at school.

The reported evaluation empirically assesses Novelette usability by involving both educators and pupils. While educators tested Novelette in a controlled environment, pupils evaluated it in a real context, at school, during their curricular activities. As a result, Novelette is very *usable* for both target groups. Qualitatively speaking, pupils appreciated the possibility to use a clear and easy-to-use tool for authoring stories, and they felt the opportunity to increase their creativity. Educators demonstrated high satisfaction by using a controlled environment to exchange materials with learners and supervise pupils' work. Moreover, Novelette resulted from an engaging approach to bring pupils closer to storytelling, both in remote and in-person activities.

In the future, we aim to overcome issues reported by educators and pupils in the evaluation reports. In terms of usability, we aim to assess its usability and ease of use in remote activities as COVID-19 taught us that it is crucial to be ready to move an activity from in-person to a remote configuration [12]. We also aim to quantify the role plaid by the collaborative dimension in terms of learning outcomes and improvement in participants' creativity, similar to the experience reported in the literature [163, 281], by also considering participants with disabilities, such as dyslexia. Finally, as good usability is a precursor of a successful learning approach [69], we aim to investigate further the extent to which Novelette succeeds in supporting pupils in learning and developing creative skills. Thus, we aim to assess the learning and creativity outcomes achieved thanks to the `Novelette` support, by first considering standard approaches used in the literature [66, 334] and, if needed, proposing novel assessment criteria. It would also be interesting to perform a longitudinal evaluation by involving the same participants in a novel evaluation to assess their retention level and the long-term engagement once the initial enthusiasm concerning novelties disappears.

Concerning the smart object design process, storytelling in general and `Novelette` in particular, seem to be a promising approach to let children describe not only the appearance of the designed object but also its interaction model. We are currently working on the proposal of a unified workflow to let pupils explore ideation by card-based and storytelling interface while supporting programming and prototyping with block or text-based programming interface, according to learners' age.

11 KNOWLEDGE GRAPHS AND VIRTUAL EXHIBITIONS

> *Virtual Reality is a way to escape the real world into something more fantastic.*
>
> – Chris Milk

Many private and public organisations, such as Galleries, Libraries, Archives, and Museums (GLAM) institutions, have invested in massive digitisation campaigns to digitise billions of resources [192] and enable knowledge transfer [148]. As a result, CH is the most successful application domain of the Semantic Web technologies [33] due to the vast amount of data published as KGs. The interest of the CH community in KGs can be justified by the fact that LOD behave as a promising approach in facing CH challenges, such as multilingualism, interlinking nature, semantic richness, and content heterogeneity [148]. Thanks to their linked format, LOD allow easy data reusability and integration [33].

Both public institutions and private organisations have invested in KGs. In some cases, data providers prefer to enrich national aggregators and discontinue KGs at the institution level. For instance, Europeana [133] incorporates most of the European KGs concerning museums, libraries, or galleries. The same trend can be observed by the Digital Public Library of America [250].

There is a substantial interest in materialising data by modelling museums, galleries, and libraries content as a KG [44, 77, 96, 134, 149, 168], defining models, mainly tailored to libraries, archives, and museums [91], and precise terminology by thesauri, such as the Art and Architecture Thesaurus [1] and the UNESCO Thesaurus [324].

Due to the extraordinary acceleration in digital transformation processes, we are experiencing in the last years [5], GLAM institutions are exploiting the plethora of available data to improve the spread of culture [5] by creating virtual versions of traditional cultural activities [174].

Virtual exhibitions represent a modern way for cultural institutions to raise public interest in CH [58]. They have been evaluated as one of the most promising means for disseminating cultural content from GLAM institutions in the digital era [58]. Virtual exhibitions may enrich, not substitute, traditional museums by overcoming any physical constraint within the limits of proposed technological solutions [42].

Nevertheless, visitors still behave as spectators [342]. To overcome this limit, my *objective* is letting visitors play the role of virtual exhibition curator. This chapter proposes a virtual reality-based virtual exhibition generator to give CH lovers the possibility to author any exhibition of interest, only limited by their imagination. The research presented in this chapter has been published in the following contribution:

> Daniele Monaco, Maria Angela Pellegrino, Vittorio Scarano, Luca Vicidomini. *Linked Open Data in Authoring Virtual Exhibitions*. *Accepted* in the Journal of Cultural Heritage. 2021. *In press.*

11.1 RELATED WORK

Recently, researchers strive to provide the CH community with interfaces and digital solutions to improve user experience [183]. They must consider the adopted technology and, consequently, the achieved level of immersion to meet the entertainment requirement. Furthermore, looking at the queried data source, developers should satisfy the heterogeneity requirement and enable customisation options. Finally, a crucial aspect is turning the user role from passive visitor to an active curator. Table 11.1 classifies related work in terms of queried data format, user role, and technical details concerning tracking mode, display option, input devices, and achieved level of immersion.

DATA SOURCE. Virtual exhibition generation heavily relies on the CH digitisation process [37, 97]. Virtual museums and exhibitions authored by GLAM institutions may rely on proprietary data, requiring a digitisation phase of owned cultural objects be-

Table 11.1: Virtual exhibition generator related work.

Author(s)	Data	User role	Tracking	Display	Input	Immersion
Wojciechowski et al. [342]	PD	visitor	No tracking	PC	Device	No
White et a. [338]	PD	visitor	No tracking	PC	Device	No
Carvajal et al. [186]	PD	visitor	No tracking/Sensor	PC & HMD	Device	No & Full
Bruno et al. [47]	PD	visitor	No tracking	Screen	Device	Semi
Hsieh et al. [145]	PD	visitor	No tracking	Projector	Hybrid	Semi
Haydar et al. [137]	PD	visitor	Optical	Screen & HMD	Device	Semi & Full
Hernández et al. [140]	PD	visitor	Hybrid	HMD	Sensor	Full
Barsanti et al. [124]	PD	visitor	Hybrid	HMD	Sensor	Full
Katsouri et al. [159]	PD	visitor	Optical	CAVE	Device	Full
Hayashi et al. [136]	OD	curator	No tracking	PC	Device	No
Scarano et al. [288]	OD	curator	No tracking	PC & HMD	Device	No & Full
Di Stefano et al. [86]	OD	visitor	Optical	PC & HMD	Device	Full
Reski et al. [269]	OD	visitor	Optical	HMD	Device	Full
Kiourt et al. [165]	OLD	curator	No tracking	PC	Device	No
Mallia et al. [195]	OLD	curator	No tracking	PC & HMD	Device	No & Full
Minelli et al. [130]	LOD	visitor	No tracking	PC	Device	No
Our proposal [216]	LOD	curator	No tracking	PC & HMD	Device	No & Full

fore creating virtual exhibitions [47, 124, 137, 145, 159, 338, 342]. However, proprietary data (PD) obstacle easy reuse by end-users.

In times of a thriving OD movement [156], the exploration of OD in an immersive VR-based environment can provide users with new insights and perspectives about exploited data sources [269]. There are several efforts to create virtual museums and exhibitions starting from open datasets [86, 136, 269, 288]. However, it is not obvious how to exploit OD while addressing heterogeneity. Users may need to merge datasets containing information of interest before authoring virtual exhibitions, for example combining information about museums and hosted paintings.

Some frameworks [165, 195] exploit multiple (open) data sources, such as Europeana and Google Images, and adequately link them. Kiourt et al. [165] referred to these data as Open Linked Data (OLD). While OLD represent heterogeneous sources ad-hoc combined, LOD represent the effort to provide structured data which are both machine and human-readable [332] attached to the license of free reuse.

Minelli et al. [130] proposed a toolkit that allows cultural institutions to edit and publish digital exhibitions. Their research represents the first attempts to exploit ontologies and semantic web technologies for editing digital and virtual exhibitions. However, it is not designed to enable visitors to author their virtual exhibitions. We propose to take advantage of LOD in VR-based virtual exhibition authoring. To the best of our knowledge, it is the first attempt in this direction.

END-USERS ROLE: VISITORS VS CURATORS. To fully involve end-users, we desire to enable CH lovers in authoring virtual exhibitions by playing the role of exhibition curators. An exhaustive effort is posed in actively engaging end-users, but only a few works [136, 165, 288] move end-users into the position of exhibition curators.

TECHNOLOGICAL ASPECTS: THE ROLE PLAID BY VIRTUAL REALITY. In the literature, augmented, virtual, and mixed reality plaid a crucial role in the CH [23, 32, 338]. Virtual reality (VR) has often been discussed as a promising tool to provide immersive data visualization and exploration [211]. Users seem to feel more satisfied when using VR data exploration tools, demonstrating VR potentialities as an engaging tool for visual data analytic [211]. VR is mainly used in educative and exploratory applications to generate virtual museums or exhibitions. Our generator is based on VR technologies to meet the *entertainment* requirement. From a technical point of view, we can classify VR projects according to the tracking aspect, the employed display, the input interface, and, consequently, the level of immersion [23]. In more detail, the level of immersion can be specialised as follows: 1) non-immersive, when users are no-tracked, and they only use a desktop as display [136, 165, 288, 342], 2) semi-immersive, when applications adopt screen or projectors [47, 137, 145], 3) fully immersive, when VR enables tracking or immersive display methods [86, 124, 137, 159, 269, 288]. The proposed generator supports a desktop interaction resulting in a non-immersive VR and the head-mounted display results in a fully immersive experience.

11.2 APPROACH AND METHODOLOGY

Due to LOD potentialities and the plethora of available data, the research at the basis of this section aims to exploit CH KGs in creating virtual exhibitions by enabling CH lovers to play the role of exhibition curator. We also aim to meet the following requirements.

ENTERTAINMENT: engaging artifacts exploitation;

CUSTOMISATION: creation of thematic virtual exhibitions where users can freely customised the experience theme;

HETEROGENEITY: easy querying of geographically distributed information.

We address the entertainment requirement by actively engaging visitors in authoring a VR-based virtual exhibition. Concerning the customisation option, users are free to choose the desired exhibition theme. About the heterogeneity component, the proposed generator starts from KGs by leading to query heterogeneous information in the LOD format. Thus, virtual exhibition curators may retrieve data of interest [136] and combine geographically distributed information [134].

However, LOD exploitation by end-users is a challenging task embroiling the syntactical challenges posed by available querying languages, such as SPARQL [75, 106]. As we cannot assume that CH lovers have skills in query languages, we want to mask technical challenges in querying KGs during the virtual exhibitions authoring process.

The proposed virtual exhibition authoring process is illustrated in Fig. 11.1. It is composed of the following phases:

DATA SELECTION Users have to extract data of interest by querying KGs. We designed this phase by making no assumption on end users' skills. Thus, we must require minimum/no technical skills in query languages.

MUSEUM CUSTOMISATION Users can personalise both the museum content and its layout. Concerning the museum content customisation, users can select data to visualize, such as images/3D models, their labels and descriptions, artifacts geo-localisation, and additional information. Concerning the museum layout customisation, users can freely choose the museum appearance that is compliant with users' preferences and objectives.

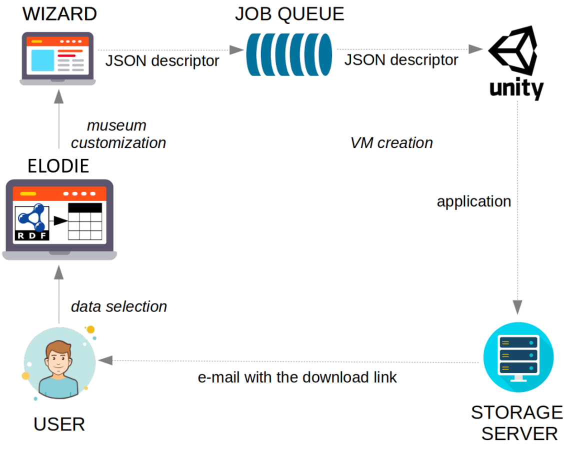


Figure 11.1: Virtual exhibition generation process.

MUSEUM CREATION A virtual exhibition is automatically created according to users' settings.

It is worth clarifying that we will further use the museum and exhibition terms as synonyms as virtual exhibitions authored by our generator resemble physical museums.

11.3 TOOLKIT: ELODIE AND VIRTUAL EXHIBITION GENERATOR

DATA SELECTION. This phase aims to guide users in querying KGs and retrieving data of interest. During the overview of the process, we referred to it as *data selection* phase. However, it is worth noticing that this phase cannot be interpreted as a mere dataset selection step but should define *dataset creation* phase. Users cannot access ready-to-use and off-the-shelf datasets. Instead, they are provided with a vast amount of data in the LOD format and decide which data they want to collect.

Users can query a KG at a time by relying on already-created interconnections made available by the KG publishers. The *output* of this phase is a dataset containing query results modelled as a tabular view. The returned dataset may have artwork of interest and related information, such as the title or a human-readable description. During the dataset creation phase, users should consider the museum layout they aim to author. For example, if users are interested in classic layout museums, the created dataset should contain, at least, a representation of the artifacts of interest, their title, and description. If users are interested in a museum with an underlying geographical map, they should also geo-coordinates to localise artifacts retrieving the geographic location of museums that host collected artifacts.

`ELODIE` implements this phase. Thanks to the NL, `ELODIE` guarantees the readability of the query mechanism. Users are guided in the query formulation step by step interactively, and the content of the queried source is automatically discovered by inspection. `ELODIE` solves the data conceptualisation issue as users have always access to suggestions without explicitly asking for them. `ELODIE` navigates available edges by path traversal queries, and retrieved results are returned as suggestions. This approach is generic enough to retrieve data from any endpoint by solving the portability issue. Only a limited number of results and suggestions, which can be manually customised by users within the interface, are computed at each iteration to address the scalability issue.

Users can select one of the available suggestions, organised in classes, predicates, and results to formalise their request. Their interactions are automatically verbalised as a controlled natural language query to demonstrate how the system interpreted users' selections. Moreover, these interactions also result in an automatically generated SPARQL query used to retrieve users' query results. Results are automatically organised in a tabular view. Thus, `ELODIE` requires neither technical skills in query languages nor awareness of the underlying data structure.

The focus mechanism of `ELODIE` allows users to interact with the natural language query and click on the query element that must be used as the insertion point for applying the following query transformations. This mechanism enables users to create more complex queries, that are not straightforward.

MUSEUM CUSTOMISATION. Once users are satisfied with the retrieved results, they can move to the museum customisation phase to customise the virtual exhibition content and layout by a guided wizard component.

Content customisation. Starting from data retrieved during the data selection phase, users can customise the authored virtual exhibition. They can expose paintings or 3D models and provide self-explanatory descriptions by attaching title, textual or multimedia details to each artefact. While the artefact representation and its title are always visible, description and further details are returned on request by interacting with the artwork. According to the museum's appearance, users might be required to add further information. For instance, if users desire to create a location-aware exhibition (by choosing the `Map` as museum template), they have to provide an artefact geo-localisation. If users want to organise the collection into several rooms (as in the `Hall` template), they have to categorise artifacts. During this phase, users are guided by a wizard to select all the required or useful details for creating their virtual exhibitions. The wizard guides users in selecting, for each field, a column of the previously created dataset. In selecting fields, users should recall that the title and the image URL will always be visible, while all the other information, such as textual descriptions or media link content, will be provided on request during the exhibition visit when visitors will interact with the artworks.

Layout customisation. Users can now customise the museum appearance and the artifacts arrangement. For example, if a classical or a modern museum is selected, artifacts will be wall aligned along corridors and rooms, while in the map layout, artifacts will be geo-localised pins located on an underlying geographic map. The provided information also influences the exhibition arrangement. For instance, if the dataset contains 3D models, they will be distributed on tables aligned along idles and rooms. Users can always recognise artifacts by labels, while further details (such as description) are provided on-demand.

Interaction mode. Finally, users can customise the interaction mode by choosing among first-person, showcase, or fly-through as play mode. The first-person mode leads users to move around in the exhibition and actively interact with artifacts. The showcase mode is an interactive 3D media gallery where the user assists in the exhibition without walking around but simply navigating forward and backward through artifacts. The fly-through mode disables user interaction and automatically moves the user avatar.

Type modes move from the more to the less interactive. Concerning the light mapping, the greater is the quality, the higher is the required building time, requiring from 10 minutes to 4 hours.

museum creation. Once the artifacts and the related settings have been provided, the exhibition can be generated by a Unity component. This step is completely decoupled from the previous ones. The Unity component receives data and settings by the wizard output, and it automatically creates the virtual exhibition. The channel between the wizard and the 3D engine is based on a job queue that collects all the artefact instances and builds them one by one. Thanks to this decoupling, it is possible to easily generate several exhibitions in parallel by employing more than one machine. This is a crucial feature in case of sudden peaks in requests. When an exhibition is created, the dataset is wrapped in it. This choice implies the efficiency loss in terms of the size of the resulting model as data are stored with the generated museum. An alternative option might be storing the dataset containing artifacts and the related information as a separate source and storing the link to the source within the virtual exhibition. However, keeping data and the virtual exhibition strongly coupled leads to an advantage in terms of autonomy and independence by external services. In the case of a lack of Internet connection, the dataset is always accessible as a local resource.

a unified approach from (linked) open data to virtual exhibitions. The main contribution of the proposed approach lies in exploiting the richness and variety of LOD in virtual exhibition authoring. The technical challenges of SPARQL are masked by `ELODIE`, an intuitive and user-friendly interface to query well-known KGs and represent retrieved results in a tabular format. It should provide users with the maximum freedom to create datasets of interest by easily querying heterogeneous data and geographically distributed information into a tabular view. Then, starting from a tabular dataset, the virtual exhibition is automatically created according to user-defined settings. The same generation workflow may be retraced by starting from an already available tabular OD. Therefore, the virtual exhibition approach can deal with both KGs and OD by only differentiating the way the dataset is either retrieved or created from scratch. Thanks to this consideration, we proposed a unified approach to exploit both OD and LOD by a similar process to author virtual exhibitions. When users decide to start from OD, they can access the dataset of interest by pasting its URI (as described by Scarano et al. [288]). Instead, when they decide to query KGs, they are scaffolded by ELODIE to query interest data and organise them as tabular data. Finally, besides the original data source format, they are guided in creating tailored virtual exhibitions.

By enabling the exploitation of LOD and KGs, users can move from homogeneous, isolated, and ad-hoc open datasets to heterogeneous data and query according to their needs. Our proposal may unlock the Semantic Web potentialities to a broader audience interested in CH KGs but threatened by challenges posed by query languages.

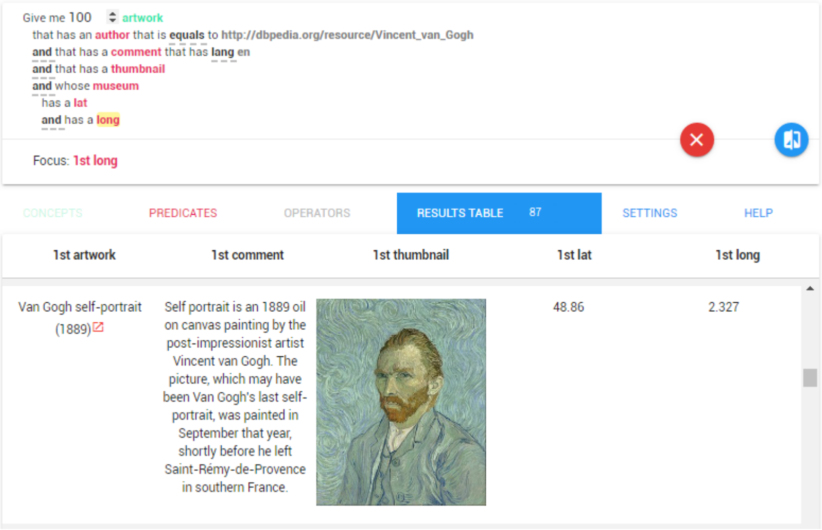
11.4 VAN GOGH'S EXPERIENCE: A USE-CASE STEP BY STEP

Supposing a user, named Alice, is a Van Gogh's lover, and she is interested in curating a virtual exhibition that collects all his paintings. First, she has to collect all the artifacts painted by Vincent Van Gogh. Furthermore, she may be interested in generating a location-aware virtual exhibition. Therefore, she has to also ask for the museums that host the retrieved paintings. Implicitly, the geographical location of the museum will provide her with the position of the hosted artifacts. Second, once she has collected Van Gogh's paintings and the related information, she can create her location-aware exhibition by selecting the map layout. We will separately go through each step of Van Gogh's use case. Both the video of the creation process[1] and the resulting exhibition tour[2] are available on YouTube.

DATA SELECTION Alice selects DBpedia as queried SPARQL endpoint and she aims to retrieve *geographical distribution of Van Gogh's paintings* according to DBpedia. Fig. 11.2 shows both the

[1] Use case - dataset creation: https://youtu.be/63Smst0_x78
[2] Use Case - Van Gogh's virtual exhibition: https://youtu.be/9LNdFY_20Jw

user query and an overview of the collected results at the last step of the reported use case.



Give me 100 ⇕ artwork
that has an author that is equals to http://dbpedia.org/resource/Vincent_van_Gogh
and that has a comment that has lang en
and that has a thumbnail
and whose museum
has a lat
and has a long

Focus: 1st long

1st artwork	1st comment	1st thumbnail	1st lat	1st long
Van Gogh self-portrait (1889)	Self portrait is an 1889 oil on canvas painting by the post-impressionist artist Vincent van Gogh. The picture, which may have been Van Gogh's last self-portrait, was painted in September that year, shortly before he left Saint-Rémy-de-Provence in southern France.		48.86	2.327

Figure 11.2: ELODIE in the *Van Gogh's experience*

Table 11.2 contains iterative queries as verbalised by ELODIE of this navigation scenario. At each step, the bold part represents the last suggestion selected by the user, and the underlined part represents the query focus.

MUSEUM CUSTOMISATION AND CREATION. Once Alice is satisfied with the created dataset, she can go on with the museum customisation. Alice opts for the *Map* layout to create a position-aware Van Gogh's exhibition. The resulting geographical aware virtual exhibition is visible in Fig. 11.3 and represents an open-space museum with an underlying worldwide map. For each artefact, a pin localises its geographical position. Users can walk around, zoom in to get closer to a pin, and click on it to get artefact details. The resulting virtual exhibition[3] can be freely downloaded and explored.

The building time is strictly related to the generation of the museum room light-maps, which requires heavy computational

[3] Link to the built virtual exhibition and related information: https://www.isislab.it/en/virtual-museum/. Requirements to run it are i3, Windows as Operating System, RAM 8GB, dedicated graphical board.

Table 11.2: A navigation scenario in ELODIE over DBpedia

Step	Query
1	Give me something
2	Give me an **artwork**
3	Give me an artwork **that has an author**
4	Give me an artwork that has an author **that is equal to dbr:Vincent_Van_Gogh**
5	Give me an artwork that has an author that is equal to dbr:Vincent_Van_Gogh
6	Give me an artwork that has an author that is equal to dbr:Vincent_Van_Gogh and **that has a comment**
7	Give me an artwork that has an author that is equal to dbr:Vincent_Van_Gogh and that has a comment **that has lang en**
8	Give me an artwork that has an author that is equal to dbr:Vincent_Van_Gogh and that has a comment that has lang en
9	Give me an artwork that has an author that is equal to dbr:Vincent_Van_Gogh and that has a comment that has lang en and **that has a thumbnail**
10	Give me an artwork that has an author that is equal to dbr:Vincent_Van_Gogh and that has a comment that has lang en and that has a thumbnail
11	Give me an artwork that has an author that is equal to dbr:Vincent_Van_Gogh and that has a comment that has lang en and that has a thumbnail and **that has a museum**
12	Give me an artwork that has an author that is equal to dbr:Vincent_Van_Gogh and that has a comment that has lang en and that has a thumbnail and whose museum **has a lat**
12	Give me an artwork that has an author that is equal to dbr:Vincent_Van_Gogh and that has a comment that has lang en and that has a thumbnail and whose museum has a lat
13	Give me an artwork that has an author that is equal to dbr:Vincent_Van_Gogh and that has a comment that has lang en and that has a thumbnail and whose museum has a lat and **that has a long**

time. By choosing the museum with the underlying map, lights are completely absent, and the museum is created in 10 minutes.

ALTERNATIVE VIRTUAL EXHIBITIONS FOR THE VAN GOGH'S EXPERIENCE The proposed virtual exhibition generator enables the possibility to reuse the dataset returned by the data selection phase to create create alternative exhibitions. Thus, Alice can reuse the same dataset containing artifacts painted by Van Gogh to author different exhibitions. For example, supposing Alice is also interested in a more traditional virtual exhibition, she can opt for a classical (as in Fig. 11.4) or modern museum (as in Fig. 11.5) during the layout customisation. Classical or modern

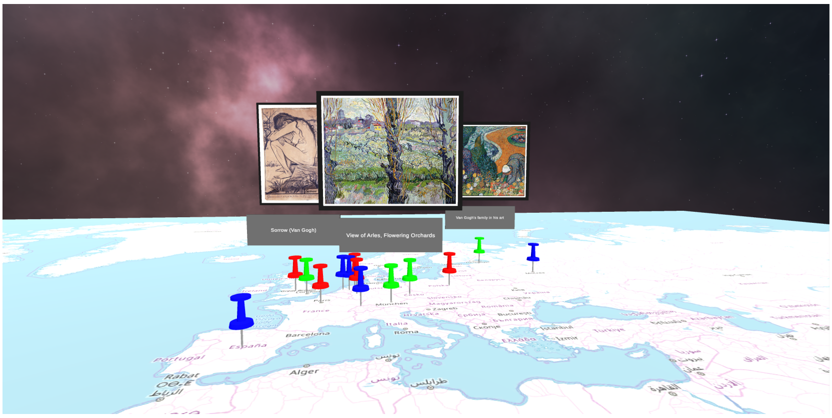


Figure 11.3: A location-aware Van Gogh virtual exhibition

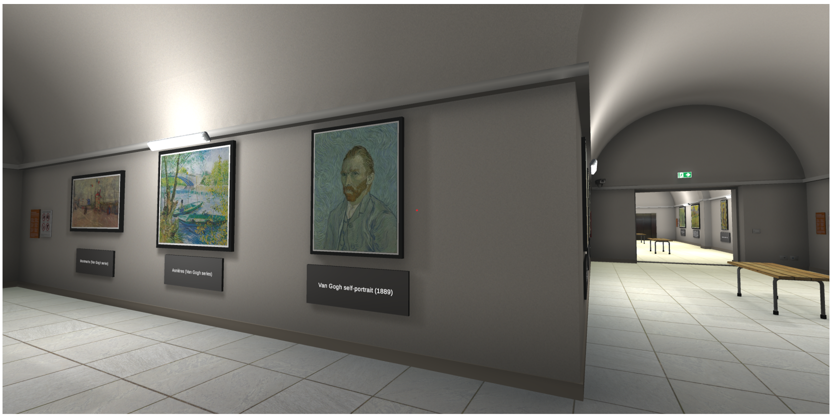


Figure 11.4: The Van Gogh's experience as a *modern* museum

museums ignore location information. Therefore, artworks listed in the dataset will be wall-aligned and dislocated in different rooms. As stated before, the building time depends on the generation of the light-maps. The museum generator creates a fixed number of lights for each room. Therefore, the building time transitively relies on the number of rooms needed to dislocate artworks. Moreover, the building time is strictly dependent on the required quality. By asking for a medium quality light-map, the building time is 10 minutes also for the traditional museum. By asking for high-quality resolution, the building time almost reaches 3 hours.

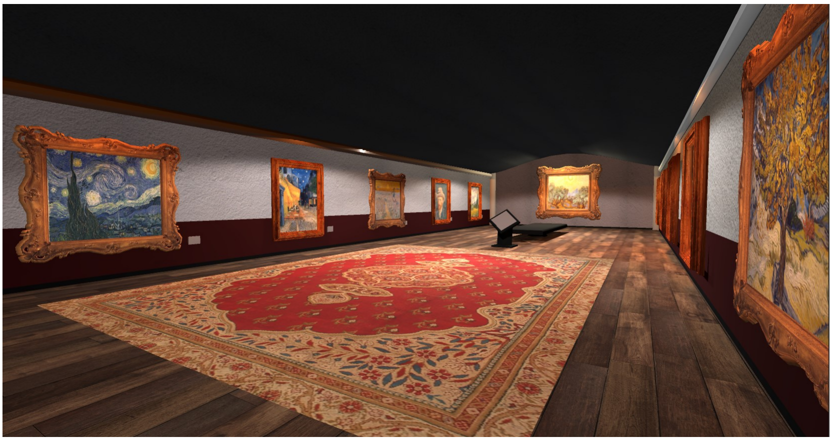


Figure 11.5: The Van Gogh's experience as a *classical* museum

11.5 DISCUSSION

11.5.1 *Toolkit features*

In the virtual exhibition generation process presented in this chapter, users are scaffolded by an intuitive user interface. In more detail, during the query formulation, they are provided with a step-by-step exploratory search. During the museum configuration phase, the selection process is clarified by a guided wizard. The readability requirement is met by explaining all the performable actions in natural language. For instance, ELODIEverbalises the user query, suggestions, and results in the user language.

Users are guided in querying heterogeneous data sources, i.e., tabular and graph-like data, by a unified approach. The proposed mechanism is data provider-independent. For instance, ELODIE guarantees the KG independence by dynamically retrieving suggestions through path traversal queries, generic enough to work on any SPARQL endpoint.

By referring to challenges posed by LOD querying, ELODIE is a query builder that automatically creates SPARQL queries according to user interactions with a faceted search interface. Thus, it succeeded in abolishing any technical barrier posed by KG query languages. Moreover, users are provided with all the available options to go on in the query formulation. As users are not required to manually look for them, ELODIE masks data conceptualisation challenges.

Technically speaking, the scalability is addressed by retrieving a limited number of results and suggestions during the dataset creation step, although users can easily customise this limit. On the museum creation side, the virtual exhibition instances queue guarantees the management of any request peak. Furthermore, our proposal is entirely based on Web standards by achieving the portability requirement.

11.5.2 *Limitations*

ELODIE queries a single KG at a time according to the SPARQL endpoint chosen by users at the beginning of the data selection phase. ELODIE relies on already defined interconnections without introducing extra mappings. Thus, while it behaves well on curated topics, it suffers KG incompleteness.

ELODIE requires zero configuration by design, meaning that it may work with any KG as it automatically discovers classes, predicates, and inspects existing node neighbourhood to go on in the query formulation. ELODIE is configured to query heterogeneous KGs, such as general-purpose KGs, such as DBpedia; CH KGs, such as ARCO; repositories of academic contributions, such as Persée. It only requires CORS enabled SPARQL endpoint URL, i.e., a specification that enables truly open access across domain boundaries.

In the current version, the virtual exhibition generation can rely on data queried by users during the dataset creation phase by ELODIE. Users may also be interested in keeping track of semantic metadata that describe the meaning of data sources (i.e., semantic annotations) or denote real-world objects by ontologies [177]. Semantic metadata may be specialised in descriptive metadata (i.e., information about who created a resource, what the resource is about, what it includes); structural metadata (i.e., concerning the ontological aspect, their relationships, and the structure they exist in); administrative metadata (information about the origin of resources, their type, and access rights). According to available metadata, users may query them by ELODIE and then opt for making them always visible (as title) or visible on demand (as description).

11.5.3 *Considerations on virtual exhibitions*

Virtual exhibitions may enhance traditional museum visits by offering the ability to adapt, expand, and personalise the artefact collections [35]. They can overcome any space, time, and location restrictions [335, 342] by permanently exposing [122] a limitless number of exhibits [35], accessible at any time and from any place [111] without additional costs [122].

Some CH objects' fragile nature and the risk of damaging expensive artifacts prevent real museum curators from making them available to the public [97, 342]. Digital artifacts do not age in virtual exhibitions and are not subject to decay in the digital space. It represents a first step in the direction of ensuring long-term preservation [122]. However, it is worth noting that long-term preservation is not given purely by the digitisation process, but it is an active and continuous work widely explored by researchers in migration and emulation of digital artefacts [135].

In traditional tours, the visitors' interaction with the exhibited artifacts is threatened by boundaries or glasses. Moreover, they usually cannot look at artifacts from all angles, compare artifacts, study them in different context [342]. On the contrary, in virtual exhibitions, visitors are completely free in interacting with CH objects without any constraint [42], of course, within the limits of the available technological solutions.

Virtual exhibitions behave as an extraordinary means of cultural dissemination for users who cannot (because of disabilities, economic restrictions, or other constraints, such as pandemics) visit real museums. They naturally guarantee full accessibility as they enable different audiences, including people with disabilities and individuals of all ages, to access and interact with vast numbers of artifacts scattered among various localities in an engaging and informative way [42, 335]. Furthermore, they represent a cheaper alternative to real tours for people that cannot afford travelling costs to enjoy CH. Moreover, GLAM institutions widely adopt it during the pandemic posed by COVID-19 to create virtual versions of cultural activities [5, 174].

Users can behave as active creators of their virtual exhibitions and leave the passive position of CH spectators [342]. Users can create a tailored exhibition or a virtual replica of real museums. They can expose tangible, abstract, and imaginary artifacts, restorations of damaged objects, and hypothetical models of real artifacts that no longer exist [111]. They can combine artifacts that are physically stored in different geographically or organisationally museums [122, 186]. Moreover, virtual exhibitions can easily guarantee multilingualism support [335].

Virtual exhibitions may enhance (not substitute) traditional museum visits by offering the ability to adapt, expand, and personalise the artefact collections [35]. Moreover, they may increase virtual visitors, which might become real visitors to cultural institutions [58].

11.6 FINAL REMARKS

Despite the huge amount of available CH data, CH lovers are rarely moved to the position of exhibition curators. They passively assist data curators in museums and exhibitions, without the possibility to play an active role. This chapter reports a prototype to generate thematic, tailored, and custom virtual exhibitions taking advantage of LOD without asking for any technical skill in query languages. Combining LOD and virtual exhibitions enables users to free their imagination and create original virtual exhibitions according to their interests and needs.

Virtual exhibitions may enhance traditional museum visits by offering customisation options, overcoming space restrictions, preserving CH integrity, letting visitors interact with art, overcome accessibility issues. Moreover, an authoring interface let CH lovers play the role of the exhibition curator, leaving the passive role of a visitor and behaving as an active museum curator.

As a future direction, we aim to extend ELODIE by enabling the possibility to merge results from several sources, such as several KGs queried concurrently or image sources (such as Google Image or International Image Interoperability Framework[4], considered a de facto standard for image-sharing within the GLAM

[4] https://iiif.io/

sector). Moreover, we aim to provide users with additional default information, such as the data source (e.g., the queried KG), and provide access to the artwork source link to let users explore the node neighbourhood in the KG as an opportunity to further contextualise the retrieved artwork.

12 KNOWLEDGE GRAPHS AND VIRTUAL ASSISTANTS

> *We are entering a new world. The technologies of machine learning, speech recognition, and natural language understanding are reaching a nexus of capability. The end result is that we will soon have artificially intelligent assistants to help us in every aspect of our lives.*
>
> – Amy Stapleton

In the last decade, public institutions and private organisations have invested in massive digitisation campaigns to create vast digital collections, repositories, and portals that allow online and direct access to billions of resources [192]. Up to now, digitisation has been adopted as a good practice to shift from in-person activities to virtual meetings and remote workstations. However, this possibility from potential means is going transform into a real need due to the COVID-19 and the universal lock-downs. The pandemic's rapid and unstoppable spread spurs any social and economic sector in moving online and adopting ways of delivering services while limiting physical contact [295].

Digitisation causes an extraordinary acceleration in digital transformation processes [5] that affected any field, from education to business models [295], from health care [161] to CH [5]. Focusing on the CH field, public and private organisations have invested in digitising any form of data to ensure its long-term preservation and support the knowledge economy [192].

The United Nations Educational, Scientific and Cultural Organization (UNESCO) defines CH as "*the legacy of physical artifacts and intangible attributes of a group or society inherited from past generations, maintained in the present and bestowed for the benefit of future generations*" [310]. CH includes *tangible* culture (such as buildings, monuments, landscapes, books, works of art, and artifacts); *intangible* culture (such as folklore, traditions, language, and knowledge), and *natural* heritage (including culturally significant landscapes, and biodiversity) [310].

Nowadays, CH has become one of the most successful application domains of the Semantic Web technologies [33]. CH as LOD improves data reusability and allows easier integration with other data sources [33]. It behaves as a promising approach in facing CH challenges, such as syntactically and semantically heterogeneity, multilingualism, semantic richness, and interlinking nature [148]. The availability of CH data in digital machine-processable form has enabled a new research paradigm called Digital Humanities [33] and aims to facilitate researchers, practitioners, and generic users to consume cultural objects [148].

However, KG exploitation is mainly affected by i) required technical competencies in generic query languages, such as SPARQL, and in understanding the semantics of the supported operators [329], too challenging for lay users [24, 75, 106, 233, 329], and ii) conceptualization issues to understand how data are modelled [24, 329]. Hence, non-expert users require interfaces (e.g., forms, QA, or keyword search tools) to access KGs [233] without experiencing the underlying syntactic complexity [106].

NL interfaces can mitigate these issues, enabling more intuitive data access and unlocking the potentialities of KGs to the majority of end-users [160]. NL interfaces may provide lay users with QA functionalities where users can adopt their terminology and receive a concise answer. Researchers argue that multi-modal communication with virtual characters using NL is a promising direction in accessing KGs [56].

Consequently, VAs have witnessed an extraordinary and increasing interest as they naturally behave as QA systems. Many companies and researchers have combined (CH) KG and VAs [11, 72, 192], but no one has provided end-users with a generic methodology to generate extensions to query KGs automatically.

To fill this gap, the *objective* of this chapter is the definition of a general-purpose approach that makes KGs accessible to all by requiring minimum-no technical knowledge in Semantic Web technologies. VAs usually give the possibility to extend their capabilities by programming new features, also referred to as VA extensions. It implies that (potentially) everyone can implement custom extensions and personalise VA behaviour. However, playing the VA extension creator's role requires programming competencies to design and implement the application logic.

Moreover, users must be aware that VA extensions are provider-dependent. Therefore, an extension implemented for Alexa is not directly reusable for other providers, such as Google Assistant.

The proposed approach empowers lay-users by letting them leave VA users' passive position and play the role of VA extensions creator by requiring little/no-technical competencies. Hence, the goal of this work can be reformulated as i) enabling QA over KGs (KGQA) by VA and ii) allowing (lay) users to automatically create ready-to-use VA extensions to query KGs by popular VAs, e.g., Amazon Alexa and Google Assistant. It results in the design of a community-shared software framework (a.k.a. generator) that enables lay users to create custom extensions for performing KGQA for any cloud provider, unlocking the potentialities of the Semantic Web technologies by bringing KGs in everyone's "*pocket*", accessible from smartphones or smart speakers.

The research presented in this chapter has been preliminary published in the following contribution:

- Maria Angela Pellegrino, Mario Santoro, Vittorio Scarano, Carmine Spagnuolo. *Automatic VA extension Generation for Knowledge Graph Question Answering*. In the Proceedings of Extended Semantic Web Conference (ESWC) (Satellite Events) 2021.

An extended version of this article containing a quantitative and qualitative analysis of CH KGs and an extensive overview of the proposed community shared software framework has been submitted as the following contribution and it is actually under the second round of evaluation:

- Maria Angela Pellegrino, Vittorio Scarano, Carmine Spagnuolo. *Move Cultural Heritage Knowledge Graphs in Everyone's Pocket*. *Submitted* to the Semantic Web Journal in March 2021.

12.1 RELATED WORK

QA systems can be classified as domain-specific (a.k.a. closed domain) or domain-independent (a.k.a. open domain). While in domain-independent QA, there is no restriction on the question domain and systems are usually based on a combination of IR and NL Processing techniques [142]; in domain-specific QA, questions are bound to a specific context [7] and developers can rely on techniques that are tailored to the domain of interest [73]. Besides the scope, they can be classified by the type of questions it can accept (e.g., facts or dialogs) and queried sources (structured vs. unstructured data) [188]. While systems querying text collections are classified as tools working on unstructured data (e.g., WEBCOOP [25]), systems querying KGs are classified as tools working on structured data. According to this classification, we propose an approach to pose factoids questions (wh-queries, e.g., who, what, and how many, and affirmation/negation questions) over semantically structured data where questions aim to be as general as possible to classify our proposal as domain-independent.

KGQA is a widely explored research field [88, 305, 355]. While it is rare to observe keyword-based questions, most of them address full NL questions. Usually, questions can be posed in English, while some tools deal with European and non-European languages [88]. There is a consistent effort in proposing domain-independent QA systems to query DBpedia and Wikidata [88, 305] by exploiting heterogeneous solutions ranging from combinatorial approaches [88] to neural networks [305], from graph-based solutions [355] to NL request mapping to SPARQL queries [94].

By focusing on CH KGQA, i.e., domain-specific systems in the CH domain, they can benefit from many standard data sources. CIDOC Conceptual Reference Model (CRM) is an example in this direction, and it is widely adopted as a base interchange format by GLAM institutions all over the world [90]. CIDOC-CRM has been identified as the knowledge reference model for the PIUCULTURA project, funded by the Italian Ministry for Economic Development, which aims to devise a multi-paradigm platform that facilitates the fruition of Italian CH sites. Within the PIUCULTURA project, Cuteri et al. [73] proposed a QA system tailored to the CH domain to query both general (e.g., online data collections) and specific (e.g., museums databases) CIDOC-compliant knowledge sources by exploiting logic-based transformation. As an alternative approach, PowerAqua [325] maps input questions to SPARQL templates under the hypothesis that the SPARQL query's overall structure is almost determined by the syntactic structure of the NL question.

KGQA by VAs is natively offered in well-known VAs, such as Google Assistant and Alexa, that provide users with content from generic KGs (Google Search and Microsoft Bing, respectively). Their main limitations are that i) they query proprietary and ii) general-purpose KGs without exploring domain-specific QA, iii) the provided mechanism can not be extended by users and ported on other KGs. Therefore, the Semantic Web community invested in increasing VA capabilities by providing QA over open KGs. Among others, Haase et al. [129] proposed an Alexa VA extension to query Wikidata by a generic approach, while Krishnan et al. [171] made the NASA System Engineering domain interoperable with VAs.

By considering CH KGQA via VAs, CulturalERICA (Cultural hERItage Conversational Agent) [192] is an intelligent conversational agent to assist users in querying Europeana [133] via NL interactions and Google Assistant technology. The authors state that CulturalERICA is database-independent and can be configured to serve information from different sources. Besides technological differences (we opt for Alexa while they opt for Google Assistant), while they enable iterative refinement of the queries, at the moment, we only provide one-step iterations. However, they only enable path traversal, while we also support more complex queries, such as sort pattern, numeric filters, class refinement. Anelli et al. [11] developed a VA to enable the exploitation of the Puglia Digital Library by delegating the speech recognition to Google Assistant. Through subsequent interactions, the VA creates and keeps the context of the request. While they enable keyword-based search, we opt for NL questions. Cuomo et al. [72] proposed an answering system and adapted it to implement a VA able to reply to questions about artworks exposed in Castel Nuovo's museum in Naples. Their proposal replies to questions about artworks, their author, and related information posed by visitors during the touristic tour. Even if it represents an interesting work in the direction of CH KGQA via VAs, it is bound to hardware devices within the museum, and it is not a solution that users can exploit everywhere. About the integration of CH KGs and chatbots, we can cite the chatbot proposed by Lombardi et al. [187] supporting users during an archaeological park visit (i.e., in Pompeii) by simulating the interaction between visitors and a real guide to improve the touristic experience by exploiting natural language processing techniques. In the same direction, Pilato et al. [262] propose a community of chatbots (with specialized or generic competencies) developed by combining the Latent Semantic Analysis methodology and the ALICE technology.

These works behave as evidence of the interest in developing KGQA by VA by promoting interesting applications to make CH KGs interoperable with VAs to accomplish the QA task, but they do not empower end-users by providing them with the opportunity to create their VA extensions. The main difference between our proposal and the ones reported so far is that the literature proposes ready-to-use VA extensions, while we are proposing a generator of VA extensions that are bound to neither any KG nor any specific VA provider. To the best of our knowledge, the proposed community-shared software framework is the first attempt to provide users without technical competencies in the Semantic Web technologies to create KGQA systems via VAs. Consequently, it represents the main *novelty* of our proposal.

12.2 KNOWLEDGE GRAPH QUESTION-ANSWERING VIA VIRTUAL ASSISTANTS

This section introduces the design methodology to make KGs compliant with VAs to address the QA task. It focuses on Amazon Alexa and its terminology without losing generality, as the same considerations can also be adapted for other customizable providers. Alexa VA extensions are named `skills`, and include the interaction model and the back-end logic. The interaction model defines the supported features referred to as `intents`, and each intent can be modelled by a set of `utterances`, i.e., phrases to invoke it. Utterances may specify a set of `slot` keywords, i.e., variables that will be instantiated according to the users' requests.

The KGQA task can be defined as follows: given an NL question Q and a KG K, the QA system produces the answer A, which is either a subset of entities in K or the result of a computation performed on this subset, such as counting or assertion replies [327].

Fig. 12.1 proposes a parallel between a general KGQA and a VA-based process.

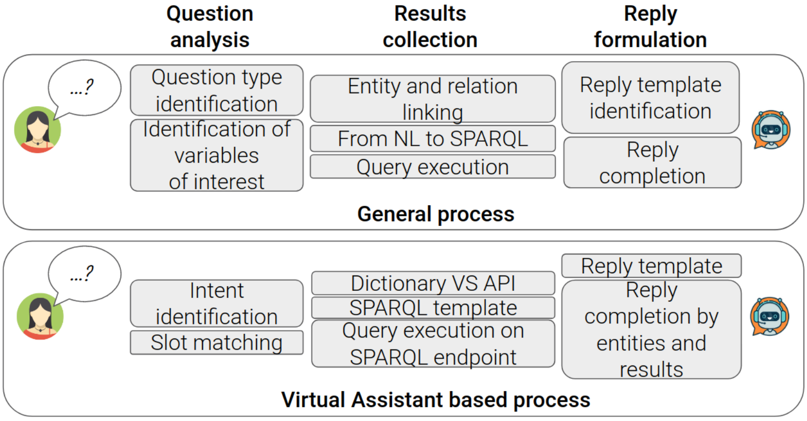

Question analysis
Results collection
Reply formulation
...?
Question type identification
Identification of variables of interest
Entity and relation linking
From NL to SPARQL
Query execution
Reply template identification
Reply completion
General process
...?
Intent identification
Slot matching
Dictionary VS API
SPARQL template
Query execution on SPARQL endpoint
Reply template
Reply completion by entities and results
Virtual Assistant based process

Figure 12.1: Parallel of a general and a Virtual Assistant-based Knowledge Graph Question Answering process

A general KGQA workflow is composed of the question analysis phase, followed by the query construction to retrieve results [87]. We also consider a final step to formulate an NL reply to verbalise the retrieved results and return them to the user. Consequently, the high-level KGQA workflow is an adaptation of the methodological approach proposed in the literature by Diefenbach et al. [87]. How this general approach has been narrowed down as a VA-based process is a proper original contribution of the proposed approach. While the general process reports a high-level approach detailing terminology commonly used in the context of KGQA, the VA-based process narrows it down to terms related to VA extensions (such as intent, slots) and reports low-levels detailed considered in implementing a KGQA via VAs. For instance, while the general phase to retrieve the entity or predicate URI attached to a NL label is usually named linking, in the VA-based process, it might be implemented by using dictionaries or calling APIs. While the general process focuses on the high-level role of each component, the VA-based process considers VA peculiarities and low-level implementation alternatives.

The *question analysis* step performs the question type identification (and consequently, the expected reply template) and the linking phase. The *query construction* phase formulates the

SPARQL query corresponding to the NL question and runs it on a SPARQL endpoint to retrieve raw results. During the *reply formulation* step, retrieved results are organised as an NL reply.

In a VA-based process, users pose a question in NL by pronouncing or typing it via a VA app or dedicated device (e.g., Alexa app/device). During the *question analysis* phase, VAs interpret the request and identify the intent that matches the user query by an NL processing component. During the intent identification, VAs also solve intent slots. For instance, suppose that we implement a VA extension representing a thesaurus to recognise questions related to term definition. It might expect requests matching the template `Can you define the term <WORD>?`, where <WORD> is the slot that needs to be completed by the user. Therefore, when the user poses the question `Can you define the term <CULTURAL HERITAGE>?`, where cultural heritage behaves as a slot value. Once retrieved slot values, the VA extension performs the linking step to retrieve the URI(s), which may correspond to the label pronounced by users. The linking phase may be performed by consulting a lookup dictionary or by calling an API service. Completed the question analysis step, we can move to the *query formulation* step. If the KGQA system behaves as a query builder, the VA extension has to recognise the SPARQL pattern that fulfils the user request and formulate the SPARQL query. The SPARQL query can be run on the SPARQL endpoint. Once returned results, the VA extension performs the *reply formulation* step by identifying the reply template corresponding to the activated intent, completing it with actual results, and returning it to users.

This section describes the proposed approach to design and implement a VA extension to enable KGQA by focusing on Amazon Alexa as a VA provider. It details the introduced concepts related to Alexa skills and the proposed implementation of a KGQA VA extension. It is not a loss of generality since it can be easily adapted to any other VA that enables custom VA extension definition, such as Google Assistant, or in bot implemented by Microsoft Azure Bot Service or Googlebot. We opt for Alexa instead of plausible alternatives as Amazon Alexa holds the record of the provider with the greatest number of sold devices. However, the architecture of the generator leads to easy integration of novel VA providers, such as Google Assistant.

AMAZON ALEXA SKILLS. As stated before, functionalities in Alexa have named skills. Among the supported types of skills, we are interested in *custom* skills where we can define the requests the skill can handle (intents), and the words users say to invoke those requests (utterances) [84]. By developing new VA extensions, you have to define: a set of intents that represent actions that users can do with your VA extension; a set of sample utterances that specify the words and phrases users can use to invoke your intents; an invocation name that identifies and wake-ups your VA extension; a cloud-based service that accepts and fulfils these intents. By mapping utterances to intents, you are defining the VA extension interaction model. Utterances can contain slots, i.e., variables bound by users when formulating their requests, that can be validated by attaching to each slot a list of valid options during the interaction model definition. The back-end code can be either an AWS Lambda function or a web service. An AWS Lambda (an Amazon Web Services offering) is a service that lets you run code in the cloud without managing servers. When the user poses a question, Alexa recognises the activated intent and communicates to your code both the recognised and slot(s) values. Then, the back-end can perform any necessary actions to collect results and elaborate a reply [84].

VIRTUAL ASSISTANTS FOR QUESTION-ANSWERING. The proposed approach models each supported SPARQL query template as an intent. The implemented intents (listed in Table 12.1) are tailored towards SPARQL constructs, and they mainly cover questions related to a single triple enhanced by the refinement of the subject or object class. More in detail, we cover `SELECT` and `ASK` queries, `class specification`, numeric `filters`, `order by` to get the superlative and path traversal. In table 12.1, we report, for each intent, an exemplary NL query that activates the intent, the intent name, an utterance by specifying slots among braces, and the related SPARQL triples. In defining utterances, we aim to separate the supported SPARQL patterns clearly to enable users to assess the query correctness generated out of their input. We also avoid utterance overlapping to ensure, as much as possible, a deterministic intent activation.

Table 12.1: List of implemented intents in the proposed community-shared software framework

Intent name	Utterance	SPARQL Triple
What is the {author} of {Mona Lisa}?		
getPropertyObject	What is the {p} of {e}?	<e><p>?
What is {cultural heritage}? Can you define {cultural heritage}?		
getDescription	What/Who is {e}?	<e><definition>?
Where is {Rome}? Where is the {Mona Lisa}?		
getLocation	Where is {e}?	<e><location>?
Show me {Paris}. Show me {Mona Lisa}.		
getImg	Show me {e}	<e><img>?
What has {Beethoven} as {author}?		
getPropertySubject	What has {e} as {p}?	? <p><e>
How many {paintings} are there?		
getClassInstances	How many {e} are there?	? <instanceof><e>
Which {pianist} were {influenced} by {Beethoven}?		
getPropertySubjectByClass	Which {c} were {p} by {e}?	? <instanceof><c>. ? <p><e>.
What has been {modifies} {in} {2020}?		
getNumericFilter	What has {p} {symbol} {val}?	? <p>?o. FILTER(?o <symbol><val>)
Which {source} has been {modified} {in} {2020}?		
getNumeriFilterByClass	Which {c} has {p} {symbol} {val}?	? <instanceof><c>. ? <p>?o. FILTER(?o <symbol><val>)
Which is the {creation} with the {maximum} {number of collaborators}?		
getSuperlative	What is the {c} with {sup} {p}?	? <p>?o. ORDER BY (?o). LIMIT 1
Can you verify if {intangible cultural heritage} as {folklore} as {narrower}?		
getTripleVerification	Can you verify if {s} has {o} as {p}?	ASK <s><p><o>
Give me all the results		
getAllResultsPreviousQuery	Give me all the results	-

When the end-user poses a question, Alexa identifies the activated intent and notifies the back-end by communicating both the activated intent and the slot(s) values. For instance, in the CH use case reported in Fig. 12.2, the user asks for Mona Lisa's painter. The VA recognises that it corresponds to the `getPropertyObject` intent with utterance *what/who is the {property} of {entity}* and attaches to the property slot the value *painter* and to the entity slot the value *Mona Lisa*.

Consequently, the entity and relation linking phase must be performed. It is worth noting that the performed task is a simplified version of the more general entity and relation linking problem. Entity linking is generally referred to as identifying in a text snippet entities and matching these to the corresponding KG entity. For instance, mapping in the question *Who is the wife of the mayor of Rome?* the textual evidence of *Rome* has to be isolated first, and then it can be mapped to the corresponding KG

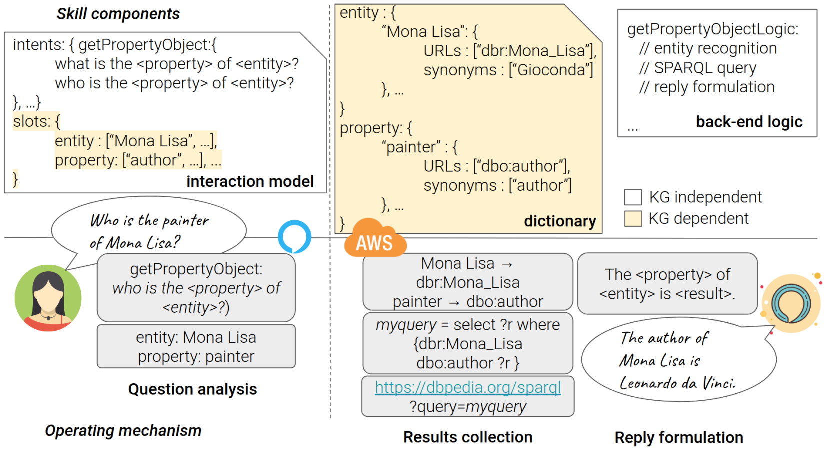

Skill components
intents: { getPropertyObject:{
what is the <property> of <entity>?
who is the <property> of <entity>?
}, ...}
slots: {
entity : ["Mona Lisa", ...],
property: ["author", ...], ...
}
interaction model
entity : {
"Mona Lisa": {
URLs : ["dbr:Mona_Lisa"],
synonyms : ["Gioconda"]
}, ...
}
property: {
"painter" : {
URLs : ["dbo:author"],
synonyms : ["author"]
}, ...
}
dictionary
getPropertyObjectLogic:
// entity recognition
// SPARQL query
// reply formulation
...
back-end logic
KG independent
KG dependent
Who is the painter of Mona Lisa?
AWS
getPropertyObject: who is the <property> of <entity>?)
entity: Mona Lisa property: painter
Question analysis
Operating mechanism
Mona Lisa → dbr:Mona_Lisa painter → dbo:author
myquery = select ?r where {dbr:Mona_Lisa dbo:author ?r }
https://dbpedia.org/sparql ?query=myquery
Results collection
The <property> of <entity> is <result>.
The author of Mona Lisa is Leonardo da Vinci.
Reply formulation

Figure 12.2: Graphical representation of VA extension components

entity. In VA-based solutions, named entity textual evidence is already detected by VAs, and we have only to map the named entity textual evidence to a KG node (like *Rome* to the node in the graph representing the city of Rome). To perform this (simplified) linking phase an alternative is performing a dictionary lookup. In such a case, we store the mapping label-URIs in a dictionary by querying KG classes, predicates, and resources URIs and the corresponding labels. The VA extension back-end exploits the dictionary to retrieve the URI(s) corresponding to NL labels. Resolved entities and predicates are used to complete the SPARQL template. The proposed approach attaches to each intent a different SPARQL query template. Consequently, any NL query posed by end-users is matched to the corresponding intent (according to the VA interaction model), and each intent corresponds to a SPARQL query template (according to our approach). To reconstruct the complete SPARQL query corresponding to each intent, you can proceed as follows: you have to introduce the SPARQL triple(s) reported in Table 12.1 with the SELECT operator and append the optional request of the label attached to the variable of interest. For instance, the triple `<e><p>?` corresponds to the SPARQL query *SELECT DISTINCT ? ?label WHERE{* `SPARQL triple` *} OPTIONAL { ? <label>?label. FILTER(LANG(?label)="en")}* (supposing that the VA extension language is English). The notation <e> means that the triple is completed by URIs attached to the label e in the dictionary. Once the query has been formulated, it can be posed to the SPARQL endpoint. We opt for running a GET query on the SPARQL endpoint and by asking for results in the JSON format. Once results are returned, the back-end formulates them as an NL reply. We attach to each intent a reply template. The back-end completes it with the resolved entities and with the retrieved results. The complete reply, i.e., the reply that includes the resolved entities, enables the end-users to inspect how the system interpreted the performed question implicitly. For instance, in the CH use case in Fig. 12.2, the end-user acknowledges that the *painter* word has been interpreted as *author*. It behaves as a step forward in the direction of the explicability of the application back-end logic.

12.3 TOOLKIT: A COMMUNITY-SHARED SOFTWARE FRAMEWORK

This section describes the designed and implemented community-shared software framework (a.k.a. generator) that enables lay users to create custom extensions for performing KGQA for any cloud provider, unlocking the potentialities of the Semantic Web technologies by bringing KGs in the "*pocket*" of everyone, accessible from smartphones or smart speakers. It overviews the architecture and implementation of the proposed software framework to automatically generate VA extensions implementing KGQA by requiring little/no-technical competencies in programming and query languages.

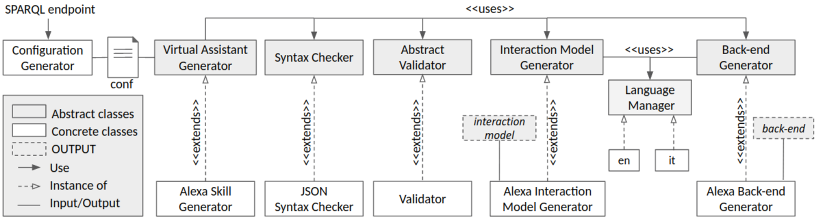


Figure 12.3: Architecture of the proposed generator of KGQA by VAs.

The proposed framework provides users with the opportunity to customise VA extension capabilities and generate ready-to-use VA extensions. Each phase is kept separate by satisfying the modularity requirement, and it is implemented as an abstract module. The proposed generator architecture is reported in Fig. 12.3. It is available on GitHub[1] with an open-source license.

The generator takes as input a configuration file containing the VA extension customization process. The configuration file is parsed to verify the syntactical correctness, the semantic validity and, if all the checks pass, both the interaction model and the back-end implementation can be generated. The syntactical correctness is verified in terms of JSON valid format in the actual implementation, but it can be substituted according to the configuration file format. The semantic validation is in charge of spotting any configuration conflict and verifying consistency. Both the validations are performed by parsing the configuration file.

Once passed these validations, the interaction model is created by extrapolating from a separated mapping file (stored in the back-end implementation as a JSON file) from each intent required by the configuration file, the corresponding set of utterances in the language configured by the end-user. It guarantees the ease in extending new supported languages, the possibility to revise utterances for each intent, and model new intents. The back-end is implemented in Node.js and maps to each intent the corresponding behaviour. It is configured according to the required user language and the SPARQL endpoint of interest. The back-end is returned as a ZIP file containing both the Node.js web hook and the implementation of the linking approach.

virtual assistant generator input: the configuration file. The VA `Generator` module takes as input a configuration file containing the VA extension customization options: the `invocation name`, i.e., the VA extension wake-up word; the list of desired `intents`, according to supported intents listed in Table 12.1; the `SPARQL endpoint` the user aims to query; the `lang`, by choosing among *en* and *it* at the moment, even though further languages can be easily introduced. Users can specify a (incomplete) dictionary mapping URIs to entities and properties labels.

[1] VA-generator source code: `https://github.com/mariaangelapellegrino/virtual_assistant_generator.git`

Users can manually create the configuration file. Otherwise, they can exploit the `Configuration Generator` module that takes as input the SPARQL endpoint of interest and automatically retrieves both classes and properties labels and URIs. It looks for used classes/properties and the ones defined according to standard approaches, such as classes defined as `owl:Class` or `rdfs:Classes`, properties defined as `rdf:Property` or `owl:DatatypeProperty`. Moreover, it also expands labels with synonyms and variations by exploiting Wordnet, e.g., nouns used as properties are expanded by their verbal or adjective forms. The configuration file is returned as output, and it can be directly used to start the VA extension generation process. Users can manually check the auto-generated configuration file before generating the VA extension to revise supported resources.

WORKFLOW & OUTPUT. Once provided the VA Generator module with the configuration file, it can start the generation workflow, i.e., i) it checks the syntactical correctness of the configuration file by the `Syntax checker`; ii) validates the semantic correctness of the configuration by the `Validator`; iii) creates the `interaction_model.json` by the `Interaction Model Generator` containing configured intents, its utterances and the slot values according to the configuration file; iv) generates the back-end code by the `Back-end generator` and it produces the `back-end` (as a ZIP file) containing the back-end logic implementation.

While the syntax checker and the validator strictly depend on the configuration file, the interaction model and the back-end generator depend on the VA provider API. As we require a JSON configuration file, the `JSON Syntax Checker` has to verify that the file is a valid JSON file, while the `Validator` checks if all the mandatory fields are provided, and the configuration is consistent. If any error occurs, the generator immediately stops and returns a message reporting the occurred error. If the configuration is properly provided, the generator returns a folder entitled as the VA extension wake-up word containing the `interaction model` as JSON file and the `back-end` Node.js code as a ZIP file. It is worth noting that the generated VA extension is ready to be used, i.e., it can automatically be uploaded on

Amazon developer[2] and Amazon AWS[3]. The generated code corresponds to manually created VA extensions but may reduce required technical competencies and development time.

EXTENSION POINTS. In the current version ($v1.0$), we support the Amazon Alexa provider. Thus, once validated the configuration file, the Alexa skills components (the JSON interaction model and the zip file implementing the VA extension back-end that can be uploaded on Amazon AWS) can be created. Thanks to the architecture modularity, it is easy to develop new VA providers' support by focusing on the `Back-end generator` implementation.

Concerning the linking phase, it is performed in a dedicated function (as reported in the documentation) to enable end-users (with competencies in programming and KG querying) to customise it, e.g., it by calling APIs. By default, the back-end exploits the dictionary to perform the linking step. If the slot value is resolved as a list of URIs, it will exploit them during the SPARQL query formulation. Otherwise, the user value is used as-is in a SPARQL query by comparing it with resource labels.

Moreover, developers may add new supported languages by translating utterances in the target language and extending the reply formulation mechanism to return replies in the desired language. At the moment, English and Italian are supported.

To add a new pattern, developers have to model the new intent as a set of utterances and extend the back-end logic to formulate the related SPARQL query and the reply.

12.4 USE CASES

This section provides an overview of the benefits and challenges in querying KGs by VAs by presenting a pool of Alexa skills for CH KGs. We overview the generator configuration options, and we show the VA extension in action. The VA extensions back end and its interaction model are freely available on GitHub[4].

[2] Links for Alexa skill deployment: http://developer.amazon.com

[3] Links for Alexa skill deployment: http://aws.amazon.com

[4] GitHub project: https://github.com/mariaangelapellegrino/virtual_assistant_generator.git

It proposes a use case for each category of the CH taxonomy. In particular, for the tangible category, we propose the Nomisma use case for the movable sub-category, and the Hungarian museum use case for the immovable one; DBTune for the intangible category; NaturalFeatures for the natural heritage category; the UNESCO thesaurus represents the terminology category.

TANGIBLE MOVABLE CATEGORY: MAPPING MANUSCRIPT MIGRATIONS. Mapping Manuscript Migrations (MMM) [297] is a semantic portal for finding and studying pre-modern manuscripts and their movements, based on linked collections of the Schoenberg Institute for Manuscript Studies, the Bodleian Libraries, and the Institute for Research and History of Texts. It models physical manuscript objects, the intellectual content of manuscripts, events, places, people and institutions related to manuscripts.

Configuration. We automatically configured the MMM VA extension by exploiting the generator configuration component. The returned configuration file is used to initialize the generator.

VA extension in action. We ask for databases aggregated by the MMM portal by posing the *How many databases are there?* question. Used resources are i) Bibale (which stands for Bibliothèque médiévale), a long-term project of the Codicological Section of the IRHT (The Institute for Research and History of Texts) in Paris; ii) Bodley, i.e., Medieval Manuscripts in Oxford Libraries, and iii) SDBM, i.e., Schoenberg Database of Manuscripts. The user request *How many databases are there?* match an utterance attached to the `getClassInstances` intent, which returns the instances of a given class (database in this case). To verify the timeliness of retrieved information, we ask *Which database has modified equals to 2020?* which corresponds to an utterance matching the `getNumericFilterByClass` intent that verifies which instance of a given class (database in our use case) has a property (modified in our case) matching a given numerical value (2020 in our case). It replies to the CH community need to verify the queried sources and the timeliness of the retrieved information.

TANGIBLE IMMOVABLE CATEGORY: HUNGARIAN MUSEUM. The Hungarian Museum [222] provides access to the Museum of Fine Arts Budapest data.

Configuration. The Hungarian museum VA extension has been manually configured by retrieving `owl:class`, used classes and triples subjects, and the used properties. Labels are mainly provided in Hungarian, without English translation. Moreover, resources often lack any label.

VA extension in action. By querying *what is the creation with the maximum value of had participant* we activated the `getSuperlative` pattern which returns the class instance (the creation in our case) corresponding to the maximum (or minimum) value of a given property (had participant in our use case). The VA extension usually refers to resources by labels. In this case, it returns the creation URL. It makes evident the consequences of lack of labels attached to resources and the difficulties in exploiting them in VA-based applications.

INTANGIBLE CATEGORY: DBTUNE CLASSICAL. DBTune classical [267] describes concepts and individuals related to the Western Classical Music canon. It includes information about composers, compositions, performers, and influence relationships.

Configuration. I automatically configured the DBTune classical VA extension by exploiting the generator configuration component. The returned configuration file is used to initialise the generator after applying basic configuration manipulation, such as identifying which relation can play the role of label predicate (`alias`[5] is exploited). This use case demonstrates developers' challenges when the KG adopts a non-standard way to attach human-readable labels to resources.

VA extension in action. *Who has Beethoven as influenced by?* activates the `getProperty Subject` intent which retrieves the subject of triples where *influenced by* is the property and *Beethoven* is the object. This use case addresses the CH community interest in retrieving curiosities about musicians and artists.

NATURAL HERITAGE CATEGORY: NATURAL FEATURES. It is part of Scotland's official statistics [293] that provides a range of statistical and geographic data about Scotland from various or-

[5] http://dbtune.org/musicbrainz/resource/vocab/alias

ganisations. In particular, we are interested in aspects concerning geodiversity, ecology, and biodiversity.

Configuration. We automatically configured the Natural feature VA extension by exploiting the generator configuration component. The returned configuration file is used to initialise the generator.

VA extension in action. *What is the relevance of terrestrial breeding birds?* activates `getPropertyObject` intent which returns the value playing the object role in triples related to *terrestrial breeding birds* as subject and *relevance* as predicate. CH lovers and experts joining the user survey on the impact and potentialities of the proposed approach in the CH domain stress that VA extensions might be useful for educational scenarios. This use case simulates the possibility of deeper domain-specific information for familiarising with the terminology or conducting researches.

TERMINOLOGY CATEGORY: THE UNESCO THESAURUS. The UNESCO Thesaurus [324] is a controlled and structured list of terms used in subject analysis and retrieval of documents and publications in education, culture, natural sciences, social and human sciences, communication, and information. Continuously enriched and updated, its multidisciplinary terminology reflects the evolution of UNESCO programs and activities. Like a thesaurus, it mainly provides access to synonyms and related concepts. As a dictionary, it provides term definitions.

Configuration. We manually configured the UNESCO VA extension by retrieving $(4,421)$ `skos:Concept` that defines all the thesaurus terms and the used properties. All the concepts are attached to a human-readable label (by `skos:prefLabel`), while we generate property labels by local names of URIs.

VA extension in action. We can ask for the term definitions, e.g., *what is intangible cultural heritage?*. It activates the `getDescription` intent, i.e., a special case of `getPropertyObject` where the property is bound to a relation modelling term description. The VA extension retrieves the description (configured as skos:scopeNote) attached to intangible CH, and it returns the term definition. We can also pose ask queries. By querying *Can you verify if intangible cultural heritage as folklore as narrower* we activate the `getTripleVerification` pattern, which model ask queries that verify if the stated triple is modelled in the KG. It replies to the interest of the CH community to clarify and use domain-specific terms properly.

12.5 EVALUATION: PERFORMANCE ASSESSMENT

This section empirically assesses the quality of the generated VA extensions and test to what extent configuration options affect the returned VA extensions. All the presented skills and the discussed results are online available on GitHub [6] while its permanent URI is `https://zenodo.org/record/4605951`.

12.5.1 *Evaluation design*

This evaluation tests the accuracy and the precision of the auto-generated VA extension as an approach to verify to what extent the configuration affects the proposed evaluation. It demonstrates that the generation of a VA extension in a single click already returns VA extensions that are as accurate as systems proposed in the literature evaluated on the same benchmark. It also demonstrates that by tuning the generator configuration, end-users can significantly improve the accuracy and precision of the auto-generated VA extension. It is relevant to assess the performance of the auto-generated VA extensions as a special case of KGQA VA compared with systems categorized as traditional KGQA.

METHODOLOGY. The following questions (Qs) guided our evaluation process:

Q1 Are the results achieved by the auto-configured VA extensions comparable with other KGQA systems?

Q2 To what extent does the manual configuration refinement affect results?

Q3 Which linking approach between the dictionary lookup and API-based approach achieve the best results?

[6] GitHub project: `https://github.com/mariaangelapellegrino/virtual_assistant_generator`

While Q1 compares the proposed approach with alternative KGQA approaches, Q2 and Q3 have been formulated and evaluated to overcome any scepticism by end-users in terms of the impact the generator configuration on the generated VA extensions has in terms of performance. Thus, they analyse to what extent the linking approach and the lookup mechanism affect the performance of auto-generated VA extensions.

DATASET & BASELINES. The evaluation is performed on a standard benchmark on KGQA systems, QALD, as it contains benchmarks for multiple well-established KGs (i.e., DBpedia and Wikidata), and it tests both simple and complex questions. The QA systems joining the challenge are used as baselines, by referring to official results published in the QALD report. While QALD-9 contains questions tailored to DBpedia, QALD-7 refers to Wikidata. As systems joining the QALD-7 challenge relied on a version of Wikidata different from the actual one, the evaluation only reports results achieved by the Wikidata VA extension generated by the proposed software framework and the updated version of the QALD-7 dataset to enable further comparisons.

SETTINGS. First, the evaluation requires the generation of the assessed VA extensions, which are DBpedia and Wikidata VA extensions. The generated VA extensions are different in configuration options (manual VS auto) and linking approach (dictionary VS APIs). Further details follow.

Manual Configured DBpedia VA extension. The manual configuration option requires end-users to perform standard queries on the SPARQL endpoint of interest to retrieve all the classes, properties, and resources and to organise them in the JSON format. As Alexa requires the specification of custom slot values in the interaction model and poses a constraint on the interaction model size (1.5MB), developers have to query a sub-graph of the KG of interest. In the sub-graph retrieval, we focus on heterogeneous macro-areas. In particular, the entities dictionary contains all the declared classes (750) and 28.5K resources, distributed as follows: 5K people; 5K cities countries, and continents, 2K rivers and mountains related to the geography field; 3K films, 2.5K musical works, and 3K books belonging to the entertainment category; 4K museums and monuments and 1.5K artworks belonging to the art field; 2.5K animals and celestial bodies, related to the scientific field. The property dictionary contains all the declared properties (5K). It exploits the first results returned by the DBpedia SPARQL endpoint without either applying any sorting option or checking the returned results' relevance. Then, we performed basic cleansing operations, such as lower-casing labels and removing codes as labels to avoid readability issues. Finally, the VA extension has been generated.

Auto-configured DBpedia VA extension. Users can opt for the auto-configuration by specifying the endpoint of interest and the `Configuration generator` automatically creates the configuration file. The configuration file contains DBpedia classes and property, while it lets VA extension users freely refer to resources, and the queried labels will be compared against KG resource labels during the query formulation step. Users can either accept the generated file or manually cleanse the configuration file before generating the VA extension. It behaves as a checkpoint to further reduce the human effort and enable end-users to control the VA extension generation process. The configuration file initialised the VA generator.

Dictionary-based WikiSkill. It is based on a Wikidata sub-graph. It results in a dictionary composed of 2K classes and 28.5K resources, obtained following the same topic distribution described for the manual configured DBpedia VA extension. The property dictionary contains all used properties (6.5K). We lowercase all the labels, remove the ones containing unreadable codes and add synonyms to entities and properties by retrieving the Wikidata `also known as` property. The VA extension has been generated and used as-is without applying any further modification.

API-driven WikiSkill. The generator back-end offers the opportunity to modify the linking approach by affecting a dedicated script to customise back-end logic functions. It is based on wikibase-sdk[7], a library to make read queries to a Wikibase in-

[7] Wikibase-sdk: https://www.npmjs.com/package/wikibase-sdk

stance (e.g., Wikidata). `searchEntities` enables the opportunity to perform entity (and property) linking by resolving labels given as input. We created the `API-driven WikiSkill` by i) modifying the linking method from the dictionary lookup to the invocation of the `searchEntities` function and ii) the SPARQL query execution with `sparqlQuery` function in the `Dictionary-based WikiSkill` back-end.

PROTOCOL. For each VA extension, the described evaluation follows these steps:

- for each question contained in the QALD testing set (QALD-7 for Wikidata and QALD-9 for DBpedia),
  - it manually checked if the posed question either matches one of the supported intents or it can be transformed into a chain of supported intents. For instance, the question "What is the time zone of Salt Lake City?" in QALD-9 on DBpedia matches the getPropertyObject intent ("What is the p of e?") where <time zone> plays the role of p and <Salt Lake City> plays the role of e. The question "What is the name of the university where Obama's wife studied?" in QALD-9 on DBpedia can be transformed into a chain of supported intents where first users can ask for "Who is the wife of Obama?" (corresponding to the getPropertyObject intent where wife is the predicate and Obama is the entity) and, then, "What is the school of Michelle Obama?" (corresponding to the getPropertyObject intent where school is the predicate and Michelle Obama is the entity). If not, the question is skipped. Otherwise, it continues the procedure.
  - the activated intent is checked and the query is formulated according to one of the supported utterances.
  - the question reformulation has been posed to the VA extension;
  - all the replies returned by our VA extension, including empty results, were stored in a JSON file.

- Finally, at the end of the evaluation, results have been computed through the official system used to evaluate the QA systems joining the QALD challenge, GERBIL [8].

For Wikidata and QALD-7, the previous procedure required updating replies in the testing set to compile the current Wikidata version (July 2020). The evaluation has been performed on the updated version of the QALD-7 training dataset[9] and the resulting dataset has been shared online to encourage further comparison.

data gathering. The evaluation considers the standard evaluation metrics for the end-to-end KGQA task, which are precision (P) and recall (R) and F-measure (F1) at a micro and macro level.

12.5.2 *Results*

configuration options, the dbpedia case. It compares results achieved by the manual and auto-configured DBpedia VA extensions[10] with the results achieved by the systems that joined the QALD-9 challenge [326] (Table 12.2). Regardless of the considered configuration approach, the DBpedia VA extensions achieve the best results in all the metrics, and they obtain results from 2 to 6 times better than the second-best result obtained by the participants in the challenge (Q1). The achieved results are justified by i) the exploitation of structured NL questions and ii) the possibility to tune the VA extension initialisation according to specific needs by a fine-grained control. End-users can add data of interest in the configuration file, for instance, resources required by the testing dataset that the previous coarse-grained entity selection has not included. While the manually configured VA extension obtains optimal results due to the user's full control, the auto-configured DBpedia VA extension provides lay-users

[8] GERBIL: http://gerbil-qa.aksw.org/gerbil

[9] QALD-7 training set updated to July 2020 Wikidata status qald-7-train-en-wikidata-July2020Version.json

[10] Manual configured and auto-configured DBpedia VA extension results, respectively: http://gerbil-qa.aksw.org/gerbil/experiment?id=202012170018 and http://gerbil-qa.aksw.org/gerbil/experiment?id=202012170019

with a good starting point to be used with or without manual refinement (Q2).

Table 12.2: DBpedia VA extensions performances over QALD-9

Tool	Micro results			Macro results		
	P	R	F1	P	R	F1
ELON	0.095	0.002	0.003	0.049	0.053	0.050
QASystem	0.039	0.021	0.027	0.097	0.116	0.098
TeBaQA	0.163	0.011	0.020	0.129	0.134	0.130
wdaqua	0.033	0.026	0.029	0.261	0.267	0.250
gAnswer	0.095	0.056	0.070	0.293	0.327	0.298
Auto-c.	**0.991**	0.197	0.328	0.369	0.358	0.354
Manual c.	0.990	**0.284**	**0.441**	**0.683**	**0.677**	**0.678**

LINKING APPROACH, THE WIKIDATA CASE. The QALD-7 training set contains 100 questions, but 4 questions cannot be more answered. We can reply to 76/96 questions, while the remaining 20 questions are not supported patterns. The `Dictionary-based` VA extension obtains the best results (see Table 12.3) due to the full control that users have in solving ambiguities and customise priority in the linking phase (Q3).

Table 12.3: The WikiSkills performances on QALD-7

	Micro-results			Macro-results		
	P	R	F1	P	R	F1
Dict. WS	**0.989**	**0.946**	**0.967**	**0.736**	**0.747**	**0.739**
API WS	0.954	0.262	0.412	0.664	0.677	0.669

Not surprisingly, the dictionary-based linking approach is more precise than the API-driven approach as a dictionary gives the possibility to tune and customize the order and the priority in the URIs list attached to the same entity or predicate. For instance, the term Paris might be attached to the French capital and VIPs whose name is Paris, such as Paris Hilton. If the VA extension is designed to be used as a museum virtual guide, the dictionary-based configuration can attach a higher priority to Paris as a city instead of other interpretations. This mechanism cannot be performed in the API-based configuration. Even if the dictionary represents a static snapshot of the KG content, it can be exploited both in the entity and relation linking task. On the contrary, it is required to verify if APIs offer both linking mechanisms. The dictionary-based linking approach is also KG-agnostic, i.e., it is independent of any external service. It only requires configuration time and extra storage in the back-end but guarantees direct and immediate (without execution time) access to URIs. Moreover, the dictionary-based solution is general enough to enable the VA extension back-end configuration with any KGs without any constraint.

12.6 evaluation: user experience

This section assesses the usability of an auto-generated Alexa skill according to delegates of HETOR[11], a CH association of the Campania region in Italy, and it behaves as a preliminary usability assessment of the proposed approach in a controlled environment. The HETOR project collects and makes available as OD the Open Heritage published by the National Institutions, such as ISTAT, MIBACT, MIUR and Campania Region (Italy), and the Open Heritage that can be created by the citizens themselves, concerning their territories, improving the quality and quantity of OD available at a local and national level. HETOR mainly collaborates with schools to study and preserve the historical and collective memory of local CH. In the context of their activities with schools, they organize co-creation sections for encouraging learners to familiarise themselves with CH, collect information about the CH to preserve and model it as tabular data by caring about the correct terminology. It requires to familiarize with terms and their definition, hierarchy of concepts, mastering synonyms and analogies. Thus, during the activities, learners usually ask mentors questions like *What is the meaning of geo-localization?*, *Can you define a point of interest?*, *What do you mean by year of foundation?* In this context, the HETOR group has

[11] HETOR: http://www.hetor.it

the real need to address a plethora of requests posed by each group to clarify terminology. The situation is even worse during the COVID-19 as activities were performed online and there was a limited possibility to clarify all the doubts due to the lesson settings and also due to the wider exploitation of asynchronous activities that required learners to do without continuous support from moderators. We proposed to the HETOR group to think about the possibility to use a VA extension generated by the proposed approach configured to query a thesaurus, in particular the UNESCO thesaurus, and test the usability of the VA extension in the first person.

12.6.1 *Evaluation design*

METHODOLOGY. The RQ at the basis of the performed evaluation is *"To what extent CH data curators perceive a VA-based mechanism usable to exploit CH data"*?

PARTICIPANTS AND SETTING. 5 delegates of the HETOR project joined the usability evaluation of the UNESCO Alexa skill generated by the proposed generator, corresponding to the one described in Section 12.4. The evaluation took place remotely due to the COVID-19. As the VA extension has not already been published on the Alexa store, we deployed the VA extension on the Alexa developer console and asked participants to interact with the textual interface. We behaved as moderators, while we asked participants to formulate questions and we collected their thought and reactions.

PROTOCOL. The performed protocol involved:

- an introductory overview of the objective of the user experience evaluation, the queried source by looking at the UNESCO Thesaurus browsing interface[12], the setting of the evaluation, and inspecting the presentation of the VA extension which introduce itself by pronouncing *"Hi!! Welcome to the UNESCO personal assistant! Ask me your curiosities, such as: Can you define digital heritage? What is the narrower of*

[12] http://vocabularies.unesco.org/browser/thesaurus/en/

cultural heritage? What is broader of churches? What is related to digitization?";

- the assignment of a collection of tasks to each participant posing questions such as *The definition of Digitisation*, *The definition of CH*, *The specializations of digital heritage*, *The generalization of digital heritage*, *The terms related to CH*. For each task, participants are encouraged to identify the pattern to pose the related questions and collect replies returned by the Alexa skill.

data collection. At the end of the evaluation, the moderator asked for the fulfillment of a final questionnaire to evaluate i) users' satisfaction based on the SUS questionnaire [181] and ii) their interest in using and proposing the tool by a BI survey. The questions of the BI survey are i) *"I will use this approach in the future"*; ii) *"I will recommend others to use the proposed approach."* and users can use a 5-point scale to reply. Moreover, the moderator annotates all the comments and observations raised during the evaluation.

12.6.2 *Results*

The proposed approach achieved a SUS score of 80, close to the higher step, interpreted as a great appreciation of the proposed tool and the propensity to propose it to others. The latter result is verified by the BI survey which achieved a mean score of 4.6.

Besides the tasks explicitly assigned by the moderator, participants started posing queries, asking for *The generalization of mosques and synagogues*, *the definition of amphitheater*, *the generalisation of Catalan or Gothic* and *painting*, and *the specialisation of fine arts*.

Question templates. It resulted naturally to pose questions according to specific templates as it is the traditional approach used to query Alexa and its VA extensions, but it requires training to learn the supported templates. Participants inquired the moderator asking for the other supported patterns, besides the ones tested in the UNESCO Alexa skill, and they were almost satisfied with the covered templates. In particular, they asked for further details on numerical filters, quantitative queries, mechanisms to retrieve images, and they assessed that the query to retrieve object proprieties is the easiest and the more natural one. However, they considered tailoring utterances according to the target user a crucial aspect. For instance, educational contexts may require further simplifying the used terminology and adopting a wider used way to formulate similar questions. Participants suggested integrating the definition intent with utterances such as *What is the meaning of X?*. We also discussed if a keyword-based search might result in a dirtier but quicker way to retrieve information. Further study should be performed to verify the expressiveness capability of a keyword-based query mechanism.

Target age. The proposed mechanism is perceived as a powerful approach above all for young people that are more and more accustomed to query VAs to perform daily tasks. Participants observed that it seems also particular adapt with very young learners, also in the pre-scholar phase, as vocal commands represent the unique approach that they can use as they cannot already write commands. Similarly, this approach might be critical for learners with disabilities that prevent them from typing questions or adopting textual interfaces, which may result too difficult for blind people or people with a limited range of motion.

The role plaid by the data source. Queried data sources play a crucial role in the effectiveness of the resulting VA extension. For instance, the UNESCO Thesaurus is too generic for domain-specific questions and CH experts also disagree with some of the reported definitions. As an example, they are surprised by the taxonomy proposed by the UNESCO for the CH concept, as they expect the well-known taxonomy based on tangible, intangible, and natural heritage. As data modeling impacts also the VA extension utterances, it is crucial to evaluate the naturalness of the resulting questions.

Application contexts. The HETOR group appreciated the proposed approach as a way to provide learners with continuous support to master terminology about CH and become familiar with related concepts. Learners are less and less accustomed to consulting a dictionary to look for the right terminology. The proposed approach might revive the interest in familiarising and disambiguating terms and enrich the personal vocabulary. As in the described activities, this proposal has interesting implications for groups of works to retrieve thematic information and images.

Discussing alternative application contexts, the proposed mechanism seems to be particularly useful in guided tours, to guarantee personalized interactions, guided by curiosities avoiding boring pre-packaged presentations of artworks and points of interest. VA extensions as virtual guides can overcome the lack of interest in the entire exhibition and too detailed descriptions, lack of personalization in terms of tour duration, interests, curiosities. It also solves the linguistic gap between visitors and personnel. It enables the possibility to perform tours to the desired speed with the possibility to repeat unclear passages without bothering others. If it may be an interesting alternative to audio guides already available in museums, it might be revolutionary for city tours to explore points of interest spread in cities or minor realities.

12.7 IMPACT AND UTILITY PER END-USERS

This section discusses the perceived impact and utility from the end-user perspective. We proposed an online survey to collect opinions and suggestions. We do not limit to experts in the CH field, but also try to involve CH lovers to take their opinion into account. Moreover, it is worth noting that we do not limit this survey to experts in the field as we are assessing the perceived impact and utility from the end-users side meaning that we need to collect opinions by inquiring potential users of the resulting VA extensions.

12.7.1 *Evaluation design*

METHODOLOGY. The RQ at the basis of the reported evaluation is "*Which is the perceived impact and utility of a VA-based CH exploitation approach according to end-users*"?

PARTICIPANTS AND SETTING. 86 people joined the online survey that was administered for one week, from September 15th to September 22nd, 2021. All the participants spontaneously joined the survey in an anonymous form. According to general information collected by the survey, 73 people are (very) interested in CH by rating their interest at least as 4 out of 5 and 24 of them are experts in CH by rating their expertise in CH at least as 4 out of 5. By looking at people that consider themselves as experts in the CH, they have limited expertise in Computer Science, stressing that it is crucial to provide the CH community with tools that do not take for granted their competencies in programming, query languages, and computer science.

DATA GATHERING AND SURVEY OUTLINE. The survey has been administered in English and in Italian and its content is described in Table 12.4 that reports questions, reply format, and the rationale behind each posed question. The survey is structured in three main sections: i) general information about participants expertise and interest in CH, the spread of VAs within the CH community interpreted as people that are either experts or interested in CH, alternative means used to query and explore CH;

Table 12.4: Impact and utility survey outline

Question	Reply format	Question role
General information		
Your interest in the CH.	1-5	User profiling
Your expertise in the CH.	1-5	User profiling
Your expertise in computer science.	1-5	User profiling
Used Virtual Assistants	None Alexa Google Assistant Others More than one	Spread of VAs within the CH community
Frequency of Virtual Assistant device usage.	Never Rarely - Less than once a week Sometimes - 3 times a week Always - Everyday	Spread of VAs within the CH community.
Have you ever looked for CH information?	Yes/No	Alternative exploitation means
If so, used device and application.	Free text	Alternative exploitation means
Impact and Utility to query Cultural Heritage by Virtual Assistants		
In which context does the proposed approach may be useful?	Library In museums as virtual guides As learning assistant at school No utility	Perceived utility in terms of application contexts
Application CH context advantaged by VA	Free text	Application context
To what extent VAs can spread the CH?	1-5	Perceived impact
Example of queries you are interested in	Free text	Intent coverage
Are there activities performable only by VAs?	Yes/No/Maybe	Perceived utility
If so, which one?	Free text	Perceived utility
Are there activities improved by VAs?	Yes/No/Maybe	Perceived utility
If so, which one?	Free text	Perceived utility
Suggestions and Comments		
Any suggestion	Free text	Collection of suggestions
Any comment	Free text	Collection of comments and feedback

ii) questions concerning the perceived utility in terms of application contexts and as an alternative to traditional CH exploitation means, the perceived impact achieved by spreading CH data by VAs while queries users are interested in are both useful to assess the intent coverage and to collect ways users naturally pose questions; iii) finally, general suggestions and comments.

12.7.2 *Results*

This paragraph reports the most common replies and comments that raised interesting considerations concerning the proposed approach in the field of CH.

current exploitation means. More than half of the participants that assessed to be interested in CH by rating their interest at least as 4 out of 5 (56%) queried and looked for CH data, at least once, by rarely exploiting websites dedicated to the CH of interest (in 2 out of 43 cases) or bibliographic sources (in 3 out of 43 cases) and by mainly googling it (in 28 out of 43 cases). It is interesting to notice that in 10 out of 43 cases, CH lovers already exploit VAs to fulfil their curiosities.

Table 12.5: Diffusion of VAs in the CH community

		Used VAs					Usage frequency			
	Tot.	None	Alexa	Google Assistant	Others	More than one	Never	Rarely	Sometimes	Always
Interested in CH	73	18	18	20	8	9	18	29	20	6
CH experts	24	6	6	6	4	2	6	8	6	4

The diffusion of VAs within CH lovers and experts are also confirmed by results reported in Table 12.5 that summarises the most used VA providers and the frequency of their usage. Most of the participants have their favourite VA and usually get stick to it without experiencing multiple providers. If a single provider is chosen, Google Assistant appears to be the preferred one. VA usage is still limited to a few days a week meaning that there are still barriers in the wide exploitation of VAs in this community and it requires overcoming scepticism, perhaps, leveraging on curiosity connected to novelty or by demonstrating to the potential users about utility and potentialities.

If compared with traditional text-based interfaces, by asking for activities that can be performed **only** with VAs, participants recall the potentialities to use VAs with users with disabilities, such as blindness, or situations that impede the usage of a keyword to type questions. It seems to be particularly useful at school during teamwork. By asking for activities that can benefit from the usage of VAs, users underlined the advantage to pose questions rapidly, interactive consultation of sources, simplify lookup operations.

application contexts. Considering the entire set of replies, independently from participants' interest and expertise in CH, just in one case, a participant cannot see the potentialities of the proposed approach, while all the other ones selected at least a CH application context that might take advantage of VAs. We first asked participants to choose among a set of options, i.e., in libraries to help to look up books, as a virtual guide in museums, and as a learning assistant at school. Fig. 12.4 reports participants' opinions who seem to be convinced that our proposal is a promising approach as a virtual guide in museums.

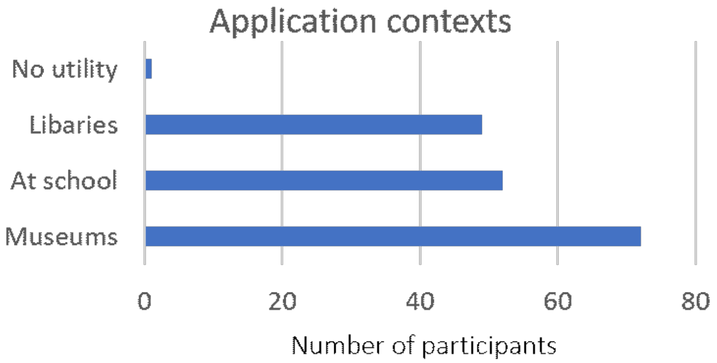


Figure 12.4: Application contexts rated by survey participants.

Furthermore, we also asked users to think about any other application context that can take advantage of VAs. In 39 out of 67 cases, users see the potentialities to adopt the proposed approach as a virtual guide not only in a museum but to guide visitors while wandering in an unknown city, above all while visiting small villages, unconventional destinations, or cities with low population density and high cultural impact, dense of archaeological parks or churches. An interesting consideration has been proposed by more than one participant that assessed that the proposed approach is particularly useful when there is no possibility to type requests, for instance, while driving. Moreover, a user also suggested thinking about the exploitation of VAs in a virtual museum by simulating a real tour also in terms of tour guide. In 14 out of 67 cases, users state that it might result in a promising individual learning tool at university to learn about terminology and clarify doubts while preparing exams or scientific contributions, at school to disambiguate terms, at home to deeper knowledge and awareness about CH, for young learners to overcome limits posed by textual interfaces and to leveraging on their curiosity. Moreover, further considerations concern the inclusiveness of the proposed approach, able to overcome disabilities related to limited usability of textual interfaces or blindness. Users also proposed VAs as support in superintendence offices, in archives to guide the lookup phase, as support in offices, and (surprisingly!) in hospitals.

UTILITY AND IMPACT. We explicitly asked users to assess using a 5-Likert scale the perceived impact of using VAs as a means to spread the culture, interest, and awareness about CH. Fig. 12.5 graphically represents the resulting perceived impact demonstrating that most of the people think that VAs have the potentiality to wider spread the interest about CH, by leveraging of curiosity or providing an immediate access to data of interest.

Analysing topics users are generally interested in addressing by the posed questions, they inquire about point of interest's curiosities and details, such as *"Which is the history of monument x?"*, *"Which museums are located in x?"*, *"When x happened?"*, *"What is the aim of x?"*, *"Who is the author of x?"*, *"What is the historical context of x?"*, *"Which artists authored any painting hosted in the museum x?"*, *"Where is x?"*, *"Which is the architectural style of x?"*

Participants expect to use the proposed approach to plan a trip. Thus, besides collecting cultural information, users are also interested in collecting practical information about points of interest, such as the accessibility posing questions like *"Which is the ticket for visiting x?"*, *"Does x support an all-inclusive ticket?"*,*"What is the opening time of x?"*, *"How do visitors rate place x?"*, *"What can*

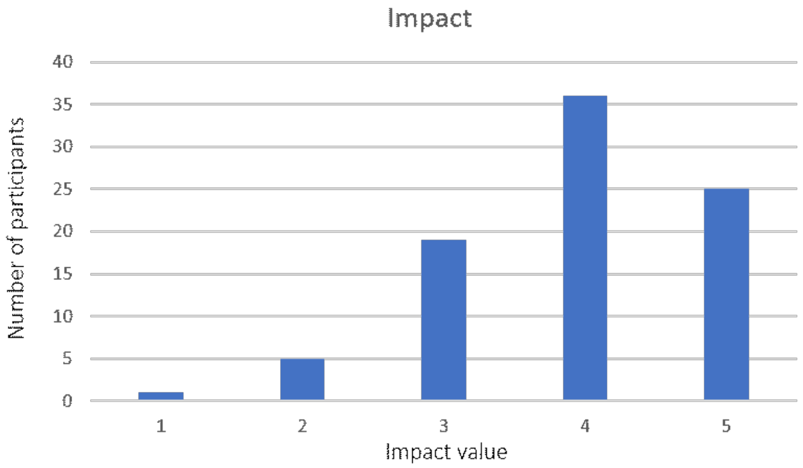


Figure 12.5: Perceived impact of the capability of VAs to spread the culture of CH.

I visit in city x?", "Which are the most important artworks hosted in x?", "Where can I visit x?", "Are there events in x?", 'Which is the most important point of interest in x?", "Which are the most famous x?", "Where is the oldest surviving monument of the world?".

Some of the participants simulate a conversation with a thesaurus, by clarifying terms and terminology. Moreover, learners can be interested either in looking for details about the queried sources to assess their reliability or retrieving the list of sources containing information about a given topic. As an alternative, according to the school level and subject, learners might be interested in specific information, such as *"Who is restoring x?", "Who is curator of x?", "Is x curated by UNESCO?", "How was called x before date d?", 'How tall is x?"*. Learners might be interested in the story of the past, such as *"Did Romans took baths at the sea?", "Who was x?", "What is authored/discovered/invented by x?"* and vice versa *"What did x author/discover/invent?", "In which occasion x has been build?", 'Why x is famous?", "Which artist influenced x?", "Which is the art movement of x?", "Which are the most important artworks authored in x?", "What characterise x?", "How many artworks have been authored by x?".*

Many participants treat the proposed mechanism as an approach to fulfill general curiosities, such as *"Is there any legend* *behind x?"*. An interesting aspect that emerged by the collected replies is that VAs interpreted as vocal assistants can easily perform storytelling and can be queried to tell a random event or curiosity about an artist or a monument, can narrate *"How was city x before event x happened?"*, *"Tell me the story of x"*, *"Can you describe x?"*, *"Talk about x."*, *"Give me further details about x"*, *"Tell me curiosities about x."*

Even if the most common questions concerns tangible CH, both movable, such as artworks, coins, and documents, and immovable, such as churches, monuments, and castles, users are also interested in curiosities related to intangible CH, such as the folklore of local traditions, such as *"What are the traditional dances of x?"*, music, such as *"Which musician has the most albums?"*, events, co-occurrences of terms in books and manuscripts, such as *"Does x discuss about x?"*, *"How many times x talk about y?"*.

While in most of the cases participants posed punctual questions, some of them used their imaginary VA extension to explore available data. For instance, *"Give me all the titles/authors belonging to the topic/category x."*, *"Give me artworks related to x"*.

Besides textual replies, users are also interested in visualizing photos and videos, such as *"Show me x"* to retrieve examples of presbyters or a building plan, or *"Show me video related to x"* where x might be an event or a monument.

In most of the cases, participants formulated complete questions, while rarely they posed commands to collect information. Looking at the way questions have been formulated, the proposed intents successfully reply to most of them. The proposed approach misses the fulfillment of complex queries, which are quite rare in this questionnaire. For instance, currently, it cannot deal with questions like *"How long did it take the building of x?"* if it implies computing the difference between the foundation and the completion date, composed questions like *"Give me artworks painted by the same author of x"* requires users to split them into two queries as demonstrated in the performance evaluation.

GENERAL COMMENTS. As general suggestions, participants noticed that even if our proposal is particularly suitable for the CH field, its usage might be hypothesis in any application context, such as PAs and hospitals. It is crucial to care about queried sources in terms of coverage of topics and reliability. As the interaction model is strictly connected to the target user group, the usability of the resulting VA extension should be carefully checked to tune the way questions can be posed and to make the interaction as natural as possible and limit the error rate. It might be interesting either to introduce a playful component or to evaluate the combination of VAs and virtual exhibitions. Participants also suggested automatically merging information from multiple sources letting users save time in querying individual sources.

Participants mainly used the comment question to compliments about the project, assessing that *the project has enormous potential, it guides digital transitions to our country, we are in a modern world and everything is going to be connected with technology, CH should not remain out of this, extremely versatile as it might be applied to any application context*.

12.8 IMPACT AND UTILITY PER DATA CURATORS

This section discusses the perceived impact and utility from the CH data curators' perspective. We proposed an online survey to collect opinions and suggestions and we administered it to two different groups of CH experts who are either modeling or are planning to model their data as KGs. This survey collects opinions and comments of potential users of the generator who might decide to propose the resulting VA extensions as data exploitation mean.

12.8.1 *Evaluation design*

METHODOLOGY. The RQ at the basis of the reported evaluation is "*Which is the perceived impact and utility of a VA-based CH exploitation approach according to CH data curators*"?

PARTICIPANTS AND SETTING. 5 people joined the online survey belonging to two different groups of CH experts. While 3 of them are delegates of the HETOR project, the other 2 researchers belong to a research group of Medieval Philosophy of the University of Salerno. The HETOR group mainly models tangible CH concerning local and national CH as tabular data releasing them according to the OD directive. They are planning to expose their data as KGs in the future. On the other side, the research group of philosophers is designing an ontology to model their collection of medieval manuscripts by representing co-occurrence of terms, philosophical concept interpretation, philosophical movements, both concerning Greek and Latin culture. They spontaneously joined the survey and represent experts or interested in CH.

DATA GATHERING AND SURVEY OUTLINE. The survey has been administered in English and in Italian and its content is described in Table 12.6 that reports questions, reply format, and the rationale behind each posed question. The survey is structured in three main sections: i) general information about participants expertise and interest in CH, the spread of VAs within the CH community interpreted as people that are either experts or interested in CH, alternative means used to query and explore CH; ii) the interest in making their data accessible by VAs; iii) finally, general suggestions and comments.

12.8.2 *Results*

This paragraph reports the most common replies and comments that raised interesting considerations related to the exploitation of the proposed approach in the field of CH.

CURRENT EXPLOITATION MEANS. The HETOR group performs analysis on their data by using query builders and data visualization approaches, mainly via `SPOD`. The used mechanism supports users in exploring, visualizing, and interpreting data, but requires expertise in data analysis and takes time to have a fast insight on available data. They feel that a VA extension might be a powerful approach to have an immediate insight of data, without limits on the dataset size or specific competencies.

The group of Medieval Philosophy are modelling history of philosophy data and related metadata by ontologies and plan to materialize the related KGs in the next future and to make them accessible to all. Even if planned exploitation tools are still under investigation, they hypothesize to exploit data in data visualization approaches to guide users in interpreting data.

Table 12.6: Interest in making CH data accessible by VAs survey outline.

Question	Reply format	Question role
General information		
Your interest in the CH.	1-5	User profiling
Your expertise in the CH.	1-5	User profiling
Your expertise in computer science.	1-5	User profiling
Used Virtual Assistants	None Alexa Google Assistant Others More than one	Spread of VAs within the CH community
Frequency of Virtual Assistant device usage.	Never Rarely - Less than once a week Sometimes - 3 times a week Always - Everyday	Spread of VAs within the CH community.
Used device and application to access CH data.	Free text	Alternative exploitation means
Impact and Utility to make CH data exploitable by Virtual Assistants		
Are you modelling data as KGs?	Yes/No/Maybe	Info about available data
Modelled data	Free text	Info about available data
Expertise in SPARQL in your working group	Yes/No/Maybe	Competences in CH groups
Do you plan to make your data accessible to others?	Yes/No/Maybe	Interest in data exploitation means
Which task do you plan to perform on your data?	Free text	Application context
Impact of VAs to spread CH data	1-5	Impact of VAs
Reaction to the proposed approach	Sceptical Suprised Euphoric Neutral Entusiastic Curious	Perceived impact
Reaction justification	Free text	Reaction to our proposal
Foreseen potentialities	Free text	Reaction to our proposal
Foreseen obstacles	Free text	Reaction to our proposal
Example of queries on your data	Free text	Intent coverage
Would you think about VAs as data exploitation means?	Yes/No/Maybe	Perceived utility
If so, which one?	Free text	Perceived utility
Suggestions and Comments		
Any suggestion	Free text	Collection of suggestions
Any comment	Free text	Collection of comments and feedback

They see potentialities to make them accessible by VAs, reacting with curiosity and enthusiasm, but mainly focusing on actual data to improve their accessibility. They are a bit skeptical about making metadata accessible by VAs as they cannot already foresee an application context of interest as only experts are usually interested in metadata, according to their opinion.

It is worth noting that in both groups, participants stated that their working groups have no competencies in querying languages, such as SPARQL. It is crucial to provide this community with CH data exploitation tools not requiring technical competencies.

APPLICATION CONTEXTS. According to data published by the HETOR group, they are interested in retrieving artworks information, such as location, author, date posing questions like *"How many* x *are in* y*?"* as a general question to quantify castles in Campania, or museums in Italy, or churches in Naples; *"Which is the construction year of* x*?"*, *"Where is* x*?"*, *"Which are artworks authored by* x*?"*. They also hypothesize to query a VA extension to obtain terms definitions and disambiguation, such as *"What is meaning of* x*?"*. Our proposal is perceived as a promising approach at school to familiarise with terms and concepts also during remote sessions, in museums, or city tours as virtual guides.

IMPACT AND UTILITY. Participants assess that they would be delighted to query data by pronouncing questions instead of datasheets and query builders. Moreover, they assess that in their opinion the impact of making CH data accessible by VAs might be very high (grade 5 out 5). They reacted with enthusiasm to our proposal and are curious about the consequent applications. They foresee great potentialities given the possibility to spread the interest and the usage to a vast public without constraints on the age and without requiring any technical competence. They only see refrains by people that are still sceptical about the extensive use of technologies, but for sure it might be useful to engage young CH lovers to deepen their awareness and expertise in CH.

GENERAL COMMENTS. Participants suggested introducing the possibility to query multiple data sources at a time; tune the interaction model according to the target group and the planned application context; carefully check the used source in terms of accuracy and reliability. They explicitly stated that they foresee potentialities in this project and it would be extremely useful in the CH field, according to their opinion.

12.9 FINAL REMARKS

We propose a general-purpose approach to perform KGQA by VA, and we embed it into a community shared software framework to generate VA extensions by requiring minimum/no programming and query language competencies. Our proposal may have a significant impact as it may unlock the Semantic Web technologies potentialities by bringing KGs in everyone "*pocket*". It is the first attempt to empower lay-users to create personalised VA extensions, ready to be published on any VA provider.

Besides its general-purpose nature, we considered it particularly suitable for the CH community for different reasons. First, the CH community heavily invested in publishing data as KGs. Consequently, we believe that it is useful to provide them with tools and approaches to easily exploit the vast amount of available data. Second, CH lovers are usually provided with tools and interfaces to explore results of data exploitation means, such as virtual exhibitions, data visualization tools, QA applications, but they are rarely moved to the position of active curator of a mechanism to access data. Thus, the proposed generator moves the CH community in the position of generating their QA tools able to query any data source of interest provided with a working SPARQL endpoint. Thus, librarians can query their book archive, musicians can pose queries on music collections, art gallery curators can provide visitors with a virtual guide able to reply to questions instead of reproducing standard tracks narrating artifacts details. It is the first attempt to empower lay-users to create personalized and ready-to-be-use VA extensions.

We propose a reusable prototype of VA extensions generator to query any KG. In its actual open-source release (*v*1.0), we allow the building of Alexa extensions, and we aim to provide support for the Google Assistant. It is important to notice that we followed all the best practices in software design (e.g., abstraction and modularity) to guarantee technical quality and make the generator fully extensible.

The proposed approach queries KGs in real-time by exploiting up-to-date data and it is entirely KG-independent. It is also a general-purpose approach and it can be easily adapted to domain-specific applications, as in the use case section. As a general process, utterances make no assumption on question interpretation and the application context. The proposed approach is general enough to be exploited both in querying a single KG than in querying multiple KGs by aggregating query results in the reply formulation step, which means improving the back-end implementation without modifying the general approach. We aim to further investigate multiple KGs in the future.

By the overviewed use cases, we demonstrate most of the intents reported in the tab. 12.1 in real settings. We verify that the proposed approach is general enough to query data concerning different categories of CH, from museums to manuscripts, from music to term definition. Moreover, we also experienced some KG properties that affect VA-based KG exploitation.

label coverage. To cope with the scarce provision of human-readable labels, they can be generated by local names of URIs, as we performed in UNESCO Thesaurus. This practice can be performed if resources have human-readable URLs, such as in DBpedia. As evidenced in the Hungarian museum use case, the lack of label provision is an obstacle to resource understanding.

multilingualism. Some KGs, such as Finland datasets, Hungarian museum, Cultura, only provide access to labels in the data provider's native language without enriching resources by English translations. This lack of multilingualism prevents the wide exploitation of modeled data.

sparql support. A technical detail must be remarked. Before implementing the intents to SPARQL queries mapping, developers have to carefully check if the queried endpoint fully supports SPARQL or omits some patterns. For instance, to use alternative predicates, we exploited the `VALUES` pattern. It is not supported by some of the queried KGs, such as Munnin and CULTURA. It affects the back-end implementation or limits the endpoints that can interface with your implementation. Moreover, there are endpoints, such as CIDOC-CRM and AAT, that do not support the `COUNT` aggregator. It affects queries as simple as *How many artifacts are hosted in the Uffizi museum*.

Part IV

KNOWLEDGE GRAPHS AND MACHINE LEARNING

13 KNOWLEDGE GRAPHS AND MACHINE LEARNING

Predicting the future is not magic, it is artificial intelligence.
– Dave Waters

While KG are graph shaped by nature, most traditional ML algorithms expect data in a vector form. To make graph elements compatible with ML requirements, they have to be transformed into vectors by a vectorization or graph embedding approach. During the last decade, several graph embedding techniques have been proposed.

A graph embedding technique takes a KG in the form of an RDF graph as its input and creates a low-dimensional feature vector representation of nodes and edges of the graph. Formally, a graph embedding technique aims to learn a function $f : G(V, E) \rightarrow \mathbb{R}^d$ which is a mapping from the graph $G(V, E)$, where V is the set of vertices and E is the set of edges, to a set of numerical embeddings for the vertices and edges, where d is the dimension of the embedding. If only nodes are embedded, the technique is defined as node embedding.

Its purpose is to represent each node and edge in a graph (or a subset of them) as a low-dimensional vector while preserving semantic properties (e.g., keeping similar entities close together) and/or topological features. A desirable property for the obtained vectors is that they would be *task-independent*, meaning that they can be reused for other applications as they were created for. Therefore, it is useful to verify how the vectors perform on different tasks to broaden the insight into the information the embedding algorithm can preserve. It is also important to know whether the vectors show very good performance on a given task while their performance degrades significantly on others. While the *extrinsic* evaluation is not the only (and probably it is not the best) way to elect the best embedding approach, this kind of evaluation is extremely useful to choose the best set of vectors according to the tasks they will be used for. Besides the evaluation and comparison, a systematic evaluation is also useful in parameter tuning. Many embedding algorithms have various parameters, which are difficult to set. Therefore, it is interesting to compare different versions of the same algorithm and check how the parameters affect extrinsic evaluations.

Systematic comparative evaluations of different approaches are scarce. Approaches are rather evaluated on a handful of often project-specific data sets. Rarely proposed graph embedding techniques show how they perform on large and less regular graphs, such as DBpedia or Wikidata.

Comparing approaches is interesting for 1) developers of new embedding techniques to verify in which cases their proposal outperforms the state-of-art and 2) consumers of these techniques in choosing the best approach according to the task(s) the vectors will be used for. The comparison could be delayed by the choice of tasks, the design of the evaluation, the selection of models, parameters, and needed datasets.

Our *objective* is the proposal of a mechanism to simplify the evaluation and the comparison of graph embedding techniques. The research presented in this chapter has been published in the following contributions:

- Maria Angela Pellegrino, Michael Cochez, Martina Garofalo, Petar Ristoski. *A Configurable Evaluation Framework for Node Embedding Techniques*. In the Proceedings of Extended Semantic Web Conference (ESWC) (Satellite Events) 2019.

- Maria Angela Pellegrino, Abdulrahman Altabba, Martina Garofalo, Petar Ristoski, Michael Cochez: *GEval: A Modular and Extensible Evaluation Framework for Graph Embedding Techniques*. In the Proceedings of Extended Semantic Web Conference (ESWC) 2020.

13.1 RELATED WORK

Software evaluation frameworks can be categorised by the supported tasks. Moreover, frameworks can be distinguished according to the expected input. In the case of embedding algorithms, an evaluation framework can take as input the model and train it before starting the evaluation. In the alternative, it could expect pre-computed vectors. The input format can influence the type of covered tasks. For example, for a fair comparison in link prediction, it is important to know the input graph used to train the model. Only by bounding the training set, it is possible to fairly test unknown edges and verify the ability of the embedding algorithm to forecast only positive edges. This section focuses on frameworks to evaluate graph embedding approaches by pointing out the covered tasks. Table 13.1 lists frameworks by reporting the publication year, the covered tasks, and if they expect the model or the pre-trained vectors.

Table 13.1: Evaluation framework comparison

	Year	**Tasks**	**Embedding technique**
Bonner et al.	2017	Topological structure	Model
GEM	2018	Classification, Clustering, Link Prediction, Network Comparison, Visualisation	Model
Rulinda et al.	2018	Clustering, Link Prediction, Visualisation	Model
OpenNE	2019	Classification, Visualisation	Model
EvalNE	2019	Link Predition	Model
AYNEC	2019	Link Prediction	-
Bogumil et al.	2019	Clustering	Model
GEval	2019	Classification, Clustering, Document Similarity, Entity Relation, Regression Semantic Analogies	Vectors

Goyal and Ferrara [125] released an open-source Python library, GEM (Graph Embedding Methods), which provides a unified interface to train many state-of-the-art embedding techniques on Zachary's Karate graph and test them on network compression, visualization, clustering, link prediction, and node classification. GEM modular implementation should help users to introduce new datasets. This library is bound to the embedding methods provided by the authors, while the introduction of new embedding approaches requires compliance with an interface defined within the library. It focuses more on the implementation of embedding approaches than on the effective evaluation workflow.

Bonner et al. [36] provide a framework to assess the effectiveness of graph embeddings approaches in capturing detailed topological structure, mainly at the vertex level. For instance, they hypothesise that a good graph embedding should be able to preserve the vertex centrality. The evaluation is based on empirical and synthetic datasets. Also, this task needs to be aware of the graph used during the training phase of the model to verify the presence of topological structure in the vectors. The authors do not state if further tasks can be added.

Rulinda et al. [278] implement a collection of graph embedding techniques and, once trained, they evaluate the resulting vectors on clustering, link prediction, and visualization. The framework focuses only on uniform graphs.

Even if OpenNE[1] is an open-source package to train and test graph embedding techniques on node classification and network visualisation. The framework has been used by different embedding algorithms, however, it is more focused on the generation phase than on the evaluation aspect.

EvalNE [199] focuses on the Link Prediction task. It starts from an incomplete training graph along with a (more) complete version of the graph to test and verify the prediction power. EvalNE interprets the link prediction task as a binary classification task and it can be extended by adding other binary classifiers.

AYNEC [16] focuses on the link prediction task. It provides some incomplete graphs as a training set. Users should train a graph embedding algorithm on these datasets and run the link prediction task on the testing datasets. AYNEC takes as input the forecast edges and evaluates them by considering the complete graph. It provides all the useful phases to evaluate, not perform, the link prediction task.

Bogumil et al. [158] focus on the clustering task and they define a divergence score that can be used to distinguish good and bad embeddings. They test a pool of embeddings of synthetic and real datasets. From their work, it appears that they plan to extend the framework to hypergraphs. They do not state how or whether the framework can be used and extended for other tasks.

13.2 TOOLKIT: GEVAL

`GEval`[252] is a software framework to perform evaluation and comparison of graph embedding techniques. It takes as input a

[1] OpenNE: https://github.com/thunlp/OpenNE

file containing pre-computed vectors. More in detail, the input file must provide pairs of an embedded node (represented by its IRI) and the related vector. For each task, ground truth is modeled as a gold standard, which will be further referred to as *gold standard datasets*. They contain the tested entities and their ground truth. Fig. 13.1 is a diagrammatic representation of the involved parts in the framework and their interactions. The starting point is the *Evaluation Manager* which is the orchestrator of the whole evaluation and it is in charge of 1) verifying the correctness of the parameters set by the user, 2) instantiating the correct data manager according to the data format provided by the user, 3) determining which task(s) the user asked for, and 4) managing the storage of the results.

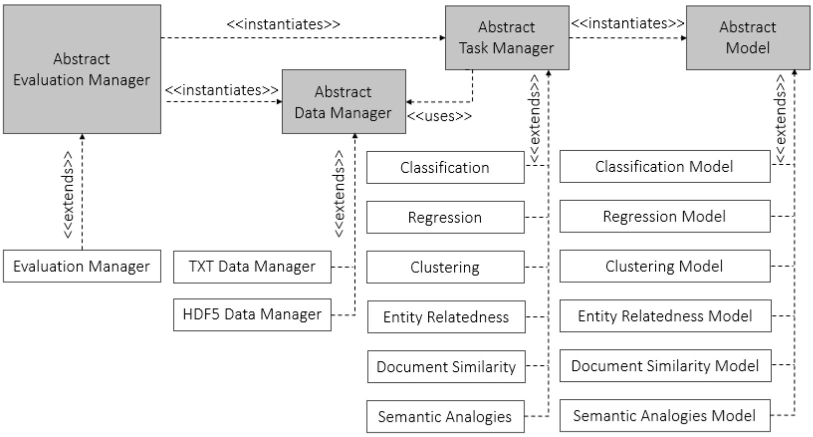


Figure 13.1: GEval architecture

13.2.1 *Running details*

GEval can be run from the command line and by APIs. As stressed before, most of the actions performed by the evaluator strictly depend on the user settings and preferences. Users can customise the evaluation settings by: i) specifying them on the command line (useful when only a few settings must be specified and the user desires to use the default value for most of the parameters); ii) organising them in an XML file (especially useful when there is the need to define most of the parameters);

iii) passing them to a function that starts the evaluation. In the `example` folder of the project on GitHub, there are examples of the different ways to provide the following parameters:

VECTORS_FILE path of the file where the embedded vectors are stored;

VECTOR_FILE_FORMAT data format of the input file;

VECTORS_SIZE length of the embedded vectors;

TASKS list of the tasks to execute;

PARALLEL task execution mode;

DEBUGGING_MODE `True` to run the tasks by reporting all the information collected during the run, `False` otherwise;

SIMILARITY_METRIC metric used to compute the distance among vectors. When an embedding technique is created, there is often also a specific distance metric which makes sense to measure similarity in the created space. This measure is a proxy for the similarity between the entities in the graph;

ANALOGY_FUNCTION function to compute the analogy among vectors. By specifying `None`, the default function is used. To customize it, the programmatically provided function handler must take 3 parameters and return a number.

```
def default_analogy_function(a, b, c){return b - a + c}
```

TOP_K it is used to look for the *right* answer among the top_k values. The vector returned by the analogy function (that will be referred to as *predicted vector*) gets compared with the *k* most similar ones. If the predicted vector is among the *k* most similar ones, the answer is considered correct;

COMPARE_WITH list of the runs to compare the results with. Each run is identified uniquely and the user can refer to specific runs to compare with by using these IDs. It is auto-generated by the framework and it corresponds to *vectorFilename_vectorSize_similarityMetric_topK* and a progressive number to disambiguate runs with the same parameters.

Table 13.2: `GEval` parameters and configuration options

Parameter	Default	Options	Mandatory	Used_by
vectors_file			✓	*
vector_file_format	TXT	TXT, HDF5		data_manager
vectors_size	200	numeric value		data_manager
tasks	_all	Class, Reg, Clu, EntRel, DocSim, SemAn		evaluation_manager
parallel	False	boolean		evaluation_manager
debugging_mode	False	boolean		*
similarity_metric	cosine	Sklearn affinity metrics[2]		Clu, DocSim
analogy_function	None	handler to function		semantic_analogy
top_k	2	numeric value		SemAn
compare_with	_all	list of run IDs		evaluation_manager

Table 13.2 details for each parameter the default value, the accepted options, if the parameter is mandatory, and which component uses it.

13.2.2 *Data management*

The input file can be provided either as a plain text (also called TXT) file or as an HDF5. In particular, the TXT file must be a white-space separated value file with a line for each embedded entity. Each row must contain the IRI of the embedded entity and its vector representation. Since most of the tasks implemented in the evaluation framework need to intersect (inner join) the data set(s) used as a gold standard and the input file, we also work with an indexed file format to speed up the merging phase. Indeed, the direct access to the entities of interest gives us the chance to save time during the merging step and also to save space since we do not read the complete vectors file into the memory. Among the available formats, we decided to work with HDF5[3]. The HDF5 vectors file must provide one group called *vectors*. In this group, there must be a dataset for each entity with the base32 encoding of the entity name as the dataset name and the embedded vector as its value. Depending on the file format, the data manager decides to read the whole content or not. For

[3] HDF5: https://www.hdfgroup.org/solutions/hdf5/

instance, the TXT file will be completely read. HDF5, instead, provides immediate access to vectors of interest.

Each data manager has to i) read the gold standard datasets, ii) read the input file, and iii) determine how to merge each gold standard dataset and the input file. The behaviour of the data manager is modelled by the `abstract data manager`, implemented by a concrete data manager based on the input file format, and refined by task data managers.

13.2.3 *Task management*

Once data have been accessed, the task(s) can be run. Each task is modelled as a pair of task manager and model. The task manager is in charge of 1) merging the input file and each gold standard file (if more than one is provided) (by exploiting the data manager), 2) instantiating and training a model for each configuration to test, and 3) collecting and storing results computed by the model. Therefore, the framework is in charge of retrieving entities of interest, i.e., entities listed in gold standard datasets, and the related vectors. Only the intersection of entities provided by the input file and the ones required by gold standard datasets will be considered in the evaluation. Each task can decide if the missing entities (i.e., the entities required in the gold standard file, but absent into the input file) will affect the final result of the task or not. According to the user preferences, tasks can be run in sequential or in parallel. The parallelization is trivially handled: by asking for the parallel execution, a new process is created for each task and it is immediately run. Once results are returned, they are collected and stored by the Evaluation Manager.

The available tasks are Classification, Regression, and Clustering that belong to the ML field, and Entity Relatedness, Document Similarity, and Semantic Analogies, more related to the semantic field. Each task is implemented as a concrete task manager that implements functionalities modelled by the `Abstract Task Manager`. Each task follows the same workflow:

1. the task manager asks data manager to merge each gold standard dataset and the input file and keeps track of both the retrieved vectors and the *missing* entities, i.e., entities

required by the gold standard dataset, but absent in the input file;

2. a model for each configuration is instantiated and trained;
3. the missing entities are managed: it is up to the task to decide if they should affect the final result or they can be simply ignored;
4. the scores are calculated and stored.

We will separately analyse each task, by detailing the gold standard datasets, the configuration of the model(s), and the computed evaluation metrics.

13.2.3.1 *Classification*

Table 13.3 contains details related to the gold standard datasets used for the Classification task, the trained models and their parameter(s) (if any), and the evaluated metric. The gold standard datasets have been designed for use in quantitative performance testing and systematic comparisons of approaches. They can be freely downloaded from the author's website[4]. The missing entities are simply ignored. The results are calculated using stratified 10-fold cross-validation.

Table 13.3: Details of the Classification task.

	Dataset	**Semantic of classes**	**Classes**	**Size**	**Source**
INPUT	Cities	Living style	3	212	Mercer
	AAUP	Salary of professors	3	960	JSE
	Forbes	Agency income	3	1,585	Forbes
	Albums	Album popularity	2	1,600	Metacritic
	Movies	Movie popularity	2	2,000	Metacritic
MODEL	**Model**		**Conf**		
	Naive Bayes (NB)		-		
	C4.5 decision tree		-		
	k-NN		k=3		
	SVM		$C \in \{10^{-3}, 10^{-2}, 0.1, 1, 10, 10^2, 10^3\}$		
OUTPUT	**Metric**		**Range**		**Optimum**
	Accuracy		[0,1]		Highest

[4] Gold standards for the Classification and Regression tasks: http://data.dws.informatik.uni-mannheim.de/rmlod/LOD_ML_Datasets/data/datasets/

13.2.3.2 *Regression*

Table 13.4 contains details related to the gold standard datasets used for the Regression task, the trained models and its parameter(s) (if any), and the evaluated metric. The gold standard datasets used for the Regression tasks are the same used for the Classification task. The missing entities are simply ignored. The results are calculated using stratified 10-fold cross-validation.

Table 13.4: Details of the Regression task.

	Dataset	**Semantic of values**	**Size**	**Source**
INPUT	Cities	Living style	212	Mercer
	AAUP	Salary of professors	960	JSE
	Forbes	Agency income	1,585	Forbes
	Albums	Album popularity	1,600	Metacritic
	Movies	Movie popularity	2,000	Metacritic
MODEL	**Model**			**Conf**
	Linear Regression			-
	M5Rules			-
	k-NN			k=3
OUTPUT	**Metric**		**Range**	**Optimum**
	Root Mean Squared Error (RMSE)		[0,1]	Lowest

13.2.3.3 *Clustering*

Table 13.5 contains details related to the gold standard datasets used for the Clustering task, the trained models and its parameter(s) (if any), and the evaluated metrics. The gold standard datasets encompass different domains:

- the *Cities, Metacritic Movies, Metacritic Albums, AAUP* and *Forbes datasets* are the datasets already used for the Classification and Regression task, here used as a single dataset. Since these datasets contain resources belonging to distinct classes (City, Music Album, Movie, University, and Company), the goal of the clustering approach on this dataset is to verify the ability to distinguish elements belonging to completely different classes. Therefore, the entities from each set are considered a member of the same cluster;

Table 13.5: Details of the Clustering task.

	Dataset	Interpretation of clusters	Clusters	Size
INPUT	Teams	{Football T., Basketball T.}	2	4,206
	Cities and Countries	{Cities, Countries}	2	4,344
	Cities, Albums, Movies, AAUP, Forbes	{Cities, Albums, Movies, Universities, Societies}	5	6,357
	Cities and Countries	{Cities, Countries}	2	11,182
MODEL	**Model**		**Conf**	
	Agglomerative Clu.		*similarity_metric*	
	Ward Hierarchical Clu.		*similarity_metric*	
	DBscan		*similarity_metric*	
	k-Means		-	
OUTPUT	**Metric**		**Range**	**Optimum**
	adjusted rand score		[-1,1]	Highest
	adjusted mutual info score		[0,1]	Highest
	Fowlkes Mallow index		[0,1]	Highest
	v_measure score		[0,1]	Highest
	homogeneity score		[0,1]	Highest
	completeness score		[0,1]	Highest

- *Cities* and *Countries* are retrieved by SPARQL queries over DBpedia, asking for all `dbo:City`[5] and `dbo:PopulatedPlace`, respectively.

- the small version of the dataset *Cities* and *Countries* is defined as before, but balancing the clusters by retrieving only 2,000 Cities. The balancing operation has been performed since the majority of clustering approaches (k-means is an example in this direction) attempt to balance the size of the clusters while minimising the interaction between dissimilar nodes [340]. Therefore, unbalanced clusters could strongly affect the final results.

- *Football* and *Basketball teams* are retrieved by SPARQL queries run against the DBpedia SPARQL endpoint, asking for all `dbo:SportsTeam` whose identifier contains respectively `football_team` or `basketball_team`.

All the models but k-Means allow to customize the distance function. Therefore, we exploit the user-defined similarity_metric given in input by the user.

[5] `dbo` is the prefix of `http://dbpedia.org/ontology/`

For each missing entity, a *singleton cluster* is created, i.e., a cluster that contains only the current entity. Further, soft clustering approaches, such as DBscan, do not cluster all entities. We call them *miss-clustered entities* and they are managed as the missing entities, i.e., we create a singleton cluster for each of them. The evaluation metrics are applied to the combination of the clusters returned by the clustering algorithm and all the *singleton clusters*.

13.2.3.4 *Entity Relatedness*

In the entity relatedness task, we assume that two entities are related if they often appear in the same context [271]. The goal of this task is to check if embedded vectors are able to preserve the semantic relatedness which can be detected from the original entities. The relatedness between vectors is brought back to the computation of the similarity metric among them.

Table 13.6: Details of the Entity Relatedness task.

	Dataset	**Structure**	**Size**
INPUT	KORE [143]	*main entity* with a *sorted list of 20 related entities*	420
	Model		**Conf**
MODEL	sim_scores = [ ] for each main entity as *me*: for each related entity as *re*: sim_scores.add(similarity(*me*, *re*)) sort(sim_scores) //from more to less similar		*similarity_metric*
	Metric	**Range**	**Interpretation**
OUTPUT	Kendall's tau correlation coefficient	[-1,1]	Extreme values: correlation Values close to 0: no correlation

Table 13.6 contains details related to the gold standard dataset used for the Entity Relatedness task, the model, and the evaluated metrics. The original version of the gold standard dataset KORE [143] consists of 420 pairs of words: for each of 21 main words, there are 20 words whose relatedness has been manually assessed. The dataset has been adapted by manually resolving each word as DBpedia entities. The main entities belong to four

distinct categories: Actors, Companies, TV series, and Videogames. Missing entities are managed as follows:

- if the main entity is missing, it is simply ignored;
- if one or more related entities attached to the same main entity are missing, first, the task computes the similarity among the available entities as reported in the model described in the Table 13.6; then, all the missing related entities are randomly put in the tail of the sorted list, and, finally, the evaluation metric is calculated on the ranking obtained by the similarity score among all the available pairs concatenated with the missing entities.

13.2.3.5 *Document similarity*

Table 13.7 contains details related to the gold standard datasets used for the Document Similarity task and the evaluated metric. The original dataset used as the gold standard is the LP50 data set [176], a collection of 50 news articles from the Australian Broadcasting Corporation. It was pairwise annotated manually by 8 to 12 different university students who evaluated the similarity among documents assigning to each pair a point in the range [1,5] where 5 means maximum similarity. To create the gold standard dataset, we worked as follows. For each pair of documents, the average of the manually assessed rates is computed. Then, we the extract the entities from the documents using the annotator xLisa[6]. The algorithm takes two documents d_1 and d_2 as its input and calculates their similarity as follows:

- For each document, the related set of entities is retrieved. The output of this step are the sets E_1 and E_2, respectively.
- For each pair of entities (i.e., for the cross product of the sets), the similarity score is computed.
- Only the maximum value is preserved for determining the document similarity evaluation. Therefore, for each entity in E_1 the maximum similarity to an entity in E_2 is kept and vice versa.

[6] Annotator xLisa: http://km.aifb.kit.edu/sites/xlisa

Table 13.7: Details of the Document Similarity task.

INPUT	**Dataset**	**Structure**	**Size**
	LP50 [176]	doc1 doc2 avg	50 docs
MODEL	**Model**		**Conf**
	it is described into the Doc. Sim. section		*similarity_metric*
OUTPUT	**Metric**	**Range**	**Interpretation**
	Pearson correlation (P_cor)	[-1,1]	Extreme: correlation Close to 0: no correlation
	Spearman correlation (S_cor)	[-1,1]	Extreme: correlation Close to 0: no correlation
	Harmonic mean of P_cor and S_cor	[-1,1]	Extreme: correlation Close to 0: no correlation

- The similarity score between the two documents is calculated by averaging the sum of all the maximum similarities.

The annotators also provided weights. Hence, the previous procedure is repeated by considering the weights to normalise the distances. The Document Similarity task simply ignores any missing entities and computes the similarity only on entities that both occur in the gold standard dataset and the input file.

13.2.3.6 *Semantic Analogies*

The Semantic Analogies task is based on quadruplets of words ($word_1, word_2, word_3, word_4$) and it checks whether it is possible to predict the last word based on the first three ones, given that the same analogy exists between $word_1$ and $word_2$ as between $word_3$ and $word_4$. A practical example [210] is the quadruplet (king, queen, man, woman). Then, one can compute X=vector(“queen")-vector(“woman")+vector(“man") and check if X is near to the embedding of “king". In Word2Vec both syntactic and semantic analogies are considered. However, in our evaluation framework, we consider only semantic analogies as KGs do generally not provide conjugated verbs, female and male nouns, singular and plural words, which are required information to perform the syntactic analogy evaluation. The original datasets used as gold standard can be freely be downloaded[7]. To create the gold standard datasets for the Semantic Analogies task, all the words have been manually substituted with DBpedia entities.

Table 13.8: Details of the Semantic Analogies task.

	Dataset	Structure	Size	Source
INPUT	Capitals and countries	ca1 co1 ca2 co2	505	Word2Vec [210]
	Currency (and Countries)	cu1 co1 cu2 co2	866	Word2Vec [210]
	Cities and State	ci1 st1 ci2 st2	2,467	Word2Vec [210]
	(All) capitals and countries	ca1 co1 ca2 co2	4,523	Word2Vec [210]
MODEL	**Model**			**Conf**
	it is described into the Sem. An. section			*analogy_function*
OUTPUT	**Metric**		**Range**	**Optimum**
	accuracy		[0,1]	Highest

The task takes the quadruplets (v_1, v_2, v_3, v_4) and works on the first three vectors to predict the fourth one. Among all the vectors, the nearest to the predicted one is retrieved, where the *nearest* is computed by the dot product. The analogy function to compute the predicted vector can be customised. The vector returned by the function (the *predicted vector*) gets compared with the *top_k* most similar ones. If the actual fourth vector is among the *top_k* most similar ones, the answer is considered correct. *top_k* can be customised by the user.

13.2.4 *Extension points*

Each task is kept separate, by satisfying the modularity requirement. By the usage of abstraction, it is easy to add new tasks and/or data managers. The `abstract data manager` defines the interface of a data manager, while `abstract task manager` and `abstract model` define the interface of a new task. Extending the framework with new data formats and/or new tasks is as simple as creating a class implementing these interfaces.

To further enrich an already implemented task, it is easy to retrieve the exact point to modify since each task is limited to its task manager and model.

[7] Gold standards for the Semantic Analogies task: https://sites.google.com/site/semeval2012task2/download

To extend the evaluation also to edges, it is enough to create a gold standard dataset containing edges and related ground truth.

Our default gold standard datasets contain DBpedia entities. However, this is not a framework requirement; it is possible to evaluate different sets of entities (and embeddings of other KGs) by adding gold standard datasets.

13.2.5 *Results storage*

For each task and each file used as a gold standard, `GEval` will create i) an output file that contains a reference to the file used as a gold standard and all the information related to evaluation metric(s) provided by each task, ii) a file containing all the *missing* entities, iii) a log file reporting extra information, occurred problems, and execution time and iv) information related to the comparison with previous runs. It reports the values effectively considered in the comparison and the ranking of the current run upon the other ones. The results of each run are stored in the directory `results/result_<starting time of the execution>` generated by the evaluation manager in the local path.

13.3 EVALUATION ON THE PERFORMANCE OF `geval`

This section reports the evaluation of the performance of `GEval` on embeddings generated by RDF2Vec using the uniform weighting technique, both in-text[8] and HDF5 [9] format. The experiments are performed on a system with an Intel(R) Xeon(R) E5-2640 v4 CPU at 2.40 GHz and 256 GB RAM. Table 13.9 reports i) the running times of each task run separately, ii) the whole task set run in sequential and iii) in parallel. The sequential time does not match the sum of individual tasks because in the sequential time the input file is read only once, while performing each task separately the reading step is repeated by each run. The times are related to a single run. However, the Classification and Regression tasks perform a 10-fold cross validation. As future

[8] RDF2Vec emdeddings [TXT]: `https://doi.org/10.5281/zenodo.1318146`

[9] RDF2Vec emdeddings [HDF5]: `https://doi.org/10.5281/zenodo.2017356`

direction, we also aim to verify its correctness by comparing our results with those returned by other tools (e.g. GEM).

Table 13.9: Execution time evaluation of GEval

File format	*Classification & Regression*	*Clustering*	*Semantic Analogies*	*Document Similarity*	*Entity Relatedness*	*Reading time*	*Sequential time*	*Parallel time*
txt	36:26	9:16	6:30	22:22	3:15:54	5:39	4:07:42	3:20:26
HDF5	25:31	7:04	0:10	0:05	3:25:04	33:49	3:56:44	3:26:00

IMPACT OF VECTOR SIZE ON EXECUTION TIME To estimate how the vector size affects the computational time, the remaining evaluation reports the impact of vector size on the execution time for the Classification and Regression tasks. The experiments are performed on a system with an Intel(R) Core(TM) i7-8700T CPU at 2.40GHz and 16 GB RAM. We evaluated vectors produced by RDF2Vec [63] and by KGloVe [62]. Here, we report only results related to KGloVe since in both cases we observed the same trend. We extrapolated only vectors required by the Classification and Regression tasks, because of memory limitations. We crop the filtered vectors by considering 10, 20, 50, 100, 150, 180, 200 as size. We perform the Classification and the Regression tasks on all the obtained vectors.

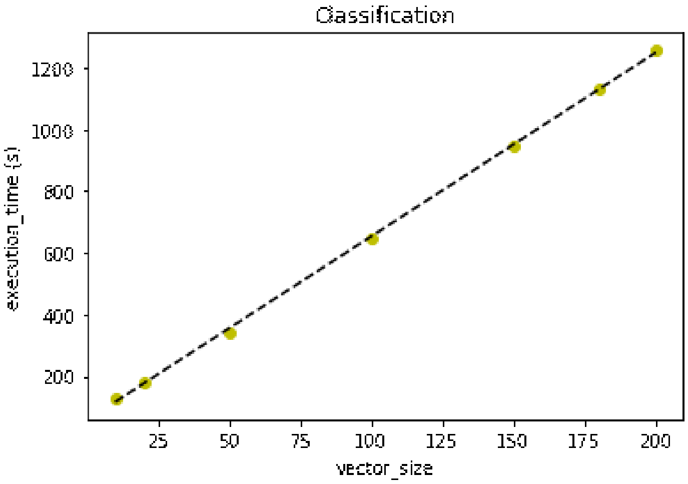


Figure 13.2: Impact of vector size on the GEval execution time in the Classification task

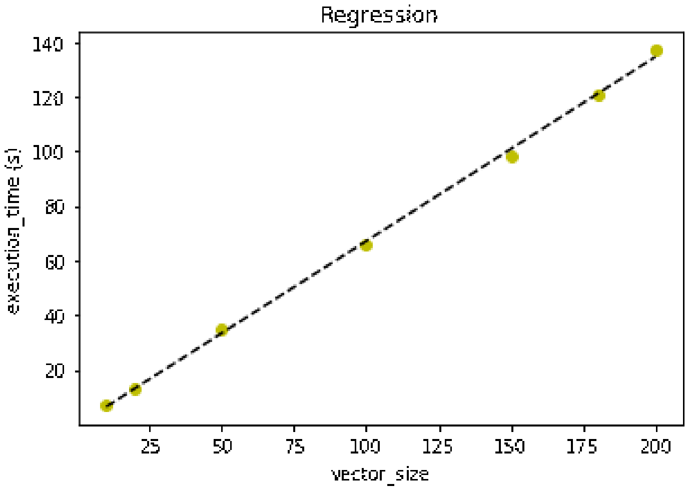


Figure 13.3: Impact of vector size on the `GEval` execution time in the Regression task

Fig. 13.2 and Fig. 13.3 report the execution times of the classification and the regression tasks, and it is evident that the execution time of Classification and Regression tasks is linearly correlated with the vector size.

13.4 USE CASE: PARAMETER TUNING

This use case focuses on parameter tuning and we will use results produced by ML tasks to detect the best combination of hyper-parameters. This evaluation considers a modified version of KGloVe [61] where the difference with the original algorithm lies in the parallel implementation (GPU based) of the underlying GloVe [9]. Our goal is to optimize KGloVe parameters to find out the values that produce vectors which gain the best results in ML tasks. In Fig. 13.4, the entire pipeline is visible. Starting from DBpedia 2016, the *graph walks* produces a co-occurrence matrix for the nodes of the graph [61]. The parameters that affect the co-occurrence matrix are α, ϵ, and the *weighting function* which is applied once on the graph (forward weighting function) and once on the graph with reversed edges (backward weighting function). The *Parallel GloVe* [9] implementation takes the co-occurrences matrix as input and trains the vectors in parallel by minimising

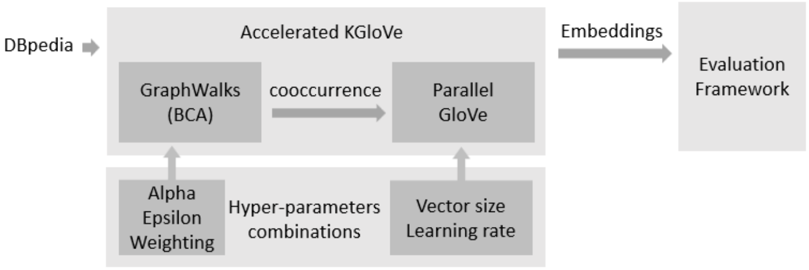


Figure 13.4: Pipeline of hyper-parameter tuning

the loss function defined by GloVe [261]. The produced embeddings are affected by GloVe parameters, i.e., the vector size and the learning rate. To reduce the employed resources in finding the optimum parameters combination, we opt for a random search. We performed the evaluation by considering a set of 105 uniformly random generated combinations: we tested

$\alpha \in \{0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95\}$;

learning rate$=0.01$;

vector size$=50$; $\epsilon \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$;

weighting functions (forward, backward)
$|\text{weighers}| \times |\text{weighers}| = 12 \text{ options } \times 12 \text{ options}$

Once produced vectors, we evaluate them on the Classification and Regression tasks implemented by `GEval`. `GEval` runs 10 times both the Classification and the Regression task and returns the average result. By considering the combination of models and their configurations (see Table 13.3), the classification task produces 10 accuracy scores, while the regression task produces 2 RMSE scores (k-NN and LR). For each run, we take the average of the results produced by the 5 datasets used as a gold standard.

Then, we rank the runs according to the 12 different scores. The average rank is taken for evaluating the corresponding parameter combination. To find out the performance according to a given parameter y, we plot the performance for each run of y (if there is a value of y which is used multiple times, we compute the average). In Figs. 13.5 and 13.6 the ranked values of α and ϵ are presented. We observe that $\alpha = 0.7$ and $\epsilon = 10^{-5}$ produce the best embeddings for ML tasks.

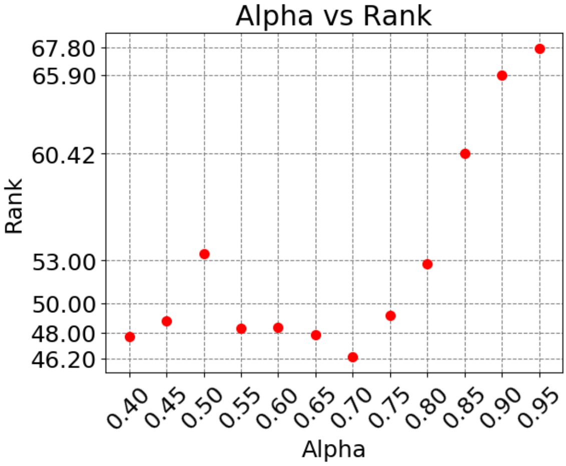


Figure 13.5: Parameter tuning of KGloVe, alpha parameter

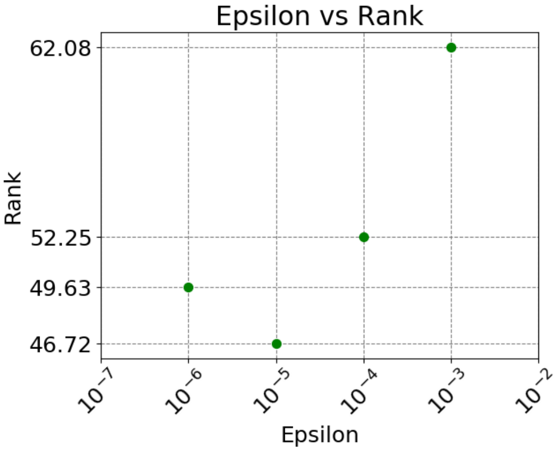


Figure 13.6: Parameter tuning of KGloVe, epsilon parameter

13.5 FINAL REMARKS

`GEval` simplifies the evaluation phase of KG embedding techniques providing tasks ranging from ML to semantic ones. To the best of our knowledge, our proposal is one of the most comprehensive frameworks to evaluate KG embedding techniques for heterogeneous graphs. GEval can be used in evaluation and comparison over multiple tasks. Moreover, it can be also used in parameter tuning, as shown in the presented use case. The modularity of GEval is achieved by keeping each task separated but still abstracting away the commonalities.

Our software framework can be used to perform benchmarks, but it is not designed as a benchmark itself. We provide the framework as a command-line tool and by APIs[10]. We do not provide server-side execution, since the computation of tasks and the memory requirements can be onerous and can not be determined apriori. In our opinion, it is more beneficial to provide the software and give the opportunity of choosing the hardware requirements adapt to the size of the managed vectors. GEval is not bounded to evaluate only node embeddings. By incorporating also edges into the gold standard datasets, it is possible to consider graph embeddings that embed both nodes and edges. Our default gold standard datasets contain DBpedia entities. However, this is not a framework requirement; it is possible to evaluate different sets of entities (and embeddings of other KGs) by adding gold standard datasets.

The framework has been published with an open-source licence to be used by the whole community. GEval is already of interest for experimentation with graph embedding techniques by the authors' institutes (Fraunhofer FIT, the RWTH Aachen University, the University of Salerno, and IBM research). Moreover, other institutes show an interest in collaborating on this project. The Télécom ParisTech is interested in extending the already available tasks to incorporate gold standard datasets related to (French) museums. We are certain that also others will benefit from this valuable resource.

[10] GEval APIs: https://pypi.org/project/evaluation-framework

Part V

CONCLUSIONS AND FUTURE DIRECTIONS

14 CONCLUSIONS

Finally, in conclusion, let me say just this.

– Peter Sellers

This section remarks on the main outcomes described in this dissertation in terms of contributions along with the peer-reviewed publications supporting findings and assertions, points out the impact of the proposed approaches and implemented toolkits, and concludes with future directions.

14.1 REMARKS

The purpose of the contributions presented in this dissertation is to support users in the entire process of data publication and exploitation. My contributions mainly focus on i) *OD quality and privacy assessment and improvement* aiming to propose approaches and toolkit to support data curators in assessing and improving data quality and privacy concerns by focusing on OD released in a tabular format; ii) *KG exploitation* aiming to propose approaches and prototype tools to express users' needs or explore available data by an NL interface to guide end-users with different interests, types of background, age, and needs to query KGs and take advantage of them without requiring technical skills in query languages. End-users with technical background should not left behind. They might be interested in performing data analysis, in general, and ML tasks on raw data. Hence, they need to perform task-oriented evaluation and comparison to detect the best graph embedding technique according to the desired end usage. Hence, we focus on iii) the analysis and evaluation of the compatibility between *KGs and ML* where KGs are graph shaped by nature and ML algorithms expect data in a vector form. It requires the definition and a systematic evaluation and comparison of graph embedding techniques.

14.1.1 *Open Data quality and privacy assessment and improvement*

Data quality gains the interest of both data producers and consumers in any application domain. In this dissertation, we mainly focus on a RPA, the Campania Region, as a data cleansing target user. It is not a loss of generality as the proposed approaches are general enough to be used by any data curator and user.

SPOD AS THE INTERNAL PLATFORM OF THE CAMPANIA REGION ADMINISTRATION. The partnership with the Campania Region resulted in the adaptation of SPOD as an internal platform for our RPA. In the article entitled *Government as a platform in a regional public administration* under evaluation at Transforming Government: People, Process and Policy, we retrace political and technological changes applied in our RPA to break down its data siloed structure by revising the data production process [15].

SPOD results from an H2020 project, ROUTE-TO-PA, coordinated by my research group, while its modification to be compliant with our RPA requests is referred to as *Campania Crea*. During this partnership, we assisted in a strategic flagship project in which the Campania Region invested in 2018 and 2019 to completely transform both its organisation and technological support. By focusing on the technical and technological changes, our RPA transformed the internal workflow to produce and publish OD and adopted Campania Crea to support RPA members in (co-)creating, refining, publishing, and exploiting OD. First, we collected requirements and needs by discussing with RPA delegates to adapt and extend SPOD according to their needs. Finally, we assessed the acceptance rate of Campania Crea by RPA members measured by the TAM questionnaire. Results are rather positive, highlighting that the Campania Region succeeded in involving RPA members in this revolutionary plan and they positively accept a social platform to collaboratively create OD by enabling the multi-disciplinary and multi-departmental G2G collaboration. It represents a step forward in breaking data siloes.

DATA QUALITY ASSESSMENT. SPOD is provided with a role-based orchestration approach to ensure data quality while dealing with large groups where roles guarantee syntactic and semantic data quality. The article entitled *Orchestrated co-creation of high-quality open data within large groups* presented in the Electronic Government (EGOV) conference in 2019 reports a motivational use case that guided the defined roles and how the co-creation process of OD has been revised to guarantee the orchestration of large groups of collaborators [109].

Furthermore, we also considered the scenario that any stakeholder may be interested in assessing the quality of the dataset under definition and spot any privacy concerns during the dataset population phase. The proposed quality and privacy assessment approach presented at the international conference on Digital Government Research (Dg.O.) 2019 [108] relies on a type inference step where for each value, and transitively for each column, the corresponding datatype is inferred. Then, *accuracy* and *completeness* are assessed while privacy concerns are detected both at an instance level if any value exposes structured sensitive information, such as an SSN or an IBAN in a textual description and, at the schema level, checking if any combination of columns exposes personal information. The proposed type inference approach is implemented by a decision-tree model that succeeds in selecting only a subset of datatypes to test achieving better performance in terms of scalability if compared with a brutal type inference approach that tests all the supported datatype, without losing in accuracy as demonstrated by the evaluation of real OD released by the Campania region.

DATA QUALITY IMPROVEMENT. Once identified quality issues, data curators should be guided in performing data quality improvement. As open datasets, such as the ones published by the Campania region, contain highly inaccurate textual geographical data where the accuracy is compromised by the use of abbreviations, inconsistent representations, misspellings, we designed a (semi) automatic approach to assess and improve textual geographical data accuracy. As a result, we proposed a clustering-based approach for assessing and improving the quality of textual geographical data corresponding to provinces and municipalities, at the instance level. Both the proposed approach and its evaluation have been presented at the ACM IKDD CODS-COMAD conference in 2021 [255]. While the assessment phase detects inaccurate values, such as typos, abbreviations, misspellings, or any other syntactical errors, the improvement phase proposes a correction for any detected error. We empirically verified that the combination of an approximate string matching and the well-known Levensthein improves (or not worsens) the results of using Levenshtein in isolation. Since the clustering-based approach returns datasets containing fewer errors if compared with a pair-wise comparison algorithm between each value in input with a dictionary of correct values to detect the corresponding entity, clustering helps in detecting and correcting inaccurate values in textual geographical data. Finally, our proposal obtains extremely better results than OpenRefine, a popular tool used by data curators to refine data quality, in terms of accuracy thanks to the exploitation of a dictionary of correct values.

DATA PRIVACY-PRESERVING. Data owners are spur in opening up their data to enable informed decision making, ensure transparency, audience engagement, and release social and commercial value. Unfortunately, data in their raw and original form could contain personal and sensitive information about individuals. Thus, data curators should perform Privacy-preserving data publishing to publish useful data without violating individuals' privacy. It requires detecting identifiers and quasi-identifiers and applying corrective actions. The proposed approach is based on a privacy issues detection step followed by an anonymity approach based on generalisation and suppression and it has been presented at the EGOV conference in 2020 [249]. The workflow starts from a user-defined dataset and automatically returns the best QID, defined as the *minimum* number of columns/attributes that leads to the disclosure of the *highest* number of singletons. If the best QID matches (year, municipality, gender), the proposed approach also provides the corresponding generalisations. The proposed anonymisation approach can be defined as a modified version of *k*-Anonymity where *k* is at least equal to 2; suppression is discouraged in favour of generalisation, and changes are applied to work locally. Tests on real datasets concerning driver licenses released by the Italian Ministry of Infrastructure and Transport demonstrate that our proposal achieves the same results of widely adopted k-anonymity in terms of privacy-preserving, while it obtains better data quality by its local recording.

14.1.2 *Knowledge Graphs exploitation*

A traditional KM process requires i) retrieving data, ii) refining them, and iii) performing data exploitation. As a data retrieval interface, we propose query builders enhanced with (controlled) NL interfaces to guide users in naturally posing questions by simulating, as much as possible, human interactions. If users have a clear objective, they can directly type or pronounce their requests. In this dissertation, we mainly focused on OD experts and PA, education, and the CH community as target groups. The proposed approach and the resulting prototypes have been presented as a Doctoral Consortium at CHItaly 2021 [251]. Vice versa, in exploratory search, users are guided in iteratively creating and refining questions. As a query builder, NL queries can be translated to SPARQL to be run over a SPARQL endpoint that is a standard way to expose KG content. Among SPARQL constructs, SELECT query results can be naturally represented as tabular data. Thus, retrieved data are modelled as tables, which can be manually or automatically refined, and finally, used in data exploitation mechanisms. It may result in textual replies or concrete artifacts, perhaps customizable and exportable, such as data visualisations, data stories, or VR-based data representations.

KNOWLEDGE GRAPHS AND DATA VISUALISATION. OD experts are usually aware of data refinement and exploitation approaches, considered their comfort zone, while they might be unaware of KG query languages. We proposed a *transitional approach* where OD experts are guided from LOD querying to their comfort zone. It resulted in `QueDI` whose interaction model and interface have been presented in Semantics 2020 [94]. `QueDI` allows users to build queries step-by-step with an auto-complete mechanism and to exploit retrieved results by exportable and dynamic visualizations. `QueDI` scaffolds users in, first, creating a tabular representation of the dataset of interest by `ELODIE`, a query builder enhanced with a controlled NL interface. `ELODIE` realises an exploratory search by organising available data in facets and supporting users in automatically retrieving both user query results and data to go on with the query formulation by querying a configured SPARQL endpoint. Second, QueDI supports a manual dataset manipulation phase where users can exploit their skills in data refinement by aggregating, sorting, filtering, and cleaning data by interacting with a form-based interface that behaves as a SQL builder. Finally, it enables the creation of an exportable and reusable visualization. Besides its accuracy, expressiveness, and scalability features, reported in the article presented at Semantics 2020 [94], we assessed its usability by involving OD experts, computer scientists, and lay users. While the usability evaluation of OD experts has been presented in the international conference Computer Supported Co-operative Work in Design (CSCWD) in 2019, the comparison of the usability according to computer scientists and lay users has been presented at Semantics 2020 [93, 94]. All the groups considered QueDI of a *good* usability level demonstrating that it results in a non-intimidating approach to exploit KGs without requiring any technical skill in query languages.

We also proposed QueDI as a KM tool in the educational context to support future citizens in going beyond the passive inspection of results returned by a search engine, and in actively searching for the data that best answer their questions. The potentialities of using QueDI at school have been presented in the Technology-enhanced learning for future children (TEL4FC) workshop at the international conference Methodologies and Intelligent Systems for Technology Enhanced Learning (MIS4TEL) 2020 [81].

KNOWLEDGE GRAPHS AND STORYTELLING. We investigated the implicit exploitation of KGs in retrieving synonyms in Novelette [2–4], a digital storytelling environment where storytellers can create stories to graphically represent tales, data stories, or media stories, as demonstrated in the poster paper presented at Information Visualization (IV) 2020 [2]. If users experience writer's block, Novelette supports a suggestion provision mechanism. Users can type the word of interest and Novelette will automatically retrieve synonyms by querying BabelNet and organising retrieved results in (navigable) word clouds. It represents a keyword-based interface to implicitly explore KGs by navigating synonyms. By testing Novelette and the suggestion provision mechanism in a real context at school, it resulted that Novelette engages children in inventing and authoring stories, as reported in the article entitled *Engaging children in digital storytelling* presented at the workshop TEL4FC@MIS4TEL 2021 [4]. Children also stated that it is *super-adapt* for them in terms of usability as reported in the article entitled *Novelette, a usable visual storytelling digital learning environment* accepted by the journal IEEE Education Society [3]. In the context of letting children familiarising with smart city concepts and design their smart object [76, 256, 258, 275], Novelette results in a non-intimidating environment to clarify terminology and exploit storytelling to model smart objects interactions, as reported in the long abstract presented at I-CITIES 2021 [254].

KNOWLEDGE GRAPHS AND THE CULTURAL HERITAGE COMMUNITY. Concerning the CH community, in the last year, virtual exhibitions have been widely adopted to enhance physical tours, but CH lovers still behave as visitors. We propose to take advantage of CH KGs in an authoring platform for VR-based virtual exhibitions by combining ELODIE and an automatic mechanism to create VR-based solutions. The proposed approach and a guided use case have been accepted in the Journal of CH and it is currently in press [216].

Moreover, we investigated how to make VAs compatible with KGs. It resulted in a community shared software framework (a.k.a. generator) that enables lay-users to create ready-to-use custom extensions for performing Question-answering over KGs. While a demonstration of the proposed generator has been presented at Extended Semantic Web Conference 2021 as demo paper [259], its working mechanism and evaluation have been submitted to the Semantic Web Journal in March 2021 and it is currently under evaluation [260]. This proposal represents a step forward in enabling direct search and lookup over KGs.

14.1.3 *Knowledge Graphs and Machine Learning*

While KGs are graph shaped by nature, most traditional ML algorithms expect data in a vector form. To transform graph elements to vectors, several graph embedding approaches have been proposed. Systematic comparative evaluations of different approaches are scarce; approaches are rather evaluated on a handful of often project-specific data sets. We propose a mechanism to simplify the evaluation and the comparison of graph embedding techniques. It results in `GEval`, a software framework to perform evaluation and comparison of graph embedding techniques. While a preliminary version of the framework has been presented at ESWC 2019 as poster paper [253], its extension has been presented at ESWC 2020 as a resource paper [252].

14.2 IMPACT

The proposed approaches and prototypes make **no assumption on users expertise** and competencies, limiting as much as possible the requirements in terms of skills, expected awareness of the underlying accessed data, and accessing mechanisms. As a result, the same tool might be adopted by heterogeneous communities to accomplish different and surprising tasks.

It results in **general-purpose** toolkits, that can be easily adapted to different requirements and application contexts. As a few examples, `Novelette` is designed to support pupils in authoring stories, but we demonstrated that it can be used to create also media and data stories. `QueDI` has been designed as a transitional approach for people familiar with tabular data manipulation, but the same toolkit has been also tested by computer scientists and stakeholders with heterogeneous backgrounds, resulting in a comparable result in terms of usability.

The designed and implemented toolkits have the potential to unlock the potentialities of (Linked) Open Data by letting lay users play the data consumers role. It is worth remarking that the offered data exploitation mechanisms do not return static outcomes, such as an image, but provide inexperienced users the possibility to **generate dynamic content**, such as always up-to-date data visualisations, digital stories that can be shared as live HTML components, working and ready-to-be-published VA extensions or shareable VR-based virtual exhibitions. The implemented authoring mechanisms do not require technical skills in programming or Semantic Web technologies, but technical challenges are masked to the final users making their experience as natural as possible. The proposed mechanism move lay users in the position of active consumers of available data, letting them opt for the preferred data exploitation means according to their needs and interests.

The implemented prototypes follow the best practices in software design (e.g., abstraction and modularity) to guarantee **technical quality** and make them easily and fully extensible. All code, prototypes, and evaluation of the proposed approaches are freely **available** on GitHub with an open-source license.

14.3 FUTURE DIRECTIONS

My research addresses several problems that any data curators and consumers experience during their data management processes, such as assessing data quality issues, detecting privacy leakages, improving data quality, and preserving individual privacy, easily querying and exploiting data according to users' needs and expertise. However, many directions can be explored to further simplify data production and exploitation mechanisms.

14.3.1 *(Linked) Open Data quality and privacy assessment and improvement*

DATA DISCOVERABILITY. The first difficulty on the data consumer side is to recognise data sources according to a topic of interest and choose the best one(s) in terms of accessibility, content, inner data quality. It requires support data users in easily *discovering* data and performing on-the-fly data quality assessment. In this direction, we are designing a search engine able to guide users in retrieving, evaluating, and comparing eligible data sources to start the exploratory phase. In particular, we are exploring how to simplify the KG discoverability process by providing users with a unified point of access to heterogeneous data sources, such as data aggregators or single resources.

LINKED OPEN DATA QUALITY ASSESSMENT AND IMPROVEMENT. While this dissertation mainly addresses the OD quality assessment and improvement, it may be interesting to further explore the analysis and the improvement of data quality directly of KGs, by applying machine learning or deep learning approaches. Several directions can be explored. As an example, KG quality can be assessed in terms of data *accessibility* or the easiness in querying data. It implies assessing the KG publication mechanisms verifying if a SPARQL endpoint or APIs have been exposed, verifying the status, and estimating the performances of the exposed accessing methods. Besides assessing the KG accessing methods quality in a given frame of time, it might be interesting to periodically check it and analyse its behaviour over time. Concerning the KG quality improvement, it may be interesting to address KG *incompleteness*. In this direction, I am currently in contact with professor Michael Cochez of the Free University of Amsterdam to design and implement a neural network approach to improve the quality of KGs in the QA task by both dealing with KG incompleteness and syntactical errors that can be committed by users in posing their questions.

14.3.2 *Knowledge Graph exploitation*

FROM TABULAR DATA TO KNOWLEDGE GRAPHS. While this dissertation mainly focuses on the improvement of data quality and privacy aspects in tabular data and the exploitation of LD while neglecting the phase to convert data from tabular to linked format, data curators and consumers should also be guided in this process. We are currently designing and prototyping a guided workflow to support users in defining an ontology modelling their domain of interest and, subsequently, converting available data in a materialised KG. In this context, I am collaborating with professor Calvanese at the University of Bozen to consider the possibility to propose a mechanism to guide lay users, such as the CH community, to transform tabular data in a virtual or a materialised KG.

MULTIPLE KNOWLEDGE GRAPH QUERYING. Usually, users are interested in querying more than one resource at a time while the proposed mechanisms in my dissertation are limited to a unique data source. We are now exploring approaches to combine replies from multiple sources to provide end-users with an exhausting reply to their questions.

BIBLIOGRAPHY

[1] AAT editors. *The Art and Architecture Thesaurus (AAT)*. [Online] Last access November 2020. 2017. URL: http://www.getty.edu/research/tools/vocabularies/aat/index.html.

[2] Agnese Addone, Renato De Donato, Giuseppina Palmieri, Maria Angela Pellegrino, Andrea Petta, Vittorio Scarano, and Luigi Serra. "Visual Storytelling by Novelette." In: *Information Visualisation (IV)*. IEEE. 2020, pp. 723–728.

[3] Agnese Addone, Renato De Donato, Giuseppina Palmieri, Maria Angela Pellegrino, Andrea Petta, Vittorio Scarano, and Luigi Serra. *Novelette, a Usable Visual Storytelling Digital Learning Environment*. Submitted to IEEE Education Society in Sep. 2021. 2021.

[4] Agnese Addone, Giuseppina Palmieri, and Maria Angela Pellegrino. "Engaging Children in Digital Storytelling." In: *Methodologies and Intelligent Systems for Technology Enhanced Learning, 11th International Conference. Workshops*. Springer International Publishing, 2022.

[5] Deborah Agostino, Michela Arnaboldi, and Antonio Lampis. "Italian state museums during the COVID-19 crisis: from onsite closure to online openness." In: *Museum Management and Curatorship* 35.4 (2020), pp. 362–372.

[6] Nikhat Akhtar, Nazia Tabassum, Asif Perwej, and Yusuf Perwej. "Global Journal of Engineering and Technology Advances." In: *Global Journal of Engineering and Technology Advances* 3.02 (2020), pp. 28–50.

[7] Ali Mohamed Nabil Allam and Mohamed Hassan Haggag. "The question answering systems: A survey." In: *International Journal of Research and Reviews in Information Sciences (IJRRIS)* 2.3 (2012).

[8] Miika Alonen, Tomi Kauppinen, Osma Suominen, and Eero Hyvönen. "Exploring the Linked University Data with Visualization Tools." In: *The Semantic Web: ESWC Satellite Events*. 2013, pp. 204–208.

[9] Abdulrahman Altabba. "Accelerating KGloVe Graph Embedding." unpublished thesis. 2019.

[10] Nagraj Alur, Reginald Joseph, Harshita Mehta, Jørgen Tang Nielsen, and Denis Vasconcelos. *IBM WebSphere Information Analyzer and Data Quality Assessment*. 2007. URL: http://www.redbooks.ibm.com/redbooks/pdfs/sg247508.pdf.

[11] Vito Walter Anelli, Tommaso Di Noia, Eugenio Di Sciascio, and Azzurra Ragone. "Anna: A Virtual Assistant to Interact with Puglia Digital Library." In: *27th Italian Symposium on Advanced Database Systems*. 2019.

[12] Carmelo Ardito, Maria Francesca Costabile, Maria De Marsico, Rosa Lanzilotti, Stefano Levialdi, Teresa Roselli, and Veronica Rossano. "An approach to usability evaluation of e-learning applications." In: *Universal access in the information society* 4.3 (2006), pp. 270–283.

[13] Marcelo Arenas, Bernardo Cuenca Grau, Evgeny Kharlamov, Sarunas Marciuska, Dmitriy Zheleznyakov, and Ernesto Jimenez-Ruiz. "SemFacet: semantic faceted search over yago." In: *The World Wide Web Conference WWW*. 2014, pp. 123–126.

[14] Hannah Arendt. *The recovery of the public world*. St. Martin's Press, 1979.

[15] Salvatore Avella, Angela Cocchiarella, Dario Fonzo, Giuseppina Palmieri, Maria Angela Pellegrino, and Vittorio Scarano. *Government as a Platform in a Regional Public Administration*. Submitted to Transforming Government: People, Process and Policy in July 2021.

[16] Daniel Ayala, Agustín Borrego, Inma Hernández, Carlos R. Rivero, and David Ruiz. "AYNEC: All You Need for Evaluating Completion Techniques in Knowledge Graphs." In: *The Semantic Web - 16th International Conference, ESWC*. Portorož, Slovenia, 2019, pp. 397–411.

[17] Ruth S Aylett, Sandy Louchart, Joao Dias, Ana Paiva, and Marco Vala. "FearNot!–an experiment in emergent narrative." In: *International workshop on intelligent virtual agents*. Springer. 2005, pp. 305–316.

[18] Carliss Y Baldwin, C Jason Woodard, et al. "The architecture of platforms: A unified view." In: *Platforms, markets and innovation* 32 (2009).

[19] Frank Bannister. "Dismantling the Silos: Extracting New Value from IT Investments in Public Administration." In: *Information Systems Journal* 11 (2001), pp. 65–84.

[20] Scott Bateman, Regan Mandryk, Carl Gutwin, Aaron Genest, David Mcdine, and Christopher Brooks. "Useful Junk? The effects of visual embellishment on comprehension and memorability of charts." In: *Conference on Human Factors in Computing Systems*. Vol. 4. 2010, pp. 2573–2582.

[21] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. "Methodologies for data quality assessment and improvement." In: *ACM computing surveys* 41.3 (2009), pp. 1–52.

[22] Amin Beheshti, Alireza Tabebordbar, and Boualem Benatallah. "iStory: Intelligent Storytelling with Social Data." In: *Proceedings of the Web Conference*. 2020, pp. 253–256.

[23] Mafkereseb Kassahun Bekele, Roberto Pierdicca, Emanuele Frontoni, Eva Savina Malinverni, and James Gain. "A Survey of Augmented, Virtual, and Mixed Reality for Cultural Heritage." In: *Journal on Computing and Cultural Heritage* 11.2 (2018), 7:1–7:36.

[24] Pierfrancesco Bellini, Paolo Nesi, and Alessandro Venturi. "Linked open graph: Browsing multiple SPARQL entry points to build your own LOD views." In: *Journal of Visual Languages & Computing* 25.6 (2014), pp. 703 –716.

[25] Farah Benamara. "Cooperative question answering in restricted domains: the WEBCOOP experiment." In: *Proceedings of the conference on question answering in restricted domains*. 2004, pp. 31–38.

[26] Martin Beno, Kathrin Figl, Jürgen Umbrich, and Axel Polleres. "Open Data Hopes and Fears: Determining the Barriers of Open Data." In: *Conference for E-Democracy and Open Government*. 2017, pp. 69–81.

[27] Margo E Berendsen, Jeffrey D Hamerlinck, and Gerald R Webster. "Digital story mapping to advance educational atlas design and enable student engagement." In: *ISPRS International Journal of Geo-Information* 7.3 (2018), p. 125.

[28] Tim Berners-Lee. *5-star OD*. Access 2020/02/22. 2006. URL: `http://5stardata.info`.

[29] Tim Berners-Lee, J. Hollenbach, Kanghao Lu, J. Presbrey, Eric Prud'hommeaux, and Monica M. C. Schraefel. "Tabulator Redux: Browsing and Writing Linked Data." In: *Proceedings of the WWW Workshop on Linked Data on the Web, LDOW*. Beijing, China, 2008.

[30] Laure Berti-Equille. "Learn2clean: Optimizing the sequence of tasks for web data preparation." In: *The World Wide Web Conference*. 2019, pp. 2580–2586.

[31] Nikos Bikakis, Melina Skourla, and George Papastefanatos. "rdf: SynopsViz - A Framework for Hierarchical Linked Data Visual Exploration and Analysis." In: *The Semantic Web: ESWC Satellite Events*. 2014, pp. 292–297.

[32] Silvia Blanco-Pons, Berta Carrión-Ruiz, José Luis Lerma, and Valentín Villaverde. "Design and implementation of an augmented reality application for rock art visualization in Cova dels Cavalls (Spain)." In: *Journal of Cultural Heritage* 39 (2019), pp. 177–185. DOI: `10.1016/j.culher.2019.03.014`.

[33] Victor de Boer, Jan Wielemaker, Judith van Gent, Michiel Hildebrand, Antoine Isaac, Jacco van Ossenbruggen, and Guus Schreiber. "Supporting LD Production for Cultural Heritage Institutes: The Amsterdam Museum Case Study." In: *Semantic Web*. 2012, pp. 733–747.

[34] Piero Andrea Bonatti, Stefan Decker, Axel Polleres, and Valentina Presutti. "Knowledge graphs: New directions for knowledge representation on the semantic web (dagstuhl

seminar 18371).” In: *Dagstuhl Reports*. Vol. 8. 9. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 2019.

[35] Bill Bonis, Spyros Vosinakis, Ioannis Andreou, and Themis Panayiotopoulos. “Adaptive virtual exhibitions.” In: *DESIDOC Journal of Library and Information Technology* 33.3 (2013), pp. 183–198.

[36] Stephen Bonner, John Brennan, Ibad Kureshi, Georgios Theodoropoulos, Andrew Stephen McGough, and Boguslaw Obara. “Evaluating the Quality of Graph Embeddings via Topological Feature Reconstruction.” In: *IEEE International Conference on Big Data*. 2017, pp. 2691–2700.

[37] Vladia Borissova. “Cultural heritage digitization and related intellectual property issues.” In: *Journal of Cultural Heritage* 34 (2018), pp. 145–150. DOI: https://doi.org/10.1016/j.culher.2018.04.023.

[38] Jessica R Botfield, Christy E Newman, Caroline Lenette, Kath Albury, and Anthony B Zwi. “Using digital storytelling to promote the sexual health and well-being of migrant and refugee young people: A scoping review.” In: *Health education journal* 77.7 (2018), pp. 735–748.

[39] P. Bourges-Waldegg, Luis Moreno, and Teresa Rojano Ceballos. “The role of usability on the implementation and evaluation of educational technology.” In: *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*. 2000, 7 pp.–. DOI: 10.1109/HICSS.2000.926722.

[40] Lucia Boxelaar, Mark Paine, and Ruth Beilin. “Community engagement and public administration: Of silos, overlays and technologies of government.” In: *Australian Journal of Public Administration* 65 (2006), pp. 113–126.

[41] Stefano Braghin, Aris Gkoulalas-Divanis, and Michael Wurst. *Detecting quasi-identifiers in datasets*. U.S. Patent 15 193 536 Jul. 2016.

[42] Jas Brooks. “Promises of the virtual museum.” In: *XRDS: Crossroads, The ACM Magazine for Students* 25 (Jan. 2019), pp. 46–50. DOI: 10.1145/3301483.

[43] Alan Brown, Jerry Fishenden, Mark Thompson, and Will Venters. "Appraising the impact and role of platform models and Government as a Platform (GaaP) in UK Government public service reform: Towards a Platform Assessment Framework (PAF)." In: *Government Information Quarterly* 34.2 (2017), pp. 167–182.

[44] Richard Brownlow, Stefano Capuzzi, Sven Helmer, Luciana Martins, Immanuel Normann, and Alex Poulovassilis. "An Ontological Approach to Creating an Andean Weaving Knowledge Base." In: *Journal on Computing and Cultural Heritage* 8.2 (2015).

[45] Hans De Bruijn and Willemijn Dicke. "Strategies for safeguarding public values in liberalized utility sectors." In: *Public administration* 84.3 (2006), pp. 717–735.

[46] Jerome Bruner. *Acts of meaning.* Harvard university press, 1990.

[47] Fabio Bruno, Stefano Bruno, Giovanna De Sensi, Maria-Laura Luchi, Stefania Mancuso, and Maurizio Muzzupappa. "From 3D reconstruction to virtual reality: A complete methodology for digital archaeological exhibition." In: *Cultural Heritage* 11.1 (2010), pp. 42–49.

[48] Angelika C. Bullinger, Anne-Katrin Neyer, Matthias Rass, and Kathrin M. Moeslein. "Community-Based Innovation Contests: Where Competition Meets Cooperation." In: *Creativity and Innovation Management* 19.3 (2010), pp. 290–303. DOI: 10.1111/j.1467-8691.2010.00565.x.

[49] Romina Cachia, Anusca Ferrari, et al. *Creativity in schools: A survey of teachers in Europe*. 2010.

[50] Coral Calero, Angélica Caro, and Mario Piattini. "An Applicable Data Quality Model for Web Portal Data Consumers." In: *World Wide Web* 11.4 (2008), pp. 465–484. DOI: 10.1007/s11280-008-0048-y.

[51] Calvin ML Chan. "From open data to open innovation strategies: Creating e-services using open government data." In: *2013 46th Hawaii International Conference on System Sciences*. IEEE. 2013, pp. 1890–1899.

[52] Chitat Chan and Carmen Yau. “Digital storytelling for social work interventions.” In: *Oxford bibliographies in social work* (2019).

[53] Bee-Chung Chen, Daniel Kifer, Kristen LeFevre, and Ashwin Machanavajjhala. “Privacy-Preserving Data Publishing.” In: *Foundations and Trends in Databases* 2 (2009), pp. 1–167.

[54] Min-Hsun Chiang. “Exploring the Effects of Digital Storytelling: A Case Study of Adult L2 Writers in Taiwan.” In: *IAFOR Journal of Education* 8.1 (2020), pp. 65–82.

[55] CIDOC CRM Special Interest Group. *CIDOC-CRM, Conceptual Reference Model*. 2006. URL: http://www.cidoc-crm.org.

[56] Philipp Cimiano and Stefan Kopp. “Accessing the Web of Data through embodied virtual characters.” In: *Semantic Web* 1 (2010), pp. 83–88.

[57] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati. “k-Anonymity.” In: *Secure Data Management in Decentralized Systems*. Springer US, 2007, pp. 323–353. DOI: 10.1007/978-0-387-27696-0_10.

[58] Cristian Ciurea and Florin Gheorghe Filip. “Virtual Exhibitions in Cultural Institutions: Useful Applications of Informatics in a Knowledge-based Society.” In: *Studies in Informatics and Control* 28.1 (2019), pp. 55–64.

[59] Fabio Clarizia, Saverio Lemma, Marco Lombardi, and Francesco Pascale. “An ontological digital storytelling to enrich tourist destinations and attractions with a mobile tailored story.” In: *Intern. Conf. on green, pervasive, and cloud computing*. 2017, pp. 567–581.

[60] J Spencer Clark, Suzanne Porath, Julie Thiele, and Morgan Jobe. *Action research*. New Prairie Press, 2020.

[61] Michael Cochez, Petar Ristoski, Simone Paolo Ponzetto, and Heiko Paulheim. “Global RDF Vector Space Embeddings.” In: *16th ISWC*. 2017, pp. 190–207.

[62] Michael Cochez, Petar Ristoski, Simone Paulo Ponzetto, and Heiko Paulheim. *KGloVe DBpedia uniform embeddings*. 2017. URL: https://doi.org/10.5281/zenodo.1320148.

[63] Michael Cochez, Petar Ristoski, Simone Paulo Ponzetto, and Heiko Paulheim. *RDF2Vec DBpedia uniform embeddings*. 2017. URL: https://doi.org/10.5281/zenodo.1318146.

[64] Mike Cohn. *User stories applied: For agile software development*. Addison-Wesley Professional, 2004.

[65] Fiona Collins. "The Use of Traditional Storytelling in Education to the Learning of Literacy Skills." In: *Early Child Development and Care* 152.1 (1999), pp. 77–108.

[66] Liliana Colodeeva. "Teaching writing: integrating digital storytelling into the classroom." In: *Buletinul ştiinţific al Universităţii de Stat" Bogdan Petriceicu Hasdeu" din Cahul, Seria" Stiinte Umanistice"* 9.1 (2019), pp. 33–42.

[67] Gennaro Cordasco et al. "Engaging Citizens with a Social Platform for Open Data." In: *Proceedings of the 18th Annual International Conference on Digital Government Research, DG.O*. ACM, 2017, pp. 242–249.

[68] Antonio Cordella and Andrea Paletti. "Government as a platform, orchestration, and public value creation: The Italian case." In: *Government Information Quarterly* 36.4 (2019), p. 101409.

[69] Maria Francesca Costabile, Maria De Marsico, Rosa Lanzilotti, Vito Leonardo Plantamura, and Teresa Roselli. "On the usability evaluation of e-learning applications." In: *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*. IEEE. 2005, 6b–6b.

[70] Lee J Cronbach. "Coefficient alpha and the internal structure of tests." In: *psychometrika* 16.3 (1951), pp. 297–334.

[71] Robert Crone. "Big Data Veracity Assessment: Improving risk assessment by adding high veracity data to existing contents insurance models." In: (2016).

[72] Salvatore Cuomo, Giovanni Colecchia, Vincenzo Cola, and Ugo Chirico. "A virtual assistant in cultural heritage scenarios." In: *Concurrency and Computation: Practice and Experience* 33.3 (2021), e5331.

[73] Bernardo Cuteri, Kristian Reale, and Francesco Ricca. "A Logic-Based Question Answering System for Cultural Heritage." In: *Logics in Artificial Intelligence*. 2019, pp. 526–541.

[74] Aba-Sah Dadzie and Matthew Rowe. "Approaches to visualising Linked Data: A survey." In: *Semantic Web* 2.2 (2011), pp. 89–124.

[75] Danica Damljanovic, Milan Agatonovic, and Hamish Cunningham. "Natural Language Interfaces to Ontologies: Combining Syntactic Analysis and Ontology-based Lookup Through the User Interaction." In: *The Semantic Web: Research and Application*. 2010, pp. 106–120.

[76] Mauro D'Angelo and Maria Angela Pellegrino. "Roobopoli: A Project to Learn Robotics by a Constructionism-Based Approach." In: *International Conference in Methodologies and intelligent Systems for Techhnology Enhanced Learning*. Springer, 2020, pp. 249–257.

[77] Marilena Daquino, Francesca Mambelli, Silvio Peroni, Francesca Tomasi, and Fabio Vitali. "Enhancing Semantic Expressivity in the Cultural Heritage Domain: Exposing the Zeri Photo Archive as Linked Open Data." In: *Journal on Computing and Cultural Heritage* 10.4 (2017).

[78] Dataguise. *DGSecure*. last access January 2019. 2018. URL: https://www.dataguise.com/detect/.

[79] Fred D Davis. "Perceived usefulness, perceived ease of use, and user acceptance of information technology." In: *Management Information Systems Quarterly* (1989), pp. 319–340.

[80] Renato De Donato, Giuseppe Ferretti, Antonio Marciano, Giuseppina Palmieri, Donato Pirozzi, Vittorio Scarano, and Luca Vicidomini. "Agile production of high quality open data." In: *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. ACM, 2018, p. 84.

[81] Renato De Donato, Martina Garofalo, Delfina Malandrino, Maria Angela Pellegrino, and Andrea Petta. "Education Meets Knowledge Graphs for the Knowledge Management." In: *Methodologies and Intelligent Systems for Technology Enhanced Learning (MIS4TEL)*. Springer. 2020, pp. 272–280.

[82] Kamil Deja. "Using machine learning techniques for Data Quality Monitoring in CMS and ALICE." In: *Proceedings of Science*. Vol. 350. 2019.

[83] Corine Deliot. "Publishing the British national bibliography as linked open data." In: *Catalogue & Index* 174 (2014), pp. 13–18.

[84] Amazon Developer. *Build Skills with the Alexa Skills Kit*. Last access March, 2021. 2014. URL: https://developer.amazon.com/en-US/docs/alexa/ask-overviews/build-skills-with-the-alexa-skills-kit.html.

[85] Charles Dhanaraj and Arvind Parkhe. "Orchestrating innovation networks." In: *Academy of management review* 31.3 (2006), pp. 659–669.

[86] Chiara Di Stefano and Federica Battisti. "Caravaggio in Rome: A qoe-based proposal for a virtual gallery." In: *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*. 2017, pp. 1–4.

[87] Dennis Diefenbach, José Giménez-García, Andreas Both, Kamal Singh, and Pierre Maret. "QAnswer KG: Designing a Portable Question Answering System over RDF Data." In: *The Semantic Web*. 2020, pp. 429–445.

[88] Dennis Diefenbach, Pedro Henrique Migliatti, Omar Qawasmeh, Vincent Lully, Kamal Singh, and Pierre Maret. "QAnswer: A Question Answering Prototype Bridging the Gap between a Considerable Part of the LOD Cloud and End-Users." In: *The World Wide Web Conference*. 2019, pp. 3507–3510.

[89] Chris Dijkshoorn, Lizzy Jongma, Lora Aroyo, Jacco Van Ossenbruggen, Guus Schreiber, Wesley Ter Weele, and Jan Wielemaker. "The Rijksmuseum collection as Linked Data." In: *Semantic Web* 9.2 (2018), pp. 221–230.

[90] Martin Doerr. "The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata." In: *AI magazine* 24.3 (2003), pp. 75–75.

[91] Martin Doerr. "Ontologies for cultural heritage." In: *Handbook on ontologies*. Springer, 2009, pp. 463–486.

[92] Till Döhmen, Hannes Mühleisen, and Peter Boncz. "Multi-Hypothesis CSV Parsing." In: *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. ACM, 2017, p. 16.

[93] Renato De Donato, Martina Garofalo, Delfina Malandrino, Maria Angela Pellegrino, Andrea Petta, and Vittorio Scarano. "Linked Data Queries by a Trialogical Learning Approach." In: *Computer Supported Cooperative Work in Design, CSCWD*. IEEE, 2019, pp. 117–122.

[94] Renato De Donato, Martina Garofalo, Delfina Malandrino, Maria Angela Pellegrino, Andrea Petta, and Vittorio Scarano. "QueDI: from Knowledge Graph Querying to Data Visualization." In: *Semantics*. 2020.

[95] Renato De Donato, Delfina Malandrino, Giuseppina Palmieri, Andrea Petta, Donato Pirozzi, Vittorio Scarano, Luigi Serra, Carmine Spagnuolo, Luca Vicidomini, and Gennaro Cordasco. "Datalet-Ecosystem Provider (DEEP): Scalable Architecture for Reusable, Portable and User-Friendly Visualizations of Open Data." In: *2017 Conference for E-Democracy and Open Government, CeDEM*. IEEE Computer Society, 2017, pp. 92–101.

[96] Mauro Dragoni, Sara Tonelli, and Giovanni Moretti. "A Knowledge Management Architecture for Digital Cultural Heritage." In: *Journal on Computing and Cultural Heritage* 10.3 (2017).

[97] Gabriela Dumitrescu, Cornel Lepadatu, and Cristian Ciurea. "Creating Virtual Exhibitions for Educational and Cultural Development." In: *Informatica economica* 18.1 (2014).

[98] Lisa Ehrlinger and Wolfram Wöß. "Towards a Definition of Knowledge Graphs." In: *SEMANTiCS (Posters, Demos, SuCCESS)* 48.1-4 (2016), p. 2.

[99] Thomas Erickson. "Design as storytelling." In: *interactions* 3.4 (1996), pp. 30–35.

[100] Hossein Estiri, Jeffrey G. Klann, and Shawn N. Murphy. "A clustering approach for detecting implausible observation values in electronic health records data." In: *BMC Medical Informatics and Decision Making* 19.1 (2019).

[101] EUROPEAN DATA PORTAL. *Protecting data and opening data*. https://www.europeandataportal.eu/en/highlights/protecting-data-and-opening-data. 2019.

[102] Regulation EU 2016/679 of the Europena parliament and of the council. *General Data Protection Regulation - GDPR*. https://eur-lex.europa.eu/eli/reg/2016/679/oj. 2016.

[103] Victoria Eyharabide, Vincent Lully, and Florentin Morel. "MusicKG: Representations of Sound and Music in the Middle Ages as Linked Open Data." In: *Semantic Systems. The Power of AI and Knowledge Graphs*. Springer International Publishing, 2019, pp. 57–63.

[104] Imane Ezzine and Laila Benhlima. "A study of handling missing data methods for big data." In: *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*. IEEE. 2018, pp. 498–501.

[105] Federal Committee on Statistical Methodology. "Statistical Policy Working Paper 22." In: *Report on Statistical Disclosure Limitation Methodology* (2005).

[106] Sébastien Ferré. "SQUALL: The expressiveness of SPARQL 1.1 made available as a controlled natural language." In: *Data & Knowledge Engineering* 94 (2014), pp. 163–188. DOI: https://doi.org/10.1016/j.datak.2014.07.010.

[107] Sebastien Ferre. "Sparklis: An expressive query builder for SPARQL endpoints with guidance in natural language." In: *Semantic Web* 8.3 (2017), pp. 405–418.

[108] Giuseppe Ferretti, Delfina Malandrino, Maria Angela Pellegrino, Donato Pirozzi, Gianluigi Renzi, and Vittorio Scarano. "A Non-prescriptive Environment to Scaffold High Quality and Privacy-aware Production of Open Data

with AI." In: *20th Annual Inter. Conf. on Digital Government Research*. 2019, pp. 25–34.

[109] Giuseppe Ferretti, Delfina Malandrino, Maria Angela Pellegrino, Andrea Petta, Gianluigi Renzi, Vittorio Scarano, and Luigi Serra. "Orchestrated Co-creation of High-Quality Open Data Within Large Groups." In: *Electronic Government - 18th IFIP WG 8.5 International Conference*. 2019, pp. 168–179.

[110] Jerry Fishenden and Mark Thompson. "Digital government, open architecture, and innovation: why public sector IT will never be the same again." In: *Journal of public administration research and theory* 23.4 (2013), pp. 977–1004.

[111] Schubert Foo. "Online Virtual Exhibitions: Concepts and Design Considerations." In: *Journal of Library and Information Technology* 28 (2008), pp. 22–34. DOI: 10.14429/djlit.28.4.194.

[112] Micro:bit Educational Foundation. *Micro:bit Educational Foundation*. https://microbit.org. Accessed: 2019-09-06. 2019.

[113] The World Wide Web Foundation. "Open Data Barometer 4th Edition." In: *Global Report* (2017).

[114] Lin Fu, Dion Hoe-Lian Goh, Schubert Shou-Boon Foo, and Jin-Cheon Na. "Collaborative querying through a hybrid query clustering approach." In: *International Conference on Asian Digital Libraries*. Springer. 2003, pp. 111–122.

[115] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. "Privacy-Preserving Data Publishing: A Survey of Recent Developments." In: *ACM Computing Surveys* 42.4 (2010), 14:1–14:53.

[116] María Jesús García Godoy, Esteban López-Camacho, Ismael Navas-Delgado, and José F. Aldana-Montes. "Sharing and executing linked data queries in a collaborative environment." In: *Bioinformatics* 29.13 (2013), pp. 1663–1670. DOI: 10.1093/bioinformatics/btt192.

[117] Anne Haugen Gausdal and Etty Ragnhild Nilsen. "Orchestrating innovative SME networks. The case of "HealthInnovation"." In: *Journal of the Knowledge Economy* 2.4 (2011), pp. 586–600.

[118] Rosella Gennari, Maristella Matera, Alessandra Melonio, Mehdi Rizvi, and Eftychia Roumelioti. "A Board Game and a Workshop for Co-Creating Smart Nature Ecosystems." In: *9th International Conference in Methodologies and Intelligent Systems for Technology Enhanced Learning (MIS4TEL)*. Springer, 2019.

[119] Rosella Gennari, Maristella Matera, Alessandra Melonio, and Eftychia Roumelioti. "A Board-Game for Co-Designing Smart Nature Environments in Workshops with Children." In: *End-User Development*. Springer, 2019, pp. 132–148.

[120] Rosella Gennari, Maristella Matera, Alessandra Melonio, and Eftychia Roumelioti. "SNaP 2: The Evolution of a Board Game for Smart Nature Environments." In: *Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*. 2019, pp. 405–411.

[121] Rosella Gennari, Maristella Matera, Alessandra Melonio, Mehdi Rizvi, and Eftychia Roumelioti. "Reflection and Awareness in the Design Process Children Ideating, Programming and Prototyping Smart Objects." In: *Multimedia Tools and Applications* (2020), pp. 1–24.

[122] Ivan Giangreco, Loris Sauter, Mahnaz Amiri Parian, Ralph Gasser, Silvan Heller, Luca Rossetto, and Heiko Schuldt. "Virtue: a virtual reality museum experience." In: *Proceedings of the 24th international conference on intelligent user interfaces: companion*. 2019, pp. 119–120.

[123] Michail N. Giannakos, Monica Divitini, and Ole Sejer Iversen. "Entertainment, engagement, and education: Foundations and developments in digital and physical spaces to support learning through making." In: *Entertainment Computing* 21 (2017), pp. 77–81.

[124] Sara Gonizzi Barsanti, Giandomenico Caruso, LL Micoli, M Covarrubias Rodriguez, Gabriele Guidi, et al. "3D visualization of cultural heritage artefacts with virtual reality devices." In: *25th International CIPA Symposium*. 2015, pp. 165–172.

[125] Palash Goyal and Emilio Ferrara. "Graph embedding techniques, applications, and performance: A survey." In: *Knowledge-Based Systems* 151 (2018), pp. 78–94.

[126] Alvaro Graves. "Creation of Visualizations Based on Linked Data." In: *3rd International Conference on Web Intelligence, Mining and Semantics*. 2013, 41:1–41:12.

[127] Ben Green, Gabe Cunningham, Ariel Ekblaw, Paul Kominers, Andrew Linzer, and Susan Crawford. "OPEN DATA PRIVACY." In: *Open Data Privacy* (2017).

[128] Kylene Guse, Andrea Spagat, Amy Hill, Andrea Lira, Stephen Heathcock, and Melissa Gilliam. "Digital storytelling: A novel methodology for sexual health promotion." In: *American Journal of Sexuality Education* 8.4 (2013), pp. 213–227.

[129] Peter Haase, Andriy Nikolov, Johannes Trame, Artem Kozlov, and Daniel M. Herzig. "Alexa, Ask Wikidata! Voice Interaction with Knowledge Graphs using Amazon Alexa." In: *International Semantic Web Conference*. 2017.

[130] Sam Habibi Minelli, Maria Natale, Paolo Ongaro, Marzia Piccininno, Rubino Saccoccio, and Daniele Ugoletti. "MOVIO: A Toolkit for Creating Curated Digital Exhibitions." In: *Procedia Computer Science* 38 (2014), pp. 28–33. DOI: 10.1016/j.procs.2014.10.006.

[131] Preben Hansen and Kalervo Järvelin. "Collaborative Information Retrieval in an Information-intensive Domain." In: *Inf. Process. Manage.* 41.5 (2005), pp. 1101–1119. DOI: 10.1016/j.ipm.2004.04.016.

[132] Andreas Harth. "VisiNav: A system for visual search and navigation on web data." In: *Semantic Web* 8.4 (2010), pp. 348–354.

[133] Bernhard Haslhofer and Antoine Isaac. "data. europeana. eu: The europeana linked open data pilot." In: *International Conference on Dublin Core and Metadata Applications*. 2011, pp. 94–104.

[134] Bernhard Haslhofer, Antoine Isaac, and Rainer Simon. "Knowledge Graphs in the Libraries and Digital Humanities Domain." In: *Encyclopedia of Big Data Technologies*. 2019.

[135] Laurence Hauttekeete, Tom Evens, Katrien De Moor, Dimitri Schuurman, Erik Mannens, and Rik Van de Walle. "Archives in motion: Concrete steps towards the digital disclosure of audiovisual content." In: *Journal of Cultural Heritage* 12.4 (2011), pp. 459–465.

[136] Masaki Hayashi, Steven Bachelder, and Masayuki Nakajima. "Automatic Generation of Personal Virtual Museum." In: *International Conference on Cyberworlds, CW 2016*. Chongqing, China, 2016, pp. 219–222.

[137] Mahmoud Haydar, David Roussel, Madjid Maidi, Samir Otmane, and Malik Mallem. "Virtual and augmented reality for cultural computing and heritage: a case study of virtual exploration of underwater archaeological sites." In: *Virtual Reality* 15.4 (2011), pp. 311–327.

[138] Marti Hearst. *Search user interfaces*. Cambridge university press, 2009.

[139] Ola Henfridsson and Bendik Bygstad. "The generative mechanisms of digital infrastructure evolution." In: *Management Information Systems Quarterly* (2013), pp. 907–931.

[140] Luis A. Hernández, Javier Taibo, David Blanco, José A. Iglesias, Antonio Seoane, Alberto Jaspe, and Rocío López. "Physically Walking in Digital Spaces — A Virtual Reality Installation for Exploration of Historical Heritage." In: *International Journal of Architectural Computing* 5.3 (2007), pp. 487–506.

[141] Mary E Hess. "A new culture of learning: Digital storytelling and faith formation." In: *Dialog* 53.1 (2014), pp. 12–22.

[142] Lynette Hirschman and Robert Gaizauskas. “Natural language question answering: the view from here.” In: *natural language engineering* 7.4 (2001), p. 275.

[143] Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. “KORE: Keyphrase Overlap Relatedness for Entity Disambiguation.” In: *Proc. of the 21st ACM CIKM*. 2012, pp. 545–554.

[144] Hajar Homayouni, Sudipto Ghosh, and Indrakshi Ray. “ADQuaTe: An automated data quality test approach for constraint discovery and fault detection.” In: *20th International Conference on Information Reuse and Integration for Data Science*. IEEE, 2019, pp. 61–68.

[145] Chun-Ko Hsieh, Wen-Ching Liao, Meng-Chieh Yu, and Yi-Ping Hung. “Interacting with the Past: Creating a Time Perception Journey Experience Using Kinect-Based Breath Detection and Deterioration and Recovery Simulation Technologies.” In: *Journal on Computing and Cultural Heritage* 7.1 (2014), 1:1–1:15.

[146] Hai Huang, Kher Hui Ng, Benjamin Bedwell, and Steve Benford. “A card-based internet of things game ideation tool for museum context.” In: *Journal of Ambient Intelligence and Humanized Computing* (2020), pp. 1–12. DOI: 10.1007/s12652-020-02627-2.

[147] Jessica Hullman, Nicholas Diakopoulos, Elaheh Momeni, and Eytan Adar. “Content, Context, and Critique: Commenting on a Data Visualization Blog.” In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work - Social Computing*. 2015, pp. 1170–1175.

[148] Eero Hyvönen. “Publishing and Using Cultural Heritage Linked Data on the Semantic Web.” In: *Synthesis Lectures on Semantic Web: Theory and Technology* 2 (2012), pp. 1–159.

[149] Eero Hyvönen, Eetu Mäkelä, Mirva Salminen, Arttu Valo, Kim Viljanen, Samppa Saarela, Miikka Junnila, and Suvi Kettula. “MuseumFinland - Finnish museums on the semantic web.” In: *Semantic Web* 3.2-3 (2005), pp. 224–241.

[150] ISO 9241-210. *Ergonomics of human-system interaction*. Last access 2021/06/03. 2019. URL: https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-2:v1:en.

[151] Steven Jacobs. "The Use of Participatory Action Research within Education–Benefits to Stakeholders." In: *World Journal of Education* 6.3 (2016), pp. 48–55.

[152] Sugandha Jain and Abhishek Srivastava. "Why Storytelling Should be the Medium of Design Education." In: *Design for Tomorrow*. Springer, 2021, pp. 895–902.

[153] Tomasz Janowski, Elsa Estevez, and Rehema Baguma. "Platform governance for sustainable development: Reshaping citizen-administration relationships in the digital age." In: *Government Information Quarterly* 35.4 (2018), S1–S16.

[154] Marijn Janssen, Yannis Charalabidis, and Anneke Zuiderwijk. "Benefits, adoption barriers and myths of open data and open government." In: *Information systems management* 29.4 (2012), pp. 258–268.

[155] Marijn Janssen, Yannis Charalabidis, and Anneke Zuiderwijk. "Benefits, Adoption Barriers and Myths of Open Data and Open Government." In: *Information Systems Management* 29.4 (2012), pp. 258–268.

[156] Marijn Janssen, Yannis Charalabidis, and Anneke Zuiderwijk. "Benefits, Adoption Barriers and Myths of Open Data and Open Government." In: *Information Systems Management* 29.4 (2012), pp. 258–268.

[157] Matthew A. Jaro. "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida." In: *J. of the American Statistical Association* 84.406 (1989), pp. 414–420.

[158] Bogumil Kaminski, Pawel Pralat, and François Théberge. "An Unsupervised Framework for Comparing Graph Embeddings." In: *CoRR* abs/1906.04562 (2019).

[159] Irene Katsouri, Aimilia Tzanavari, Kyriakos Herakleous, and Charalambos Poullis. "Visualizing and Assessing Hypotheses for Marine Archaeology in a VR CAVE Environment." In: *Journal on Computing and Cultural Heritage* 8.2 (2015), 10:1–10:18.

[160] Esther Kaufmann and Abraham Bernstein. "How useful are natural language interfaces to the semantic web for casual end-users?" In: *The Semantic Web*. 2007, pp. 281–294.

[161] Sirina Keesara, Andrea Jonas, and Kevin Schulman. "Covid-19 and Health Care's Digital Revolution." In: *New England Journal of Medicine* 382.23 (2020), pp. e82.

[162] Fariza Khalid and Tewfiq El-Maliki. "Teachers' Experiences in the Development of Digital Storytelling for Cyber Risk Awareness." In: *International Journal of Advanced Computer Science and Applications* 11 (Jan. 2020). DOI: 10.14569/IJACSA.2020.0110225.

[163] Ferit Kılıçkaya. "Learners' perceptions of collaborative digital graphic writing based on semantic mapping." In: *Computer Assisted Language Learning* 33.1-2 (2020), pp. 58–84.

[164] J.R. Kim, G.W. Shin, S.T. Hong, and D.W. Kim. "Study on data cleansing algorithms for outliers in water supply system." In: *International Conference on Big Data Analytics, Data Mining and Computational Intelligence*. 2019, pp. 242–244.

[165] Chairi Kiourt, Anestis Koutsoudis, and George Pavlidis. "DynaMus: A fully dynamic 3D virtual museum framework." In: *Journal of Cultural Heritage* 22 (2016), pp. 984–991.

[166] M. Kiran Kumar and J. Divya Udayan. "A survey of machine learning techniques for cancer disease prediction and diagnosis." In: *Indian J. of Public Health Research and Development* 10.4 (2019), pp. 157–162.

[167] Jakub Klímek, Jirí Helmich, and Martin Necaský. "Payola: Collaborative Linked Data Analysis and Visualization Framework." In: *The Semantic Web: ESWC Satellite Events, Montpellier, France*. Vol. 7955. Springer, 2013, pp. 147–151.

[168] Craig A. Knoblock et al. "Lessons Learned in Building Linked Data for the American Art Collaborative." In: *the International Semantic Web Conference*. 2017.

[169] Matthew J Koehler and Punya Mishra. "Teachers learning technology by design." In: *Journal of computing in teacher education* 21.3 (2005), pp. 94–102.

[170] Petros Kostagiolas, Anastasia Margiola, and Anastasia Avramidou. "A library management response model against the economic crisis." In: *Library Review* (2011).

[171] Jitin Krishnan, Patrick Coronado, and Trevor Reed. "SEVA: A Systems Engineer's Virtual Assistant." In: *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*. 2019.

[172] Georg Krogh and Eric Hippel. "Open Source Software and the Private-Collective Innovation Model: Issues for Organization Science." In: *Organization science* (2003), pp. 209–223.

[173] Tien Fabrianti Kusumasari et al. "Data profiling for data quality improvement with OpenRefine." In: *2016 international conference on information technology systems and innovation (ICITSI)*. IEEE. 2016, pp. 1–6.

[174] Elżbieta Kużelewska and Mariusz Tomaszuk. "European Human Rights Dimension of the Online Access to Cultural Heritage in Times of the COVID-19 Outbreak." In: *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique* (2020), pp. 1–13.

[175] SAM Labs. *SAM Labs*. `https://samlabs.com/`. Accessed: 2021-03-20. 2020.

[176] Michael D. Lee and Matthew Welsh. "An Empirical Evaluation of Models of Text Document Similarity." In: *XXVII Annual Conference of the Cognititive Science Society*. 2005, pp. 1254–1259.

[177] Yuangui Lei, Victoria Uren, and Enrico Motta. "A framework for evaluating semantic metadata." In: *Proceedings of the 4th international conference on Knowledge capture*. 2007, pp. 135–142.

[178] Vladimir I Levenshtein. "Binary codes capable of correcting deletions, insertions, and reversals." In: *Soviet physics doklady*. Vol. 10. 8. 1966, pp. 707–710.

[179] James R. Lewis. "IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use." In: *International Journal of Human–Computer Interaction* 7.1 (1995), pp. 57–78.

[180] James R Lewis. "IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use." In: *International Journal of Human-Computer Interaction* 7.1 (1995), pp. 57–78.

[181] James R. Lewis and Jeff Sauro. "The Factor Structure of the System Usability Scale." In: *Proceedings of the 1st International Conference on Human Centered Design, San Diego, CA*. Springer-Verlag, 2009, pp. 94–103. DOI: 10.1007/978-3-642-02806-9_12.

[182] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity." In: *IEEE 23rd International Conference on Data Engineering*. 2007, pp. 106–115.

[183] Yuelin Li and Chang Liu. "Information Resource, Interface, and Tasks as User Interaction Components for Digital Library Evaluation." In: *Information Processing and Management* 56.3 (2019), pp. 704–720.

[184] Ling Lin and Jinshan Su. "Anomaly detection method for sensor network data streams based on sliding window sampling and optimized clustering." In: *Safety Science* 118 (2019), pp. 70–75.

[185] He Liu, Xiaohui Wang, Shuya Lei, Xi Zhang, Weiwei Liu, and Ming Qin. "A rule based data quality assessment architecture and application for electrical data." In: *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing*. 2019, pp. 1–6.

[186] Daniel Alejandro Loaiza Carvajal, Maria Mercedes Morita, and Gabriel Mario Bilmes. "Virtual museums. Captured reality and 3D modeling." In: *Journal of Cultural Heritage* 45 (2020), pp. 234–239. DOI: 10.1016/j.culher.2020.04.013.

[187] Marco Lombardi, Francesco Pascale, and Domenico Santaniello. "An application for Cultural Heritage using a Chatbot." In: *2nd International Conference on Computer Applications Information Security (ICCAIS)*. 2019, pp. 1–5.

[188] Vanessa Lopez, Victoria Uren, Marta Sabou, and Enrico Motta. "Is Question Answering fit for the Semantic Web?: A survey." In: *Semantic Web* 2 (2011), pp. 125–155.

[189] The Open Knowledge Foundation Ltd. *Library Messytables link*. last access January 2019. 2013. URL: https://messytables.readthedocs.io/en/latest.

[190] Zakaria Maamar, Mohamed Sellami, Noura Faci, Emir Ugljanin, and Quan Z. Sheng. "Storytelling Integration of the Internet of Things into Business Processes." In: *Business Process Management Forum*. Springer, 2018, pp. 127–142.

[191] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. "l-Diversity: Privacy beyond k-Anonymity." In: *ACM Trans. Knowledge Discovery Data* 1.1 (2007).

[192] Octavian-Mihai Machidon, Aleš Tavčar, Matjaž Gams, and Mihai Duguleană. "CulturalERICA: A conversational agent improving the exploration of European cultural heritage." In: *Journal of Cultural Heritage* 41 (2020), pp. 152–165.

[193] Stephen Madigan. *Narrative therapy.* American Psychological Association, 2011.

[194] Delfina Malandrino, Ilaria Manno, Giuseppina Palmieri, Andrea Petta, Donato Pirozzi, Vittorio Scarano, Luigi Serra, Carmine Spagnuolo, Luca Vicidomini, and Gennaro Cordasco. "An Architecture for Social Sharing and Collaboration around Open Data Visualisations." In: *Proceedings of the 19th ACM Conference on Computer Supported*

Cooperative Work and Social Computing, CSCW. ACM, 2016, pp. 357–360.

[195] Michele Mallia, Marcello Carrozzino, Chiara Evangelista, and Massimo Bergamasco. "Automatic Creation of a Virtual/Augmented Gallery Based on User Defined Queries on Online Public Repositories." In: *Virtual Reality Technologies in Cultural Heritage (VRTCH)*. 2019, pp. 135–147. DOI: 10.1007/978-3-030-05819-7_11.

[196] Stanislav Malyshev, Markus Krötzsch, Larry González, Julius Gonsior, and Adrian Bielefeldt. "Getting the Most Out of Wikidata: Semantic Technology Usage in Wikipedia's Knowledge Graph." In: *The Semantic Web - 17th ISWC, Proceedings, Part II*. 2018, pp. 376–394.

[197] Ellen B. Mandinach and Edith S. Gummer. "A Systemic View of Implementing Data Literacy in Educator Preparation." In: *Educational Researcher* 42.1 (2013), pp. 30–37.

[198] Huina Mao, Xin Shuai, and Apu Kapadia. "Loose Tweets: An Analysis of Privacy Leaks on Twitter." In: *Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society*. WPES. ACM, 2011, pp. 1–12.

[199] Alexandru Mara, Jefrey Lijffijt, and Tijl De Bie. "EvalNE: A Framework for Evaluating Network Embeddings on Link Prediction." In: *Reproducibility in Machine Learning, ICLR*. New Orleans, Louisiana, United States, 2019.

[200] Gary Marchionini. "Exploratory search: from finding to understanding." In: *Communications of the ACM* 49.4 (2006), pp. 41–46.

[201] Stacy C. Marsella, W. Lewis Johnson, and Catherine M. Labore. "Interactive pedagogical drama for health interventions." In: *Artificial Intelligence in Education*. 2003, pp. 341–348.

[202] Sébastien Martin, Muriel Foulonneau, Slim Turki, and Madjid Ihadjadene. "Open data: Barriers, risks and opportunities." In: *Proceedings of the 13th European Conference on eGovernment: ECEG*. 2013, pp. 301–309.

[203] Arkady Maydanchik. *Data quality assessment*. Technics publications, 2007.

[204] Emanuela Mazzone, Netta Iivari, Ruut Tikkanen, Janet C Read, and Russell Beale. "Considering context, content, management, and engagement in design activities with children." In: *Proceedings of the 9th international conference on interaction design and children*. 2010, pp. 108–117.

[205] Samantha McAleese and Jennifer M. Kilty. "Stories Matter: Reaffirming the Value of Qualitative Research." In: *The Qualitative Report* 24 (2019), pp. 822–845.

[206] Keegan McBride, Gerli Aavik, Maarja Toots, Tarmo Kalvet, and Robert Krimmer. "How does open government data driven co-creation occur? Six factors and a perfect storm; insights from Chicago's food inspection forecasting model." In: *Government Information Quarterly* 36.1 (2019), pp. 88–97.

[207] John P. McCrae. *The LOD Cloud*. Access 2021/04/21. 2007. URL: http://lod-cloud.net.

[208] Ines Mergel, Alexander Kleibrink, and Jens Sörvik. "Open data outcomes: U.S. cities between product and process innovation." In: *Government Information Quarterly* 35.4 (2018), pp. 622–632.

[209] Microsoft. *MakeCode*. https://makecode.microbit.org. Accessed: 2019-09-06. 2019.

[210] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. "Distributed Representations of Words and Phrases and their Compositionality." In: *27th Annual Conference on Neural Information Processing Systems*. 2013, pp. 3111–3119.

[211] Patrick Millais, Simon L. Jones, and Ryan Kelly. "Exploring Data in Virtual Reality: Comparisons with 2D Data Visualizations." In: *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems - CHI EA*. 2018, pp. 1–6.

[212] Wookhee Min, Megan H. Frankosky, Bradford W. Mott, Eric N. Wiebe, Kristy Elizabeth Boyer, and James C. Lester. "Inducing Stealth Assessors from Game Interaction Data." In: *Artificial Intelligence in Education*. Springer International Publishing, 2017, pp. 212–223.

[213] Silvia Mirri, Marco Roccetti, and Paola Salomoni. "Collaborative design of software applications: the role of users." In: *Human-centric Computing and Information Sciences* 8 (2018), p. 6.

[214] Konstantin Mitgutsch and Narda Alvarado. "Purposeful by design? A serious game design assessment framework." In: *Proceedings of the International Conference on the foundations of digital games*. 2012, pp. 121–128.

[215] Abu Shamim Mohammad Arif, Jia Tina Du, and Ivan Lee. "Examining Collaborative Query Reformulation: A Case of Travel Information Searching." In: *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, Gold Coast, Queensland, Australia*. ACM, 2014, pp. 875–878. DOI: 10.1145/2600428.2609463.

[216] Daniele Monaco, Maria Angela Pellegrino, Vittorio Scarano, and Luca Vicidomini. *The role of Linked Open Data in Authoring Virtual Exhibitions*. Accepted to Journal of Cultural Heritage in Nov. 2021. 2021.

[217] Ji-Won Moon and Young-Gul Kim. "Extending the TAM for a World-Wide-Web context." In: *Information & Management* 38.4 (2001), pp. 217–230.

[218] Hamzeh Moradi and Hefang Chen. "Digital Storytelling in Language Education." In: *Behavioral Sciences* 9.12 (2019), p. 147.

[219] Peter Morville and Jeffery Callender. *Search patterns: design for discovery*. O'Reilly Media, Inc., 2010.

[220] Rajeev Motwani and Ying Xu. "Efficient algorithms for masking and finding quasi-identifiers." In: *Proceedings of the Conference on Very Large Data Bases*. 2007, pp. 83–93.

[221] Juan Manuel Muñoz-González, Esther M Vega-Gea, Cristobal Ballesteros-Regaña, and María Dolores Hidalgo-Ariza. "Psychometric Study of a Scale of Measurement of the Digital Stories Creation Using Utellstory." In: *Sustainability* 12.8 (2020), p. 3204.

[222] Museum of Fine Arts Budapest. *Open Linked data from the Museum of Fine Arts Budapest*. last access March, 2021. 2016. URL: http://data.szepmuveszeti.hu.

[223] Ariffin Abdul Mutalib, Nurulnadwan Aziz, and Zatul Amilah Shaffiei. "Digital storytelling makes reading fun and entertaining." In: *International Journal of Computer Applications* 18.1 (2011), pp. 20–26.

[224] Brad A. Myers, Andrew J. Ko, Chris Scaffidi, Stephen Oney, YoungSeok Yoon, Kerry Chang, Mary Beth Kery, and Toby Jia-Jun Li. "Making End User Development More Natural." In: *New Perspectives in End-User Development*. Springer International Publishing, 2017, pp. 1–22.

[225] Goutam Mylavarapu, Johnson P Thomas, and K Ashwin Viswanathan. "An Automated Big Data Accuracy Assessment Tool." In: *4th International Conference on Big Data Analytics*. IEEE, 2019, pp. 193–197.

[226] C. Nalini and J. Sudeeptha. "Missing data imputation in high dimensional data set using local similarity." In: *International Journal of Recent Technology and Engineering* 8.3 (2019), pp. 8070–8074.

[227] Arvind Narayanan and Vitaly Shmatikov. "Robust De-anonymization of Large Sparse Datasets." In: *IEEE Symposium on Security and Privacy*. 2008, pp. 111–125. DOI: 10.1109/SP.2008.33.

[228] Arvind Narayanan and Vitaly Shmatikov. "Myths and Fallacies of Personally Identifiable Information." In: *Communications* 53.6 (2010), pp. 24–26.

[229] Felix Naumann. "Data profiling revisited." In: *ACM SIGMOD Record* 42.4 (2014), pp. 40–49.

[230] Roberto Navigli and Simone Paolo Ponzetto. "BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network." In: *Artificial Intelligence* 193 (2012), pp. 217–250.

[231] Zahra Nematzadeh, Roliana Ibrahim, Ali Selamat, and Vahdat Nazerian. "The synergistic combination of fuzzy C-means and ensemble filtering for class noise detection." In: *Engineering Computations* (2020).

[232] Mehmet Ercan Nergiz, Christopher Clifton, and Ahmet Erhan Nergiz. “Multirelational K-Anonymity.” In: *IEEE Transaction on Knowledge and Data Engineering* 21.8 (2009), pp. 1104–1117.

[233] Axel-Cyrille Ngonga Ngomo, Lorenz Bühmann, Christina Unger, Jens Lehmann, and Daniel Gerber. “Sorry, i Don’t Speak SPARQL: Translating SPARQL Queries into Natural Language.” In: *The Web Conference WWW*. 2013, pp. 977–988.

[234] Lesley-Ann Noel and Tsailu Liu. “Using Design Thinking to create a new education paradigm for elementary level children for higher student engagement and success.” In: *Design Research Society DRS*. 2016. DOI: 10.21606/drs.2016.200.

[235] Ikujiro Nonaka and Hirotaka Takeuchi. *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*. Oxford University Press, New York, 1995.

[236] Donald A Norman. *User centered system design: New perspectives on human-computer interaction*. CRC Press, 1986.

[237] Barack Obama. “Transparency and open government.” In: *Memorandum for the heads of executive departments and agencies* (2009).

[238] Humphrey O Obie, Caslon Chua, Iman Avazpour, Mohamed Abdelrazek, John Grundy, and Tomasz Bednarz. “Authoring Logically Sequenced Visual Data Stories with Gravity.” In: *Journal of Computer Languages* (2020), p. 100961.

[239] Jaclyn Ocumpaugh. “Baker Rodrigo Ocumpaugh monitoring protocol (BROMP) 2.0 technical and training manual.” In: *New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences* 60 (2015).

[240] Carla O’Dell and C. Jackson Grayson. “If Only We Knew What We Know: Identification and Transfer of Internal Best Practices.” In: *California Management Review* 40.3 (1998), pp. 154–174.

[241] Jason B Ohler. *Digital storytelling in the classroom: New media pathways to literacy, learning, and creativity*. Corwin Press, 2013.

[242] OKF. *Open Data*. last access May, 2020. URL: https://okfn.org/opendata/.

[243] Open Data Charter. *Open Data Charter Web Site*. Last accessed on May 23, 2018. 2015. URL: https://opendatacharter.net.

[244] Open Knowledge International. *The free data management platform to publish and register datasets*. https://old.datahub.io. [Online], Last access in November 2021. 2006.

[245] Tim O'Reilly. "Government as a Platform." In: *Innovations: Technology, Governance, Globalization* 6.1 (2011), pp. 13–40.

[246] Organisation for Economic Co-operation and Development (OECD). *CORE FOUNDATIONS FOR 2030*. Accessed 2020/02/21. 2019. URL: http://www.oecd.org/education/2030-project/teaching-and-learning/learning/.

[247] Organisation for Economic Co-operation and Development (OECD). *Open Government Data*. last access October 2019. 2019. URL: https://www.oecd.org/gov/digital-government/open-government-data.htm.

[248] Sami Paavola and Kai Hakkarainen. "The Knowledge Creation Metaphor – An Emergent Epistemological Approach to Learning." In: *Science & Education* 14.6 (2005), pp. 535–557. DOI: 10.1007/s11191-004-5157-0.

[249] Matteo Pastore, Maria Angela Pellegrino, and Vittorio Scarano. "Detecting and Generalizing Quasi-Identifiers by Affecting Singletons." In: *EGOV-CeDEM-ePart-**. 2020, pp. 327–335.

[250] Robin Peek. "Digital public library of America." In: *Information Today* 29.2 (2012), pp. 24–24.

[251] Maria Angela Pellegrino. "Knowledge graphs within everyone?s means." In: *CHItaly*. 2021.

[252] Maria Angela Pellegrino, Abdulrahman Altabba, Martina Garofalo, Petar Ristoski, and Michael Cochez. "GEval: A Modular and Extensible Evaluation Framework for Graph Embedding Techniques." In: *European Semantic Web Conference (ESWC)*. 2020, pp. 565–582.

[253] Maria Angela Pellegrino, Michael Cochez, Martina Garofalo, and Petar Ristoski. "A Configurable Evaluation Framework for Node Embedding Techniques." In: *The Semantic Web: ESWC 2019 Satellite Events - ESWC 2019 Satellite Events*. Portorož, Slovenia, 2019, pp. 156–160.

[254] Maria Angela Pellegrino and Mauro D'Angelo. "Engaging Children in Smart Thing Ideation via Storytelling." In: *I-CITIES*. 2021.

[255] Maria Angela Pellegrino, Luca Postiglione, and Vittorio Scarano. "Detecting Data Accuracy Issues in Textual Geographical Data by a Clustering-based Approach." In: *8th ACM IKDD CODS and 26th COMAD*. 2021, pp. 208–212.

[256] Maria Angela Pellegrino, Eftychia Roumelioti, Mauro D'Angelo, and Rosella Gennari. "Engaging Children in Remotely Ideating and Programming Smart Things." In: *Biannual Conference of the Italian SIGCHI Chapter, CHItaly*. 2021, 20:1–20:5. DOI: 10.1145/3464385.3464728.

[257] Maria Angela Pellegrino, Eftychia Roumelioti, Mauro D'Angelo, and Rosella Gennari. "Engaging Children in Remotely Ideating and Programming Smart Things." In: *14th Biannual Conference of the Italian SIGCHI Chapter (CHItaly)*. ACM, 2021.

[258] Maria Angela Pellegrino, Eftychia Roumelioti, Rosella Gennari, and Mauro D'Angelo. "Smart City Design as a 21st Century Skill." In: *Methodologies and Intelligent Systems for Technology Enhanced Learning, 11th International Conference*. Springer, 2022, pp. 271–280.

[259] Maria Angela Pellegrino, Mario Santoro, Vittorio Scarano, and Carmine Spagnuolo. "Automatic Skill Generation for Knowledge Graph Question Answering." In: *Extended Semantic Web Conference (ESWC)*. 2021.

[260] Maria Angela Pellegrino, Vittorio Scarano, and Carmine Spagnuolo. *Move Cultural Heritage Knowledge Graphs in Everyone's Pocket*. Submitted to Semantic Web Journal in Mar. 2021. 2021.

[261] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation." In: *Proc. of the Conference on Empirical Methods in Natural Language Processing*. 2014, pp. 1532–1543.

[262] Giovanni Pilato, Giorgio Vassallo, Agnese Augello, Maria Vasile, and Salvatore Gaglio. "Expert Chat-Bots for Cultural Heritage." In: *IX Convegno della Associazione Italiana Intelligenza Artificiale Proc. of. Workshop Interazione e Comunicazione Visuale nei Beni Culturali*. 2004, p. 15.

[263] Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. "Data Quality Assessment." In: *Communication* 45.4 (2002), pp. 211–218.

[264] C Christine Porter. "De-identified data and third party data mining: the risk of re-identification of personal information." In: *Law, Commerce and Technology Journal* 5 (2008), p. 1.

[265] Cynthia Putnam, Melisa Puthenmadom, Marjorie Ann Cuerdo, Wanshu Wang, and Nathaniel Paul. "Adaptation of the System Usability Scale for User Testing with Children." In: *Extended Abstracts of the Conference on Human Factors in Computing Systems*. 2020, pp. 1–7.

[266] Wu Qiongli. "Commercialization of digital storytelling: An integrated approach for cultural tourism, the Beijing Olympics and wireless VAS." In: *International Journal of Cultural Studies* 9.3 (2006), pp. 383–394.

[267] Yves Raimond. *DBTune classical*. last access March, 2021. 2007. URL: http://dbtune.org/classical/.

[268] Thomas C. Redman. *Harvard business review*. https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year. [Online]. 2016.

[269] Nico Reski and Aris Alissandrakis. "Open data exploration in virtual reality: a comparative study of input technology." In: *Virtual Reality* 24.1 (2020), pp. 1–22.

[270] Laurens Rietveld and Rinke Hoekstra. "YASGUI: Not Just Another SPARQL Client." In: *The Semantic Web: Extended Semantic Web Conference ESWC 2013 Satellite Events*. 2013, pp. 78–86.

[271] Petar Ristoski, Jessica Rosati, Tommaso Di Noia, Renato De Leone, and Heiko Paulheim. "RDF2Vec: RDF graph embeddings and their applications." In: *Semantic Web* 10.4 (2019), pp. 721–752.

[272] Bernard R Robin. "Digital storytelling: A powerful technology tool for the 21st century classroom." In: *Theory into practice* 47.3 (2008), pp. 220–228.

[273] Gianni Rodari. *Grammatica della fantasia. Introduzione all'arte d'inventare storie*. 1973.

[274] Lior Rokach and Oded Maimon. "Clustering methods." In: *Data mining and knowledge discovery handbook*. Springer, 2005, pp. 321–352.

[275] Eftychia Roumelioti, Maria Angela Pellegrino, Rosella Gennari, and Mauro D'Angelo. "What Children Learn in Smart-Thing Design at a Distance: an Exploratory Investigation." In: *Methodologies and Intelligent Systems for Technology Enhanced Learning (MIS4TEL)*. 2021.

[276] Jonathan Rowe, Lucy Shores, Bradford Mott, and James Lester. "Integrating Learning, Problem Solving, and Engagement in Narrative-Centered Learning Environments." In: *International Journal Artificial Intelligence in Education* 21 (Jan. 2011), pp. 115–133.

[277] Jennifer Rowley. "The wisdom hierarchy: representations of the DIKW hierarchy." In: *Journal of Information Science* 33.2 (2007), pp. 163–180.

[278] Janvier Rulinda, Jean de Dieu Tugirimana, Antoine Nzaramba, Felix Oduor Aila, and Gilbert Kipkirui Langat. "An Integrated Platform to Evaluate Graph Embedding." In: *International Journal of Scientific and Engineering Research* 9 (8 2018).

[279] Tony Russell-Rose and Tyler Tate. *Designing the search experience: The information architecture of discovery*. Newnes, 2012.

[280] Alistair Russell. “NITELIGHT: A Graphical Editor for SPARQL Queries.” In: *The International Conference on Posters and Demonstrations - V. 401*. 2008, pp. 110–111.

[281] Carolina Beniamina Rutta, Gianluca Schiavo, Massimo Zancanaro, and Elisa Rubegni. “Collaborative comic-based digital storytelling with primary school children.” In: *Interaction Design and Children Conference*. 2020, pp. 426–437.

[282] Lindy Ryan. *The Visual Imperative: Creating a Visual Culture of Data Discovery*. Morgan Kaufmann Publishers Inc., 2016, pp. 1–293.

[283] Mubashrah Saddiqa, Lise Lykke Le Maire Munksgaard Rasmussen, Rikke Magnussen, Birger Larsen, and Jens Myrup Pedersen. “Bringing Open Data into Danish Schools and its Potential Impact on School Pupils.” In: *Proc. of the 15th International Symposium on Open Collaboration*. 2019.

[284] Shazia Sadiq and Marta Indulska. “Open data: Quality over quantity.” In: *International Journal of Information Management* 37.3 (2017), pp. 150–154.

[285] Pierangela Samarati. “Protecting Respondents’ Identities in Microdata Release.” In: *IEEE Transactions on Knowledge and Data Engineering* 13.6 (2001), pp. 1010–1027.

[286] Elizabeth B-N Sanders and Pieter Jan Stappers. “Co-creation and the new landscapes of design.” In: *Co-design* 4.1 (2008), pp. 5–18.

[287] Nadia Sansone, Donatella Cesareni, and Maria Beatrice Ligorio. “The Trialogical Learning Approach to innovate teaching.” In: *Italian Journal of Educational Technology* 24.2 (2016), p. 82. DOI: 10.17471/2499-4324/892. URL: https://ijet.itd.cnr.it/article/view/892.

[288] Vittorio Scarano, Roberto Andreoli, Daniele Monaco, Alberto Negro, Gianluca Santangelo, and Luca Vicidomini. “Re-using Open Data by Automatically Building Immersive Virtual Reality Worlds as Personal Museums.” In: *20th Annual International Conference on Digital Government Research - DG.O*. 2019, pp. 297–305.

[289] Jesse Schell. *The Art of Game Design: A Book of Lenses*. Morgan Kaufmann Publishers Inc., 2008.

[290] Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. "Adoption of the Linked Data Best Practices in Different Topical Domains." In: *Proceedings of International Semantic Web Conference ISWC*. Springer, 2014, pp. 245–260.

[291] Alexander Schmoelz. "Enabling co-creativity through digital storytelling in education." In: *Thinking Skills and Creativity* 28 (2018), pp. 1–13.

[292] Douglas Schuler and Aki Namioka. *Participatory design: Principles and practices*. CRC Press, 1993.

[293] Scottish Governmentl. *Open access to Scotland's official statistics*. last access March, 2021. 2010. URL: https://statistics.gov.scot.

[294] LLC Securosis. "Understanding and Selecting a Data Loss Prevention Solution." In: *Securosis, LLC* (2010).

[295] Priya Seetharaman. "Business models shifts: Impact of Covid-19." In: *International Journal of Information Management* 54 (2020), pp. 102173.

[296] Edward Segel and Jeffrey Heer. "Narrative Visualization: Telling Stories with Data." In: *IEEE transactions on visualization and computer graphics* 16 (2011), pp. 1139–48. DOI: 10.1109/TVCG.2010.179.

[297] Semantic Computing Research Group (SeCo). *Mapping Manuscript Migrations*. last access March, 2021. 2017. URL: https://mappingmanuscriptmigrations.org/.

[298] Valerie Sessions and Marco Valtorta. "Towards a method for data accuracy assessment utilizing a bayesian network learning algorithm." In: *J. of Data and Information Quality* 1.3 (2009), pp. 1–34.

[299] Xinhuan Shu, Jiang Wu, Xinke Wu, Hongye Liang, Weiwei Cui, Yingcai Wu, and Huamin Qu. "DancingWords: exploring animated word clouds to tell stories." In: *Journal of Visualization* 24 (2020), pp. 1–16.

[300] Martin G. Skjæveland. "Sgvizler: A JavaScript Wrapper for Easy Visualization of SPARQL Result Sets." In: *Proc. of ESWC*. 2012, pp. 361–365.

[301] Brett Smith and Javier Monforte. “Stories, new materialism and pluralism: Understanding, practising and pushing the boundaries of narrative analysis.” In: *Methods in Psychology* 2 (2020), p. 100016.

[302] Vincent S Smith. “Data publication: towards a database of everything.” In: *BMC research Notes* 2.1 (2009), pp. 1–3.

[303] Ola Söderström, Till Paasche, and Francisco Klauser. “Smart cities as corporate storytelling.” In: *City* 18.3 (2014), pp. 307–320.

[304] Jin-yu Song, Quan Yu, and Ruo-yu Bao. “The Detection Algorithms for Similar Duplicate Data.” In: *6th International Conference on Systems and Informatics (ICSAI)*. IEEE. 2019, pp. 1534–1542. DOI: 10.1109/ICSAI48974.2019.9010154.

[305] Daniil Sorokin and Iryna Gurevych. “End-to-End Representation Learning for Question Answering with Weak Supervision.” In: *Semantic Web Challenges*. 2017, pp. 70–83.

[306] John F Sowa. “Semantic networks.” In: *Encyclopedia of Cognitive Science* (1987).

[307] Ahmet Soylu et al. “OptiqueVQS: A visual query system over ontologies for industry.” In: *Semantic Web* 9.5 (2018), pp. 627–660.

[308] Chris speed and Arthi Kanchana Manohar. “Storytelling within an Internet of Things.” In: *Interactive Storytelling*. Springer, 2010, pp. 295–296.

[309] Besiki Stvilia, Les Gasser, Michael B Twidale, and Linda C Smith. “A framework for information quality assessment.” In: *Journal of the American society for information science and technology* 58.12 (2007), pp. 1720–1733.

[310] Ann Marie Sullivan. “Cultural Heritage & New Media: A Future for the Past, 15 J. Marshall Rev. Intell. Prop. L. 604 (2016).” In: *UIC Review of Intellectual Property Law* 15.3 (2016), p. 11.

[311] *Survey on creativity - Results*. Last access 2021/05/11. 2020. URL: https://www.schooleducationgateway.eu/en/pub/viewpoints/surveys/survey-on-creativity.htm.

[312] Latanya Sweeney. *Simple Demographics Often Identify People Uniquely*. Working paper. 2000. URL: http://dataprivacylab.org/projects/identifiability/.

[313] Latanya Sweeney. "Achieving k-anonymity privacy protection using generalization and suppression." In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002), pp. 571–588.

[314] Kathryn Szoka. "A guide to choosing the right chart type." In: *IEEE Transactions on Professional Communication* PC-25.2 (1982), pp. 98–101.

[315] Casey N Ta and Chunhua Weng. "Detecting systemic data quality issues in electronic health records." In: *Studies in Health Technology and Informatics* 264 (2019), pp. 383–387.

[316] Welderufael B. Tesfay, Jetzabel Serna, and Sebastian Pape. "Challenges in Detecting Privacy Revealing Information in Unstructured Text." In: *Proceedings of the 4th Workshop on Society, Privacy and the Semantic Web - Policy and Technology (PrivOn) co-located with 15th International Semantic Web Conference*. 2016.

[317] Nutthawut Thawanthaleunglit and Kunwadee Sripanidkulchai. "Sweeper: Automated data quality processing and model generation for data classification." In: *Proceedings of the 3rd International Conference on Software and e-Business*. 2019, pp. 17–23.

[318] Katja Thoring and Roland M Müller. "Understanding the creative mechanisms of design thinking: an evolutionary approach." In: *2nd Conference on Creativity and Innovation in Design*. 2011, pp. 137–147.

[319] Qi Tian, Mengzhou Liu, Lingtong Min, Jiye An, Xudong Lu, and Huilong Duan. "An automated data verification approach for improving data quality in a clinical registry." In: *Computer Methods and Programs in Biomedicine* 181 (2019), p. 104840.

[320] Chao Tong, Richard Roberts, Rita Borgo, Sean Walton, Robert S Laramee, Kodzo Wegba, Aidong Lu, Yun Wang, Huamin Qu, Qiong Luo, et al. "Storytelling and visual-

ization: An extended survey." In: *Information* 9.3 (2018), p. 65.

[321] Yavuz Topkaya and Yakup Doğan. "The Effect of Educational Comics on Teaching Environmental Issues and Environmental Organizations Topics in 7th Grade Social Studies Course: A Mixed Research." In: *Egitim ve Bilim* 45.201 (2020).

[322] Daniel Tunkelang. "Dynamic category sets: An approach for faceted search." In: *ACM SIGIR*. Vol. 6. Citeseer, 2006.

[323] Barbara Ubaldi. *Rebooting Public Service Delivery-How can Open Government Data help drive innovation*. 2016.

[324] UNESCO. *UNESCO thesaurus*. http://vocabularies.unesco.org/. [Online] Last access May 2020. 1977.

[325] Christina Unger, Lorenz Bühmann, Jens Lehmann, Axel-Cyrille Ngonga Ngomo, Daniel Gerber, and Philipp Cimiano. "Template-based question answering over RDF data." In: *Proceedings of the 21st international conference on World Wide Web*. 2012, pp. 639–648.

[326] Ricardo Usbeck, Ria Hari Gusmita, Axel-Cyrille Ngonga Ngomo, and Muhammad Saleem. "9th Challenge on Question Answering over Linked Data (QALD-9)." In: *Proc. of the 9th Question Answering over Linked Data challenge co-located with 17th International Semantic Web Conference (ISWC)*. 2018, pp. 58–64.

[327] Svitlana Vakulenko, Javier David Fernandez Garcia, Axel Polleres, Maarten de Rijke, and Michael Cochez. "Message passing for complex question answering over knowledge graphs." In: *Proc. of the 28th ACM International Conference on Information and Knowledge Management*. 2019, pp. 1431–1440.

[328] Hernán Vargas, Carlos Buil Aranda, and Aidan Hogan. "RDF Explorer: A Visual Query Builder for Semantic Web Knowledge Graphs." In: *The Internationala Semantic Web Conference ISWC*. 2019, pp. 229–232.

[329] Hernán Vargas, Carlos Buil Aranda, Aidan Hogan, and Claudia Lópex. "RDF Explorer: A Visual SPARQL Query Builder." In: *The Semantic Web - 18th International Semantic Web Conference ISWC*. Vol. 11778. Springer, 2019, pp. 647–663. DOI: 10.1007/978-3-030-30793-6_37.

[330] Prokopia Vlachogianni and Nikolaos Tselios. "Perceived usability evaluation of educational technology using the System Usability Scale (SUS): A systematic review." In: *Journal of Research on Technology in Education* (2021), pp. 1–18.

[331] V. Vostrovsky and Jan Tyrychtr. "Consistency of Open Data as Prerequisite for Usability in Agriculture." In: *Scientia Agriculturae Bohemica* 49.4 (2018), pp. 333–339.

[332] W3C. *LinkedData*. Last access March, 2021. 2016. URL: https://www.w3.org/wiki/LinkedData.

[333] W3Consortium. *CORS enabled*. Last access 2021/06/04. 2015. URL: https://www.w3.org/wiki/CORS_Enabled.

[334] Eka Wahjuningsih, Asih Santihastuti, I Kurniawati, and UM Arifin. "Storyboard That Platform to Boost Students' Creativity: Can It Become Real?" In: *IOP Conference Series: Earth and Environmental Science*. Vol. 485. 1. 2020.

[335] Krzysztof Walczak, Wojciech Cellary, and Martin White. "Virtual museum exbibitions." In: *Computer* 39.3 (2006), pp. 93–95.

[336] Ke Wang and Benjamin C. M. Fung. "Anonymizing Sequential Releases." In: *12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*. ACM, 2006, pp. 414–423. DOI: 10.1145/1150402.1150449.

[337] Yan Wang, Hao Zhang, Yaxin Li, Deyun Wang, Yanlin Ma, Tong Zhou, and Jianguo Lu. "A Data Cleaning Method for CiteSeer Dataset." In: *Web Information Systems Engineering*. 2016, pp. 35–49.

[338] Martin White et al. "ARCO — An Architecture for Digitization, Management and Presentation of Virtual Exhibitions." In: *Computer Graphics International Conference* (2004), pp. 622–625.

[339] Ryen W White and Resa A Roth. “Exploratory search: Beyond the query-response paradigm.” In: *Synthesis lectures on information concepts, retrieval, and services* 1.1 (2009), pp. 1–98.

[340] Scott White and Padhraic Smyth. “A Spectral Clustering Approach To Finding Communities in Graph.” In: *Proc. of the SIAM International Conference on Data Mining*. Newport Beach, CA, USA, 2005, pp. 274–285.

[341] William Winkler. “String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.” In: *Survey Research Methods* (1990).

[342] Rafal Wojciechowski, Krzysztof Walczak, Martin White, and Wojciech Cellary. “Building virtual and augmented reality museum exhibitions.” In: *Proceedings of the ninth international conference on 3D Web technology*. 2004, pp. 135–144.

[343] Raymond Chi-Wing Wong, Jiuyong Li, Ada Wai-Chee Fu, and Ke Wang. “(α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing.” In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2006, pp. 754–759.

[344] Jing Wu and Der-Thanq Victor Chen. “A systematic review of educational digital storytelling.” In: *Computers & Education* 147 (2020), p. 103786.

[345] Pei-Fen Wu, Kuang-Yi Fan, and Yi-Ting Liao. “Developing and assessing the usability of digital manipulative storytelling system for school-age children.” In: *3rd International Conference on Systems and Informatics (ICSAI)*. 2016, pp. 465–470.

[346] Pei-Fen Wu, Hui-Jiun Hu, Feng-Chu Wu, and Kuang-Yi Fan. “The Evaluation on the Usability of Digital Storytelling Teaching System in Teaching.” In: *Learning and Collaboration Technologies. Technology in Education*. Springer International Publishing, 2017, pp. 473–487.

[347] Xuefang Xu, Yaguo Lei, and Zeda Li. “An Incorrect Data Detection Method for Big Data Cleaning of Machinery Condition Monitoring.” In: *IEEE Transactions on Industrial Electronics* 67.3 (2020), pp. 2326–2336.

[348] Savita Yadav, Pinaki Chakraborty, and Prabhat Mittal. “User Interface of a Drawing App for Children: Design and Effectiveness.” In: *International Conference on Innovative Computing and Communications*. Springer Singapore, 2021, pp. 53–61.

[349] Yu-Feng Diana Yang. “Multimodal composing in digital storytelling.” In: *Computers and Composition* 29.3 (2012), pp. 221–238.

[350] Syeda Sana e Zainab, Muhammad Saleem, Qaiser Mehmood, Durre Zehra, Stefan Decker, and Ali Hasnain. “FedViz: A Visual Interface for SPARQL Queries Formulation and Execution.” In: *International workshop on Visualizations and User Interfaces for Ontologies and Linked Data*. 2015.

[351] Ruojing Zhang, Marta Indulska, and Shazia Sadiq. “Discovering data quality problems.” In: *Business & Information Systems Engineering* 61.5 (2019), pp. 575–593. DOI: 10.1007/s12599-019-00608-0.

[352] Liang Zhao, Zhikui Chen, Zhennan Yang, Yueming Hu, and Mohammad S. Obaidat. “Local Similarity Imputation Based on Fast Clustering for Incomplete Data in Cyber-Physical Systems.” In: *IEEE Systems Journal* 12.2 (2018), pp. 1610–1620.

[353] Jinling Zhou, Xinchun Diao, and Jianjun Cao. “Holistic data accuracy assessment using search & scored-based bayesian network learning algorithms.” In: *3rd Intern. Conf. on Information Management*. 2017, pp. 432–436.

[354] Jack Zipes. *Creative storytelling: Building community/changing lives*. Routledge, 2013.

[355] Lei Zou, Ruizhe Huang, Haixun Wang, Jeffrey Xu Yu, Wenqiang He, and Dongyan Zhao. “Natural Language Question Answering over RDF: A Graph Data Driven Approach.” In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 2014, pp. 313–324.

[356] Anneke Zuiderwijk and Marijn Janssen. "The negative effects of open government data-investigating the dark side of open data." In: *Proceedings of the 15th Annual International Conference on Digital Government Research*. 2014, pp. 147–152.

[357] Anneke Zuiderwijk, Marijn Janssen, Geerten van de Kaa, and Kostas Poulis. "The wicked problem of commercial value creation in open data ecosystems: Policy guidelines for governments." In: *Information Polity* 21.3 (2016), pp. 223–236.