**Università degli Studi di Salerno**

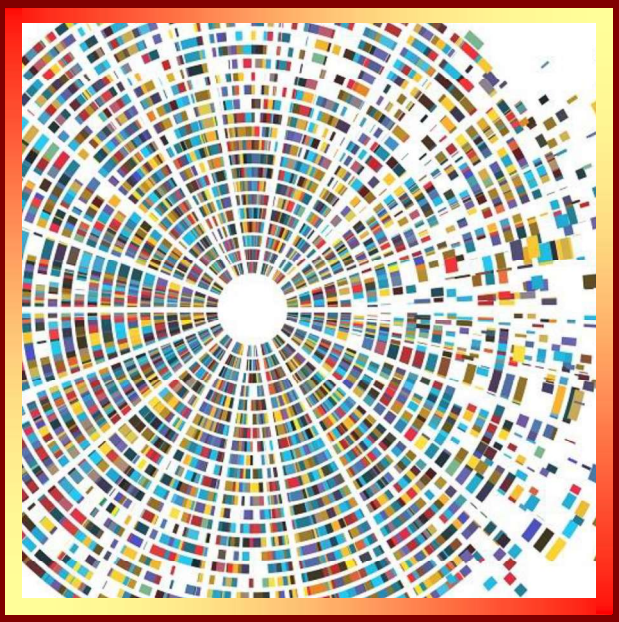
**Dipartimento di Informatica**


**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

## 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.