



# Accounting for quality in data integration systems: a completeness-aware integration approach

Cinzia Daraio<sup>1</sup> · Simone Di Leo<sup>1</sup> · Monica Scannapieco<sup>2</sup>

Received: 27 April 2021 / Accepted: 3 January 2022  
© The Author(s) 2022

## Abstract

Ensuring the quality of integrated data is undoubtedly one of the main problems of integrated data systems. When focusing on multi-national and historical data integration systems, where the “space” and “time” dimensions play a relevant role, it is very much important to build the integration layer in such a way that the final user accesses a layer that is “by design” as much complete as possible. In this paper, we propose a method for accessing data in multipurpose data infrastructures, like data integration systems, which has the properties of (i) relieving the final user from the need to access single data sources while, at the same time, (ii) ensuring to maximize the amount of the information available for the user at the integration layer. Our approach is based on a completeness-aware integration approach which allows the user to have ready available all the maximum information that can get out of the integrated data system without having to carry out the preliminary data quality analysis on each of the databases included in the system. Our proposal of providing data quality information at the integrated level extends then the functions of the individual data sources, opening the data infrastructure to additional uses. This may be a first step to move from data infrastructures towards knowledge infrastructures. A case study on the research infrastructure for the science and innovation studies shows the usefulness of the proposed approach.

**Keywords** Data and information quality · Data integrated system · Longitudinal data · Multinational data · Data infrastructures · Research infrastructures · Knowledge infrastructures

---

✉ Cinzia Daraio  
daraio@diag.uniroma1.it

Simone Di Leo  
dileo@diag.uniroma1.it

Monica Scannapieco  
scannapi@istat.it

<sup>1</sup> DIAG, Sapienza University of Rome, Rome, Italy

<sup>2</sup> ISTAT, Rome, Italy

## Introduction

In the current big data era in which we live, the problems of data integration, harmonization and above all data quality have increased rather than reduced (Ekbja et al., 2015). Paradoxically, in this context it appears more complex to identify criticalities in data and information, and profiling research infrastructures capable of showing the shortcomings and potential of the various existing data sources (Borgman, 2015). Information quality, which is more than simply accuracy, calls for an increasing interest on other significant dimensions such as completeness, consistency, and currency (Batini & Scannapieco, 2016). The quality of data is context-dependent and an appropriate quality of a single dataset, for a specific purpose, is not enough. The linkages between different datasets are relevant as well. The compatibility, interchangeability and the connectability of a given dataset with other related data are fundamental aspects which need to be taken into account (Daraio & Glänzel, 2016). Quality is also a relevant dimension, a kind of overarching principle, to keep into account when designing models of metrics (Daraio, 2017). Data integration is the activity of joining data located in diverse sources, to offer the user a cohesive outlook of these data. *Interoperability* is the consistent and meaningful exchange of information among heterogeneous databases avoiding inaccuracy and lack of coordination causing duplication of efforts and of resources (Parent & Spaccapietra, 2000). Parent and Spaccapietra (2000) distinguish between a *lowest level* of interoperability in which there is no integration and a *higher level* of interoperability with the aim of providing a comprehensive system on the existing databases, to offer the wanted level of integration, including a *intermediary level* of interoperability in which there is a certain degree of interoperability but the system does not guarantee consistency across data sources.

Different levels of interoperability have been proposed in the literature. Tolk and Muguira (2003) suggest a detailed set of levels of conceptual interoperability that goes from *isolated systems* (no integration, Level 0), to a *common conceptual model* (maximum level of integration based on semantic consistency, Level 4), including Level 1 based on the existence of documentation on data, Level 2 based on the alignment of meta data and Level 3 based on open source and dynamically aligned data.

According to the quality framework of the OECD (2011), *data quality* is defined as “fitness for use” with respect to user needs, and it has seven dimensions: (i) *relevance* grades the ability of data to address their purposes; (ii) *accuracy* measures how the data correctly describes the features they are designed to assess; (iii) *credibility* accounts for the confidence and trust of users in the data and their objectivity; (iv) *timeliness* expresses the length of time between data availability and the phenomenon described by data; (v) *accessibility* gauges how readily the data can be located and accessed; (vi) *interpretability* relates the easiness with which the user may understand and properly use and analyse the data; (vii) *coherence* refers to the degree to which data are logically connected and mutually consistent”. An important data quality aspect that is not explicitly reported in the OECD (2011) framework but very often encountered in the practical data analysis is *completeness*. For each variable, dimension and data set, completeness evaluates the number of missing values (with the meaning relevant to completeness, i.e. unavailable or temporarily unavailable) that are present.

Data quality is a very complex topic, in which the theory and practice often differ. In practice, data quality does play an important role in the design of data architectures. All the data quality efforts must start from a solid understanding of high-priority use cases, and use that insight to navigate various trade-offs to optimize the quality of the final output.

The followings are trade-offs related to data quality: Should we select data for cleaning based on the cost of cleaning effort or based on how frequently the data is used or based on its relative importance within the data models consuming it? or a combination of those factors? What sort of combination? Is it a good idea to improve data accuracy by getting rid of incomplete or erroneous data? While removing some data, how do we ensure that we do not introduce distortions or bias? Data integration systems are often the result of a huge effort that has to be paid to integrate highly heterogeneous data sources: schema harmonization, record linkage and historical data management are only some of the most common activities that these systems require in real application scenarios. Among such activities, ensuring the quality of integrated data is undoubtedly one of the main problems of integrated data systems. To address the quality problem some shared practices are there: for instance, ensuring data consistency at the integration layer is a mandatory approach in any sound data integration systems. However, when it comes to data completeness, different solutions are possible, depending also on the “completeness” requirement by the users: if it is reasonable to say that no user would like to have inconsistent data, instead different degrees of completeness can be made available depending on how the data integration layer is built. When focusing on multi-national and historical data integration systems, where the “space” and “time” dimensions play a relevant role, it is very much important to build the integration layer in such a way that the final user accesses a layer that is “by design” as much complete as possible. In this paper we address the relevance and challenges of the characterization of quality in a longitudinal and multinational data integration system. We propose a data quality approach, based on the maximization of the available information at the level of integrated infrastructure, that could be the first step, towards the building of a knowledge infrastructure.

The paper unfolds as follows. In the next "[Aim and contribution](#)" section we describe the main goal of the paper and its contribution to existing literature. "[Related studies](#)" Section outlines existing studies related to the topic addressed in the paper while "[Method](#)" section describes the proposed methodology. "[Case study](#)" section illustrates the case study on the RISIS data integrated system, while "[Discussion and conclusions](#)" section discusses the main results and concludes the paper.

## **Aim and contribution**

The aim of this work is to propose a method to characterize the quality of the information contained in a multipurpose data infrastructure characterized by historical and multinational heterogeneous data systems. We propose an approach that investigates the integration level of the overall system and is based on a completeness-aware method for maximizing the amount of information available in a data integration system. We choose completeness with respect to the target coverage defined by the integration layer because it is a fundamental data quality property that should be checked and on which we can build further to extend the functionality of existing data systems integrated in a data infrastructure. The aim of this investigation at the integrated level is to highlight opportunities of data harmonization and exploitation that were not available to the potential user of individual databases before. This investigation offers additional relevant information to the user and extends the functions of the individual data sources, opening the data infrastructure to additional uses not foreseen by the single data systems. Our approach may be considered as a first step from data infrastructures towards knowledge infrastructures.

Our proposal is focused on completeness, being it a data quality dimension that plays a relevant role in data integration systems of the type considered in this paper, namely historical and multinational ones. From an intuitive perspective, this is because both time and space gaps can be present in data, but by making them explicitly represented, the final user can profitably use this “awareness” for her data needs. A similar approach could, in principle, be applied for other data quality dimensions. Let’s think for instance to data accuracy: dealing with data accuracy at the integrated layer can be done, for instance, by making the integrated layer resulting from the highest accuracy “pieces” of data extracted from the sources.

The existing literature on this topic, namely the analysis of the quality of the integrated system built on historical and multinational sources, is scant. However these systems exhibit a significant complexity: multi-nationality is typically characterized by high heterogeneity, while historical data imply that time consistency is carefully checked and ensured at the integration layer. We believe that the development of this approach may be of considerable importance, not only from a scientific point of view but also from an applied perspective, as it allows us to provide additional functionality indications for users of the integrated data system.

The methodology proposed will be applied in a case study on data coming from the platform on research, higher education and innovation, maintained and developed within the European project H2020 Research Infrastructure for the Science and Innovation Studies (RISIS). We will show the importance of considering data and information quality at the integrated level as an ingredient to move from a data infrastructure to a knowledge infrastructure.

The contribution that this work offers consisting in a data quality analysis that will be developed on the integrated level of the data infrastructure sources, provides a set of information available to data users to decide which variables and levels of analysis present higher levels of quality and under what conditions of use.

## Related studies

The literature on the analysis of the quality of the integrated system built on historical and multinational sources is limited.

Quality-driven data integration systems are data integration systems that return an answer to a global query posed on the integrated layer by explicitly taking into account the quality of data provided by local sources. Some relevant examples of such systems are briefly described below:

- *FusionPlex* (Motro & Anokhin, 2005) is a data integration system assuming instance inconsistency, meaning that the same instance of the real world can be represented differently in the various local sources due to errors. In order to deal with such instance-level inconsistencies, Fusionplex introduces a set of quality metadata, called features, about the sources to be integrated.
- *DaQuinCIS* (Scannapieco et al., 2004) is a framework with an underlying data integration system where the sources are characterized by quality metadata that are exploited in the query answering phase. User queries, posed to the integration layer, are processed so that the “best quality” answer is returned as a result, i.e. when retrieving data

from the sources, data are compared and a best quality copy can be either selected or constructed.

- *QP-alg* (Naumann et al., 1999) specifies the mapping between local sources and the global schema is specified by means of query correspondence assertions (QCAs). Three classes of data quality dimensions, called information quality criteria (IQ criteria), are defined: Source-specific criteria, defining the quality of a whole source, QCA-specific criteria, defining the quality of specific QCAs, User-query specific criteria, measuring the quality of the source with respect to the answer provided to a specific user query. These criterias are used in the query answering phase.

Differently from the above cited systems, our approach does not base the query answering on quality metadata specified as part of the data integration system, instead the integration layer is built *by-design* to maximize the completeness. A detailed description of the proposed approach is reported in "[A completeness-aware integration approach](#)" section. It is based and uses an ontology-based data management (OBDM) approach described at length in "[Introduction on OBDM](#)" section. Lenzerini and Daraio (2019) discuss the main challenges, approaches and solutions available for integrating data on research, higher education and innovation, consolidating existing research on the topic, including Daraio et al. (2016a) which introduce *Sapientia*, the ontology of multidimensional assessment of research and Daraio et al. (2016b) that highlighted and discussed the main advantages on an OBDM approach residing in the openness, interoperability and data quality. Recently, Angelini et al. (2020) showed the usefulness of *Sapientia* and OBDM combined with visual analytics to develop general models of performance indicators.

## Method

In this section, we first illustrate the proposed method and later we present the RISIS case study that shows the application of the method to a real case. In particular, we will describe our proposal for building a data integration system with explicit quality annotations. We will first give an overview of the used data integration approach, namely OBDM; then, we will focus on our proposal to explicitly represent data quality of the integration layer, so to have a full governance of the quality of the data provided by the data integration system.

### Introduction on OBDM

Ontology-based data management was introduced about a decade ago as a new way for modeling and interacting with a collection of data sources (see Lenzerini, 2011). According to such paradigm, the client of the information system is freed from being aware of how data are structured in concrete resources (databases, software programs, services, etc.), and interacts with the system by expressing her queries and goals in terms of a conceptual representation of the domain of interest, called ontology.

More precisely, an OBDM system is an information management system maintained and used by a given organization (or, a community of users), whose architecture has the same structure of a typical data integration system, with the following components: an Integration Layer with an ontology, a Source Layer with a set of data sources, and the mapping between the two (see Fig. 1). In particular (see Aracri et al., 2018):

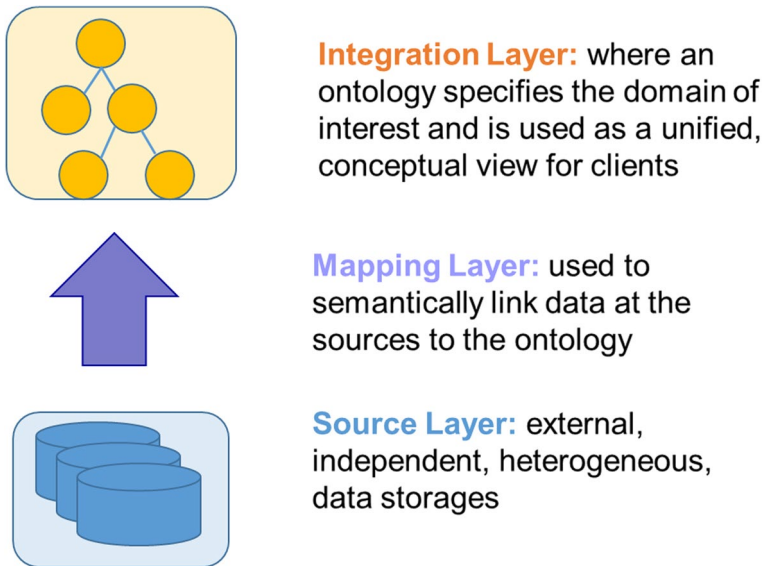


Fig. 1 OBDM layers

- *Integration Layer*, with an ontology, i.e. a conceptual, formal and shared representation of the domain of interest of the organization. An ontology includes concepts, attributes of concepts, relationships between concepts, and logical assertions formally describing the domain knowledge.
- *Source Layer*, where there are data sources, storing data by the organization. In the general case, such databases are numerous, heterogeneous, each one managed and maintained independently from the others.
- *Mapping Layer*, with the mapping as a precise specification of the correspondence between the data contained in the data sources and the elements of the ontology, i.e. concepts, attributes, or relationships.

The three layers above described define a knowledge representation system that can be managed and reasoned upon with the help of automated reasoning techniques. For example, queries expressed over the ontology can be processed by means of sophisticated algorithms automatically translating the query in terms of the data sources using the mapping (Calvanese et al., 2007). Among the many services that an OBDM system is able to provide, data quality assessment (Batini & Scannapieco, 2016) is one notable example.

### A completeness-aware integration approach

When integrating data sources being multi-national and historical, a relevant dimension to consider is the *completeness with respect to the target coverage defined by the integration layer*.

We have then to introduce a new concept of *completeness* with respect to a coverage target defined at the integration level. This target can be not fully reached by integrating the sources and is in general dependant on the way in which the sources are integrated.

Assuming that we would like to have an integrated system in which the completeness of the data available to the final users is maximized, we can reason on building the integrated system with this target in mind, as explained below.

Two intuitive examples of completeness are *geographical completeness* and *time completeness*.

Let the *Source Layer* be a set of data sources  $\{S_1 \dots S_N\}$ . Let us suppose that each source provides a set of (relational) tables, i.e.  $S_1 = \{R_{11} \dots R_{1k}\} \dots S_N = \{R_{n1} \dots R_{nk}\}$ .

Let the *Integration Layer* be defined as a set of relational tables  $\{I_1, \dots, I_m\}$ .

Let us assume, for the sake of simplicity and without loss of generality, that we are in a setting with only two sources, each one consisting of one relational table, namely:  $S = \{S_1, S_2\}$ , with  $S_1 = \{R_{11}\}$  and  $S_2 = \{R_{21}\}$ .

Let us also assume, without loss of generality, that both  $R_{11}$  and  $R_{21}$ , in the following referred to respectively as  $R_1$  and  $R_2$  for the sake of simplicity of the notation, have one single attribute for the territorial dimension  $A_{\text{territory}}$  (e.g. country) and one single attribute for the temporal dimension  $A_{\text{time}}$  (e.g. year), and similarly  $R_2$ .

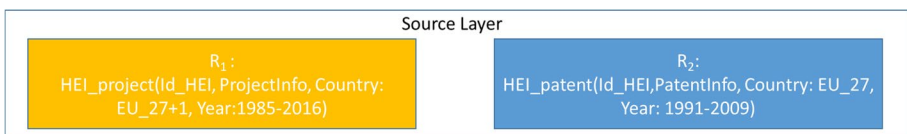
**Example 1** Looking at Fig. 2, the source level  $S$  consists of  $R_1$ , related to research and development projects of Higher Educational Institutions (HEIs) and of  $R_2$ , related to patents released by of HEIs. For  $R_1$ ,  $A_{\text{territory}} = \text{Country} = \text{EU}_{-27+1}$  (meaning EU countries plus UK), while for  $R_2$ ,  $A_{\text{territory}} = \text{Country} = \text{EU}_{-27}$ .

Instead, for  $R_2$ ,  $A_{\text{time}} = \text{Year} = 1985-2016$  (meaning the interval of years from 1985 to 2016) while for  $R_2$ ,  $A_{\text{time}} = \text{Year} = 1991-2009$ .

In this setting, the Integration Layer can be defined in order to take explicitly into account the completeness dimension, in order to give the final users the possibility to access to information at the integration layer by maximizing the amount of information they can access.

To such a scope, the Integration Layer will be composed by a set of relations  $\{I, I_1, I_2\}$ , such that:

- (1)  $I = (R_1 \cap R_2)_{\text{territory} \cap \text{time}}$  that (i) will consists of all the tuples present in both  $R_1$  and  $R_2$ , and (ii) will have  $A_{\text{time}}$  and  $A_{\text{space}}$  defined on the intersection of the domains of the two attributes in the originating sources, namely  $R_1(A_{\text{time}}) \cap R_2(A_{\text{time}})$  and  $R_1(A_{\text{space}}) \cap R_2(A_{\text{space}})$ .
- (2)  $I_1 = (R_1 - R_2)$  that (i) will consists of all the tuples present in  $R_1$  but not in  $R_2$ , and (ii) will have  $A_{\text{time}}$  and  $A_{\text{space}}$  defined as in  $R_1$ .
- (3)  $I_2 = (R_2 - R_1)$  that (i) will consists of all the tuples present in  $R_2$  but not in  $R_1$ , and (ii) will have  $A_{\text{time}}$  and  $A_{\text{space}}$  defined as in  $R_2$ .



**Fig. 2** Example of source layer in a data integration system with specific space–time features

**Example 2** Looking at Fig. 3:

- $I$  results from the same HEIs (i.e. those with the same identifiers) shared by  $R_1$  and  $R_2$ . In addition, the domain of  $A_{time}$  is intersection of the 2 years intervals [1985, 2016] and [1991–2009] and that of  $A_{space}$  is the intersection of EU\_27 and EU\_27 + 1.
- $I_1$  consists of all the tuples of  $R_1$  that are not in  $R_2$  and the domains of  $A_{space}$  and  $A_{time}$  are the same of  $A_{space}$  and  $A_{time}$  in  $R_1$
- $I_2$  consists of all the tuples of  $R_2$  that are not in  $R_1$  and the domains of  $A_{space}$  and  $A_{time}$  are the same of  $A_{space}$  and  $A_{time}$  in  $R_2$ .

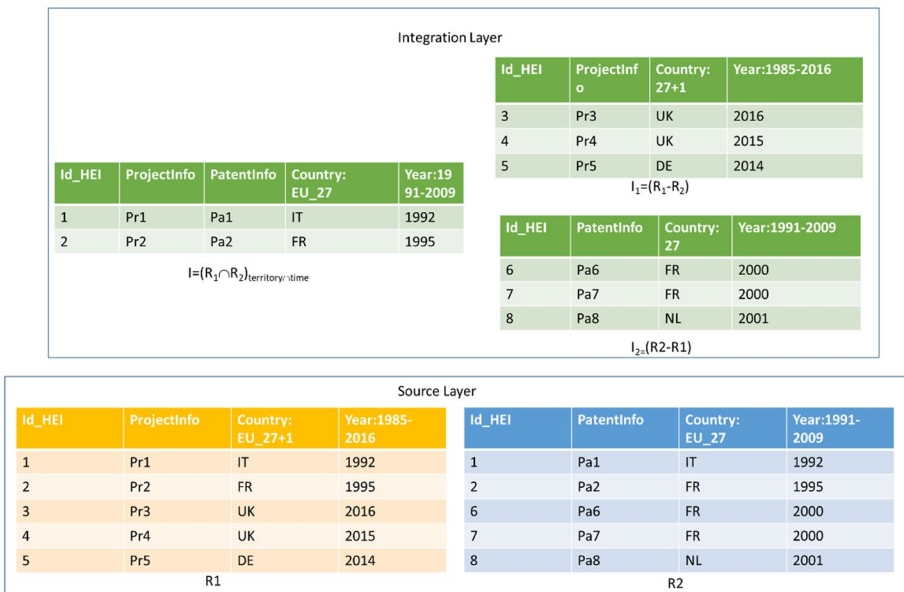
We can now define the *notion of completeness of the integration layer* as:

- $I\_Completeness$  This is the notion of completeness that provides the highest information value on a specific entity with a *given space–time view*.  $I\_completeness$  is maximum when the user queries the  $I$  relation of the Integration layer.

In the example in Fig. 3, if the user is interested to have all the information that the sources have on HEIs, then she has to access to the relation  $I$ , which indeed have both ProjectInfo and PatentInfo of HEIs.

- $S\_Completeness$  This is a notion of completeness that provides the highest information value on a specific entity with a *given attribute selection of  $S_1$  (respectively  $S_2$ )*.  $S\_Completeness$  is maximum when the user accesses  $I \cup I_1$  (respectively  $I \cup I_2$ ).

In the example in Fig. 3, if the user is only interested to ProjectInfo of HEIs, by querying both relations  $I$  and  $I_1$ , she is able to obtain ProjectInfo for all the HEIs of  $S_1$ .



**Fig. 3** Instances of relations at the *Source Layer* and at the *Integration Layer*



*Note 1* We focus on the space and time attributes of the sources as they are the ones that are typically and mandatorily shared by the sources; indeed, in order to perform a proper integration it is necessary to define the space–time scope of the population underlying the integrated datasets. Of course, it can be the case that other attributes are shared by the sources. In such a case, the shown approach can easily be extended to such attributes as well.

*Note 2* The notion of S\_Completeness allows characterizing completeness of a source at the integration layer. The question could arise: why not accessing the source directly at the source layer? The answer is: because the user will see only the integration layer and will benefit from an homogeneous representation of all the data at the sources according to a common global representation.

## Case study

### RISIS infrastructure

Research infrastructure for research and innovation policy studies (RISIS) is an infrastructure for science, technology and innovation (STI) studies (<https://www.risis2.eu/>). The databases included in the RISIS infrastructure are listed below.

- *Cheetah*, is a database featuring geographical, industry and accounting information on three cohorts of mid-sized firms that experienced fast growth during the periods 2008–2011, 2009–2012 and 2010–2013
- *The CIB/CinnoB—Corporate Invention and Innovation Boards*, is a database about the largest R&D performers and their subsidiaries worldwide, providing patenting and other indicators.
- *The CWTS publication database*, is a full copy of Web of Science (WoS) dedicated to bibliometric analyses, with additional information e.g. on standardised organisation names and other enhancements.
- *ESID*, is a comprehensive and authoritative source of information on social innovation projects and actors in Europe and beyond.
- *EUPRO*, is a unique dataset providing systematic and standardized information on R&D projects of different European R&D policy programmes.
- *JoREP 2.0*, is a database on European trans-national joint R&D programmes, storing a basic set of descriptors on the programmes and agencies participating in the programmes.
- *Mobility survey of the higher education sector (MORE)*, is a comprehensive empirical study of researcher mobility in Europe.
- *The Nano S&T dynamics database (Nano)*, collects publications and patents between 1991 and 2011 about Nano S&T.
- *ProFile*, is a longitudinal study focusing on doctoral candidates and their postdoctoral professional careers at German universities and funding organisations.
- *RISIS Patent*, offers an enriched and cleaned version of the PATSTAT database, with a focus on standardised organisation names and geolocalisation.
- *RISIS-ETER*, represents an extension by additional indicators in terms of research activities of the European Tertiary Education Register database.

- *Science and innovation policy evaluations repository (SIPER)*, is a rich and unique database and knowledge source of science and innovation policy evaluations world-wide.
- *VICO*, is a database comprising geographical, industry and accounting information on start-ups that received at least one venture capital investment in the period 1998–2014.

Besides the databases of RISIS, we considered also the public facility *OrgReg*, used by the RISIS project for the harmonization of the various institutions in the various databases. *OrgReg* (<https://risis-eter.orgreg.joanneum.at/about/data-download>) is a public facility, which provides a comprehensive register of public-sector research and higher education organizations in European countries. *OrgReg* covers organizations that are not exclusively market-oriented in all 27 + 1 (UK) European Union member states, EEA-EFTA countries (Iceland, Liechtenstein, Norway and Switzerland), as well as candidate countries (FYRM, Montenegro, Serbia and Turkey). It is a public resource whose main function is to allow integrating different RISIS datasets at the level of actors through the definition of a common list of organizations and the use of organizational IDs (*OrgReg\_Id*) that are used consistently in the RISIS datasets providing data at the level of organizational actors. Private (market-oriented) organizations are covered by parallel firms register (*FirmReg*).

## Experimental validation of the approach

The proposed methodology was applied to some of the RISIS project datasets presented above. In particular, we focus on databases containing Higher Education Institution (HEIs)'s information, though the approach is general enough to be applied to other databases as well.

A conceptual integration scheme for HEIs is available in Appendix 1. To facilitate the reading of the schema (Fig. 13), Appendix 1 reports in Fig. 12. the legend of the Graphol language including predicate and constructor nodes (Console et al., 2014; Lembo et al., 2016, 2018) used to model the domain. Details of the used datasets are given below:

- *ETER* (see Appendix 2, Fig. 14 shows the number of organizations in *ETER* by Year and Country), taking all the institutions' information in the dataset for the period 2011–2017 (full temporal coverage of the dataset). All institutions with *org\_Id* within *ETER* are mapped geographically (the *ETER\_Countries* entity in the scheme in Appendix 2). For more precise information, the geographical coverage of the data used is EU 27, UK, Montenegro, Albania, Serbia, Norway, Iceland, Turkey, Lichtenstein, Macedonia. The used dataset contains:
  - 17,652 record
  - 3205 Organizations with ID
- *CWTS*, provided by the Centre for Science and Technology Studies, includes data on academic publications from 2011 to 2017 from different countries in the world. The used dataset contains:
  - 10,086,029 records
  - 4,478,874 Unique articles
  - 3579 Organizations with ID

– *RISIS PATENT* (see Appendix 2, Fig. 15 presents the number of Patents mapped in RISIS Patent by Year and Fig. 16 shows the number of Institutions mapped in RISIS Patent by Year), thanks to the support of the Université Paris-Est Marne-la-Vallée (UPEM), with data on patents from 2011 to 2016 (last year of reference in RISIS Patent dataset). The dataset used for the case study contains:

- 57,114 records
- 1471 Organizations with ID
- 48,666 Patents
- 37 countries
- 0 null rows “orgId” in the dataset

**RISIS ETER and CWTS integration**

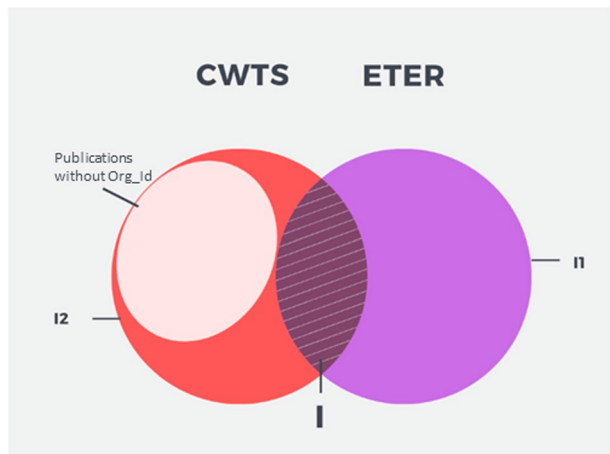
This integration task combines HEIs with related publications.

Starting from the source layers, data integrations have been performed following the methodology presented in "Aim and contribution" section (see Appendix 1 for the results), considering  $R_1$  as ETER and  $R_2$  as CWTS:

- (1)  $I = (R_1 \cap R_2)_{\text{territory} \cap \text{time}}$  = Creation of the intersection of org\_Id and years between datasets  $R_1$  and  $R_2$
- (2)  $I_1 = (R_1 - R_2)$  = Creation of the subtraction table between one dataset versus the other referenced in the previous operation will have  $A_{\text{time}}$  and  $A_{\text{space}}$  defined as in  $R_1$  ( $R_1 = \text{ETER}$ ).
- (3)  $I_2 = (R_2 - R_1)$  = Creation of the subtraction table between a dataset compared to the other dataset referred to in the previous operation will have  $A_{\text{time}}$  and  $A_{\text{space}}$  defined as in  $R_2$  ( $R_2 = \text{CWTS}$ ). See Fig. 4.

From this data, applying the methodology described above, the following results were obtained:

**Fig. 4** Representation of the integration scheme of ETER and CTWS



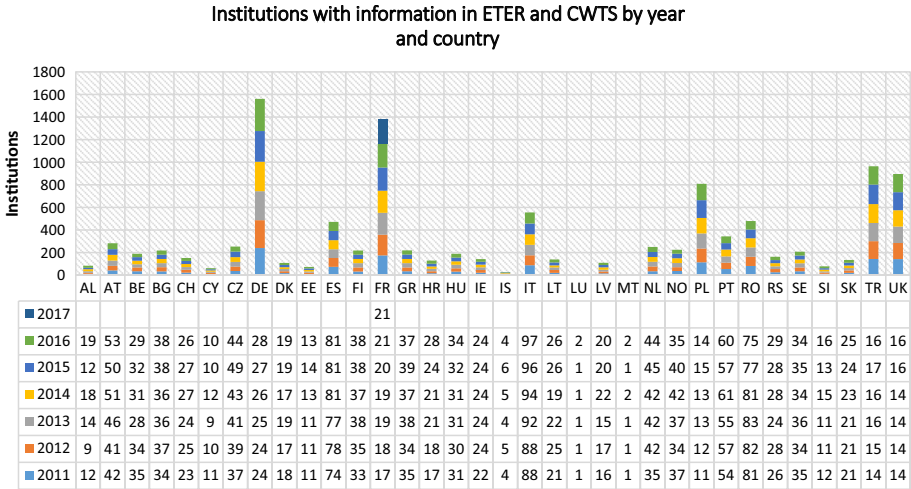


Fig. 5 Number of institutions in  $I$  by year and country (Institutions with information in ETER and CWTS by year and country)

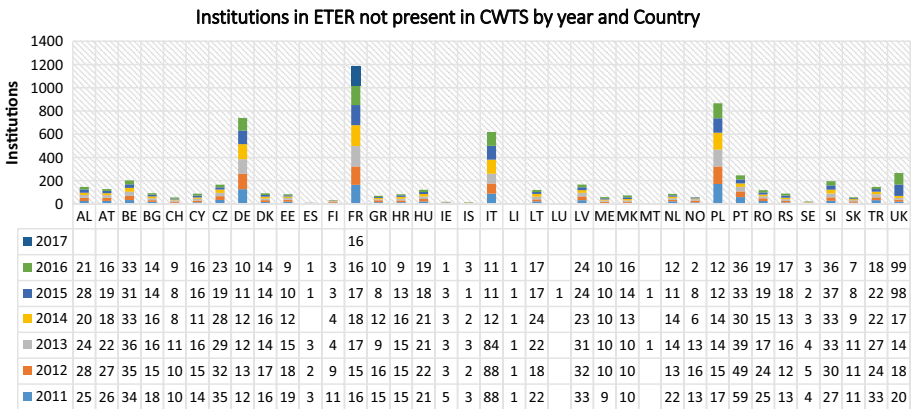


Fig. 6 Number of institutions in  $I_1$  by year and country Relation  $I_2$  (Institutions in ETER not present in CWTS by year and country)

– Relation  $I$ :

- The relation  $I$  has 6,429,051 records, which corresponds to the number of publications with information about the referenced institutions;
- $I$  contains 3,451,451 different articles, 2199 unique organizations from 34 different Countries. See Fig. 5.

– Relation  $I_1$ :

- The relation  $I_1$  has 6522 records, which correspond to the number of institutions in ETER without information in CWTS (for the period 2011–2017);

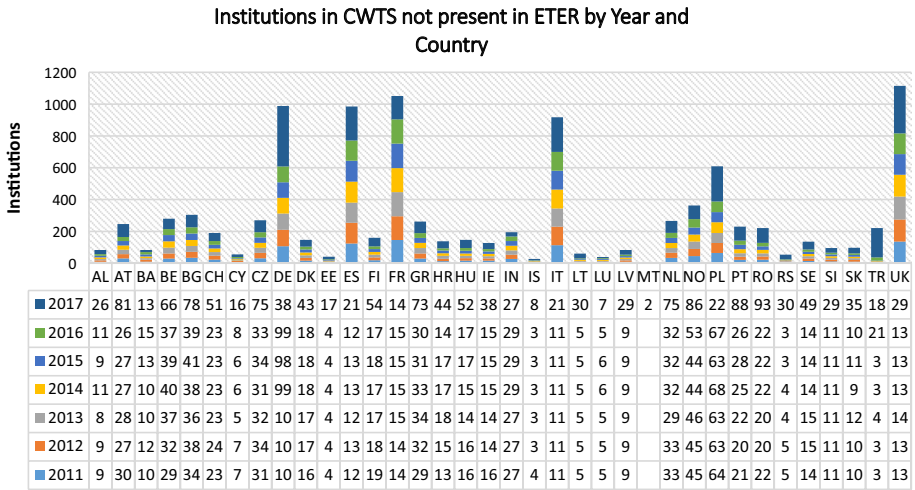


Fig. 7 Number of institutions in  $I_2$  by year and country (Institutions in CWTS not present in ETER by year and country)

Table 1 Number of articles in CWTS without Org\_Id value by year

Year	Articles without Org_Id
2011	20373
2012	20506
2013	21322
2014	23859
2015	24911
2016	26637
2017	26747
Total	164355

- $I_1$  contains information on 1006 different institutions from 37 different countries. See Fig. 6.
- In the  $I_2$  Domains there are 3,492,623 publications without Org\_Id in a certain Year for a certain institution.
- $I_2$  articles come from 1380 institutions not mapped in CWTS but not in ETER (institutions within the ETER Countries group). See Fig. 7.

In addition to these results, 164,355 records from CWTS without Org\_Id, i.e. unmapped. See Table 1.

Thanks to this approach, it is possible to highlight the completeness of the information. In each relation ( $I, I_1$  and  $I_2$ ) the  $I\_Completeness$  is equal to 1, and specifically for  $I$ , the relation has the complete information from two different sources.

The opposite approach to the proposed one involves the use of a single union table between the information in ETER and CWTS, which is composed of 10,092,551 rows.

Considering the total number of rows with complete information (6,429,051 rows) and the total rows of the report, we can calculate the  $I\_Completeness$ :  $\frac{6429051}{10092551} = 0.64$ .

This shows the relevance of our approach in maximizing the completeness and relieving the final users from receiving partially empty tables as results of their queries.

### RISIS ETER and RISIS PATENT integration

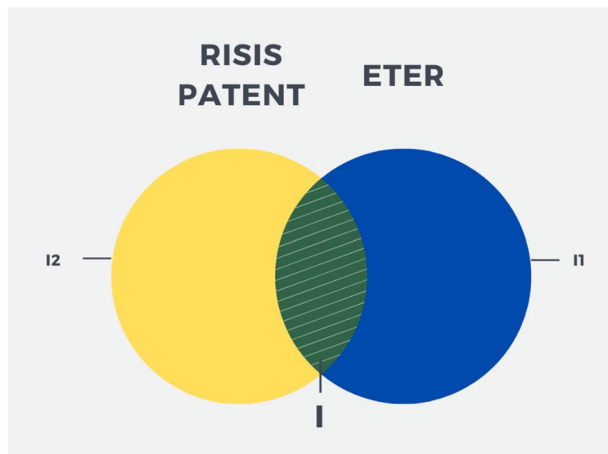
Starting from the source layers, data integrations have been performed following the methodology presented in "Aim and contribution" section (see Appendix 1 for the results), considering  $R_1$  as ETER and  $R_2$  as RISIS Patents:

- (1)  $I = (R_1 \cap R_2)_{territory \cap time}$  = Creation of the intersection of Org\_Id and years between datasets  $R_1$  and  $R_2$  (2011–2016).
- (2)  $I_1 = (R_1 - R_2)$  = Creation of the subtraction table between one dataset versus the other referenced in the previous operation will have  $A_{time}$  and  $A_{space}$  defined as in  $R_1$  ( $R_1 = ETER$ ).
- (3)  $I_2 = (R_2 - R_1)$  = Creation of the subtraction table between a dataset compared to the other dataset referred to in the previous operation will have  $A_{time}$  and  $A_{space}$  defined as in  $R_2$  ( $R_2 = RISIS Patents$ ). See Fig. 8.

From this data, applying the methodology described above, the following results were obtained:

- Relation  $I$ :
  - The relation  $I$  has 32,027 records, which correspond to the Institutions with information about the Patents.
  - $I$  contains 30,165 different patents, 829 unique organizations from 33 different Countries over the period 2011–2016 (RISIS patent last year of reference). See Fig. 9.

**Fig. 8** Representation of the integration scheme of ETER and RISIS PATENT



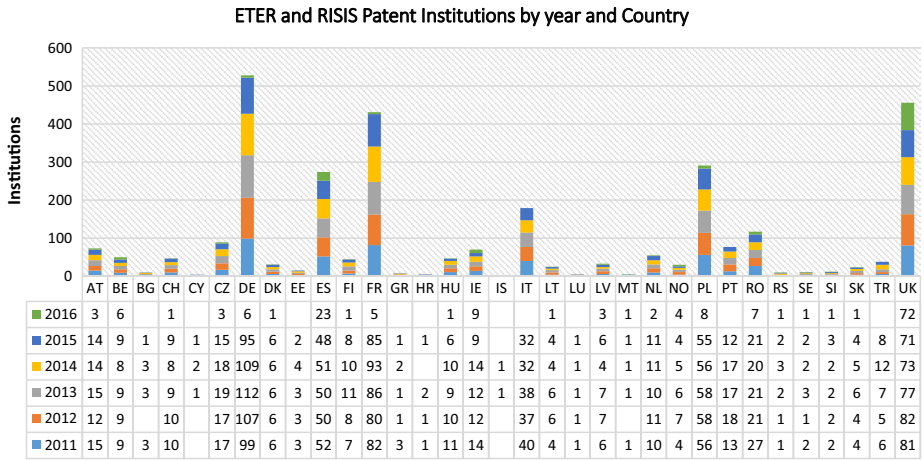


Fig. 9 Number of institutions in  $I$  by year and country (ETER and RISIS Patent Institutions)

– Relation  $I_1$

- The relation  $I_1$  has 14,177 records (for the period 2011–2016);
- $I_1$  contains information on 3060 different institutions. See Fig. 10.

– Relation  $I_2$

- In the  $I_2$  Domains there are 25,087 projects id without Org\_Id Linked in ETER in a certain year for certain institutions.
- $I_2$  refers to 664 institutions mapped in RISIS Patents but not in ETER (institutions within the ETER Countries group). See Fig. 11.

### Institutions in ETER not present in RISIS Patent by year and Country

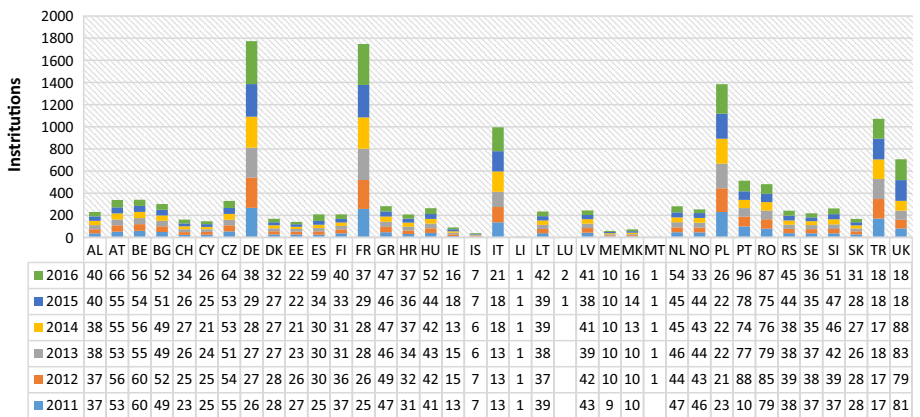


Fig. 10 Number of institutions in  $I_1$  by year and country (Institutions in ETER not present in RISIS Patent by year and Country)

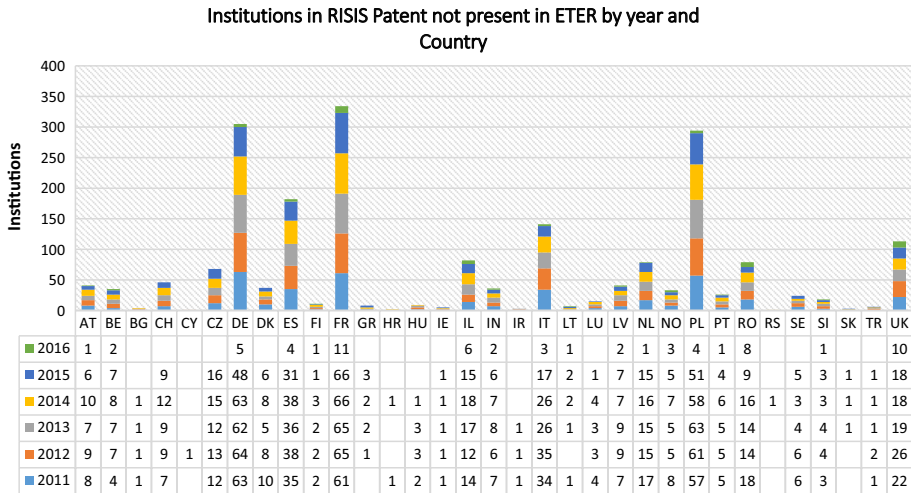


Fig. 11 Number of institutions in  $I_1$  by year and country (Institutions in RISIS Patent not present in ETER)

Thanks to this approach, it is possible to highlight the completeness of the information. In each relation ( $I$ ,  $I_1$  and  $I_2$ ) the  $I\_Completeness$  is equal to 1, and specifically for  $I$ , the relation has the complete information from two different sources for the period 2011–2016.

An approach alternative to the proposed one could involve the use of a single union table between the information in ETER and RISIS Patent, which is composed of 71,291 rows.

Considering the total number of rows with complete information (32,027 rows) and the total rows of the report, we can calculate the  $I\_Completeness$ :  $\frac{32027}{71291} = 0.45$

Hence, in this alternative approach the completeness value would be quite low.

### Impact on user

Thanks to the results above, it is possible to highlight how the use of the proposed methodology has a considerable impact on the user. By dividing the information into sub-relationships  $I$ ,  $I_1$  and  $I_2$ , the information content is maximized with an  $I\_Completeness = 1$  for each relation. The high value of completeness allows the user to know even before each query the possible amount of partial or complete information available.

Besides, the proposed approach moves the workload of finding unlinked values, incomplete information or other data cleaning operation from the user to the database manager, so that it makes easier access to the data for the user.

The opposite approach to the one proposed shows, instead, that there is a higher workload in data checking and cleaning operations by the user and that the user has no prior knowledge of the complete information contained in the dataset, but must necessarily analyze the dataset obtained from this perspective.

Evidence for these claims is shown below by contextualizing them in the results of the two case studies shown above. In the case of CWTS and ETER, it is possible to estimate that only 64% of the rows are complete. As a consequence, the user, once obtained the



dataset, will have to analyze and/or eliminate the remaining 3,663,500 rows. In the case of RISIS Patent and ETER, the situation is even more interesting. The results show that 45% of the rows are complete, leading the final user to manipulate, according to his needs, 39,264 rows, about 55% of the total.

It is important to specify that the proposed results and numbers may be subject to errors due to the quality of the dataset used. In particular, it is conceivable the presence of HEIs not mapped in ETER but mapped in the RISIS patent and CWTS datasets as these datasets contain also HEIs that are not universities.

## Discussion and conclusions

The consideration of the quality of data is an extremely important and current topic in the contemporary big data era, characterized by the paradox of the ever greater increase of available data which, however, are not accompanied by an adequate development of techniques capable of providing more information for users. Indeed, users are often overwhelmed by data and are unable, except with extreme difficulty and after several data cleaning and harmonization works, to understand what information is actually available for their empirical analyses. The fact that users are overloaded with excess data is often overlooked by data infrastructure managers who continue to add heterogeneous sources without considering their consistency and completeness.

In this paper, we propose an approach to account for quality in data integration systems. It is a completeness-aware integration approach that works at the integrated system level. The case study illustrated on European Higher Education Institutions data (included in the ETER database), integrated with bibliometric data (coming from the CWTS database) and patent data (included in the RISIS Patents database), shows the importance of the proposed approach for providing data with high level of completeness, relieving final users from the need to post-processing data in order to have adequate levels of data quality.

The proposed data quality approach offers different potentialities beyond the case study illustrated in the previous section that we briefly report below.

- (i) Designing information quality-aware methods at the integrated system level.

We proposed a data quality approach led by the maximization of information available at the integrated system layer. Our integration approach is led by the maximization of completeness at the integrated layer and can be further extended to other data quality dimensions and applied to different databases. In particular, accuracy and currency look like the most interesting quality dimensions to consider. In addition, the issue of how combining the different quality dimensions at the integrated level is worthy to be investigated in future research.

- (ii) Putting the users' needs at the center of the scene providing useful knowledge.

We proposed a *user oriented* approach that permits to reduce the workload in data checking and cleaning operations of the user and that allows the user to grasp the knowledge about the overall information available without any prior operations on the data contained in each dataset. Our approach moves the workload of finding unlinked values, incomplete information or other data cleaning operation from the user to the database manager, so that it makes easier accessing the relevant information for the user.

## (iii) A first step from data infrastructure to knowledge infrastructure.

Our approach is able to contribute to the extension of the individual data sources functions, opening the data infrastructure to additional uses. This may be a first step to move from data infrastructures towards knowledge infrastructures, whereby considering data and information quality at the integrated level play a relevant role.

The management of data at the integrated level is part of data governance and should include also a certain data literacy (Koltay, 2016). Most data can in principle be considered as infrastructural resources, as they are “shared means to many ends” that satisfy all three criteria of infrastructure resources highlighted by Frischmann (2012): 1. *Data are non-rivalrous goods* that can be consumed in principle an unlimited number of times. While it is widely accepted that social welfare is maximised when a pure rivalrous good is consumed by the person who values it the most, and that the market mechanism is generally the most efficient means for rationing such goods and for allocating resources needed to produce such goods, this is not always true for non-rivalrous goods (Frischmann, 2012). Social welfare is not maximised when the good is consumed only by the person who values it the most, but by everyone who values it. Maximising access to the non-rivalry good will in theory maximise social welfare, as every additional private benefit comes at no additional cost.

2. *Data are capital goods—Data are not a consumption good, or an intermediate good.* In most cases, data can be classified as capital goods. The System of National Accounts (SNA, 2008) defines a consumption good or service as “one that is used [...] for the direct satisfaction of individual needs or wants or the collective needs of members of the community”.

3. *Data are general-purpose inputs.* As Frischmann (2012) explains, “infrastructure resources enable many systems (markets and non markets) to function and satisfy demand derived from many different types of users”. They are not inputs that have been optimised for a special limited purpose, but “they provide basic, multipurpose functionality”. Data may often be collected for a particular purpose, and in the case of personal data the ex-ante specification of the purpose. However, there is theoretically no limitation on what purposes data can be used for, and in fact many of the benefits of data sharing arise from the reuse of data in ways that were or could not be anticipated when the data were collected. In addition, the reuse of data created in one domain may lead to further insights when applied in another. Edwards (2010) defined knowledge infrastructures as “robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds.” Nielsen (2012) argues that we are living at the dawn of the most dramatic change in science in more than 300 years. This change is being driven by powerful new cognitive tools, enabled by the internet, which are greatly accelerating scientific discovery. In his book “Reinventing Discovery” Nielsen describes an unprecedented new era of networked science. According to OECD (2015b), open data are “data that can be used by anyone without technical or legal restrictions. The use encompasses both access and reuse.” OECD (2015b, p. 7). According to OECD (2015b), open science refers to “efforts by researchers, governments, research funding agencies or the scientific community itself to make the primary outputs of publicly funded research results—publications and the research data—publicly accessible in digital format with no or minimal restriction

as a means for accelerating research; these efforts are in the interest of enhancing transparency and collaboration, and fostering innovation. [...] Three main aspects of open science are: open access, open research data, and open collaboration enabled through ICTs. Other aspects of open science—post-publication peer review, open research notebooks, open access to research materials, open source software, citizen science, and research crowdfunding are also part of the architecture of an open science system” (OECD, 2015b, p. 7). Vicente-Sáez and Martínez-Fuentes (2018), after a systematic review proposes the following broad definition of *open science* as the “transparent and accessible knowledge that is shared and developed through collaborative networks”. Daraio and Bonaccorsi (2017) show that a smart integration of existing data in an open-linked data platform may permit the construction of new and timely indicators able to address a variety of user requirements without the necessity to design indicators on a custom basis.

The quality of data and of related information is crucial to add value and improve the awareness and better exploitation of the available data, enhancing data quality-aware empirical investigations when heterogeneous data sources, included in data infrastructures, have to be integrated in Knowledge Infrastructures (KI). It has been observed that the knowledge sharing has direct impacts and interaction effects, in combination with IT infrastructure and enhance institutions and firms’ ability to innovate (Cassia et al., 2020; OECD, 2015a).

Among the most urgent research questions to address about KI recently discussed we have the following:

- (i) *Investing in Knowledge Infrastructures that enrich and develop* scholarly communication. There is an evident contradiction at the political and institutional level. There are strong pressures on researchers to share and store their data, but there is a lack of investment in rigorous and suitable infrastructures for this purpose. Scientific data has the characteristic of being heterogeneous in terms of type, volume, funding sources, instrumentation, standards and other factors. These features make it burdensome and difficult for researchers to maintain and open these data, without appropriate infrastructure available. (Borgman, 2020).
- (ii) *Developing more inclusive Knowledge Infrastructures* to foster open-minded participation. Borgman et al., (2012) discuss sustainable infrastructure development based on the participation of users in the planning and designing of the systems, including citizen science, community-based science, street science, and community research. Issues that remain to be further studied include the investigation on the nature of that participation, the difficulties and capabilities of marginalized populations, and the methods to include users in design and/or operationalization of KIs. The existing literature is divided between those that demonstrate that KIs can benefit and empower communities and citizens, particularly when are combined with open data initiatives, and those that point out how existing KIs did not help communities and citizens to address their community concerns and problems (Yoon, 2020).
- (iii) *Maximizing the scientific return of archival data*. This is a relevant question to address considering the high development rates of tools, technologies, and techniques for generating and diffusing scientific knowledge (Smith, 2020).

- (iv) *Bridging diverse Knowledge Infrastructures*. This is another crucial question to address. One of the main challenges to be faced is to find ways, bridges, to integrate and complement parts of KI that are opposite, independent and lagging behind each other (Faniel, 2020).

We have highlighted the transition from data infrastructures to knowledge infrastructures, in which the importance of considering the quality of data and information at the integrated level plays an important role. Among the most urgent research questions to be addressed for the development of the knowledge infrastructures listed above, the one on “Maximizing the scientific return of archival data” is the most related to our work. Indeed, it is not possible to achieve better exploitation of available data without seeking data quality at the global system level.

We are well aware that the road to building knowledge infrastructures on top of existing data infrastructures is still a long way to go. The approach that we have presented in this paper, and illustrated on the real case of RISIS, represents a very encouraging first step to continue the path just undertaken.

## Appendix

### Appendix 1 Data integration schema for Higher Education Institutions in Graphol

See Figs. 12, 13.
















Symbol	Name	Symbol	Name	Symbol	Name
	Concept node		Role node		Attribute node
	Value-domain node		Individual/Value node	<b>Restriction type</b> 	Domain restriction node
<b>Restriction type</b> 	Range restriction node		Intersection node		Union node
	Inverse node		One-of node		Complement node
	Chain node				
Symbol	Name	Symbol	Name	Symbol	Name
	Inclusion edge		Input edge		

Fig. 12 Graphol predicate and constructor nodes ( *Source* Console et al., 2014, p. 4)

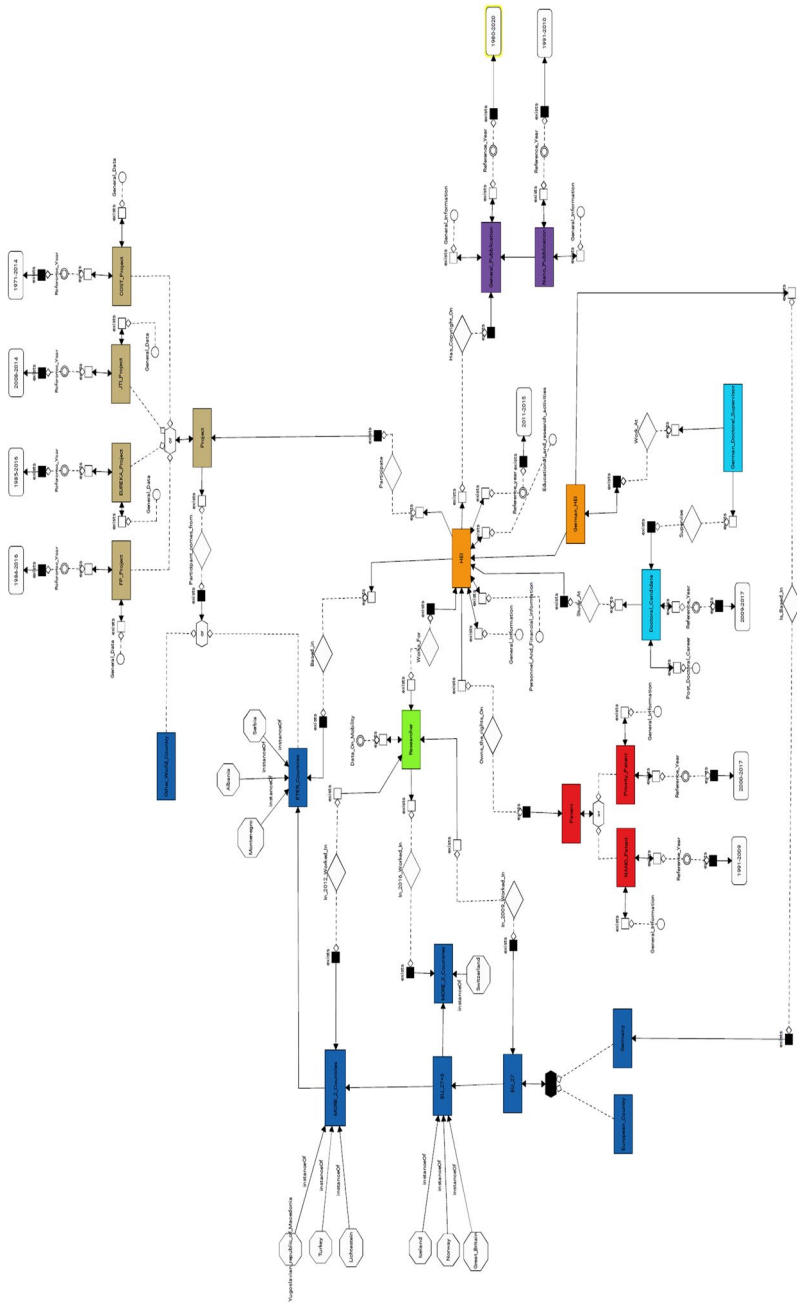


Fig. 13 Data integration model for Higher Education Institutions in Graphol

## Appendix 2 Additional figures

See Figs. 14, 15, 16.

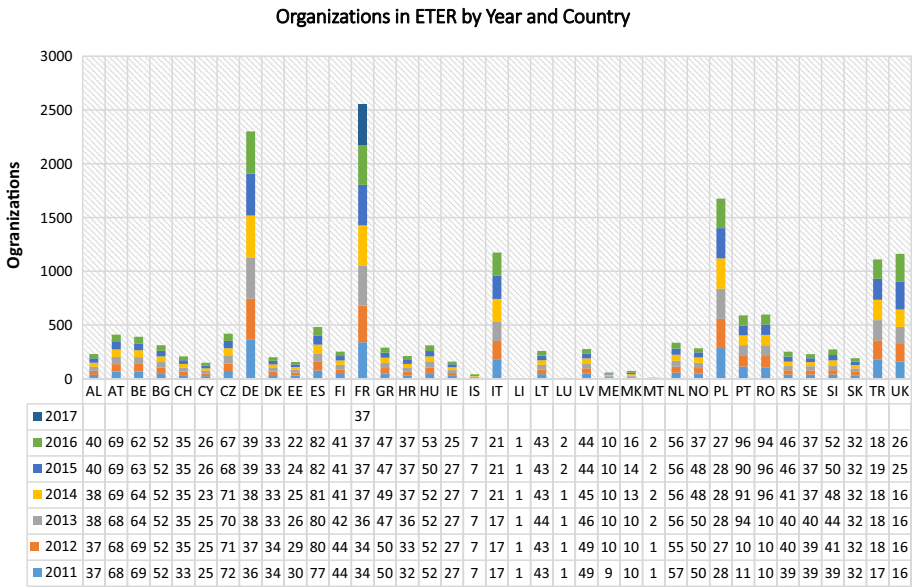


Fig. 14 Organizations in ETER by year and country

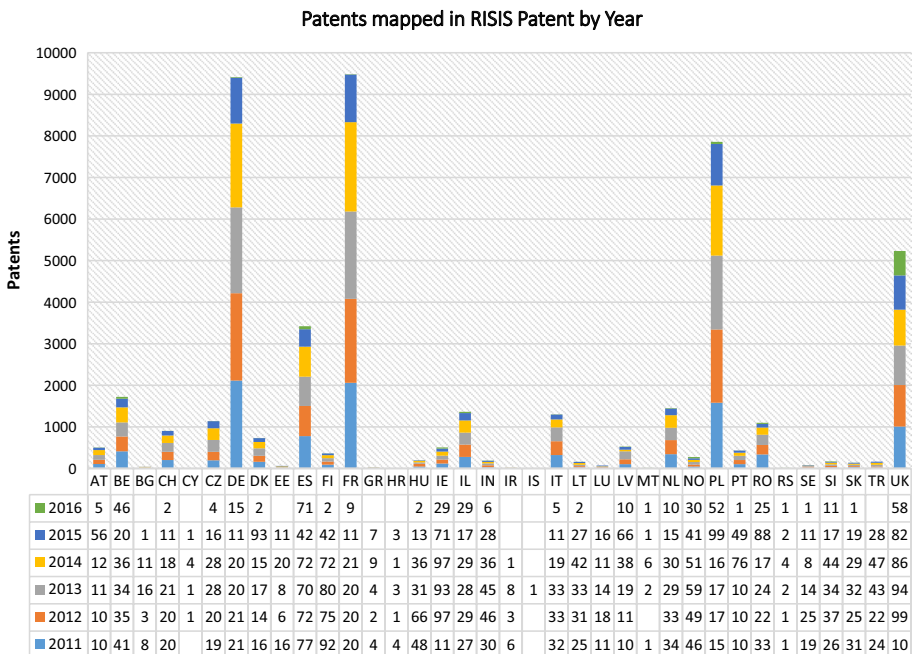
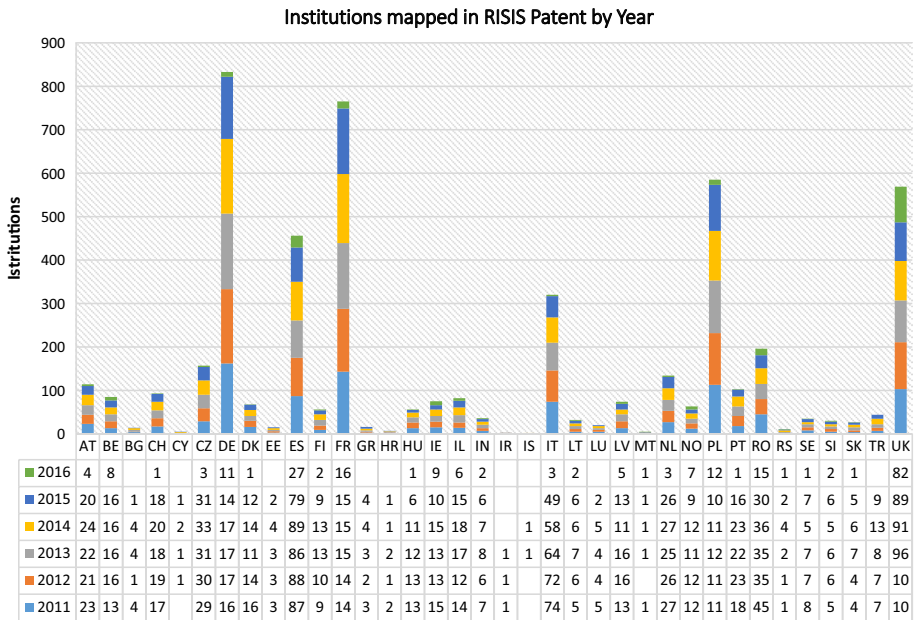


Fig. 15 Patents mapped in RISIS Patent by year



**Fig. 16** Institutions mapped in RISIS Patent by year

**Acknowledgements** The financial support of Sapienza University of Rome (through the Sapienza Awards no. RM11916B8853C925), and of the EU Horizon 2020 RISIS2 Project (grant agreement N 824091) is gratefully acknowledged. We thank Barbara Heller-Schuh, Patricia Laurens and Thomas Scherngell for providing the data used in the case study.

**Declarations**

**Conflict of interest** The first author (Cinzia Daraio) is a member of the Board of *Scientometrics*.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

**References**

Angelini, M., Daraio, C., Lenzerini, M., Leotta, F., & Santucci, G. (2020). Performance model’s development: A novel approach encompassing ontology-based data access and visual analytics. *Scientometrics*, *125*, 865–892.

Aracri, R. M., Bianco, A. M., Radini, R., Scannapieco, M., Tosco, L., Croce, F., Savo, D. F., & Lenzerini, M. (2018). On the experimental usage of ontology-based data management for the Italian integrated system of statistical registers: Quality issues. In *The 9th European Conference on Quality in Official Statistics (Q2018)*.



- Batini, C., & Scannapieco, M. (2016). *Data and information quality*. Springer.
- Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. MIT press.
- Borgman, C. L. (2020). *Knowledge infrastructures in past, present, and future tense*. UCLA, Center for Knowledge Infrastructures.
- Borgman, C. L., Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Ribes, D., et al. (2012). Knowledge infrastructures: Intellectual frameworks and research challenges. Report of a workshop sponsored by the National Science Foundation and the Sloan Foundation University of Michigan School of Information, 25–28 May 2012.
- Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., & Rosati, R. (2007). Tractable reasoning and efficient query answering in description logics: The DL-Lite family. *Journal of Automated Reasoning*, 39(3), 385–429.
- Cassia, A. R., Costa, I., da Silva, V. H. C., & de Oliveira Neto, G. C. (2020). Systematic literature review for the development of a conceptual model on the relationship between knowledge sharing, information technology infrastructure and innovative capability. *Technology Analysis & Strategic Management*, 32(7), 801–821.
- Console, M., Lembo, D., Santarelli, V., & Savo, D. F. (2014). Graphol: Ontology representation through diagrams. In 27th International Workshop on Description Logics (Vol. 1193, pp. 483–495). CEUR-WS.org.
- Daraio, C. (2017). A framework for the assessment of Research and its Impacts. *Journal of Data and Information Science*, 2(4), 7–42.
- Daraio, C., & Bonaccorsi, A. (2017). Beyond university rankings? Generating new indicators on universities by linking data in open platforms. *Journal of the Association for Information Science and Technology*, 68(2), 508–529.
- Daraio, C., & Glänzel, W. (2016). Grand challenges in data integration—State of the art and future perspectives: An introduction. *Scientometrics*, 108(1), 391–400.
- Daraio, C., Lenzerini, M., Leporelli, C., Moed, F. H., Naggar, P., Bonaccorsi, A., & Bartolucci, A. (2016b). Data integration for research and innovation policy: An ontology-based data management approach. *Scientometrics*, 106(2), 857–871.
- Daraio, C., Lenzerini, M., Leporelli, C., Naggar, P., Bonaccorsi, A., & Bartolucci, A. (2016a). The advantages of an Ontology-based Data Management Approach: Openness, interoperability and data quality. *Scientometrics*, 108(1), 441–455.
- Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. MIT Press.
- Ekbja, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., et al. (2015). Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology*, 66(8), 1523–1545.
- Faniel, I. M. (2020). *Knowledge infrastructures: A research agenda thought piece*. UCLA, Center for Knowledge Infrastructures.
- Frischmann, B. M. (2012). *Infrastructure: The social value of shared resources*. Oxford University Press.
- Koltay, T. (2016). Data governance, data literacy and the management of data quality. *IFLA Journal*, 42(4), 303–312.
- Lembo, D., Pantaleone, D., Santarelli, V., & Savo, D. F. (2016). Eddy: A graphical editor for OWL 2 ontologies. In 25th International Joint Conference on Artificial Intelligence, IJCAI 2016 (Vol. 2016, pp. 4252–4253). AAAI Press/International Joint Conferences on Artificial Intelligence.
- Lembo, D., Pantaleone, D., Santarelli, V., & Savo, D. F. (2018). Drawing OWL 2 ontologies with Eddy the editor. *AI Communications*, 31(1), 97–113.
- Lenzerini, M. (2011). Ontology-based data management. In Proceedings of CIKM 2011.
- Lenzerini, M., & Daraio, C. (2019). Challenges, approaches and solutions in data integration for research and innovation. In W. Glänzel, H. F. Moed, H. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 397–420). Springer.
- Motro, A., & Anokhin, P. (2005). Fusionplex: Resolution of data inconsistencies in the data integration of heterogeneous information sources. *Information Fusion*, 7, 176.
- Naumann, F., Leser, U., & Freytag, J. C. (1999). Quality-driven integration of heterogeneous information systems. In Proceedings of VLDB'99, Edinburgh, UK
- Nielsen, M. (2012). *Reinventing discovery: The new era of networked science*. Princeton University Press.
- OECD. (2011). *Quality framework and guidelines for OECD statistical activities*. OECD Publishing.
- OECD. (2015a). *Data-driven Innovation for Growth and Well-being*. OECD Publishing.
- OECD. (2015b). *Making open science a reality. OECD science, technology and industry policy Papers No. 25*. OECD Publishing.

- Parent, C., & Spaccapietra, S. (2000). Database integration: The key to data interoperability. In *Advances in Object-Oriented Data Modeling* (Vol. 221).
- Scannapieco, M., Virgillito, A., Marchetti, C., Mecella, M., & Baldoni, R. (2004). The DaQuinCIS architecture: A platform for exchanging and improving data quality in cooperative information systems. *Information Systems*, 29(7), 551–582.
- Smith, A. (2020). *Space Telescope Science Institute as a knowledge infrastructure*. UCLA, Center for Knowledge Infrastructures.
- SNA (2008). The System of National Accounts, ISBN 978-92-1-161522-7. <https://unstats.un.org/unsd/nationalaccount/docs/sna2008.pdf>.
- Tolk, A., & Mugira, J. A. (2003). The levels of conceptual interoperability model. In Proceedings of the 2003 fall simulation interoperability workshop (Vol. 7, pp. 1–11).
- Vicente-Sáez, R., & Martínez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*, 88, 428–436.
- Yoon, A. (2020). *Knowledge infrastructure workshop thought piece*. UCLA, Center for Knowledge Infrastructures.