ISSN 2281-4299

DIPARTIMENTO DI INGEGNERIA INFORMATICA
AUTOMATICA E GESTIONALE ANTONIO RUBERTI

SAPIENZA
UNIVERSITÀ DI ROMA

Accounting for Quality in Data Integration Systems: a Completeness-aware Integration Approach

Cinzia Daraio
Simone Di Leo
Monica Scannapieco

Technical Report n. 3, 2021

Accounting for Quality in Data Integration Systems: a Completeness-aware Integration Approach

Cinzia Daraio[1]*, Simone Di Leo*[1] *and Monica Scannapieco*[2]

[1]DIAG, Sapienza University of Rome, Italy, Email: daraio@diag.uniroma1.it; dileo@diag.uniroma1.it

[2]ISTAT, Rome, Italy Email: scannapi@istat.it

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 (RISIS) shows the usefulness of the proposed approach.

Keywords: data and information quality, data integrated system, longitudinal data, multinational data, data inftrastructures, research infrastructures

1. **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 (Ekbia, 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 and 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 and Glanzel, 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 unified view of these data. According to Parent and Spaccapietra (2000), interoperability is the way in which heterogeneous databases talk to each other and exchange information in a meaningful way. Parent and Spaccapietra (2000) propose three levels of interoperability:

(i) *lowest level of interoperability* in which there is no integration;

(ii) *intermediary level of interoperability* in which the system does not assure consistency across database borders;

(iii) *higher level of interoperability* in which the goal is to develop a global system on top of existing databases, to deliver the wanted level of integration.

There are different levels of conceptual interoperability proposed in the existing literature. Tolk and Muguira (2003) propose a detailed set of levels of conceptual interoperability that goes from the limited case of no integration, Level (0), which corresponds to an isolated systems (constituted by system specific data) to Level (4) which corresponds to the maximum level of integration and is based on the existence of a common conceptual model (constituted by harmonized data and processes with a conceptual model). Intermediary levels include: Level (1) which is characterized by the existence of documentation of data and interfaces (basically documented data), Level (2) which corresponds to the use of common reference models/common ontology (consisting in aligning static data through Meta Data Management) and Level (3) which corresponds to the existence of a common system approach and/ or open source code (consisting of aligned dynamical 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 section we describe the main goal of the paper and its contribution to existing literature. Section 3 outlines existing studies related to the topic addressed in the paper while Section 4 describes the proposed methodology. Section 5 illustrates the case study on the RISIS data integrated system, while Section 6 discusses the main results and concludes the paper.

2. **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.

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 RISIS (Research Infrastructure for the Science and Innovation Studies). 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.

3. 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 et al., 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 query correspondence assertions, 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 Section 4.2. It is based and uses an Ontology-Based Data Management (OBDM) approach described at length in Section 4.1. 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.

4. Method

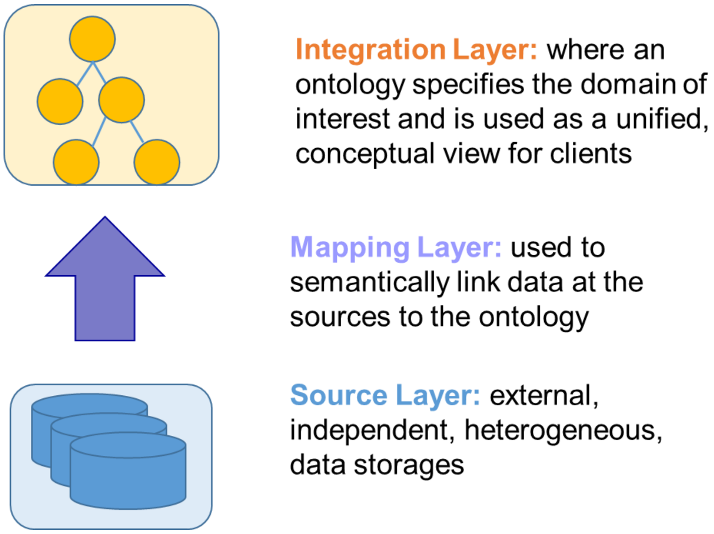
In this section, we first illustrate the proposed method and later we present the RISIS case study that shows the application of the method to a real case. In particular, we will describe our proposal for building a data integration system with explicit quality annotations. We will first give an overview of the used data integration approach, namely OBDM (Ontology-Based Data Management); 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.

4. 1 Introduction on OBDM

Ontology-Based Data Management (OBDM), 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 Figure 1). In particular:

- *Integration Layer*, with an ontology, i.e. a conceptual, formal description of the domain of interest of the organization, expressed in terms of relevant concepts, attributes of concepts, relationships between concepts, and logical assertions formally describing the domain knowledge.
- *Source Layer*, where there are data sources, which are the repositories accessible by the organization where data concerning the domain are stored. In the general case, such repositories 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. Here element means concept, attribute, or relationship.



Integration Layer: where an ontology specifies the domain of interest and is used as a unified, conceptual view for clients

Mapping Layer: used to semantically link data at the sources to the ontology

Source Layer: external, independent, heterogeneous, data storages

Figure 1: OBDA layers

We observe that the above three layers constitute a sophisticated knowledge representation system that can be managed and reasoned upon with the help of automated reasoning techniques. For example, suitable algorithms allow queries expressed over the ontology to be answered by automatically translating the query in terms of the data sources using the mapping (Calvanese et al. 2007). Although the problem of answering queries over the ontology has been the main focus in the last years, there are several other services that an OBDM system should provide. Data quality assessment (Batini and Scannapieco 2016) is one notable example.

4.2 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* $\{S_1 \ldots S_N\}$. Let us suppose that each source provides a set of (relational) tables, i.e. $S_1=\{R_{11} \ldots R_{1k}\} \ldots S_N=\{R_{n1} \ldots R_{nk})\}$.

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

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 $A_{territory}$ (e.g. country) and one single attribute for the temporal dimension A_{time}(e.g. year), and similarly R_2.

Example 1.

Looking at Figure 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 $R_{1,}$ $A_{territory}$ =Country= EU_{27+1} (meaning EU countries plus UK), while for $R_{2,}$ $A_{territory}$ =Country= EU_{27}.

Instead, for $R_{2,}$ A_{time} =Year= 1985-2016 (meaning the interval of years from 1985 to 2016) while for $R_{2,}$ A_{time} =Year= 1991-2009.

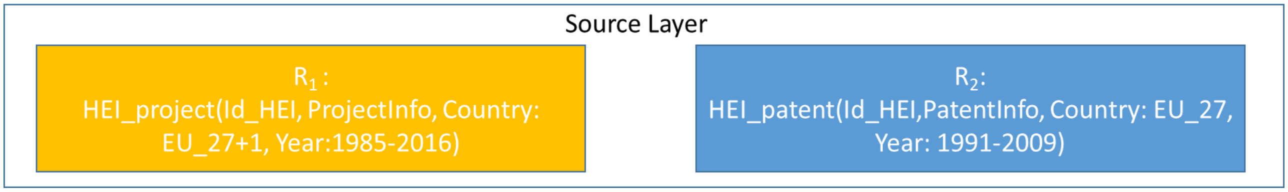

Source Layer
R_1 :
HEI_project(Id_HEI, ProjectInfo, Country: EU_27+1, Year:1985-2016)
R_2:
HEI_patent(Id_HEI,PatentInfo, Country: EU_27, Year: 1991-2009)

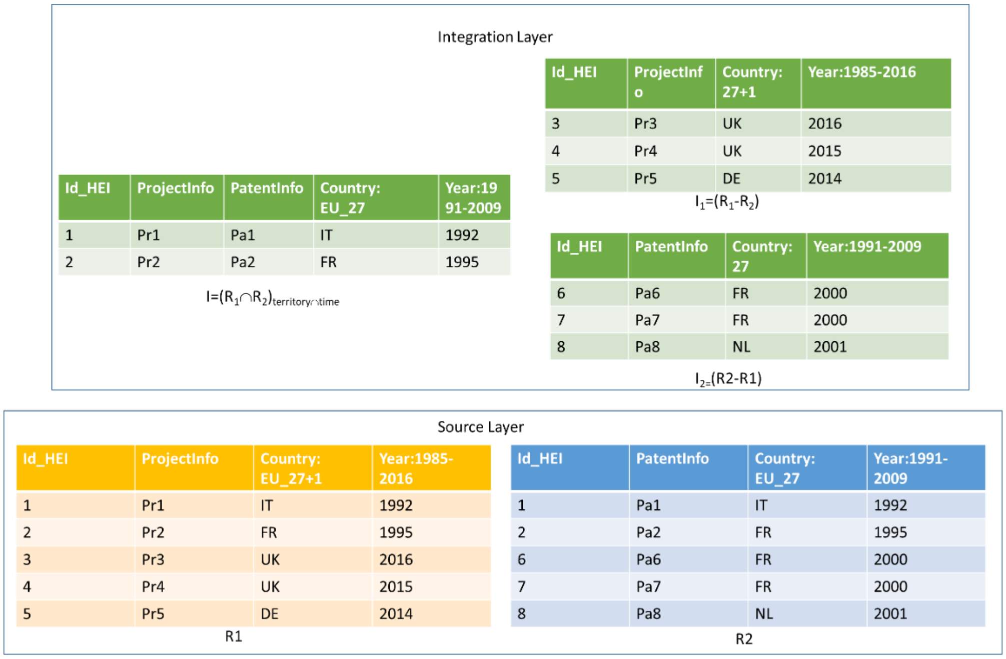
Figure 2: Example of source layer in a data integration system with specific space-time features.

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. $I= (R_1 \cap R_2)_{territory \cap time}$. that (i) will consists of all the tuples present in both R_1 and R_2, and (ii) will have A_{time} and A_{space} defined on the intersection of the domains of the two attributes in the originating sources, namely $R_1(A_{time}) \cap R_2(A_{time})$ and $R_1(A_{space}) \cap R_2(A_{space})$.
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_{time} and A_{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_{time} and A_{space} defined as in R_2.

Example 2.



Integration Layer

Id_HEI	ProjectInfo	Country: 27+1	Year:1985-2016
3	Pr3	UK	2016
4	Pr4	UK	2015
5	Pr5	DE	2014

$I_1=(R_1-R_2)$

Id_HEI	ProjectInfo	PatentInfo	Country: EU_27	Year:1991-2009
1	Pr1	Pa1	IT	1992
2	Pr2	Pa2	FR	1995

$I=(R_1 \cap R_2)_{territory \cap time}$

Id_HEI	PatentInfo	Country: 27	Year:1991-2009
6	Pa6	FR	2000
7	Pa7	FR	2000
8	Pa8	NL	2001

$I_{2=}(R2-R1)$

Source Layer

Id_HEI	ProjectInfo	Country: EU_27+1	Year:1985-2016
1	Pr1	IT	1992
2	Pr2	FR	1995
3	Pr3	UK	2016
4	Pr4	UK	2015
5	Pr5	DE	2014

R1

Id_HEI	PatentInfo	Country: EU_27	Year:1991-2009
1	Pa1	IT	1992
2	Pa2	FR	1995
6	Pa6	FR	2000
7	Pa7	FR	2000
8	Pa8	NL	2001

R2

Figure 3: Instances of relations at the Source Layer and at the Integration Layer

Looking at Figure 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 two 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 R1

- I_2 consists of all the tuples of R_2 that are not in R_1 and 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 Figure 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 Figure 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 S1.

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.

5. Case Study

5.1 RISIS Infrastructure

RISIS (Research infrastructure for research and innovation policy studies) 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.
- *MORE (Mobility Survey of the Higher Education Sector)*, 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 worldwide.
- *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).

5.2 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, Appendix 1 reports in Fig A1. the legend of the Graphol language including predicate and constructor nodes (Console et al. 2014, Lembo et al. 2016 and 2018) used to model the domain. Details of the used datasets are given below:

- *ETER* (see Appendix 2, Fig. A3 shows 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:
    - 17652 record
    - 3205 Organizations with ID

- *CWTS*, thanks to the support of the "Centre for Science and Technology Studies", with data on academic publications from 2011 to 2017 from different countries in the world. The used dataset contains:
    - 10086029 records
    - 4478874 Unique Articles
    - 3579 Organizations with ID

- *RISIS PATENT*(see Appendix 2, Fig. A4 presents Patents mapped in RISIS Patent by Year and Figure A5 shows 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 proposal contains:
    - 57114 records
    - 1471 Organizations with ID
    - 48666 Patents
    - 37 countries
    - 0 null rows "orgId" in the dataset

5.2.1 RISIS ETER and CWTS Integration

This integration task combines HEIs with related publications.

Starting from the source layers, data integrations have been performed following methodology presented in section 2 (see appendix 1 for the results), considering R_1 as ETER and R_2 as CWTS:

1) $I= (R_1 \cap R_2)_{territory \cap time}$ =Creation of the intersection of org_Id and years between datasets R_1 and R2

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 Asp A_{space} ace 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=CWTS).

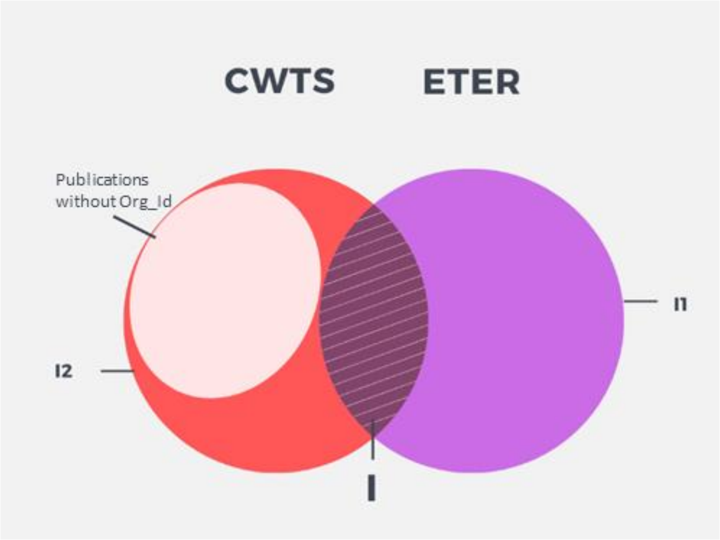
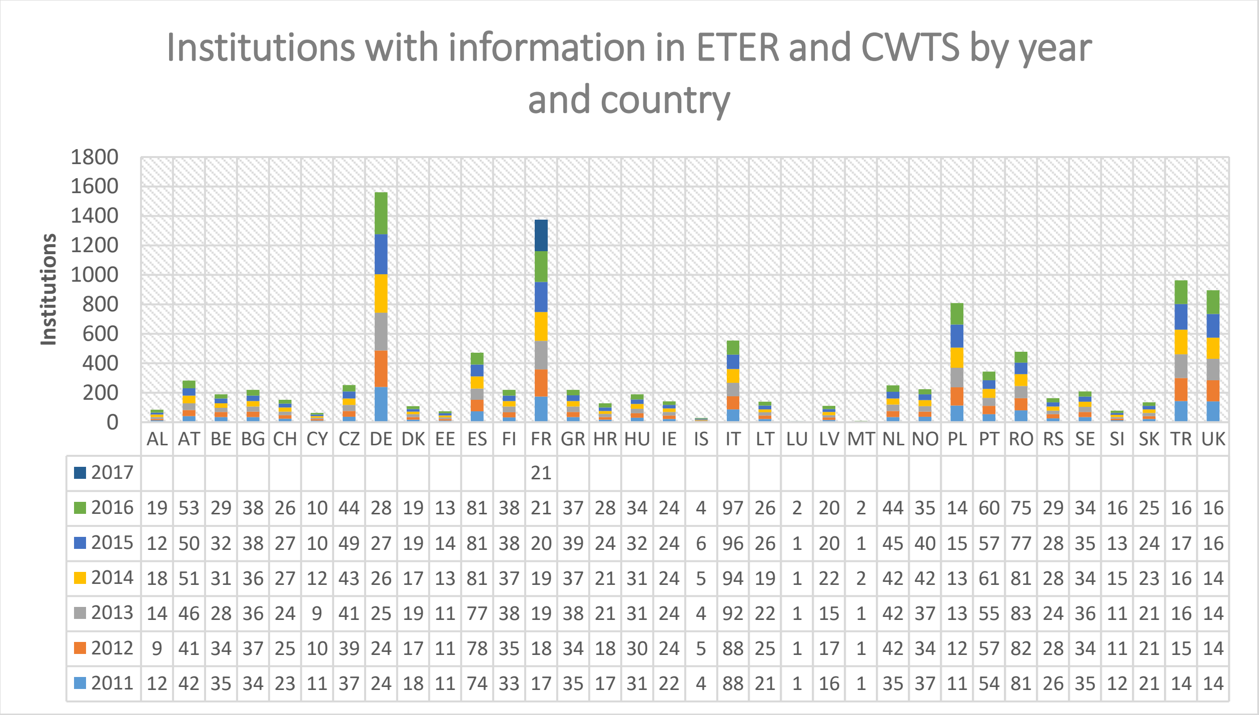


Figure 4: Representation of the integration scheme of ETER and CTWS

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

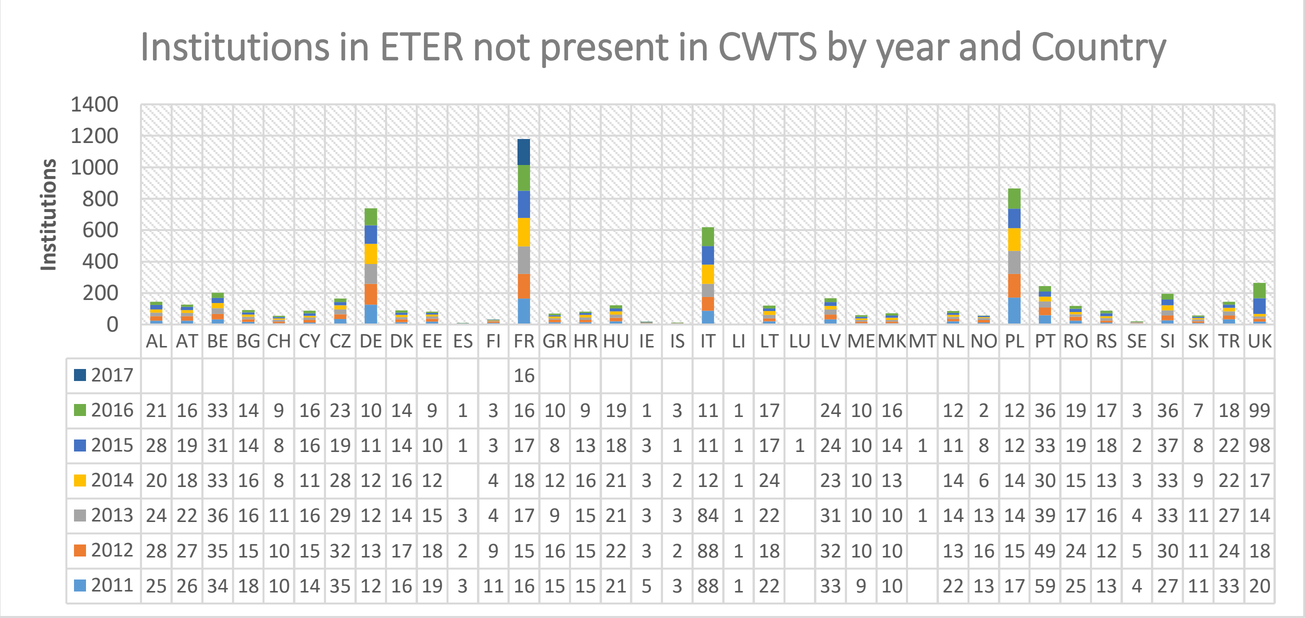
- Relation I:
  - The relation I has 6429051 records, which corresponds to the number of publications with information about the referenced institutions;
  - I contains 3451451 different articles, 2199 unique organization from 34 different Countries.



	AL	AT	BE	BG	CH	CY	CZ	DE	DK	EE	ES	FI	FR	GR	HR	HU	IE	IS	IT	LT	LU	LV	MT	NL	NO	PL	PT	RO	RS	SE	SI	SK	TR	UK
2017													21																					
2016	19	53	29	38	26	10	44	28	19	13	81	38	21	37	28	34	24	4	97	26	2	20	2	44	35	14	60	75	29	34	16	25	16	16
2015	12	50	32	38	27	10	49	27	19	14	81	38	20	39	24	32	24	6	96	26	1	20	1	45	40	15	57	77	28	35	13	24	17	16
2014	18	51	31	36	27	12	43	26	17	13	81	37	19	37	21	31	24	5	94	19	1	22	2	42	42	13	61	81	28	34	15	23	16	14
2013	14	46	28	36	24	9	41	25	19	11	77	38	19	38	21	31	24	4	92	22	1	15	1	42	37	13	55	83	24	36	11	21	16	14
2012	9	41	34	37	25	10	39	24	17	11	78	35	18	34	18	30	24	5	88	25	1	17	1	42	34	12	57	82	28	34	11	21	15	14
2011	12	42	35	34	23	11	37	24	18	11	74	33	17	35	17	31	22	4	88	21	1	16	1	35	37	11	54	81	26	35	12	21	14	14

Figure 5: Number of institutions in I by year and country (Institutions with information in ETER and CWTS by year and country)

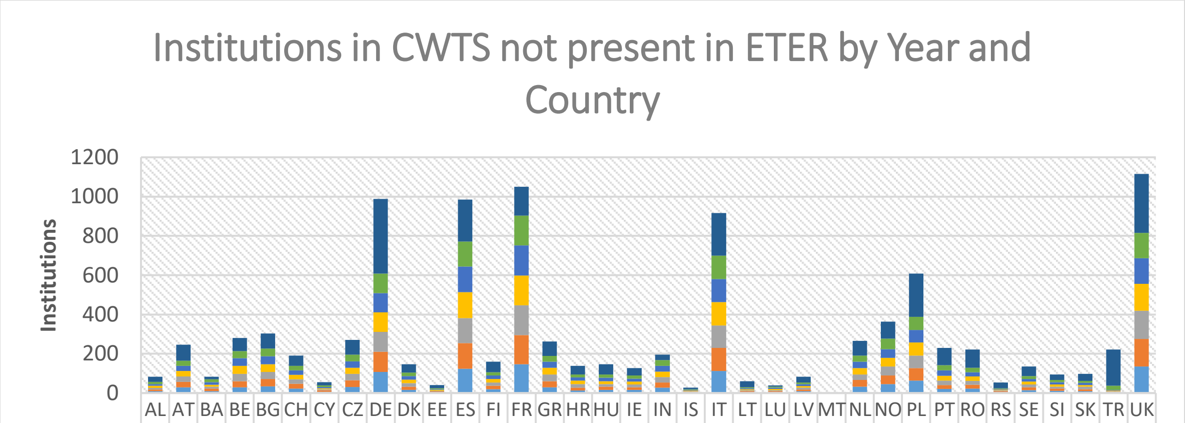
- Relation I$_1$
  - ○ The relation I$_1$ has 6522 records, which correspond to the number of institutions in ETER without information in CWTS (for the period 2011-2017);
  - ○ I$_1$ contains information on 1006 different institutions from 37 different countries.



	AL	AT	BE	BG	CH	CY	CZ	DE	DK	EE	ES	FI	FR	GR	HR	HU	IE	IS	IT	LI	LT	LU	LV	ME	MK	MT	NL	NO	PL	PT	RO	RS	SE	SI	SK	TR	UK
2017													16																								
2016	21	16	33	14	9	16	23	10	14	9	1	3	16	10	9	19	1	3	11	1	17		24	10	16		12	2	12	36	19	17	3	36	7	18	99
2015	28	19	31	14	8	16	19	11	14	10	1	3	17	8	13	18	3	1	11	1	17	1	24	10	14	1	11	8	12	33	19	18	2	37	8	22	98
2014	20	18	33	16	8	11	28	12	16	12		4	18	12	16	21	3	2	12	1	24		23	10	13		14	6	14	30	15	13	3	33	9	22	17
2013	24	22	36	16	11	16	29	12	14	15	3	4	17	9	15	21	3	3	84	1	22		31	10	10	1	14	13	14	39	17	16	4	33	11	27	14
2012	28	27	35	15	10	15	32	13	17	18	2	9	15	16	15	22	3	2	88	1	18		32	10	10		13	16	15	49	24	12	5	30	11	24	18
2011	25	26	34	18	10	14	35	12	16	19	3	11	16	15	15	21	5	3	88	1	22		33	9	10		22	13	17	59	25	13	4	27	11	33	20

Figure 6: Number of institutions in I1 by year and country Relation I2 (Institutions in ETER not present in CWTS by year and Country)

- In the I_2 Domains there are 3492623 publications without org id in a certain Years 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).



	AL	AT	BA	BE	BG	CH	CY	CZ	DE	DK	EE	ES	FI	FR	GR	HR	HU	IE	IN	IS	IT	LT	LU	LV	MT	NL	NO	PL	PT	RO	RS	SE	SI	SK	TR	UK
■ 2017	26	81	13	66	78	51	16	75	38	43	17	21	54	14	73	44	52	38	27	8	21	30	7	29	2	75	86	22	88	93	30	49	29	35	18	29
■ 2016	11	26	15	37	39	23	8	33	99	18	4	12	17	15	30	14	17	15	29	3	11	5	5	9		32	53	67	26	22	3	14	11	10	21	13
■ 2015	9	27	13	39	41	23	6	34	98	18	4	13	18	15	31	17	17	15	29	3	11	5	6	9		32	44	63	28	22	3	14	11	11	3	13
■ 2014	11	27	10	40	38	23	6	31	99	18	4	13	17	15	33	17	15	15	29	3	11	5	6	9		32	44	68	25	22	4	14	11	9	3	13
■ 2013	8	28	10	37	36	23	5	32	10	17	4	12	17	15	34	18	14	14	27	3	11	5	5	9		29	46	63	22	20	4	15	11	12	4	14
■ 2012	9	27	12	32	38	24	7	34	10	17	4	13	18	14	32	15	16	14	27	3	11	5	5	9		33	45	63	20	20	5	15	11	10	3	13
■ 2011	9	30	10	29	34	23	7	31	10	16	4	12	19	14	29	13	16	16	27	4	11	5	5	9		33	45	64	21	22	5	14	11	10	3	13

Figure 7 Number of institutions in I2 by year and country (Institutions in CWTS not present in ETER by Year and Country)

In addition to these results, 164355 records from CWTS without org_Id, i.e. unmapped.

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

Table I Number of articles in CWTS without Org_Id value by year

Thanks to this approach, is possible to highlight the completeness of the information. In each relation (I, I1 and I2) 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 10092551 rows.

Considering the total number of rows with complete information (6429051 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 tamples as results of their queries.

5.2.2 RISIS ETER and RISIS PATENT Integration

Starting from the source layers, data integrations have been performed following methodology presented in Section 2 (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).

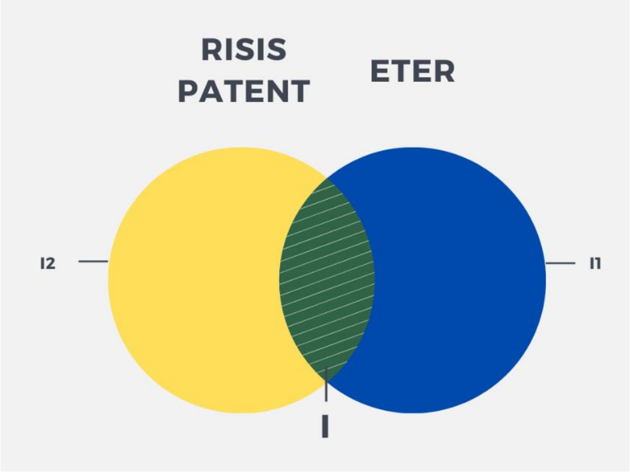
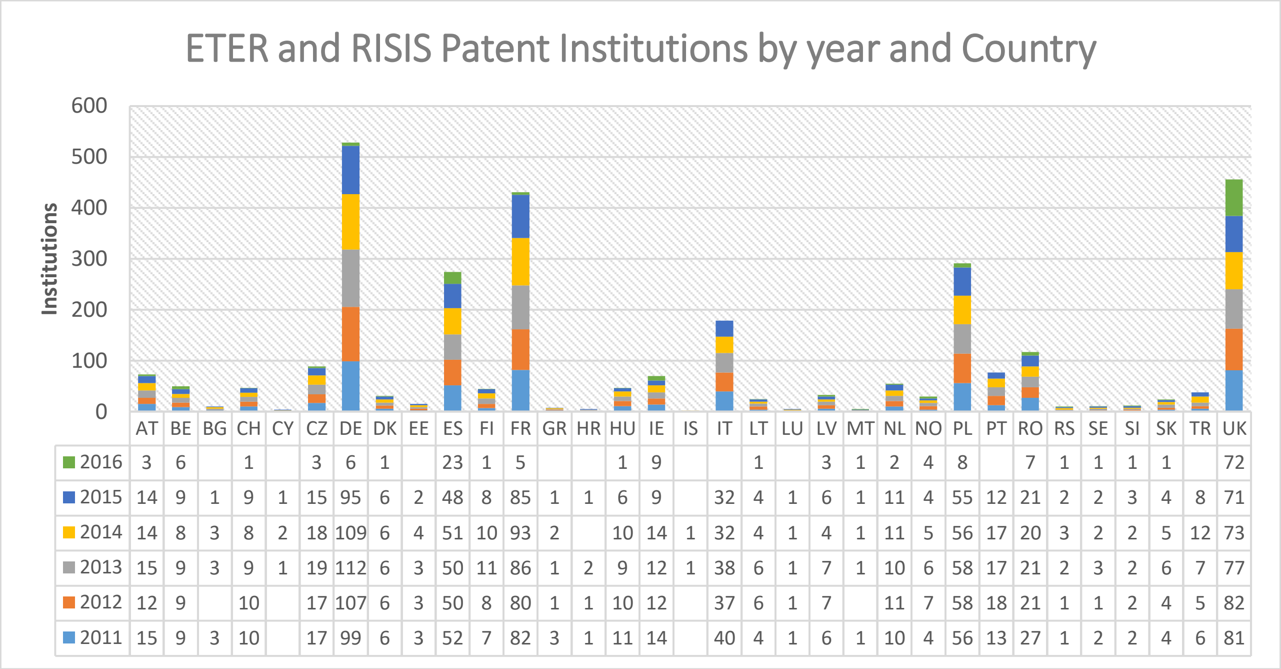


Figure 8 Representation of the integration scheme of ETER and RISIS PATENT

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

- Relation I:
    - The relation I has 32027 records, which corresponds to the Institution with information about the Patents.

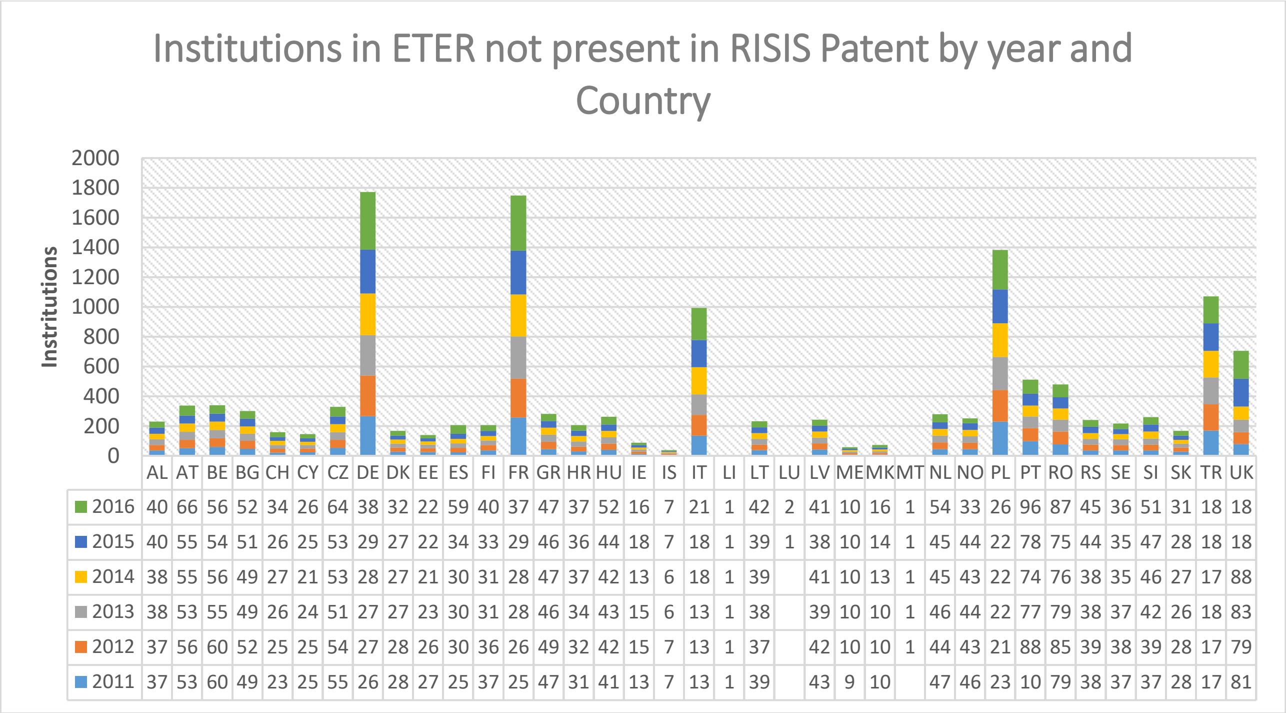
- I contain 30165 different patents, 829 unique organization from 33 different Countries and the period 2011-2016 (Risis patent last year of reference)



	AT	BE	BG	CH	CY	CZ	DE	DK	EE	ES	FI	FR	GR	HR	HU	IE	IS	IT	LT	LU	LV	MT	NL	NO	PL	PT	RO	RS	SE	SI	SK	TR	UK
🟩 2016	3	6		1		3	6	1		23	1	5			1	9			1		3	1	2	4	8		7	1	1	1	1		72
🟦 2015	14	9	1	9	1	15	95	6	2	48	8	85	1	1	6	9		32	4	1	6	1	11	4	55	12	21	2	2	3	4	8	71
🟨 2014	14	8	3	8	2	18	109	6	4	51	10	93	2		10	14	1	32	4	1	4	1	11	5	56	17	20	3	2	2	5	12	73
■ 2013	15	9	3	9	1	19	112	6	3	50	11	86	1	2	9	12	1	38	6	1	7	1	10	6	58	17	21	2	3	2	6	7	77
🟧 2012	12	9		10		17	107	6	3	50	8	80	1	1	10	12		37	6	1	7		11	7	58	18	21	1	1	2	4	5	82
🟦 2011	15	9	3	10		17	99	6	3	52	7	82	3	1	11	14		40	4	1	6	1	10	4	56	13	27	1	2	2	4	6	81

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

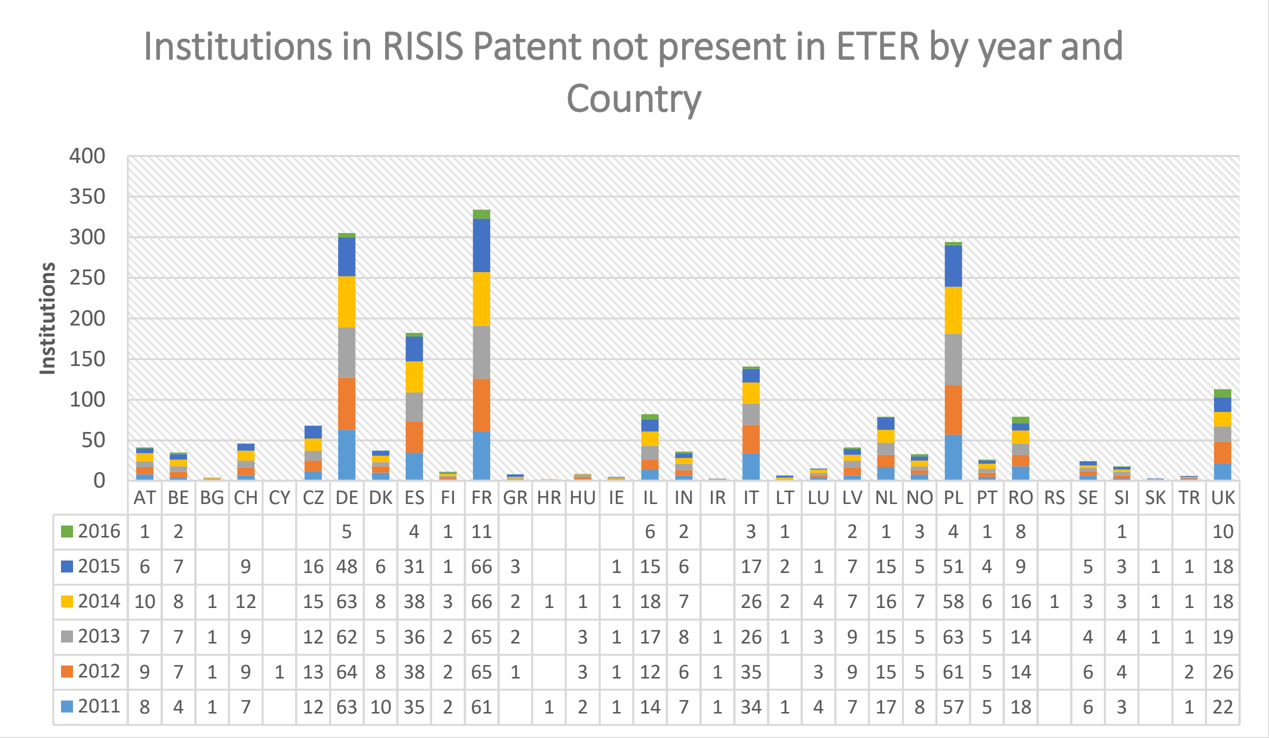
- Relation I_1
  - The relation I_1 has 14177 records (for the period 2011-2016);
  - I_1 contains information on 3060 different institutions.



	AL	AT	BE	BG	CH	CY	CZ	DE	DK	EE	ES	FI	FR	GR	HR	HU	IE	IS	IT	LI	LT	LU	LV	ME	MK	MT	NL	NO	PL	PT	RO	RS	SE	SI	SK	TR	UK
🟩 2016	40	66	56	52	34	26	64	38	32	22	59	40	37	47	37	52	16	7	21	1	42	2	41	10	16	1	54	33	26	96	87	45	36	51	31	18	18
🟦 2015	40	55	54	51	26	25	53	29	27	22	34	33	29	46	36	44	18	7	18	1	39	1	38	10	14	1	45	44	22	78	75	44	35	47	28	18	18
🟨 2014	38	55	56	49	27	21	53	28	27	21	30	31	28	47	37	42	13	6	18	1	39		41	10	13	1	45	43	22	74	76	38	35	46	27	17	88
■ 2013	38	53	55	49	26	24	51	27	27	23	30	31	28	46	34	43	15	6	13	1	38		39	10	10	1	46	44	22	77	79	38	37	42	26	18	83
🟧 2012	37	56	60	52	25	25	54	27	28	26	30	36	26	49	32	42	15	7	13	1	37		42	10	10	1	44	43	21	88	85	39	38	39	28	17	79
🟦 2011	37	53	60	49	23	25	55	26	28	27	25	37	25	47	31	41	13	7	13	1	39		43	9	10		47	46	23	10	79	38	37	37	28	17	81

Figure 10 Number of institutions in I1 by year and country (Institutions in ETER not present in RISIS Patent by year and Country)

- Relation I_2
  - In the I_2 Domains there are 25087 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).



	AT	BE	BG	CH	CY	CZ	DE	DK	ES	FI	FR	GR	HR	HU	IE	IL	IN	IR	IT	LT	LU	LV	NL	NO	PL	PT	RO	RS	SE	SI	SK	TR	UK
🟩 2016	1	2					5		4	1	11					6	2		3	1		2	1	3	4	1	8			1			10
🟦 2015	6	7		9		16	48	6	31	1	66	3			1	15	6		17	2	1	7	15	5	51	4	9		5	3	1	1	18
🟨 2014	10	8	1	12		15	63	8	38	3	66	2	1	1	1	18	7		26	2	4	7	16	7	58	6	16	1	3	3	1	1	18
⬜ 2013	7	7	1	9		12	62	5	36	2	65	2		3	1	17	8	1	26	1	3	9	15	5	63	5	14		4	4	1	1	19
🟧 2012	9	7	1	9	1	13	64	8	38	2	65	1		3	1	12	6	1	35		3	9	15	5	61	5	14		6	4		2	26
🟦 2011	8	4	1	7		12	63	10	35	2	61		1	2	1	14	7	1	34	1	4	7	17	8	57	5	18		6	3		1	22

Figure 11 Number of institutions in I1 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 71291 rows.

Considering the total number of rows with complete information (32027 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.

5.2.3 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 3663500 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, 39264 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.

6. Discussion and Conclusions

The consideration of the quality of data is an extremely important and current topic in the current 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.

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 adeguate 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 leyer and can be further extended to other data quality dimensions and applied to different databases.

(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.
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 principal 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 UN (2008) System of National Accounts (SNA) 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 on “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 the intelligent integration of existing data may lead to an open-linked data platform which permits the construction of new indicators. The power of the approach derives from the ability to combine heterogeneous sources of data to generate indicators that address a variety of user requirements without the need 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. It has been observed that the knowledge sharing has direct impacts and interaction effects, in combination with IT infrastructure and enhance firms’ ability to innovate (OECD, 2015a, Cassia et al. 2020).

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

i) *Investing in knowledge infrastructures* that *enhance* scholarly communication. Despite the political pressures and institutional requirements for university researchers to share and to retain their data, investments in knowledge infrastructures to sustain access to those data resources are relatively few. Scientific data are heterogenous in type, volume, funding sources, instrumentation, standards, and other factors, making them difficult to sustain (Borgman, 2020).

ii) *Developing more inclusive knowledge infrastructure* by fostering opportunities for fair participation. User participation in the planning and designing of tools/systems to create sustainable infrastructure development has been discussed in Edwards et al., 2013. Extremely important is the different user participation/contribution models in existing Knowledge Infrastructures (KI), such as citizen science, community-based science, street science, and community research. However, the nature of that participation, the demands and abilities of marginalized populations, and methods to reflect inclusivity in design and/or operationalization of KIs for knowledge creation should be further investigated. Many studies have already demonstrated how KIs can benefit and empower communities and citizens, especially when combined with numerous open data initiatives through existing knowledge and data infrastructures by providing access to new information and knowledge and teaching new technical skills. However, literatures have also pointed out how existing KIs did not help communities and citizens address their immediate community concerns and problems (Yoon, 2020).

iii) Maximizing the scientific return of archival data in the coming decades, especially with a fast-moving ecosystem of tools, technologies, and techniques for generating scientific knowledge (Smith, 2020).

iv) Urgent questions to address about KI include: How can parts of KI that are opposing, independent, and lagging be bridged? Which bridges facilitate success under these different circumstances? When in the life of KI is bridging more or less successful? (Faniel, 2020).

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.

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.

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.

Batini, C., Scannapieco, M. (2016). *Data and information quality*. Cham, Switzerland: Springer International Publishing.

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. Retrieved from https://escholarship.org/uc/item/5v73333z

Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., and 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*, Vol. 2 No. 4, 2017 pp 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., Naggar P., Bonaccorsi A. & Bartolucci, A. (2016b). The advantages of an Ontology-based Data Management Approach: openness, interoperability and data quality. *Scientometrics*, 108 (1), 441-455.

Daraio, C., Lenzerini, M., Leporelli, C., Moed, F. H., Naggar, P., Bonaccorsi, A. & Bartolucci, A. (2016a). Data integration for research and innovation policy: An Ontology-Based Data Management approach. *Scientometrics*, 106 (2), 857-871.

Edwards, P. N. (2010). *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming. Cambridge*, MA: MIT Press.

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., ... & Sugimoto, C. R. (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. Retrieved from https://escholarship.org/uc/item/3sq2x711

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. and Daraio C. (2019), Challenges, Approaches and Solutions in Data Integration for Research and Innovation, in *Springer Handbook of Science and Technology Indicators* edited by Glänzel W., Moed H.F., Schmoch H. and Thelwall M., 397-420, ISBN 978-3-030-02511-3.

Lenzerini, M. (2011), Ontology-based data management. In Proc. of CIKM 2011.

Motro, A., and Anokhin, P. Fusionplex, (2005). Resolution of Data Inconsistencies in the Data Integration of Heterogeneous Information Sources. *Information Fusion*.

Naumann, F., Leser, U., and Freytag, J. C. (1999). Quality-driven Integration of Heterogenous Information Systems. In Proc. 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, Paris.

OECD (2015a), Data-driven Innovation for Growth and Well-being. OECD Publishing, Paris.

OECD (2015b). Making Open Science a Reality. OECD Science, Technology and Industry Policy Papers No. 25, OECD Publishing, Paris.

Parent, C., Spaccapietra, S. (2000) Database Integration: The Key to Data Interoperability. *Advances in Object-Oriented Data Modeling*, 221.

Scannapieco, M., Virgillito, A., Marchetti, C., Mecella, M., and 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. Retrieved from https://escholarship.org/uc/item/85r357k7

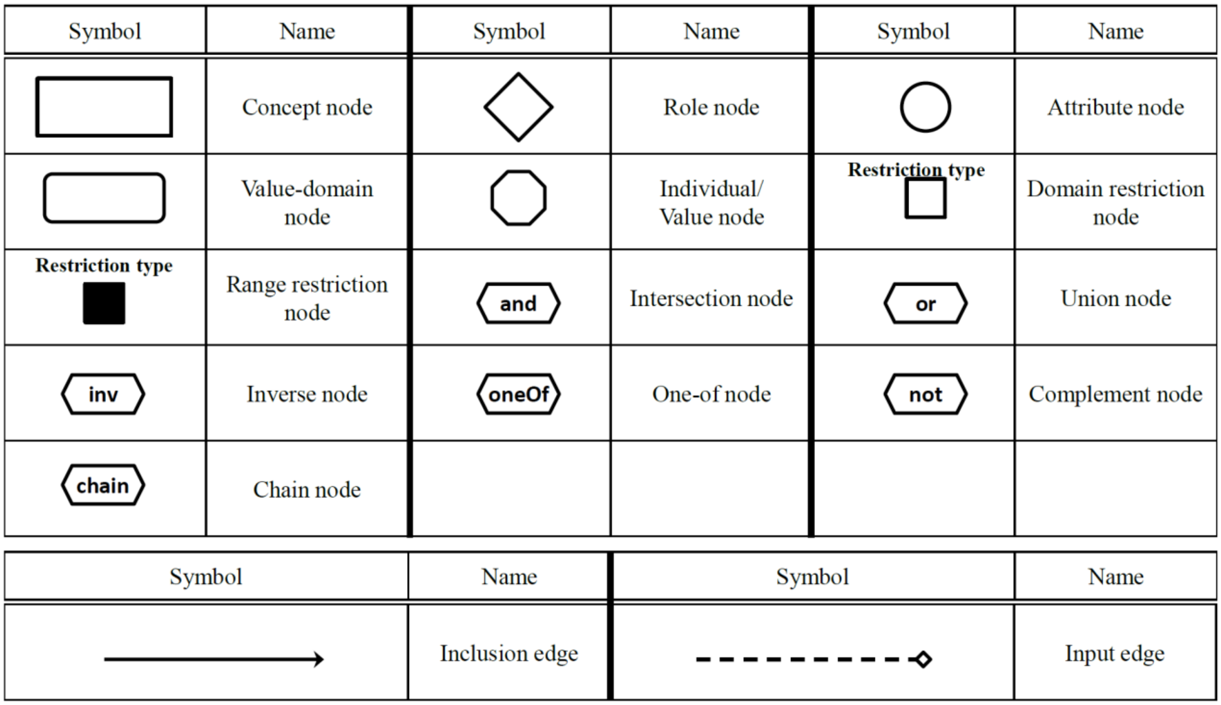
Tolk, A., Muguira, J. A. (2003). The levels of conceptual interoperability model. In *Proceedings of the 2003 fall simulation interoperability workshop* (Vol. 7, pp. 1-11).

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. Retrieved from https://escholarship.org/uc/item/0fc3g08v

Appendices

Appendix 1 Data integration schema for Higher Education Institutions in Graphol



Symbol	Name	Symbol	Name	Symbol	Name
▭	Concept node	◇	Role node	○	Attribute node
▢	Value-domain node	⯃	Individual/ Value node	Restriction type □	Domain restriction node
Restriction type ■	Range restriction node	and	Intersection node	or	Union node
inv	Inverse node	oneOf	One-of node	not	Complement node
chain	Chain node				

Symbol	Name	Symbol	Name
⟶	Inclusion edge	- - - - - - - - - -◇	Input edge

Fig A1. Graphol predicate and constructor nodes (Source: Console et al. 2014, p. 4)

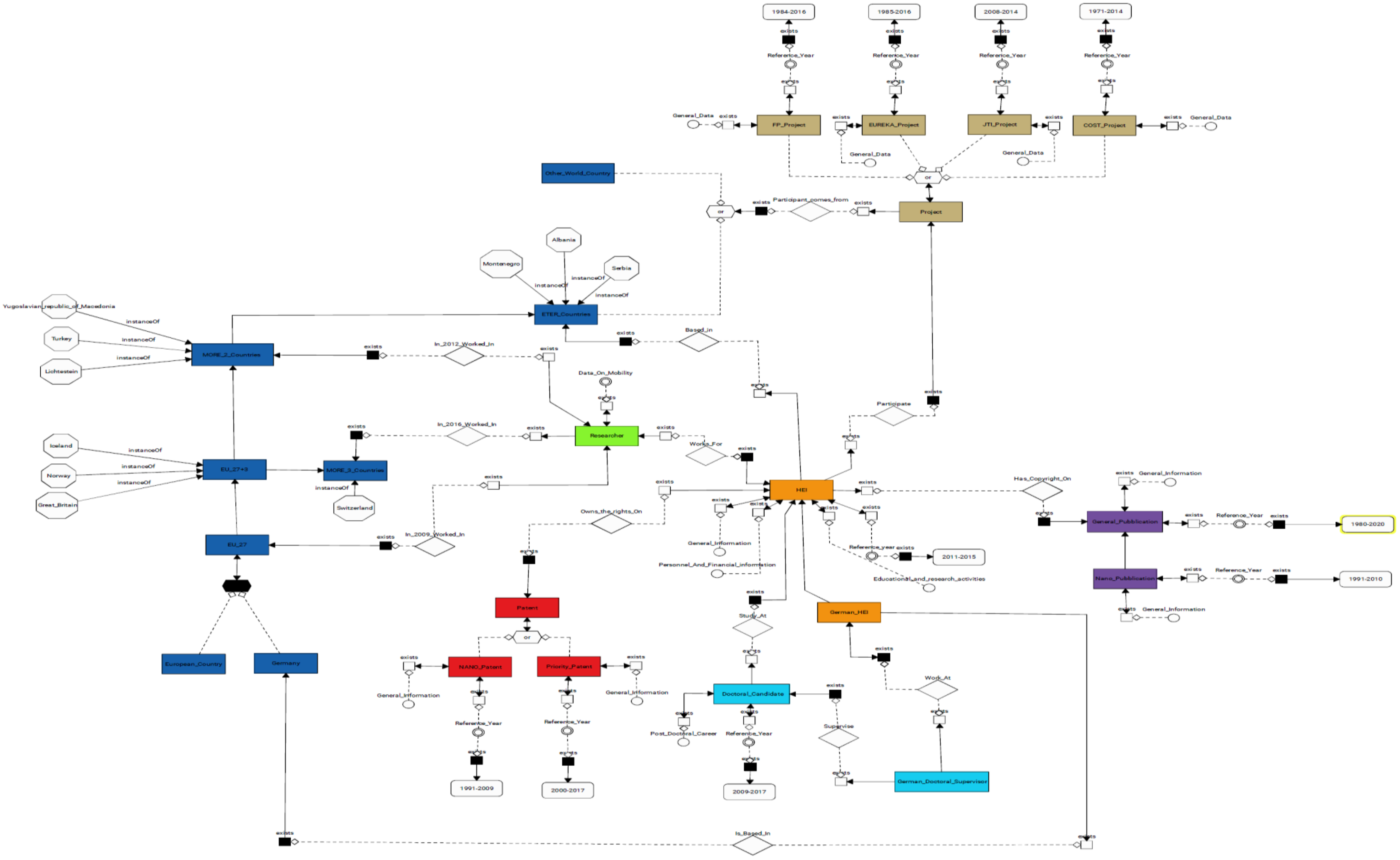
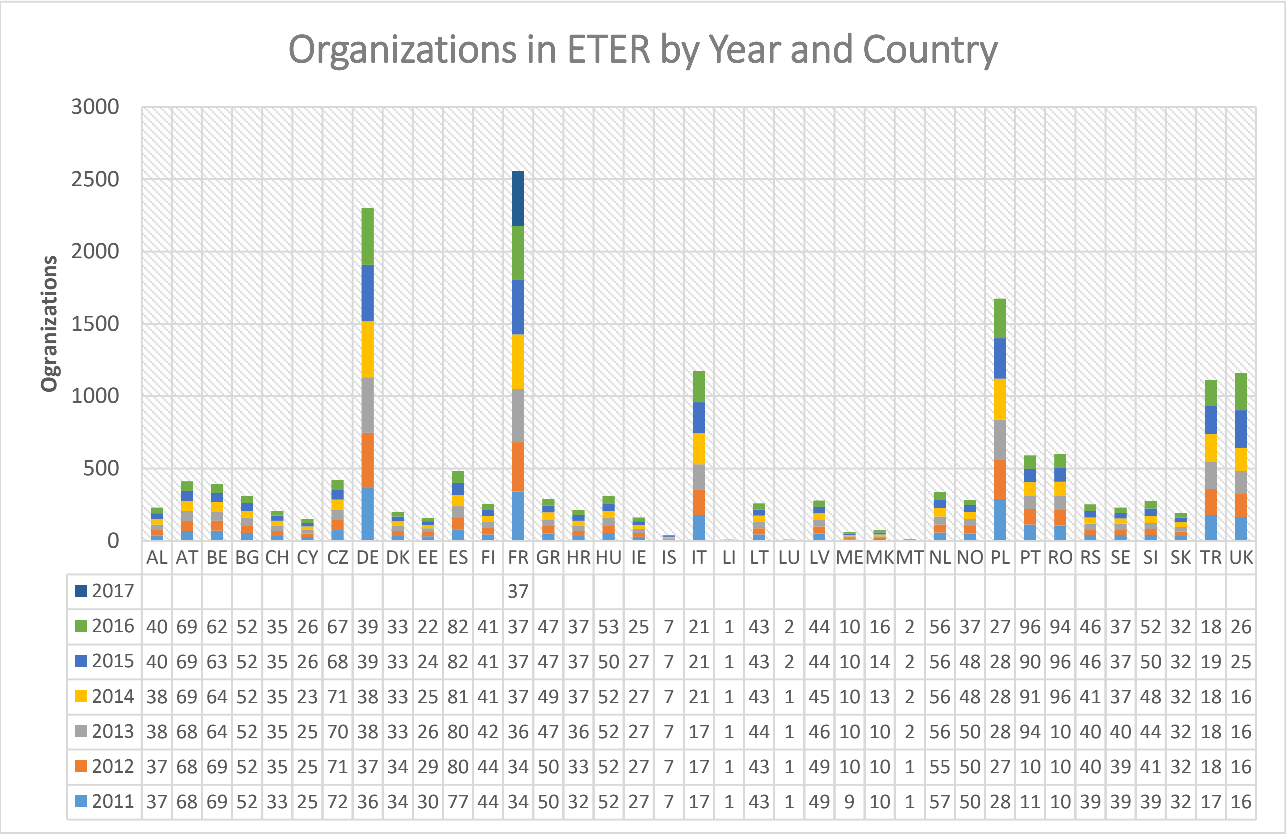


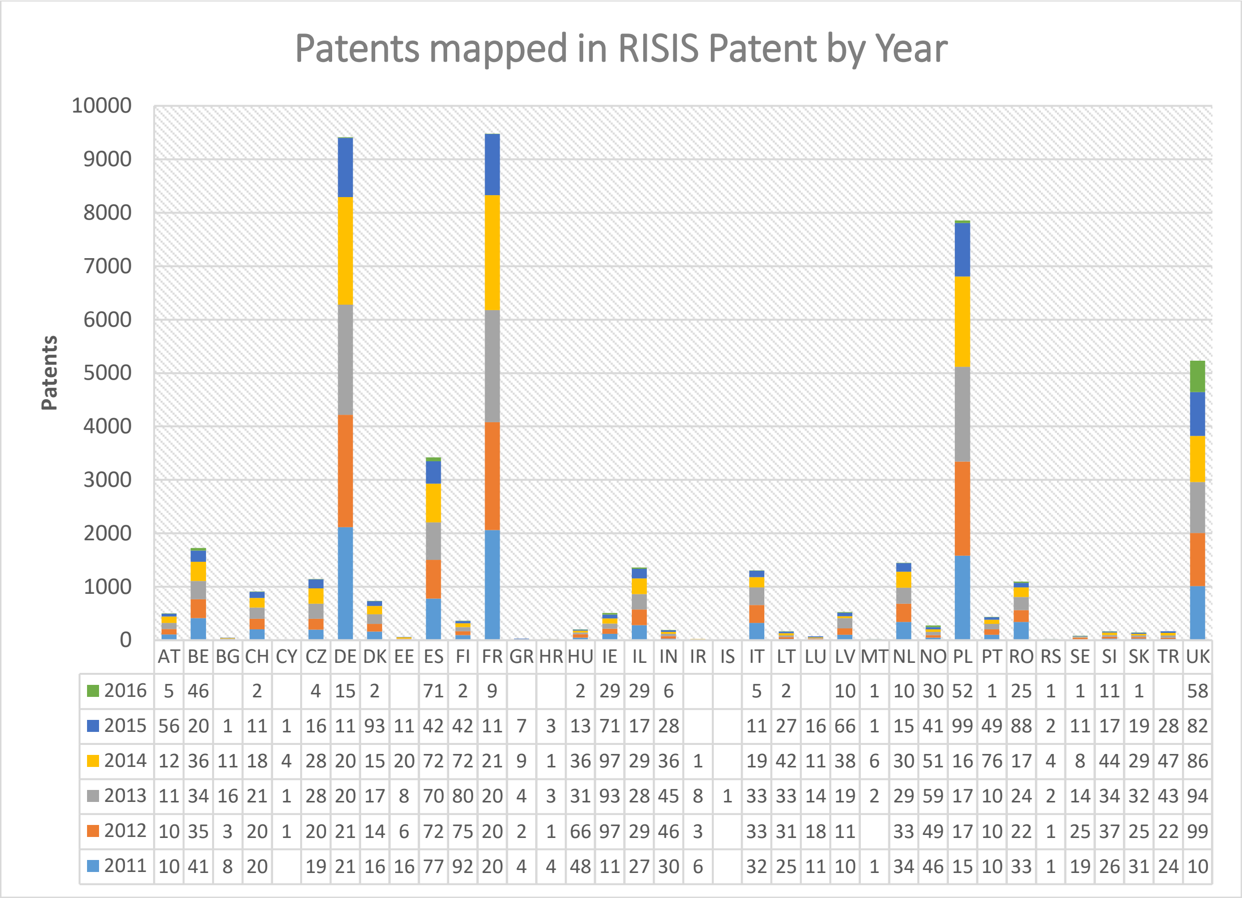
Fig A2. Data integration model for Higher Education Institutions in Graphol

Appendix 2 Additional Figures



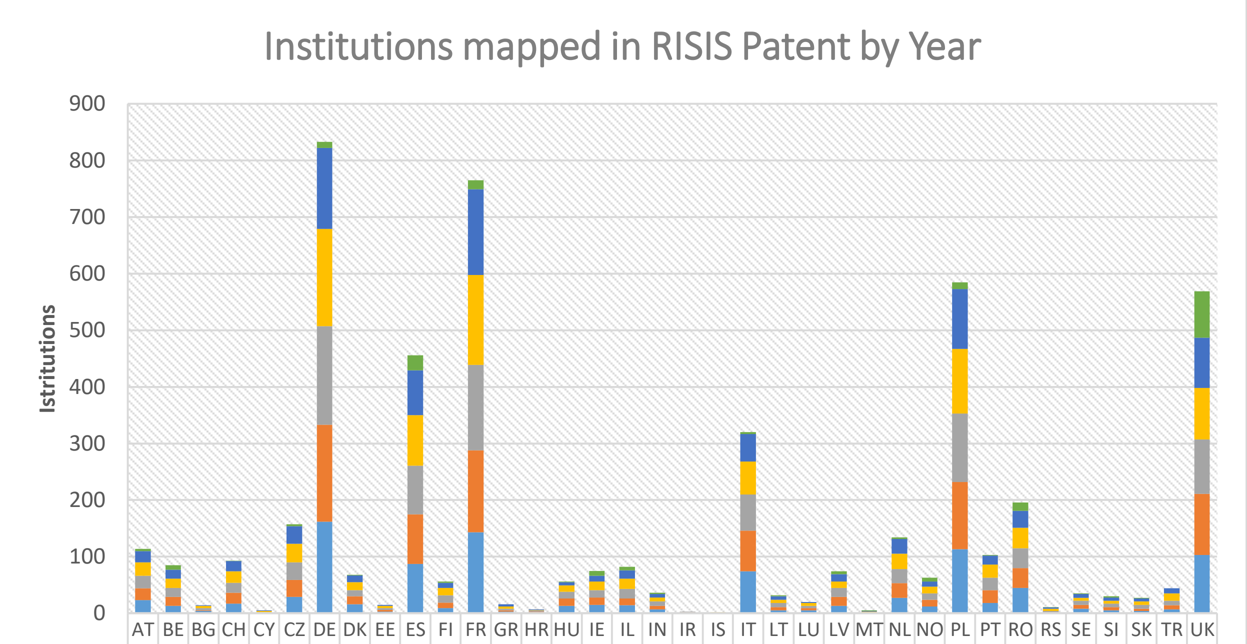
	AL	AT	BE	BG	CH	CY	CZ	DE	DK	EE	ES	FI	FR	GR	HR	HU	IE	IS	IT	LI	LT	LU	LV	ME	MK	MT	NL	NO	PL	PT	RO	RS	SE	SI	SK	TR	UK
🟦 2017													37																								
🟩 2016	40	69	62	52	35	26	67	39	33	22	82	41	37	47	37	53	25	7	21	1	43	2	44	10	16	2	56	37	27	96	94	46	37	52	32	18	26
🟦 2015	40	69	63	52	35	26	68	39	33	24	82	41	37	47	37	50	27	7	21	1	43	2	44	10	14	2	56	48	28	90	96	46	37	50	32	19	25
🟨 2014	38	69	64	52	35	23	71	38	33	25	81	41	37	49	37	52	27	7	21	1	43	1	45	10	13	2	56	48	28	91	96	41	37	48	32	18	16
⬜ 2013	38	68	64	52	35	25	70	38	33	26	80	42	36	47	36	52	27	7	17	1	44	1	46	10	10	2	56	50	28	94	10	40	40	44	32	18	16
🟧 2012	37	68	69	52	35	25	71	37	34	29	80	44	34	50	33	52	27	7	17	1	43	1	49	10	10	1	55	50	27	10	10	40	39	41	32	18	16
🟦 2011	37	68	69	52	33	25	72	36	34	30	77	44	34	50	32	52	27	7	17	1	43	1	49	9	10	1	57	50	28	11	10	39	39	39	32	17	16

Fig. A3 Organizations in ETER by Year and Country



	AT	BE	BG	CH	CY	CZ	DE	DK	EE	ES	FI	FR	GR	HR	HU	IE	IL	IN	IR	IS	IT	LT	LU	LV	MT	NL	NO	PL	PT	RO	RS	SE	SI	SK	TR	UK
🟩 2016	5	46		2		4	15	2		71	2	9			2	29	29	6			5	2		10	1	10	30	52	1	25	1	1	11	1		58
🟦 2015	56	20	1	11	1	16	11	93	11	42	42	11	7	3	13	71	17	28			11	27	16	66	1	15	41	99	49	88	2	11	17	19	28	82
🟨 2014	12	36	11	18	4	28	20	15	20	72	72	21	9	1	36	97	29	36	1		19	42	11	38	6	30	51	16	76	17	4	8	44	29	47	86
⬜ 2013	11	34	16	21	1	28	20	17	8	70	80	20	4	3	31	93	28	45	8	1	33	33	14	19	2	29	59	17	10	24	2	14	34	32	43	94
🟧 2012	10	35	3	20	1	20	21	14	6	72	75	20	2	1	66	97	29	46	3		33	31	18	11		33	49	17	10	22	1	25	37	25	22	99
🟦 2011	10	41	8	20		19	21	16	16	77	92	20	4	4	48	11	27	30	6		32	25	11	10	1	34	46	15	10	33	1	19	26	31	24	10

Fig. A4 Patents mapped in RISIS Patent by Year



	AT	BE	BG	CH	CY	CZ	DE	DK	EE	ES	FI	FR	GR	HR	HU	IE	IL	IN	IR	IS	IT	LT	LU	LV	MT	NL	NO	PL	PT	RO	RS	SE	SI	SK	TR	UK
🟩 2016	4	8		1		3	11	1		27	2	16			1	9	6	2			3	2		5	1	3	7	12	1	15	1	1	2	1		82
🟦 2015	20	16	1	18	1	31	14	12	2	79	9	15	4	1	6	10	15	6			49	6	2	13	1	26	9	10	16	30	2	7	6	5	9	89
🟨 2014	24	16	4	20	2	33	17	14	4	89	13	15	4	1	11	15	18	7		1	58	6	5	11	1	27	12	11	23	36	4	5	5	6	13	91
⬜ 2013	22	16	4	18	1	31	17	11	3	86	13	15	3	2	12	13	17	8	1	1	64	7	4	16	1	25	11	12	22	35	2	7	6	7	8	96
🟧 2012	21	16	1	19	1	30	17	14	3	88	10	14	2	1	13	13	12	6	1		72	6	4	16		26	12	11	23	35	1	7	6	4	7	10
🟦 2011	23	13	4	17		29	16	16	3	87	9	14	3	2	13	15	14	7	1		74	5	5	13	1	27	12	11	18	45	1	8	5	4	7	10

Fig. A5 Institutions mapped in RISIS Patent by Year