The advantages of an Ontology-Based Data Management approach: openness, interoperability and data quality

Cinzia Daraio[1], Maurizio Lenzerini[1], Claudio Leporelli[1], Paolo Naggar[2], Andrea Bonaccorsi[3], Alessandro Bartolucci[2]

[1] daraio@dis.uniroma1.it *(corresponding author);* lenzerini@dis.uniroma1.it, leporelli@dis.uniroma1.it;
Department of Computer, Control and Management Engineering Antonio Ruberti (DIAG), Sapienza University of Rome, via Ariosto, 25 00185 Rome (Italy)

[2] *paolo.naggar@gmail.com; alessandro_bartolucci@fastwebnet.it*
Studiare Ltd., Rome (Italy)

[3] *a.bonaccorsi@gmail.com*
DISTEC, University of Pisa (Italy)

Abstract

We illustrate the usefulness of an Ontology Based Data Management (OBDM) approach to develop an open information system, allowing for a deep level of interoperability among different databases, and accounting for additional dimensions of data quality compared to the standard dimensions of the OECD (2011) Quality Framework. Recent advances in engineering in computer science provide promising tools to solve some of the crucial issues in data integration for Research and Innovation (R&I).

Keywords: *Data integration, Open data, Comparability, Standardization, Modularization, Interoperability, Data quality, Research and Innovation*.

1. Introduction

According to an estimate of the world's technological capacity to store, communicate, and compute information, based on sixty analog and digital technologies in 2007, humanity was able to store "2.9 × 1020 optimally compressed bytes, communicate almost 2 × 1021 bytes, and carry out 6.4 × 1018 instructions per second on general-purpose computers. General-purpose computing capacity grew at an annual rate of 58%. The world's capacity for bidirectional telecommunication grew at 28% per year, closely followed by the increase in globally stored information (23%)" (Hilbert and López, 2011).
Making data widely available is very important for scientific research as it relates to the responsibilities of the research community toward transparency, standardization, and data archiving. However, to make data available, researchers have to face the huge amount, complexity, and variety of the data that are being produced (Hanson, Sugden, & Alberts, 2011). Moreover, the availability of data is not homogeneous for all disciplines and the cases of "Little data" and "No data" are not exceptions (Borgman, 2015).
In the last years there has been an extraordinary development of open access repositories all over the world (Pinfield et al. 2014). Open data[1] initiatives have been placed on the agenda of policy makers worldwide (Huijboom & Van den Broek, 2011) as a means for improving the efficiency and effectiveness of government, transparency and participation. At the European level, the eGovernment Action Plan (2011-2015) committed EU's member states to maximise "the value of re-use of public sector information (PSI), e.g. by making raw data and documents available for re-use in a wide variety of formats (including machine-readable ones) and languages and by setting up PSI portals" (European Commission, 2010).
And in a broader perspective, are open data sufficient for the realization of the open science, aiming at improving efficiency in science, increasing transparency and quality, speeding the transfer of knowledge, increasing knowledge spillovers to the economy and society, addressing global challenges more effectively, promoting citizens' engagement in science?
Data and information produced by public institutions and publicly funded projects are a kind of global public good affected by two contrasting trends. Firstly, Internet offers new opportunities for overcoming geographic limitations and the promise of unprecedented open access to public information for research on a global basis generating positive externalities and network effects. Secondly, there are growing restrictions on the availability and use of public data and information arising from the privatization and commercialization of such sources. The critical questions on how to encourage access to and sharing of such public scientific resources without unduly restricting new opportunities for commerce or the rights of authors, and how should commercial activities in the private sector be promoted without significantly compromising the availability of data and information in the public domain or through open access for global public good purposes are not new (National Research Council, 2004). However, despite the long discussion on these issues, definitive or convincing solutions on these matters have not yet been found (Borgman, 2015). In addition, these trends show specific features for developing countries and may have implications for their economic development (National Research Council, 2012).
Hence, the exploitation of the Big data potentials is strictly linked, among other factors, to issues of privacy, security and consumer welfare. As shown by Kshetri (2014), the costs, benefits and externalities associated to the use of Big data vary according to the specificities of users and their technological propensity. There is not a "one size fits all" model for the management of Big data.

Moed (2016) proposes the set of creative ideas developed by Nielsen (2012) as a framework or "architecture of attention" within which altmetrics can be positioned and further explored: "His work represents a thorough, systematic account of the potential of online tools in the research process, and, in this way, articulates the practical realization of the ethos of science and scholarship in the computerized or digital age" (Moed, 2016).
This new foundation of altmetrics as a sign of the computerization of the research process opens to new ways of including these additional elements of the research process and of the scholarly communication in the activities of research assessment and science policy.

According to Floridi (2014) we are living the fourth information revolution, living as interconnected informational organisms (inforgs), sharing with biological agents and engineered artefacts a global

[1] According to OECD (2015), open data are "data that can be used by anyone without technical or legal restrictions. The use encompasses both access and reuse." OECD (2015, p. 7).

environment ultimately made of information, the infosphere. Within this global environment, the current development of the Information and Communication Technology (ICT) is really offering new opportunities for the creation, organization and diffusion of new knowledge, that will lead us towards *designed serendipity* (Nielsen, 2012), and, one step further, towards the "open science[2]" imperative.

In a related paper (Daraio et al., 2016), we introduce an OBDM approach to coordinate, integrate and maintain the data needed for science, technology and innovation policy and illustrate its potentials for specifying STI indicators and developing science of science policies. Therein we outlined the main advantages of OBDM with respect to the traditional sylos-based approach to data integration, namely: *conceptual access to the data*, *re-usability*, *documentation and standardization*, *flexibility*, *extensibility and opening of the system*.
In this paper we focus our analysis on three main advantages of OBDM in data integration for research and innovation analysis, which encompass and further expand those listed in the earlier paper, namely openness, interoperability and data quality.

The paper is organized as follows. In the next section we describe the usefulness and the modularity of an open OBDM system. Next, the main building blocks of an open OBDM system are described. The following sections illustrate the deep degrees of interoperability and the additional dimensions of data quality allowed by an OBDM system, while the final section concludes the paper.

2 Motivations, usefulness and modularity of an open OBDM system[3]

An ontology is a formal representation of a domain of interest relevant to an organization, expressed in terms of objects, concepts (or object classes), links between objects and relationships between concepts. The ontology of the domain of interest of an organization can perform various functions. In the context of the current paper, the most interesting function is providing a shared vision of the structure of the domain of interest. It is therefore a key element for sharing the knowledge of the domain with all operating actors and to interact with users.
An OBDM system builds the explicit representation of the domain (ontology) and connects it in a formal way with the data sources through so-called mappings. A system of OBDM is therefore made of three layers: 1) the representation of the conceptual domain, called *ontology*, 2) the correspondence between the data sources and the concepts and relations of the ontology, called *mapping*, and 3) the data sources, described through schemes and related information.
It is the responsibility of the administrators of the system to keep all components of the system up to date, in particular the ontology and the mapping. The complexity of building and maintaining the ontology mappings obviously implies a cost, but the basic principle is that this cost is actually an investment, given the existence of effective software tools that allow users to make use of information services through the ontology, without taking into account the constraints and idiosyncrasies of the heterogeneous data sources.

Research on ontologies and their use in the context of databases integration conducted in recent years has produced a series of relevant scientific results (Poggi et al. 2008; Calvanese et al. 2009; Lenzerini, 2011) including a set of instruments that actually make this paradigm used in practice.
The paradigm of the OBDM includes: 1) a methodology, 2) a set of formal languages to express the artifacts that comprise it (ontology, mapping, sources), 3) a set of software tools that support the methodological

[2] According to OECD (2015), 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, 2015, p. 7).
[3] The presentation of the OBDM in this section and and in the next one follows the lines of Poggi et al. 2008; Calvanese et al. 2009; Lenzerini, 2011; Calvanese et al., 2011; Civili et al. 2013.

processes, including automated reasoners based on logic[4] and an overall management system and documentation which integrates all aspects and represents an access point for both analysts and managers of the system, and for users of data services.

Motivations for adopting an OBDM approach

There are several reasons for adopting an OBDM approach.

An OBDM approach is suitable when each group of users has a clear understanding only of particular portions of the domain, adopts its own (informal) representation of it, and refers to common concepts with a specific terminology. This results in the lack of a shared (and formalized) specification of the knowledge on the overall knowledge domain.

Data in this framework are managed in various systems, which underwent several modifications over the years, often to serve specific application needs, so that they have lost the original shape and modelling, often without an adequate documentation, and are now easily accessible only by few experts of the systems, whereas current databases in use are essentially incomprehensible for the users' domain.

An OBDM approach is required when the integrity constraints on data are not forced in the systems, or are easily circumvented, so that their quality is compromised.

Finally, an OBDM approach is convenient for the cases of updating and extensions of the information system. Each time a new information need arises, managers of the system would have to launch a new complicate process that would typically require much more additional work to be accomplished.

Usefulness of an OBDM approach

An OBDM approach is suitable for Research and Innovation (R&I) information systems, where data governance and data access can be greatly enhanced by the use of the ontology as a representation of a *common language* of the domain.

Scientific data management, would be enhanced in those R&I fields in which ontologies are available as unified representations of relevant meta-data.

In the public administration of R&I and government data management, the OBDM paradigm can be *the enabling technology* for information sharing and semantic interoperability.

An OBDM facilitate the *open data publishing process*, because the ontology can help to determine what to publish and which strategy to follow in order to enrich the data with useful meta-data.

Investment and modularity of the system

Following a real options approach in investment theory (Li and Johnson, 2002), we conceive a data platform as an asset allowing repeated use. In this context, investment costs are made by front-up costs for the platform, maintenance costs and recurring costs for projects. The revenues instead are the gains from better decisions in policy making (e.g. the possible use for performance-based allocation of public resources; the possible use for strategic priorities in S&T; or to set up public subsidies to firms for industrial R&D). A real options analysis in this context should follow a modular engineering design perspective (Baldwin and Clark, 2000) in which a quantitative model to describe the economic forces that push a design towards modularization and the consequences of modularity on the business environment are described. In this context, value creation is the goal of the modularization process and real options theory offers a natural framework to evaluate the modularization of the design of the system. There are also criteria to assess the decomposition of systems into modules (Parnas, 1972).

Modularity is a property of quasi-decomposition of hierarchical systems, based on the minimization of the interdependence of sub-systems (see Simon, 1962).

The modification of sub-systems does not require the re-design of the entire system. Making the design of products modular requires a large front up investment in conceptual design. The standardization of interfaces is necessary. However, the design of successive versions of the product and/or re-design becomes cheaper.

In an OBDM approach, the modular design and its implementation requires an initial large scale investment into the formal definition of the main relevant concepts (and relationships among them) of the domain of interest, but is facilitated by suitable graphical tools (that we will see below) which allow an easy modularization and updates of the relevant domain.

[4] An automated reasoner based on logic is a software able to derive logical consequences from a given set of axioms in an automatic way.

3 Key components of an Open OBDM Information Infrastructure

Supporting the management of OBDM applications requires to provide effective tools for[5] (i) allowing both expert and non-expert users to analyze the OBDM specification, (ii) collaboratively documenting the ontology, (iii) exploiting OBDM services, such as query answering and automated reasoning over ontologies, e.g., to support data quality check, and (iv) tuning the OBDM application towards optimized performances. To fulfil these requirement, the system called MASTRO STUDIO, based on a tool for automated reasoning over ontologies: MASTRO (Methods And Systems for Tractable Reasoning over Ontologies) reasoner (Calvanese et al. 2011), enhanced with a suite of tools and optimization facilities for managing OBDM applications, has been proposed (Civili et al. 2013).
In the following we briefly describe the main tools of the MASTRO STUDIO technology.

Description logic DL lite approximator

To the best of our knowledge, MASTRO is, along with ONTOP (developed by the University of Bolzano, also derived from MASTRO), the only system worldwide that can perfectly respond, in logical terms, to SPARQL[6] query on ontology expressed in Ontology Web Language (OWL) connected to a data layer (data sources managed by external systems, accessible via a Structured Query Language –SQL- endpoint) by mapping. MASTRO can achieve this functionality with computational costs similar to those of relational databases because it is based on logics of the DL-Lite family of Description Logics (Baader et al. 2007), well-known for providing a good trade-off between expressivity and reasoning computational complexity,.

The fundamental characteristic of this family of logics (including DL-LiteA) is that all the problems of ontology reasoning expressed in this logic are computationally tractable, or solvable in polynomial time (as opposed to OWL, in which these problems have all exponential complexity). In addition, calculate the answer to the query SPARQL than ontologies expressed in DL-Lite has the same complexity of calculation of the response to the SQL query in the relational data base, a unique feature in the languages for ontologies (other fragments treatable of OWL have indeed more complexity, and about OWL 2 its decidability has not yet been demonstrated).

The algorithm used by MASTRO has revolutionized the field of ontological reasoning systems. It is in fact the first algorithm based on the technique of *rewriting*. The original query is first translated to another query SPARQL taking into account the axioms of the ontology. The obtained query is then rewritten on the basis of the mapping, in a SQL query that is evaluated on the sources.

The DL-Lite approximator comprised in the MASTRO STUDIO environment is then a service that takes in inputs an ontology expressed in OWL 2 and allows an automatic reasoning on the ontology with acceptable computational costs and minimizing the loss of original information.

Graphical representation through Graphol

In MASTRO STUDIO ontologies are specified and represented by means of a graphical language, Graphol (http://www.dis.uniroma1.it/~graphol/), aiming both at making them accessible to non-experts of logical and ontology formalisms, and at capturing the main modelling features of OWL. This effectively supports the definition and the analysis of the ontology.
Using an editor, it is possible to construct the corresponding chart ontology expressed in Graphol, which can be automatically translated (with a suitable translator downloaded from the website of Graphol) in a superset of OWL, or in a set of axioms OWL, possibly with the addition of some axioms that are not directly expressible in OWL (such as those of identification and denial of DL-lite). The graph expressed by Graphol illustrates and highlights the relationship between the various concepts. The purpose of the graph is to offer a schematic view of the ontology, to focus attention on the concepts and how they are mutually linked in the

[5] This presentation follows the lines of Calvanese et al. (2011) and Civili et al. (2013).
[6] SPARQL is a semantic query language for databases.

representation. The usefulness of this language and its tools has been tested in many national and international projects and is witnessed by the fact that several institutions and companies (including the Italian National Statistical Office (ISTAT), Italian Ministry of Economics and Finance (MEF), Telecom Ltd) are adopting it. Graphol is also used to the development of *Sapientia*, the Ontology of Multidimensional Research Assessment, that we will describe below, in the following of the section.

Wiki-like documentation and open data as a simple query over the ontology in an open OBDM system realized with MASTRO STUDIO

MASTRO STUDIO provides the capability to equip the ontology with a wiki-like documentation that, for every ontology element (concept, attribute or role) (i) specifies its meaning (in natural language) and (ii) reports on the ontology and mappings assertions in which it is involved.
By adopting an OBDM approach it is possible to access the relevant data and let them openly available simply by expressing a query over the ontology and using MASTRO to translate the original query, based on the mapping between sources and ontology, in a query on the source that extracts exactly the requested content. This ensures that the extraction of the same portion of knowledge for different data set is carried out in the same way, because synthesized by an automatic system on the basis of a specific logic. At the same time, the query produced by the system, accessing also the meta-data directly managed by MASTRO STUDIO, can extract the meta-data relevant to describe the contents of the dataset, once again according to an automatic procedure.

Web-based information system

MASTRO STUDIO is a web-based system that, through the use of MASTRO and the other support tools (approximator, Graphol viewer, editor, etc.), and through the management of appropriate meta-data, allows to define a specific Ontology-based Data Management (OBDM) system, inspect it, share it, document it, and, verify the consistency, interrogate and produce data sets obtained with processes of extraction from the sources. As illustrated above, MASTRO STUDIO offers a *collaborative*, wiki-style environment, for publication and documentation of specific OBDM and datasets. For these purposes, it is based on some features provided by the open source content management system Drupal (https://www.drupal.org/), appropriately extended and integrated with the OBDM services.
In particular, MASTRO STUDIO, from a specific OWL ontology, automatically generates a structure of a wiki in which each predicate (concept, attribute, relationship) of the ontology is associated with a web page where you can enter descriptions in natural language, enriched by information extracted automatically from the specific ontology. For example, in the page associated to a concept, the description is accompanied by the list of attributes and key roles, the specializations and generalizations of the concept, and from the list of OWL axioms involving the concept. In addition, MASTRO STUDIO allows access the pages associated with the predicates of the ontology directly working on them through the graphic representation provided by the ontology diagrams (realized with Graphol). Wikis are available also to specify, inspect and document the mapping, and to inspect the sources of some data. In case of update of the specific OBDM, for example due to changes of the ontology, MASTRO STUDIO allows you to align in a semi-automatic way the documentation contained in the wiki pages to the new version of the specification.
The collaborative process of drafting the documentation is supported by the possibility of defining different categories of users and editors, and of distinguishing between the content published and those under development or approval. This process is based on the functionality and content management offered by specific Drupal modules, integrated with the process of updating the specification and subsequent alignment of the documentation mentioned earlier.
The specification of the mappings in MASTRO STUDIO is conform to R2RML, the W3C standard to define mappings between relational databases and RDF (Resource Description Framework –RDF- is a standard model for data interchange on the Web) datasets .
In particular, MASTRO STUDIO allows specification of SPARQL queries on the ontology, and calculate the answers using MASTRO. Thanks to this feature, *the management of data* provided by MASTRO STUDIO can be operated *by external systems*, generally independent to the system itself, and does not need to be materialized in a format that conforms to the specific ontology.

MASTRO STUDIO offers additional services related to data governance, as the verification of the provenance and of the data quality. Data quality aspects will be dealt in more details in a next section of the paper.
Finally, MASTRO STUDIO is equipped with a semi-automatic tuning mechanism, aiming at optimizing OBDM applications.

The next Figure 1 illustrates the main components of the OBDM system described above while Table 1 summarizes its main functionalities.

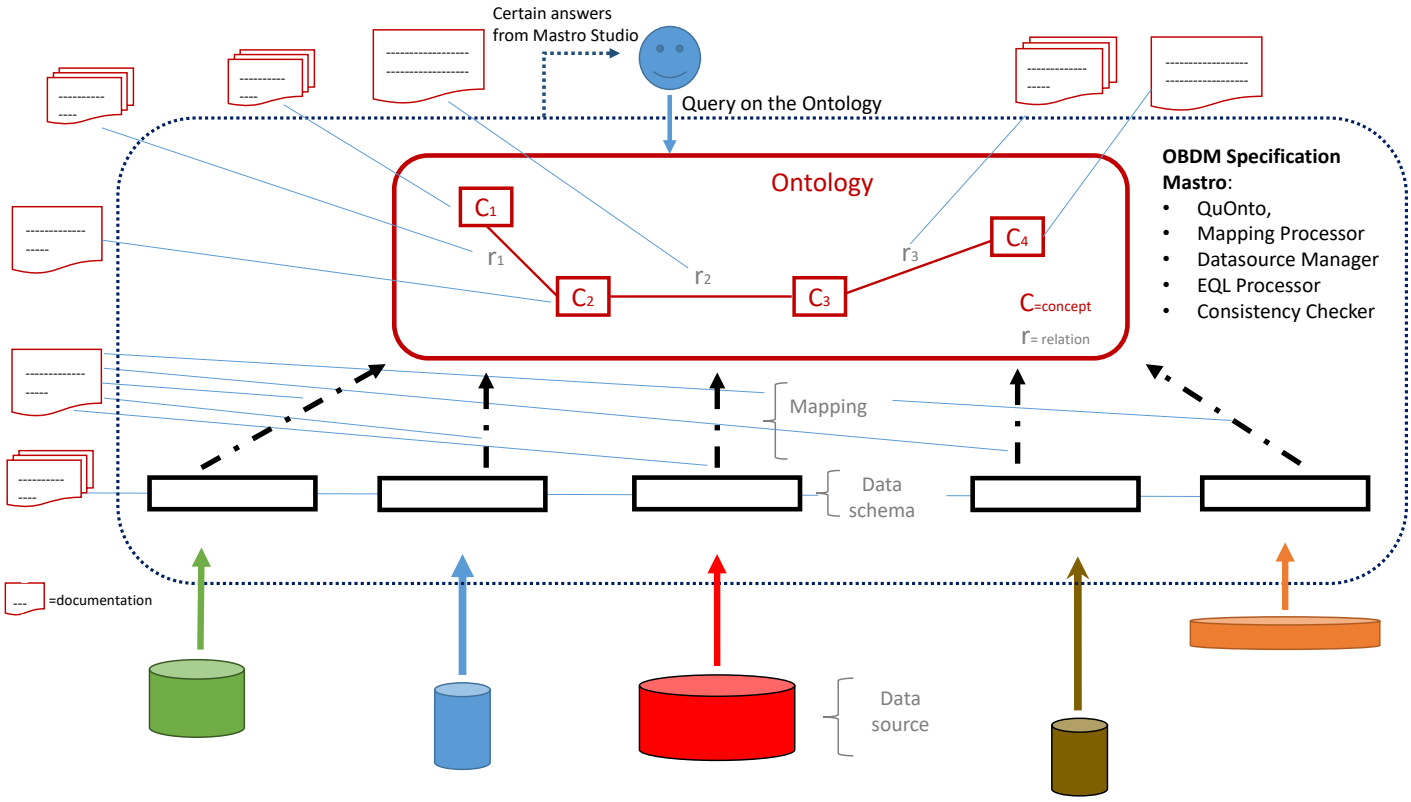

Figure 1. Illustration of an open OBDM information system

Functionality	Description
Access to the ontology documentation	By browsing the ontology wiki-like documentation provided by MASTRO STUDIO users can be introduced to the overall ontology semantics, as well as to the semantics of each ontology element. Users can test and join the collaborative semi-automatic process supporting the production of such documentation.
Analysis of the ontology	The users can experience the richness of the ontology through the reasoning facilities offered by MASTRO STUDIO. In particular, the diagrammatic representation is a means to get an easy access to the ontology and to disclose it to users not used to complex formalizations.
Analysis of the mappings	Users and interested people can have a look at the mappings to have an insight of the "cognitive distance" between the ontology and the sources, i.e., the huge difference between the data schema of the sources and the conceptualization of the domain, and will provide at the same time a mechanism to understand the sources in the light of the ontology, offering a valid documentation means.
Check of the data quality	Users are able to identify unsatisfiable ontology assertions, and retrieve source data that violate them. This shows that MASTRO STUDIO allows us to localize inconsistencies in the data, thus resulting in a valid support for data quality management (see also next section for more details on these aspects).
Querying the system	Users can issue queries over the ontology, and for each query, to access, besides its results, both the ontology rewriting and the mapping rewriting. By analysing the ontology rewriting, they will discover the kind of reasoning at the ontology level which is automatically performed by the system to produce the result. Furthermore, by analysing the mapping rewriting, they will be able to see how and from which sources results to specific queries come from.
Tuning the system	MASTRO STUDIO allows a tuning of the system, which is able to learn from previous processing of queries in order to avoid to execute some reasoning steps it already performed.

Table 1. Main functionalities of an open OBDM information system (Source: our adaptation of Civili et al. 2013, p.4)

An open OBDM system at work

MASTRO STUDIO has been successfully implemented in different contexts. For modelling the Italian public debt for the Italian Ministry of Economy and Finance (Civili, et al.2013); Selex Sistemi Integrati (SELEX-SI) a Finmeccanica Company, leader in the provision of integrated defence and air traffic control systems, Monte dei Paschi bank; Network inventory systems in the telecommunication context (Calvanese et al. 2011).

The OBDM approach has started to be implemented in a research project funded by the University of Rome La Sapienza in 2013-2015. The main output of this project has been *Sapientia*, the Ontology of

Multidimensional Research Assessment (see Daraio et al. 2016). *Sapientia* models all the activities which are related to the evaluation of the research and its impacts.[7]
The current version of *Sapientia*, version 2.0, includes 11 modules that are organized according to Figure 2.

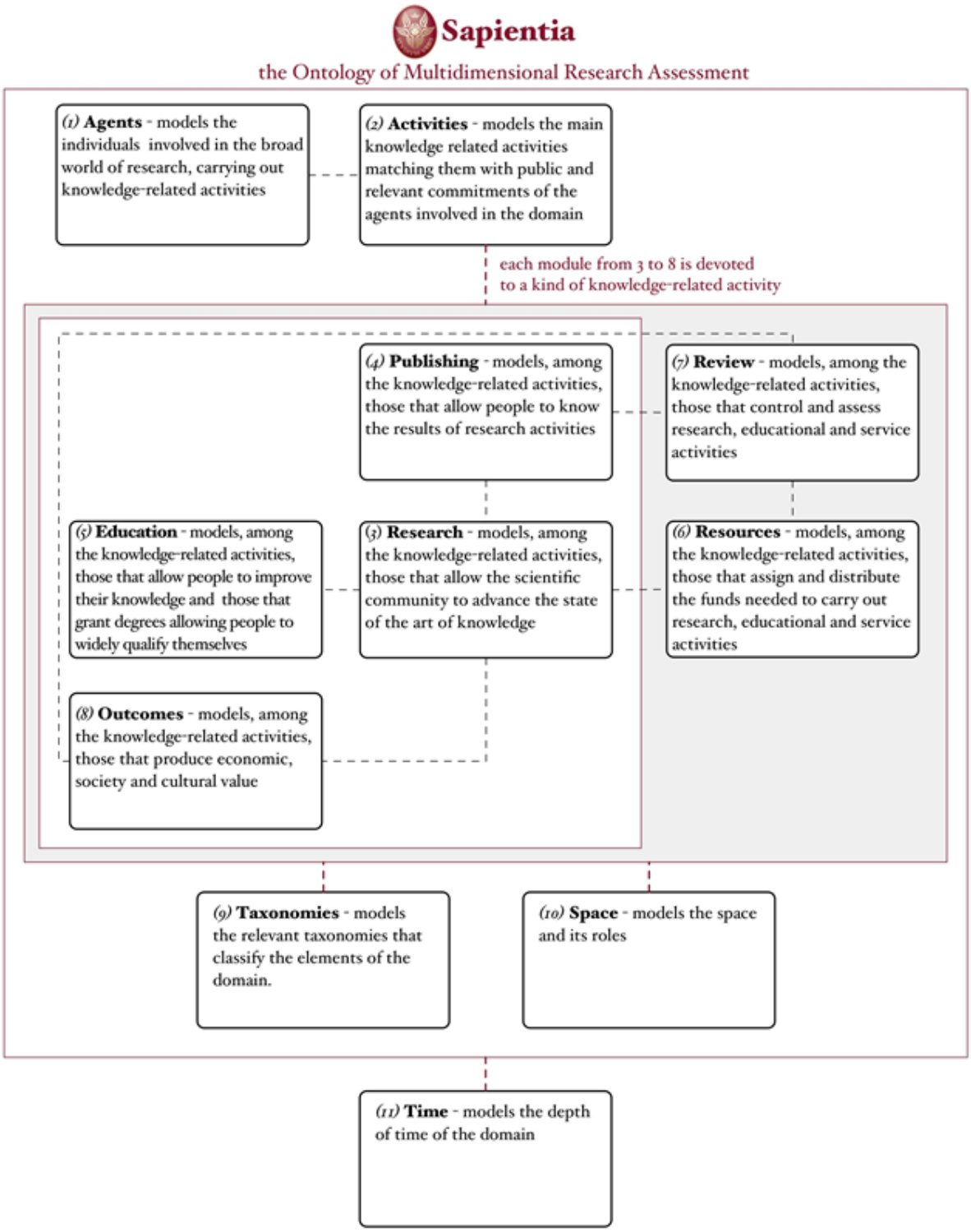

Figure 2. The 11 Modules of *Sapientia 2.0*: the Ontology of Multidimensional Research Assessment.

[7] Sapientia 1.0 was closed on the 22nd of December 2014, and was organized in 14 Modules, including around 350 symbols (concepts, relations and attributes). It has been presented at the Workshop of the 20 February 2015 held at Sapienza University of Rome (see Daraio, 2015).

Figure 2, shows the organization of the modules of *Sapientia* and their contents. *Sapientia* models the main activities (Module 2) carried out by the agents (Module 1). The ontology includes a core set of modules which are Research (Module 3), Education (Module 4) and production, including services and other third mission activities (Module 8). These activities are part of an extended set of modules which include an ancillary module of Research (Module 4 Publishing) and other two modules containing relevant activities to foster the relationships among the core set of modules (i.e., Modules 6 Resources, including funding and projects, and Module 7 Review).

4. Interoperability and levels of representations of the domain in an OBDM system

Data integration is the problem of combining data residing at different sources, and providing the user with unified view of these data (Lenzerini, 2002). According to Parent and Spaccapietra (2000), *interoperability* is the way in which heterogeneous systems talk to each other and exchange information in a meaningful way. They identified three levels of interoperability, from the lowest level (no integration), to an intermediary level (the system does not guarantee consistency across database boundaries) to an higher level that has the goal of developing a global system on top of existing system, to provide the desired level of integration of the data sources.
Some of the theoretical issues that are relevant for data integration have been identified in modeling a data integration application, processing queries in data integration, dealing with inconsistent data sources, and reasoning on queries (Lenzerini, 2002).
Several levels of conceptual interoperability have been identified in the specialized literature. For instance, Tolk and Muguira (2003) propose the following 5 levels of conceptual interoperability:

- Level 0: System specific data (isolated systems);
- Level 1: Documented data (documentation of data and interfaces);
- Level 2: Aligned static data through Meta Data Management (use of common reference models/common ontology);
- Level 3: Aligned dynamical data and "Implemented processes" (common system approach/open source code);
- Level 4: Harmonized data and processes, conceptual model, intend of use (common conceptual model/semantic consistency).

The formal and precise means to achieve level 4 of interoperability (harmonized data and processes) is a *logic-oriented ontology language*. This is exactly what the OBDM approach described in the previous sections provides. As a matter of fact, an open OBDM approach offers different levels of representations of the domain, that are:

-*Level of representation 1*: through a glossary, which contains a list of terms denoting concepts, attributes and relationships relevant to the domain; associated with these terms are considered also their descriptions or definitions in natural language.
-*Level of representation 2*: in which synonyms, homonyms, hyponymy, and so on, are added to obtain a more complete formulation of the glossary, called structured glossary.
-*Level of representation 3 (light ontology description: DL-lite*): In this level the basic elements to define a conceptualization of the domain are added. The terms of the previous level become symbols to denote concepts, properties (also called attributes) and relationships between concepts. The links that occur between these elements are specified by basic logic axioms such as: axioms of specialization and generalization (subtyping / supertyping), domain and range axioms relations (object-property typing), axioms of type and co-domain for attributes (value-property typing), axioms of disjointness, axioms functionality (functionality) and axioms identification of concepts (identification).
-*Level of representation 4* (representation in OWL), exploiting all the potentials of both DL-lite and OWL to formally conceptualize the domain.
-*Level of representation 5* (enriched ontology): ontology expressed in level 4 of representation may be finally enriched with other logical axioms that are not part of the repertoire of OWL and DL-lite.

As appears from the levels of representations, the OBDM approach is based on a logic description of the domain since the level 3.

By combining what said above with the content of the previous section we can conclude that the technology MASTRO STUDIO, which implements the OBDM approach, is able to ensure *full harmonization* and *interoperability* between heterogeneous databases in an open information system environment.

5. Data quality

An open OBDM framework like the one illustrated in Figure 1 offers the possibility to carry out a "formal" approach to data quality which goes beyond the OECD (2011) Quality Framework, based on the seven dimensions represented by: relevance; accuracy; credibility; timeliness; accessibility; interpretability and coherence.
Console and Lenzerini (2014) show that by adopting an OBDM approach it is possible to define a broader concept of *data consistency* and present algorithms and complexity analysis for several relevant tasks related to data quality consistency checks. They report the example of *satisfiability checking*, that consists in checking whether there are patterns in the data contradicting the axioms in the ontology. In MASTRO STUDIO, satisfiability can be reduced to query answering, based on the fact that to each ontology axiom we can associate a query aiming at identifying the existence of patterns representing violations in the data (see Console and Lenzerini, 2014).

Generally speaking, MASTRO STUDIO offers two kind of reasoning services:
-*intensional services.* These services are based on the ability, given a formula, to check whether it is logically implied by the axioms constituting the ontology. These services concern the reasoning over the ontology without taking into account the mappings with the data sources.
-*extensional services.* These services concern the mappings and the data sources. They are based on query answering and traditional data quality checking.

As a consequence, an OBDM approach offers the possibility to extend the traditional framework and dimensions of data quality by checking the quality of data both at the *extensional* level (content of the data sources) and at the *intensional* level (schema of the data sources and its connection with the ontology).
Within this broader framework, traditional data quality analysis is enriching by the comparison of the data at the sources with the axioms of the ontology.

In an OBDM system, the following data quality analyses may be carried out.

- *Analysis of the semantics of the sources and their description in terms of the ontological domain.* The objective of this analysis is to produce the formal description of the content of the sources of the data in terms of the representation of the domain of interest and of description of integrity constraints intra- and inter-sources. It is based on how the sources have been defined in the corresponding systems.
- *Analysis of the elements of the sources and identification of the mappings between these elements and the representation of the domain***.**
This analysis has the aim of formally characterize the meaning of the sources. In the OBDM methodology this is achieved by formally specifying the content of the sources in terms of the representation of the domain. If the source is structured or semi-structured, the correspondence with the ontology is realized through the definition of a set of *mapping assertions*. Intuitively, a mapping assertion determines that the extracted data from the sources by a certain query correspond to instances of a certain set of elements of the ontology, specified through a pattern of queries on the ontology.
It is important to note that the correct specification of the mapping between the sources and the ontology also requires to deal with what is in the literature called "data deduplication" (also called entity resolution, record linkage, etc.), or a step whose purpose is to identify elements of the sources that are referred to the same object or the same phenomenon of the reality.
-*Analysis of the quality of sources*
The purpose of this analysis is to conduct a formal review of the quality of the sources, and suggest appropriate activities to improve their level of quality. The distinguishing feature of this approach is to

consider the representation of the domain as the reference for assessing the quality of sources, as recalled above.

-*Analysis of the quality of the data*

The dimensions of data quality which can be considered are various and depend on the objective of the analysis. For instance, we may consider: accuracy (degree of adherence to the mathematically representation), consistency , completeness (degree of coverage of knowledge on the phenomena to be represented), originality (if the data is genuine or derived from other sources) and degree of timeliness. The main feature of this analysis is that, in addition to the classical data quality tools, the verification of the quality dimensions is also performed using both specific assertions on the ontology (through formal procedures, essentially based on logic) and the comparison between the sources and the representation of the domain (mappings and data sources).

-*Analysis of the quality of the source schemas*

With regard to the verification of the quality of source schemas, an OBDM approach allows for a systematic comparison of the patterns of the sources and the representation of the domain, in order to assess the completeness of the schemas (also called coverage), their minimality and their adequacy (other dimensions include readability or normalization that are important for the governance of the information system). To carry out these analysis, the use of reasoners, especially MASTRO, is crucial to perform formal verifications and to support analysts of the platform with automatic tasks.

6. Conclusions

In a related paper (Daraio et al. 2016) we introduce the OBDM idea to integrate heterogeneous data in the field of Research and Innovation (R&I). In this paper we show that the OBDM is a technology, not merely an idea, and it can be *the enabling technology* for *information sharing and semantic interoperability*. We summarize its main building blocks. To the best of our knowledge, this is the first technology capable of handling three concepts that are essential in informetrics, but that have not yet received sufficient attention, namely *openness*, *interoperability* and *data quality*. It handles these aspects in a fully wherein and structured manner.

We believe that the application of this approach of data integration and management in the study of science and innovation could reveal very promising to address important informetrics open issues and hence this new area of research deserves to be further explored.

Acknowledgments

The helpful and precious comments and suggestions of Henk F. Moed are warmly acknowledged. Research support from the Award Project 2015 no. C26H15XNFS of the Sapienza university of Rome is gratefully acknowledged.

References

Baader F., D. Calvanese, D. McGuinness, D. Nardi, P. F. Patel-Schneider, (eds) (2007). The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press, 2nd edition.

Baldwin, C. Y., Clark, K. (2000), Design Rules - The Power of Modularity, MIT Press, Cambridge.

Borgman, C. L. (2015). Big data, little data, no data: scholarship in the networked world. MIT Press.

Calvanese D., De Giacomo G., Lembo D., Lenzerini M., Poggi A., Rodriguez-Muro M, Rosati R.: Ontologies and Databases: The DL-Lite Approach. Reasoning Web 2009: 255-356.

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. J. Autom. Reasoning 39(3): 385-429.

Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M. Rosati, R. (2009), Ontology-Based Data Access and Integration. Encyclopedia of Database Systems, Springer.

Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., Rodriguez-Muro, M., Rosati, R., Rizzi, M., Savo, D.F. (2011). The Mastro system for ontology-based data access. Semantic Web, 2(1), 43-53.

Civili, C., Console, M., De Giacomo, G., Lembo, D., Lenzerini, M., Lepore, L., .. Santarelli, V. (2013). Mastro Studio: Managing ontology-based data access applications. Proceedings of the VLDB Endowment, 6(12), 1314-1317.

Console M., Lenzerini M. (2014): Data Quality in Ontology-based Data Access: The Case of Consistency. AAAI 2014: 1020-1026.

Daraio C. (2015), (Eds.), Efficiency, Effectiveness and Impact of Research and Innovation, Proceedings of the Workshop of the 20 February 2015 DIAG, Sapienza University of Rome, Efesto Edizioni, Rome, ISBN 9788899104306.

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

Floridi, L. (2014). The fourth revolution: How the infosphere is reshaping human reality. OUP Oxford.

Hanson, B., Sugden, A., Alberts, B. (2011). Making data maximally available. Science, 331(6018), 649-649.

Hilbert, M. López, P. (2011). The world's technological capacity to store, communicate, and compute information. Science, 332(6025), 60-65.

Huijboom, N., Van den Broek, T. (2011). Open data: an international comparison of strategies. European journal of ePractice, 12(1), 4-16.

Kshetri, N. (2014). Big data′ s impact on privacy, security and consumer welfare. Telecommunications Policy, 38(11), 1134-1145.

Lenzerini M. (2002), Data Integration: A Theoretical Perspective. PODS 2002: 233-246.

Li, X., Johnson, J. D. (2002). Evaluate IT investment opportunities using real options theory. Information Resources Management Journal, 15(3), 32-47.

Moed, H. F. (2016). Altmetrics as Traces of the Computerization of the Research Process. In C. R. Sugimoto (Ed.), *Theories of Informetrics and Scholarly Communication*. A Festschrift in Honor of Blaise Cronin (pp. 360–371). Berlin: De Gruyter.

National Research Council (2004). Open Access and the Public Domain in Digital Data and Information for Science: Proceedings of an International Symposium. Washington, DC: The National Academies Press.

National Research Council (2012). The Case for International Sharing of Scientific Data: A Focus on Developing Countries. National Academies Press, Washington, D.C.

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 (2015). Making Open Science a Reality. OECD Science, Technology and Industry Policy Papers No. 25, OECD Publishing, Paris. http://dx.doi.org/10.1787/5jrs2f963zs1-en

Parent, C., Spaccapietra S. (2000) "Database integration: the key to data interoperability." In Advances in Object-Oriented Data Modeling, Papazoglou M.P., Zari Z. Eds. (2000), The MIT press, 221-253.

Parnas, D. L. (1972). On the criteria to be used in decomposing systems into modules. Communications of the ACM, 15(12), 1053-1058.

Pinfield, S., Salter, J., Bath, P. A., Hubbard, B., Millington, P., Anders, J. H., Hussain, A. (2014). Open-access repositories worldwide, 2005–2012: Past growth, current characteristics, and future possibilities. Journal of the Association for Information Science and Technology.

Poggi A., D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, R. Rosati. (2008). Linking data to ontologies. J. on Data Semantics, X:133–173.

Simon, H. A. (1962). The architecture of complexity. Proceedings of the American Philosophical Society, 106, 467-82.

Tolk, A., Muguira, J. A. (2003, September). The levels of conceptual interoperability model. In Proceedings of the 2003 Fall Simulation Interoperability Workshop (Vol. 7).