# 15. Challenges, Approaches and Solutions in Data Integration for Research and Innovation

Maurizio Lenzerini, Cinzia Daraio

In order to be implemented by policy makers, science, technology, and innovation (STI) policies and indicator building need data. Whenever we need data, we need a method for data management, and in the era of big data, a crucial role is played by data integration. Therefore, STI policies and indicator development need data integration. Two main approaches to data integration exist, namely procedural and declarative. In this chapter, we follow the latter approach and focus our attention on the ontology-based data integration (OBDI) paradigm. The main principles of OBDI are:

(i) Leave the data where they are.
(ii) Build a conceptual specification of the domain of interest (ontology), in terms of knowledge structures.
(iii) Map such knowledge structures to concrete data sources.
(iv) Express all services over the abstract representation.
(v) Automatically translate knowledge services to data services.

We introduce the main challenges of data integration for research and innovation (R&I) and show that reasoning over an ontology connected to data may be very helpful for the study of R&I. We also provide examples by using *Sapientia*, an ontology specifically defined for multidimensional research assessment.





## 15.1 The Role of Data Integration for Research and Innovation

In the last years, the amount of data available for research and innovation (R&I) is growing, in particular thanks to data collections and other initiatives of international and national organizations. While the availability of data stored in current information systems and the processes making use of such data are exponentially increasing, turning this data into information and governing both data and processes are still great challenges in the context of information technology (IT) [15.1–3]. These issues arise from the proliferation of data sources and services both within a single organization and in cooperating environments. Data integration and data interoperability, which have been important in the last decades, are even more important today, in the big data era (see, for instance, two recent books on big data integration [15.4, 5]). Some of the theoretical issues that are relevant for data integration and data interoperability are modeling a data integration application, extract-


W. Glänzel et al. (Eds.), *Springer Handbook of Science and Technology Indicators*, Springer Handbooks,
https://doi.org/10.1007/978-3-030-02511-3_15

ing and exchanging data from relevant sources, dealing with inconsistent data sources, and processing and reasoning on queries [15.6].

According to *Parent* and *Spaccapietra* [15.7], interoperability is the way in which heterogeneous systems talk to each other and exchange information in a meaningful way. They recognized three stages of interoperability, from the lowest based on no integration, to an intermediary stage in which the system does not guarantee consistency across database boundaries, to a higher stage, which has the objective of developing a global system embracing the existing systems, in order to deliver the desired level of integration of the data sources.

Two main approaches to data integration exist, namely procedural and declarative. In the procedural approach, also called bottom-up approach, for every *information need*, one figures out which data are needed and how they can be accessed, and the goal is to design and realize the corresponding service. On the other hand, in the declarative approach, also called top-down approach, one defines a global representation structure that is valid for the domain of interest underlying the data sources, links this structure to the actual data, lets the user use this structure to specify the *information needs*, and the goal is to automatically extract the right data from the sources. In this chapter, we follow the latter approach and focus our attention to the ontology-based data integration (OBDI) paradigm, which is a recently introduced declarative paradigm for data integration and governance. OBDI uses knowledge representation and reasoning techniques for a new way of integrating and governing data. The principles at the basis of OBDI can be summarized as follows:

(i) Leave the data where they are.
(ii) Build a conceptual specification of the domain of interest, in terms of knowledge structures; such a conceptual representation is called ontology.
(iii) Map such knowledge structures to concrete data sources.
(iv) Express all services over the abstract representation.
(v) Automatically translate knowledge services to data services.

An OBDI system is thus constituted by three main components: the ontology, which represents a conceptual description of the domain; the data sources, where the actual data are, and the mappings that link the data sources to the ontology. Additionally, the ontology is expressed in a form that is both computational and logical. The computational form allows the ontology not only to be understood by humans, but also to be manipulated by the computer, to aid human and machine agents in their performance of tasks within that domain. The logical form is instrumental in enabling additional properties to be inferred by logical reasoning. More generally, reasoning can be used for different goals, such as verification, validation, analysis, synthesis, and exploitation of the latest development in automated reasoning.

The main benefits of the OBDI approach for R&I, as we will see in the examples reported in the following, are related to the opportunity of reasoning over the conceptual structure of the domain (the ontology), reasoning over the mappings of the data sources to the ontology, and reasoning over the data and indicators, for their consistency analysis, their validation, and their data quality assessment (see also the conclusions for an extended summary).

Data integration for R&I is a challenging issue because R&I activities are complex and their assessment is complex too. This is because it requires a systemic approach in which research activities are considered together with education and innovation activities. Moreover, the development of models of indicators or metrics requires a comprehensive framework that includes the specification of the underlying theory, methodology, and data dimensions. Models of metrics are necessary to assess the meaning, validity, and robustness of metrics [15.8]. The complexity of R&I assessment also arises from the consideration of the *implementation* problem according to this three-dimensional framework (see [15.8] for more details).

A workshop organized during the *15th International Conference on Scientometrics and Informetrics* held in Istanbul (Turkey) on 29 June–4 July 2015 discussed the Grand Challenges in Data Integration for Research and Innovation (R&I) Policy. The grand challenges identified were: handling big data, coping with quality issues and anticipating new policy needs.

The analysis of data integration for R&I policy was framed on a groundwork scheme composed by four main areas of intervention and a list of critical issues [15.9]. The main four areas of intervention are:

- Data collection/project initiatives
- Open data, linked data, and platforms for STI
- Monitoring performance evaluation
- Stakeholders, actions, options, costs and sustainability.

The identified critical issues, without being fully comprehensive, are:

- Data quality (considered as *fitness for use* with respect to user needs, see [15.10]) issues – completeness, validity, accuracy, consistency, availability and timeliness

- Comparability problems related to heterogeneous definitions of the variables, data collection practices and databases
- Lack of standardization
- Lack of interoperability
- Lack of modularization
- Problems of classification
- Difficulties in the creation of concordance tables among different classification schemes
- Problems and costs of the extensibility of the system
- Problems and costs of the updating of the system.

Interestingly, many of these issues were already discussed in a Special Session of the ISSI Conference in 1995 in Chicago [15.11]. Moreover, the need for harmonization and standardization of data in R&I is discussed in *Glänzel* and *Willems* [15.12]. It seems that the need for "a clear and unambiguous terminology and specific standards" [15.13, p. 176] is still relevant and timely nowadays.

As described in *Daraio* and *Glänzel* [15.9], the complexity of R&I systems requires a continuous information exchange. This process is due to the communication and interaction process among all actors and agencies involved in the production, processing, and application of knowledge. All data entries, all processing, development, and application of data relevant for research, technology, and innovation have their own rules and standards.

Figure 15.1 shows some elementary rules of interferences expressed in terms of data definition and standard setting in the process of data integration for different purposes, including process monitoring, input–output monitoring, and ex-ante and ex-post evaluation. The application of appropriate standards and data harmonization is crucial to achieve the interoperability of heterogeneous sources of data.

The chapter unfolds as follows. The next section presents the problem of data integration and data governance in a general way. Section 3 describes a formal framework for ontology-based data integration. Section 4 presents the ontology *Sapientia* and the OBDI system developed for multidimensional research assessment. Section 5 presents some illustrative examples of reasoning over *Sapientia*, while Section 6 summarizes the main points and concludes the chapter.

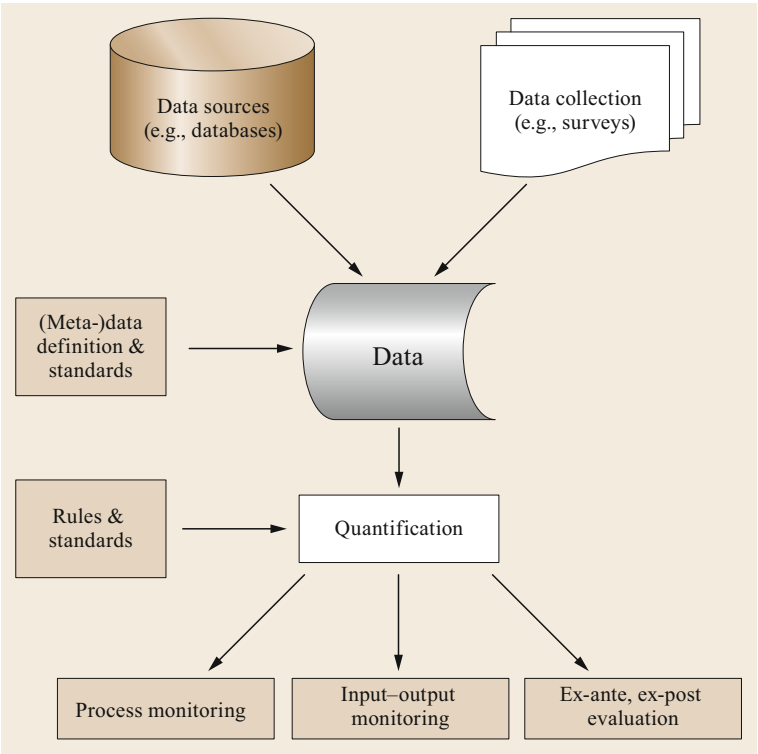



**Fig. 15.1** Data integration in use for different purposes with interference points for standardization (after *Daraio* and *Glänzel* [15.9])

## 15.2 The Problem of Data Integration and Data Governance

Big data management and analysis form a key technology for the competitive advantage of today's enterprises and for shaping the future data-driven society. However, after years of focus on technologies for big data storing and processing, many observers are pointing out that making sense of big data cannot be done without suitable tools for conceptualizing, repairing, and integrating data (http://www.dbta.com/). A common opinion in technology observers is that big data are ready for analysis; one should simply access, select, and load data from big data sources, and magically gain insight, discover patterns, and extract useful knowledge. As pointed out in [15.14], this is not the case; loading a big data platform with quality data with enough structure to deliver value is a lot of work. Thus, it is not surprising that data scientists spend a comparatively large amount of time in the data preparation phase of a project. Whether you call it data wrangling, data preparation, or data integration, it is estimated that 50–80% of a data scientists' time is spent on preparing data for analysis. If we consider that in any IT organization, data governance is also essential for tasks other than data analytics, we can conclude that the challenge of identifying, collecting, retaining, and providing access to all relevant data for the business at an acceptable cost, is huge. If we specialize the above observation to the domain of interest in this chapter, namely R&I, we can state that one of the most challenging tasks for carrying out quantitative studies in the realm of R&I is to identify, collect, integrate, organize, govern and access all relevant data.

Although data integration is one of the oldest problems in data management, the above observations show that it is a major challenge today. As we said in Sect. 15.1, in principle, there are two main approaches to this problem: procedural and declarative. In the procedural approach, sometimes called bottom-up approach, whenever an information need arises that requires accessing the integrated data, a specific program is coded, so as its execution produces the required information. In some sense, with this approach, integration is achieved on a query-by-query basis. In the declarative approach (top-down), one defines a priori an integration database structure, and in order to satisfy an information need one can simply pose a query over such a structure (Fig. 15.2 illustrates these notions relative to R&I). So, with this approach, integration is achieved independently from a specific query (or a specific set of queries).

In this chapter, we focus on the declarative approach and we refer to the typical architecture underlying this approach, which is based on three components [15.6, 16]: the global schema, the sources, and the mapping between the two. The sources represent the repositories

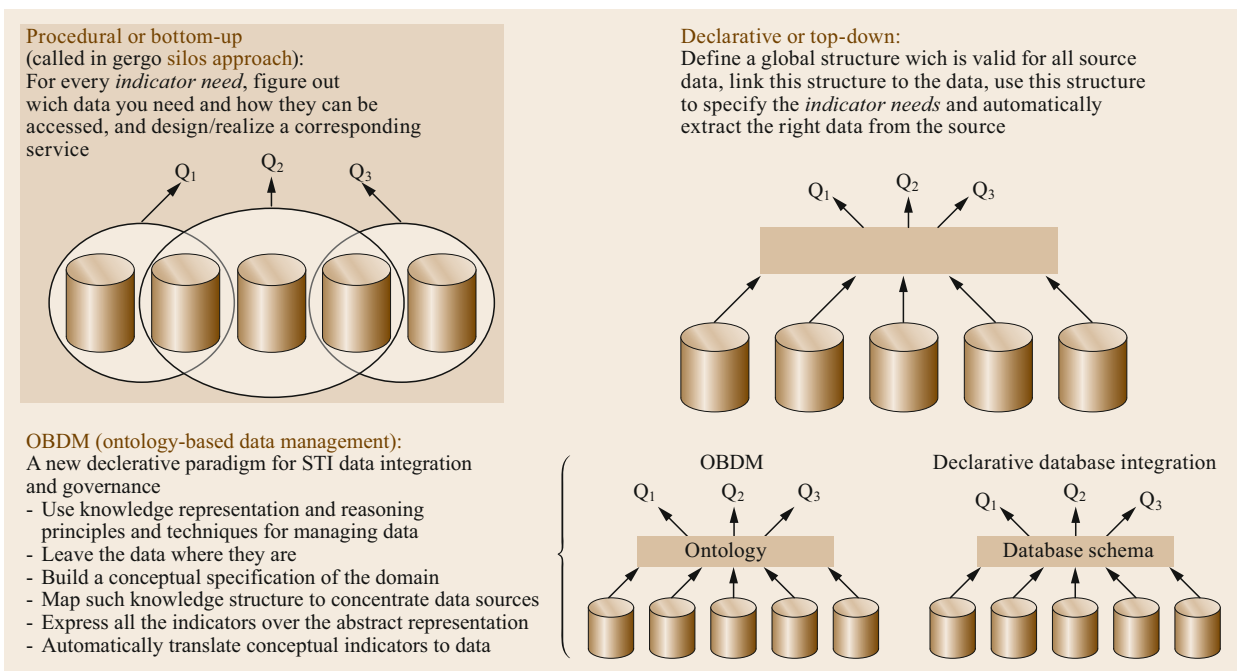



**Fig. 15.2** Approaches to data integration for STI (after [15.15])

where the data are; the global schema, also called mediated schema, represents the unified structure presented to the clients; the mapping relates the source data with the global schema. The most important service provided by the integration system is query answering, i. e., computing the answer to a query posed in terms of the global schema. Such computation involves accessing the sources, collecting the relevant data, and packaging such data in the final answer.

Formal, declarative approaches to data integration started in the 1990s [15.6, 16–18]. Since then, many aspects of the general problems have been the subject of detailed investigations both in academia and in industry.

Among them, in this chapter, we want to focus on the idea of using semantics for making data integration more powerful. As illustrated in [15.14], using semantics here means conceiving data integration systems where the semantics of data is explicitly specified and is taken into account to devise all the functionalities of the system. Over the past two decades, this idea has become increasingly crucial to a wide variety of information-processing applications and has received much attention in the artificial intelligence, database, web, and data mining communities [15.19].

As we said before, we concentrate on a specific paradigm for semantic data integration, OBDI. This new paradigm was introduced about a decade ago, as a new way for modeling and interacting with a data integration system [15.20–23]. According to such paradigm, the client of the information system is freed from being aware of how data and processes 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. An OBDI system is an information management system maintained and used by a given organization (or a community of users), whose architecture has the same structure as a typical data integration system, with the following components: an ontology, a set of data sources, and the mapping between the two. The ontology approach as outlined in the following paragraphs was originally published in [15.24].

The ontology is a conceptual, formal description of the domain of interest to the organization, expressed in terms of relevant concepts, attributes of concepts, relationships between concepts, and logical assertions formally describing the domain knowledge. The data sources are the repositories accessible by the organization where data concerning the domain are stored. In the general case, such repositories are numerous and heterogeneous, each one managed and maintained independently from the others. It may be even the case that some of the data sources are not under the control of the organization and can be accessed remotely, e. g., via the web. The mapping is a precise specification of the correspondence between the data contained in the data sources and the elements of the ontology, where by element we here mean concepts, attributes, and relationships. The main purpose of an OBDI system is to allow information consumers to query the data using the elements in the ontology as predicates. In the special case where the organization manages a single data source, the term *ontology-based data access* (OBDA) system is used instead of the OBDI system.

The notions of OBDA and OBDI were introduced in [15.20–22] and originated from several disciplines, in particular, information integration, knowledge representation and reasoning, and incomplete and deductive databases. OBDI can be seen as a sophisticated form of information integration, where the usual global schema is replaced by an ontology describing the domain of interest. The main difference between OBDI and traditional data integration is that in the OBDI approach, the integrated view that the system provides to information consumers is not merely a data structure accommodating the various data at the sources, but a semantically rich description of the relevant concepts in the domain of interest, as well as the relationships between such concepts. In general, such a description is formally defined in logic and enriches, generalizes, and relates the vocabularies of different data sources, thus providing a *common ground* for the domain knowledge. Also, the distinction between the ontology and the data sources reflects the separation between the conceptual level, the one presented to the client, and the logical/physical level of the information system, the one stored in the sources, with the mapping acting as the reconciling structure between the two levels. Notably, this separation is also instrumental for recovering the possibility of access to legacy systems, which are often excluded from the possibility of being used in interesting analyses carried out by the organization on its own data.

From all the above observations one can easily see that the central notion of OBDI is the ontology, and, therefore, reasoning over the ontology is at the basis of all the tasks that an OBDI system has to carry out. By reasoning, we mean the ability to derive all the implicit knowledge on the domain that is logically implied by the explicitly asserted facts in the ontology. In particular, the axioms of the ontology allow one to derive new facts from the source data, and these inferred facts greatly influence the set of answers that the system should compute during query processing. In the last

decades, research on ontology languages and ontology inferencing has been very active in the area of knowledge representation and reasoning. Description logics (DLs) [15.25] are widely recognized as appropriate logics for expressing ontologies and are at the basis of the W3C standard ontology language OWL. These logics permit the specification of a domain by providing the definition of classes and by structuring the knowledge about the classes using a rich set of logical operators. They are decidable fragments of mathematical logic, resulting from extensive investigations on the trade-off between expressive power of knowledge representation languages and computational complexity of reasoning tasks. Indeed, the constructs appearing in the DLs used in OBDI have been carefully chosen taking into account such a trade-off.

As we said before, the axioms in the ontology can be seen as semantic rules that are used to complete the knowledge given by the raw facts determined by the data in the sources. In this sense, the source data of an OBDI system can be seen as an incomplete database, and query answering can be seen as the process of computing the answers logically deriving from the combination of such incomplete knowledge and the ontology axioms. Therefore, at least conceptually, there is a connection between OBDI and the two areas of incomplete information [15.26] and deductive databases [15.27]. The new aspect of OBDI is related to the kind of incomplete knowledge represented in the ontology, which differs both from the formalisms typically used in databases under incomplete information (e. g., Codd tables) and from the rules expressible in deductive database languages (e. g., logic programming rules).

Once we are able to reason about the ontology, we can take advantage of the OBDI system in many interesting and relevant ways. Some of these are referred to below:

1. As already noticed, we can take into account all inferences over the ontology in processing the queries. In other words, the quality of the query answering service is potentially much higher than in the traditional setting, because all the knowledge about the domain represented by the ontology is exploited in computing the answers.
2. We can carry out the task of data source profiling again by relying on the knowledge about the domain. This allows the data designer to describe, maintain, and document the content of the various data sources in a much richer way than in traditional systems, because (s)he can now specify the characteristics of the sources in terms of the vocabulary and the metadata sanctioned by the ontology.
3. We can check and assess the quality of the data sources by comparing their content and their structure with the ontology, and, therefore, singling out inconsistencies, incompleteness, and inaccuracies of the data sources with respect to the domain knowledge.
4. We can set up new interesting services realized through the integration system. One notable example is open data publishing. The presence of the ontology makes it simple and effective to annotate the published data with the concepts and the relationships that are relevant in the domain of interest, so as to provide a conceptual description and a meaningful context of the published datasets.



## 15.3 Formal Framework for OBDI

The previous section introduced the notion of OBDI as a new paradigm for data integration. In this section, we provide the fundamental elements for a formalization of OBDI, illustrating the form of an OBDI specification, and presenting the semantics of an OBDI system. We closely follow the exposition in [15.14]. In the formal framework, we assume that the OBDI system can access the data sources through a single SQL (Structured Query Language) interface, which presents the various data as if they were in a unique database. In other words, we talk about a single data source, which is a database obtained as an abstraction for a variety of (possibly heterogeneous) data sources. This is not a real limitation, because in practice, such a database might be obtained through the use of an off-the-shelf data federation tool, which presents the sources through a schema of a single database as a wrapping of the source schemas expressed in terms of a unique format.

Given this assumption, an *OBDI specification* I is as a triple $\langle O, S, M \rangle$, where O is an ontology, S is a relational schema, called source schema, and M is a mapping from S to O. As already stated, O represents the general knowledge about the domain (i. e., relative to classes and relationships, rather than to specific objects that are instances of concepts), expressed in some logical language. Typically, O is a lightweight DL TBox [15.20], i. e., it is expressed in a language ensuring both semantic richness and efficiency of reasoning, and in particular of query answering. The mapping

SubClassOf(*Dean Professor*)
SubClassOf(*University Organization*)
ObjectPropertyDomain(*advisor Student*)
ObjectPropertyDomain(*headOf Dean*)
ObjectPropertyDomain(*takesCourse Student*)
FunctionalObjectProperty(*headOf* )
EquivalentClasses(*Person* ObjectUnionOf(*Student Professor*))
SubClassOf(*Student* ObjectSomeValuesFrom(*takesCourse Course*))
SubClassOf(*Professor* ObjectSomeValuesFrom(*worksFor University*))
SubClassOf(*Professor* ObjectSomeValuesFrom(*teacherOf Course*))
SubClassOf(*Dean* ObjectSomeValuesFrom(*headOf College*))
DisjointClasses(*Dean* ObjectSomeValuesFrom(*teacherOf Course*))

**Fig. 15.3** The ontology of the example

faculty(*UNIVERSITY_CODE, CODE, DESCRIPTION*)
students(*ID, FNAME, SNAME, DOB, ADDRESS*)
course(*FACULTY_CODE, CODE, DESCRIPTION*)
assignment(*COURSE_CODE, PROFESSOR, YEAR*)
professor(*CODE, FNAME, SNAME, ADDRESS, PHONE*)
exam(*STUD_ID, COURSE_CODE, DATE, RATING*)
career(*STUD_ID, ACADEMIC_YEAR, FACULTY_CODE*)
degree(*STUD_ID, YEAR, PROF_ID, TITLE*)

**Fig. 15.4** Relational tables of the source schema of the example

M is a set of mapping assertions, each one relating a query over the source schema to a query over the ontology.

*Example.* We now present an example of an OBDI system extracted from a real integration experiment involving data from different sources in use at Sapienza University of Rome. The ontology is defined by means of the OWL 2 assertions shown in Fig. 15.3.

It is, in fact, a portion of the Lehigh University Benchmark (LUBM) ontology, an ontology that is commonly used for testing ontology-based applications in the Semantic Web. In particular, the global schema contains the classes *Person*, *Student*, *Professor*, *Organization*, *College*, *Dean*, and *Course*, and the object properties *headOf*, *worksFor*, *takesCourse*, and *advisor*. For the sake of simplicity, we do not report in this example assertions involving data properties (i. e., attributes), but they are obviously allowed in our framework. The source schema is a set of relational tables resulting from the federation of several data sources of the School of Engineering of the Sapienza University of Rome, and the portion that we consider in this example is constituted by the relational tables shown in Fig. 15.4.

As for the mapping, referring to the global and source schemas presented above, we provide some sample mapping assertions in Fig. 15.5.

The mapping assertion $M_1$ specifies that the tuples from the source table *students* provide the information needed to build the instances of the class *Student*. In particular, the SQL query in the body of $M_1$ retrieves the code for students whose date of birth is before 1990; each such code is then used to build the object identifier for the student by means of the unary function symbol "st". Similarly, the mapping $M_2$ extracts data from the table *degree*, containing information on the student's Master's degree, such as the year of the degree, the title of the thesis, and the code of the advisor. The tuples retrieved by the query in the body of $M_2$, involving only degree titles earned after 2000, are used to build instances for the object property (relationship) advisor; the instances are constructed by means of the function symbols "pr" and "st". Finally, the mapping assertion $M_3$ contributes to the construction of the domain of the advisor, taking from the source table exam only codes of students who have passed the exam of courses that were not assigned to any professor before 1990.

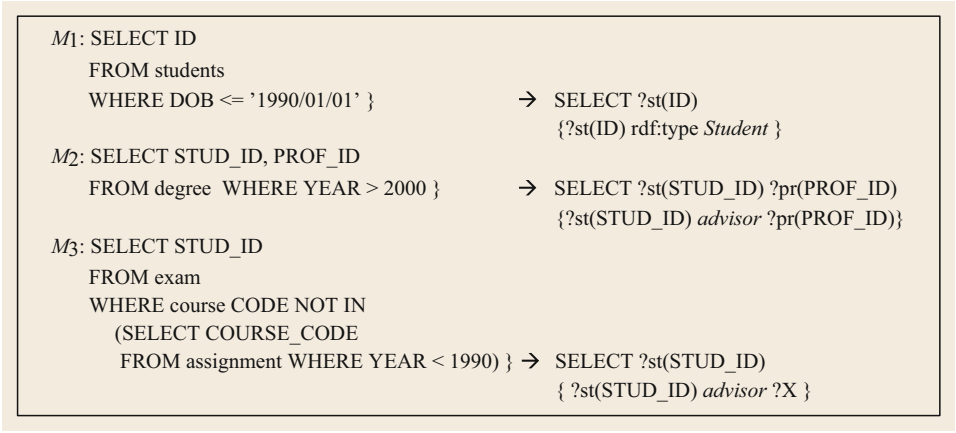



**Fig. 15.5** Mapping assertions of the example

An *OBDI system* is a pair (I, D), where I is an OBDI specification and D is a database for the source schema S, called source database for I. The semantics of (I, D) is given in terms of the logical interpretations that are models of O (i. e., satisfy all axioms of O, and satisfy M with respect to D). The notion of mapping satisfaction depends on the semantic interpretation adopted on mapping assertions. Commonly, such assertions are assumed to be sound, which intuitively means that the results returned by the source queries occurring in the mapping are a subset of the data that instantiate the ontology. The set of models of I with respect to D is denoted with ModD(I).

As we said before, in OBDI systems, the main service of interest is query answering, i. e., computing the answers to user queries, which are queries posed over the ontology. Such service amounts to return the so-called certain answers, i. e., the tuples that satisfy the user query in all the interpretations in ModD(I). Query answering in OBDI is thus a form of reasoning under incomplete information and is much more challenging than classical query evaluation over a database instance.

It is well known that carrying out inference tasks, such as answering queries in OBDI systems, may be computationally expensive. From the computational perspective, query answering depends on:

1. The language used for the ontology
2. The language used for user queries, and
3. The language used to specify the queries in the mapping.

In the following, we consider a particular instantiation of the OBDI framework, in which we choose each such language in such a way that query answering is guaranteed to be tractable with respect to the size of the data.

From the general framework we obtain a computationally tractable one by choosing appropriate languages as follows:

- The ontology language is *DL-Lite*$_A$ or its subset *DL-Lite*$_R$ [15.22].
- The mapping language follows the *global-as-view* (GAV) approach [15.6].
- The user queries are unions of conjunctive queries [15.16].

In the following, we discuss each of the above choices.

### 15.3.1 Ontology Language

*DL-Lite*$_A$ [15.22] is essentially the maximally expressive member of the *DL-Lite* family of lightweight DLs [15.20]. In particular, its subset *DL-Lite*$_R$ has been adopted as the basis of the OWL 2 QL profile of the W3C standard OWL. As usual in DLs, *DL-Lite*$_A$ allows for representing the domain of interest in terms of *concepts*, denoting sets of objects, and *roles*, denoting binary relations between objects. In fact, *DL-Lite*$_A$ also considers *attributes*, which denote binary relations between objects and values (such as strings or integers), but for simplicity, we do not consider them in this chapter. From the expressiveness point of view, *DL-Lite*$_A$ is able to capture essentially all the features of entity-relationship diagrams and UML (Unified Modeling Language) class diagrams, except for completeness of hierarchies. In particular, it allows for specifying ISA ("is a") and disjointness between either concepts or roles, mandatory participations of concepts into roles, and the typing of roles. Formally, a *DL-Lite*$_A$ TBox is a set of assertions obeying the syntax

| | |
|---|---|
| $B_1 \sqsubseteq B_2$ | (positive concept inclusion) |
| $B_1 \sqsubseteq \neg B_2$ | (negative concept inclusion) |
| $R_1 \sqsubseteq R_2$ | (positive role inclusions) |
| $R_1 \sqsubseteq \neg R_2$ | (negative role inclusions) |
| (funct R) | (role functionalities) |

where:

- $B_1$ and $B_2$ are basic concepts, i. e., expressions of the form A, $\exists P$, or $\exists P^-$.
- R, $R_1$, and $R_2$ are a basic roles, i. e., expressions of the form P, or $P^-$.
- A and P denote an *atomic concept* and an *atomic role*, respectively, i. e., a unary and binary predicate from the ontology alphabet, respectively.
- $P^-$ is the *inverse* of an atomic role P, i. e., the role obtained by switching the first and second components of P.
- $\exists P$ (or $\exists P^-$ ), called existential unqualified restriction, denotes the projection of the role P on its first (or second) component.
- $\neg B_2$ (or $\neg R_2$) denotes the negation of a basic concept (or role).

Assertions of the form (funct R) are called role functionalities and specify that an atomic role, or its inverse, is functional. *DL-Lite*$_A$ poses some limitations on the way in which positive role inclusions and role functionalities interact. More precisely, in a *DL-Lite*$_A$ TBox an atomic role that is either functional or inverse functional cannot be specialized, i. e., if (funct P) or (funct $P^-$) are in the TBox, no inclusion of the form $R \sqsubseteq P$ or $R \sqsubseteq P^-$ can occur in the TBox. *DL-Lite*$_R$ is the subset of *DL-Lite*$_A$ obtained by removing role functionalities altogether.

A *DL-Lite*$_A$ *interpretation* $J = (\Delta^J, \cdot^J)$ consists of a non-empty set constituting the *interpretation domain* $\Delta^J$ and an *interpretation function* $\cdot^J$ that assigns to each atomic concept A a subset $A^J$ of $\Delta^J$, and to each atomic role P a binary relation $P^J$ over $\Delta^J$. In particular, for the constructs of *DL-Lite*$_A$, we have (symbol $\setminus$ is used to denote set difference):

- $A^J \subseteq \Delta^J$
- $P^J \subseteq \Delta^J \times \Delta^J$
- $(P^-)^J = \{(o_2, o_1) | (o_1, o_2) \in P^J\}$
- $(\exists R)^J = \{o | \exists o'.(o, o') \in R^J\}$
- $(\neg B)^J = \Delta^J \setminus B^J$
- $(\neg R)^J = (\Delta^J \times \Delta^J) \setminus R^J$

Let C be either a basic concept B or its negation $\neg B$. An interpretation J satisfies a concept inclusion $B \sqsubseteq C$ if $B^J \subseteq C^J$, and similarly for role inclusions. Also, J satisfies a role functionality (funct R) if the binary relation $R^J$ is a function, i. e., $(o, o_1) \in R^J$ and $(o, o_2) \in R^J$ implies $o_1 = o_2$.

### 15.3.2 Mapping Language

The mapping language in the tractable framework allows mapping assertions of the following the forms:

- $\varphi(x) \rightarrow A(f(x))$
- $\varphi(x) \rightarrow P(f_1(x_1), f_2(x_2))$

where $\varphi(x)$ is a domain-independent first-order query (i. e., an SQL query) over S, with free variables x, A and P are as before variables in $x_1$ and $x_2$ also occur in x, and f, possibly with subscripts, is a function.

Intuitively, the mapping assertion of the first form, called the concept mapping assertion, specifies that individuals that are instances of the atomic concept A are constructed through the use of the function f from the tuples retrieved by the query $\varphi(x)$. Similarly for the mapping assertion of the second form called the role mapping assertion. Each assertion is of type GAV [15.6], i. e., it associates a view over the source (represented by $\varphi(x)$) to an element of ontology. However, differently from traditional GAV mappings, the use of functions is crucial here, since we are considering the typical scenario in which data sources do not store the identifiers of the individuals that instantiate the ontology, but only maintain values. Thus, functions are used to address the semantic mismatch existing between the extensional level of S and O [15.22].

Formally, we say that an interpretation J satisfies a mapping assertion $\varphi(x) \rightarrow A(f(x))$ with respect to a source database D, if for each tuple of constants t in the evaluation of $\varphi(x)$ on D, $(f(t))^J \in A^J$, where $(f(t))^J \in \Delta^J$ is the interpretation of f(t) in J that is, f(t) acts simply as a constant denoting an object. Satisfaction of assertions of the form $\varphi P(f_1(x_1), f_2(x_2))$ is defined analogously. We also point out that *DL-Lite*$_A$ adopts the unique name assumption (UNA), that is, different constants denote different objects, and thus different ground terms of the form f(t) are interpreted with different elements in $\Delta^J$.

### 15.3.3 User Queries

In our tractable framework for OBDI, user queries are conjunctive queries (CQs) [15.16] or unions thereof. With q(x), we denote a CQ with free variables x. A Boolean CQ is a CQ without free variables. Given an OBDI system (I,D) and a Boolean CQ q over I, i. e., over the TBox of I, we say that q is *entailed by* (I, D), denoted with (I, D) |= q, if q evaluates to true in every

interpretation $J \in Mod_D(I)$. When the user query q(x) is non-Boolean, we denote with $cert_D(q(x), I)$ the *certain answers* to q with respect to (I, D), i. e., the set of tuples t such that $(I, D)| = q(t)$, where q(t) is the Boolean CQ obtained from q(x) by substituting x with t.

### 15.3.4 Query Answering

Although query answering in the general framework may soon become intractable or even undecidable, depending on the expressive power of the various languages involved, the tractable framework has been designed to ensure tractability of query answering. We end this section by illustrating the basic idea to achieve tractability.

In the tractable OBDI framework previously described, one can think of a simple chase procedure [15.28] for query answering, which first retrieves an initial set of concept and role instances from the data source through the GAV mapping, and then, using the ontology axioms, *expands* such a set of instances deriving and materializing all the logically entailed concept and role assertions; finally, queries can be evaluated on such an expanded set of instances. Unfortunately, in *DL-Lite*$_A$ (and in *DL-Lite*$_R$ already) the instance materialization step of the above technique is not feasible in general, because the set of entailed instance assertions starting from even very simple OBDA specifications and small data sources may be infinite.

As an alternative to the above materialization strategy, most of the approaches to query answering in OBDI are based on query rewriting, where the aim is to first compute a query $q'$ representing a reformulation of a query q with respect to an OBDI specification I, and then evaluate $q'$ over the source database. Actually, the above described OBDI framework allows for modularizing query rewriting. Indeed, the current techniques for OBDI consist of two phases, namely the phase of *query rewriting with respect to the ontology*, and the phase of *query rewriting with respect to the mapping*:

1. In the first phase, the initial query q is rewritten with respect to the ontology, producing a new query $q_1$, still over the ontology signature; intuitively, $q_1$ *encodes* the knowledge expressed by the ontology that is relevant for answering the query q.
2. In the second phase, the query $q_1$ is rewritten with respect to the mapping M, thus obtaining a query $q_2$ to be evaluated over the source data. Thus, the mapping assertions are used for reformulating the query into a new one expressed over the source schema signature.

Thus, following the above method, computing the certain answers of a query Q over I is reduced to a simple evaluation of a suitable query over the source database, thus relying on the technology of relational databases, including the possibility of adopting well-established query optimization strategies.

## 15.4 *Sapientia* and OBDI for Multidimensional Research Assessment



*Sapientia* is the ontology of multidimensional research assessment developed by an interdisciplinary group of scholars funded by Sapienza University of Rome in the framework of its Research Awards (two main research projects). It models all the activities relevant for the evaluation of research and for assessing its impacts. For impact, in a broad sense, we mean any effect, change or benefit, to the economy, society, culture, public policy or services, health, the environment, or quality of life, beyond academia.

The first version of *Sapientia* was closed on the 22 December 2014. It consisted of 14 modules of around 350 concepts, roles and attributes (Fig. 15.6). Figure 15.7 illustrates the organization of the main components of an OBDI system including *Sapientia*.

### 15.4.1 *Sapientia*'s Philosophy and its Main Principles

*Sapientia*'s mission is being comprehensive and try to model everything related to the evaluation of research and its impacts. As we stated in Sect. 15.1, the evaluation of research, is a complex activity. It requires a systematic view and has its base on the interplay between theory, methodology, and data. To accomplish its mission, we designed *Sapientia* to be at the heart of a flexible knowledge infrastructure ("robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds" [15.29]) for the multidimensional assessment of research and its impacts. By flexible we mean a knowledge infrastructure characterized by being an open and evolving infrastructure.

The principles followed in the design and development of *Sapientia* are the following:

- We started with a top-down modeling approach, with subsequent bottom-up refinements and cyclical improvements. We describe and model the domain from a conceptual point of view, without considering the existing data and its specificity.

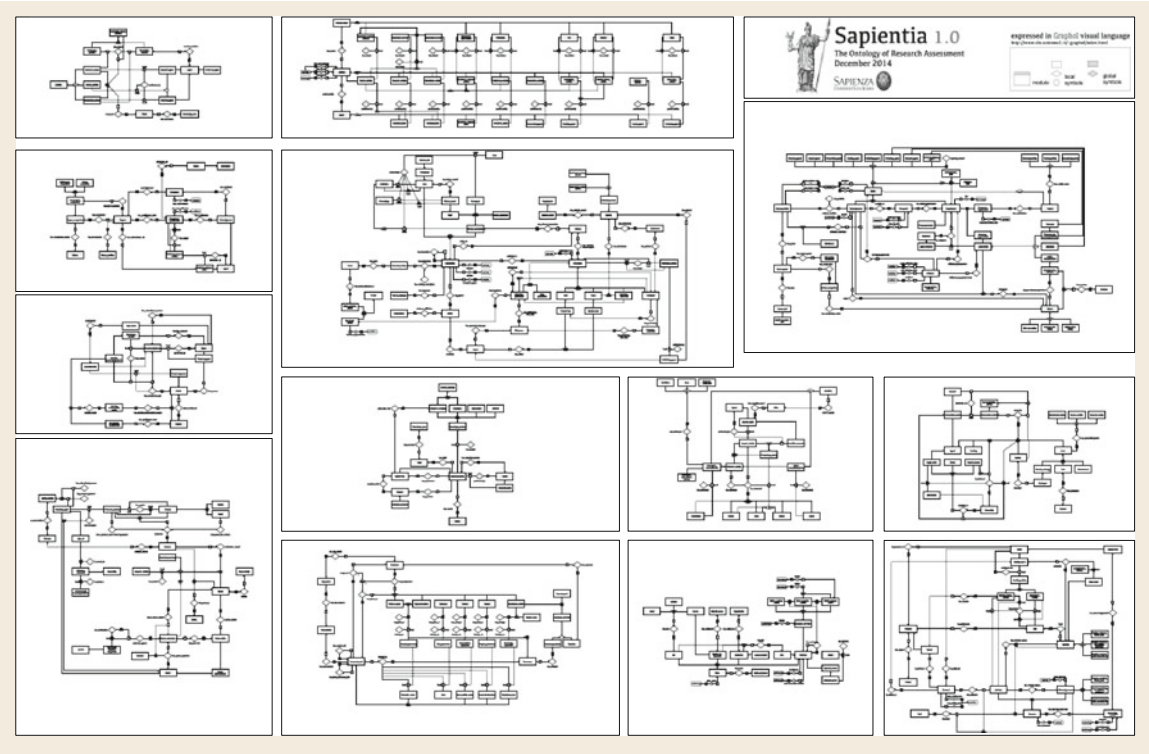



**Fig. 15.6** An illustration of *Sapientia* (the ontology of multidimensional research assessment) 1.0

- We left outside the scope of the ontological commitment all the methodological consideration about choice of the methods for the assessment of research. This is because we want that our ontology being the common ground for experimenting and testing different methods and approaches.
- We left outside the scope of the ontological commitment the implementation problem and the consequences of evaluation. Again, this is for keeping our ontology as a common ground, a shared language or vocabulary, to build a cooperative and open discussion about evaluation approaches considering the interaction of different stakeholders with different points of view and interests.
- We pursued a modeling approach based on processes, which are conceived as collections of activities. A process is composed by inputs and outputs.
- Individuals and activities are the main pillars of the ontology.
- We followed a modeling approach based on a modularization of the system. Our ontology is organized in modules. As we shall see later, we have two kind of modules: functional modules and structural modules. By functional modules we mean modules that model the main agents and activities of our domain (namely Agents, Activities, R&D, Publishing, Education, Resources and Review). By structural modules, we mean those modules that represent the constituent elements of the ontology to ensure its long lasting and general-purpose functionality (namely, Taxonomies, Space, Representations, and Time).

We consider the building of descriptive, interpretative, and policy models of our domain as a distinct step with respect to the building of the domain ontology. In the following part of this section, we will reproduce the findings that we previously published in *Daraio*, *Lenzerini*, et al. [15.24]. However, the ontology will intermediate the use of data in the modeling step and should be rich enough to allow the analyst the freedom to define any model she considers useful to pursue her analytic goal.

Obviously, the actual availability of relevant data will constrain both the mapping of data sources on the ontology and the actual computation of model variables and indicators of the conceptual model. However, the analyst should not refrain from proposing the models that she considers the best suited for her purposes and to express, using the ontology, the quality requirements, the logical, and the functional specification for her ideal model variables and indicators. This approach has many merits, in particular:

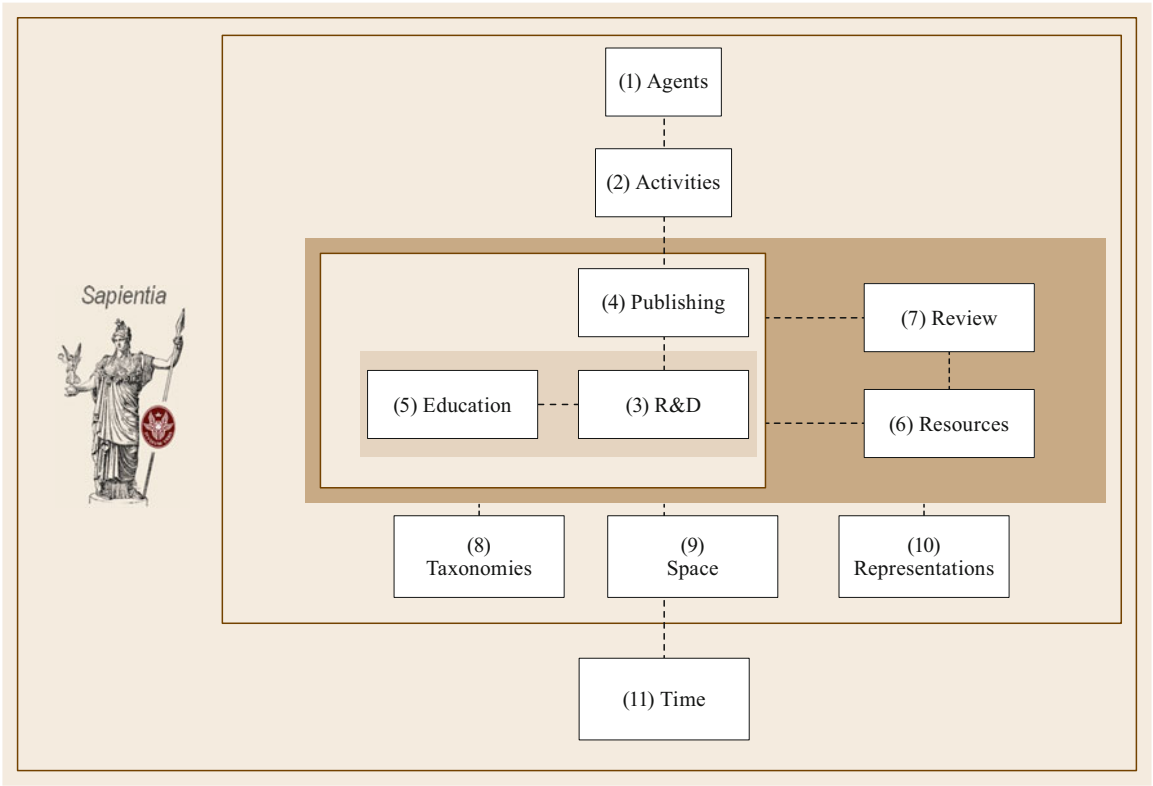



**Fig. 15.7** Modules of Sapientia 3.0

- It allows the use of a common and stable ontology as a platform for different models.
- It addresses the efforts to enrich data sources, and verify their quality.
- It makes transparent and traceable the process of approximation of variables and models when the available data are less than ideal; it makes use of every source at the best level of aggregation, usually the atomic one.

More generally, this approach is consistent with the effort of avoiding "the harm caused by the blind symbolism that generally characterizes a hasty mathematization" put forward by *Georgescu-Roegen* in his seminal work on production models and on methods in economic science [15.30–32]. In fact, one can verify the logical consistency of the ontology and compute answers to unambiguous logical queries.

Moreover, the proposed ontology allows us to follow the Georgescu-Roegen approach also in the use of the concept of process. We can analyze the knowledge production activities at an atomic level, considering their time dimension and such funds as the cumulated results of previous research activities, both those available in relevant publications and those embodied in the authors' competences and potential, the infrastructure assets, and the time devoted by the group of authors to current research projects. Similarly, we can analyze the output of teaching activities, considering the joint effect of funds such as the competence of teachers, the skills and initial education of students, and educational infrastructures and resources. Thirdly, service activities of research and teaching institutions provide infrastructural and knowledge assets that act as a fund in the assessment of the impact of those institutions on the innovation of the economic system. The perimeter of our domain should allow us to consider the different channels of transmission of that impact: mobility of researchers, career of alumni, applied research contracts, joint use of infrastructures, and so on. In this context, different theories and models of the system of knowledge production could be developed and tested.

### 15.4.2 Requested Investment and Modularity of the System

In an OBDI 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 (which we will see below) that allow an easy modularization and updates of the relevant domain. The following paragraph is taken from *Daraio*, *Lenzerini*, et al. [15.33].

Following a real options approach in investment theory [15.34], we can 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 [15.35] 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. In modular design, the main criteria to assess the decomposition of systems into modules are those of cohesion and coupling. The principal rule to assess the quality of the modularization of a system, attributed to *Parnas* [15.36] even if the paper does not contain the terms cohesion and coupling, is of *high cohesion within modules and loose coupling between modules*.

*Cohesion* refers to the degree to which the elements of a module belong together and, hence, it is a measure of how strongly related each piece of a module is. Modules with high cohesion tend to be preferable, because high cohesion is associated with several desirable properties including robustness, reliability, reusability, and understandability.

*Coupling* is the manner and degree of interdependence between modules; a measure of how closely connected two modules are. Low coupling is often a sign of a well-structured system and a good design, and when combined with high cohesion, supports the general goals of high readability and maintainability. Modularity is a property of quasi-decomposition of hierarchical systems, based on the minimization of the interdependence of subsystems [15.37].

The modification of subsystems 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.

The current version of *Sapientia*, *Sapientia* 3.0, includes around 600 concepts, roles, and attributes and is organized in 11 modules (Fig. 15.7). Another module on skills is going to be included. Its aim is to model all the competences involved in the assessment of research and its impacts.

The *central* modules of *Sapientia* are Agents (no.1) and their Activities (no.2), which are expanded into the five *main* process-based modules: 3 R&D, 4 Publishing (ancillary module of R&D), 5 Education, 6 Resources, 7 Review. These are connected to the four *auxiliary* modules (also defined as *structural* modules): 8 Taxonomies, 9 Space, 10 Time and 11 Representations, which support the modeling of the main modules from 1 to 7.

The module Agents (1) models the subjects involved in the world of research, carrying out the activities described in module 2 (Activities). Activities models, overall, the relevant actions carried out by agents and their products. Module 3 (R&D) models Research and Development (R&D) activities, those that allow the scientific community to advance the state of the art of knowledge. The module Publishing (4) models the publishing activities, those that allow people to communicate (and disseminate) the results of the R&D activities carried out. Education (5) models the educational activities, those that allow people to improve their knowledge and to acknowledge the improvements made by other people. Resources (6) models the resources and the activities carried out for their management. Review (7) models the reviewing activities for assessing the R&D activities and their results. Taxonomies (8) models the nomenclatures that classify the several elements of the domain. Space (9) models the regions of space where agents and activities are located and their roles. Representations (10) models the representations of the objects of the domain according to a Source. A data source is a possible source, but also other sources, such as a theory may be a source. Time (11) models the depth of time of the domain and cut across all the other modules.

## 15.5 Reasoning over *Sapientia*: Some Illustrative Examples

*Sapientia*, the ontology introduced in the previous section, is an important element of an OBDI system introduced in the previous sections (Fig. 15.8).

An OBDI system allows us to reason on each component of the system, i. e., on the ontology, on the data sources, and on the mappings. In the following, we show some examples.

### 15.5.1 Reasoning over the Ontology

In this section, we illustrate an example of reasoning over the ontology *Sapientia* analyzing some extracts from the Module 3 (R&D). The R&D module aims at modeling the research activities of researchers and their products. The central concept of the module is the R&D activity that is linked to two important concepts, i. e., research product and research outcome.

In *Sapientia*, any research activity has:

- Its direct output (has_output), available without the contributions of any other activity.
- Its outcome (has_outcome): an output of any activity (not necessarily a research activity) participating a value chain where the research activity has an enabling role (i. e., without the research activity, that specific output would not be generated).

In *Sapientia*, a Publication aims at reporting empirical or theoretical work and describes the results obtained in some knowledge field. Publications are described in the Module Publishing (4), which concerns the activity that allows people knowing the results of research. The output of a publishing activity is a publication, which is a way to represent a content through some media. There are four kinds of contents in *Sapientia*: paper-like content (a content structured as for being published paper); book-like content (a content structured as for being published as monographs or edited chapters), patent-like content (a content structured as for being published as patent applications) and Project-like content (a content structured as being suitable to apply for a call).

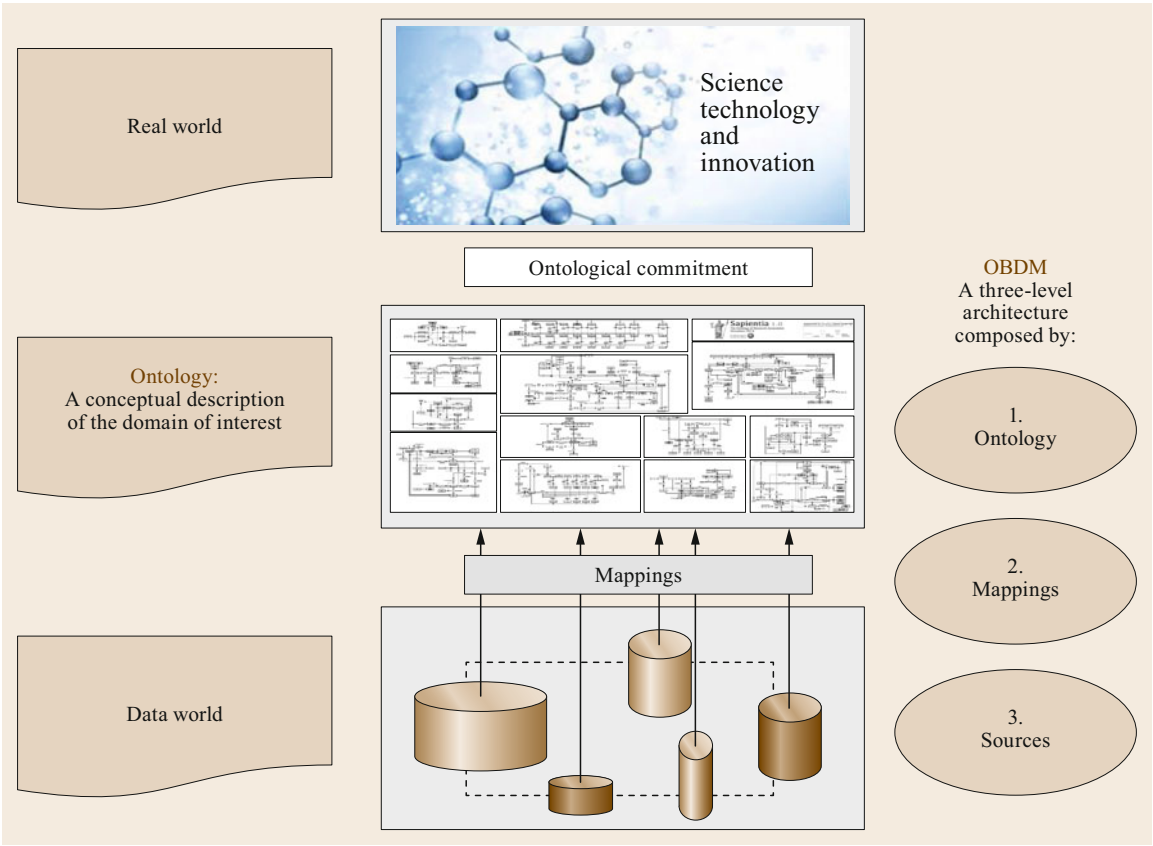



**Fig. 15.8** Key components of an OBDI information system (after [15.15])

We point out that publication of research work is not considered an output of that research, because the output of the research work is the content of the publication. Publications, in fact, are the main concept of another module, the Module 4 (Publishing).

There are three kinds of agents involved in a publication:

- *Author*: An author of a publication is an agent that has contributed in writing the content of the publication (for instance, reporting the results of some research (s)he has carried out).
- *Editor*: An editor of a complex publication (where contributions of different authors need to be verified, harmonized, and combined) is an agent that oversees and coordinates the publication.
- *Publisher*: A publisher of a publication is the agent that provides some media to deliver and display a publication.

In *Sapientia* there are three kinds of publications:

- *Atomic publications*: A publication resulting from a unique, indivisible act of writing by one or more authors.
- *Collections*: A publication disseminating a group of atomic publications in a unique impulse, during a limited and short period of time.
- *Series*: Each disseminating a group of atomic publications during a long and (perhaps) unlimited period of time.

In *Sapientia*, a patent application is a possible publication, and is the output of an applied research. Notice that a patent is different from the other types of atomic publication; it is a right granted by a state which may concern a research output, not an output itself. A patent application follows its own path within the three levels of publications:

- It is an atomic publication itself.
- It is published in an issue (a collection).
- That issue appears in an intellectual property law journal (a series).

Note that there are no constraints between contents and publications where they can be published (for example, a patent_like_content can be placed in a part of a paper).

In Figs. 15.9 and 15.10, we show an example extracted from the Module R&D of *Sapientia* 3.0. Figures 15.9 and 15.10 display the path from Researcher to Publication in *Sapientia*. Figure 15.11 reproduces a legend to interpret the symbols used in the previous Figs. 15.9 and 15.10. Table 15.1 describes the main concepts and relations showed in Figs. 15.9 and 15.10. The language Graphol (http://www.dis.uniroma1.it/~graphol/) developed at the Sapienza university and implemented in the software Eddy [15.38, 39] is used in Figs. 15.9 and 15.10.

Graphol permits the expression of the axioms of an ontology as described in Sections 2 and 3 but in a more readable manner, in the form of a graph. This graphical representation of an ontology is very practical and useful to those who do not have a thorough knowledge of mathematical logic. Using the editor Eddy, 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 and the various reports. 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 representation.

The examination of Figs. 15.9 and 15.10 allows us to reason over the ontology about the path from researcher to publication. It clearly appears that the path from researcher to R&D activities to publication goes through *content*.

### 15.5.2 Reasoning over the Mappings

The mappings in an OBDI system play a crucial role. They bind the data sources to the ontology. The connection is made through the *materialization or staging* phase (Fig. 15.12). Different levels of materialization are possible, and the refreshing of the materialization can be different according to the data source. The realization of the mappings in the context of *Sapientia* and its OBDI system is an interesting peculiar case, because the problem of *disambiguation* is a very important problem in bibliometrics and affects the assessment of the research and its impacts.

The solution proposed so far for the implementation of the mappings of *Sapientia* to its data sources is based on the *balance* between the level of computation and the level of materialization. Therefore, *quality checks* on data sources can be carried out.

The connection with the entity resolution (ER) occurs at the *computation* level through *modularization*. This allows a high degree of flexibility; ER algorithms can be replaced and/or updated without modifying the overall system. This is a *computational* flexibility

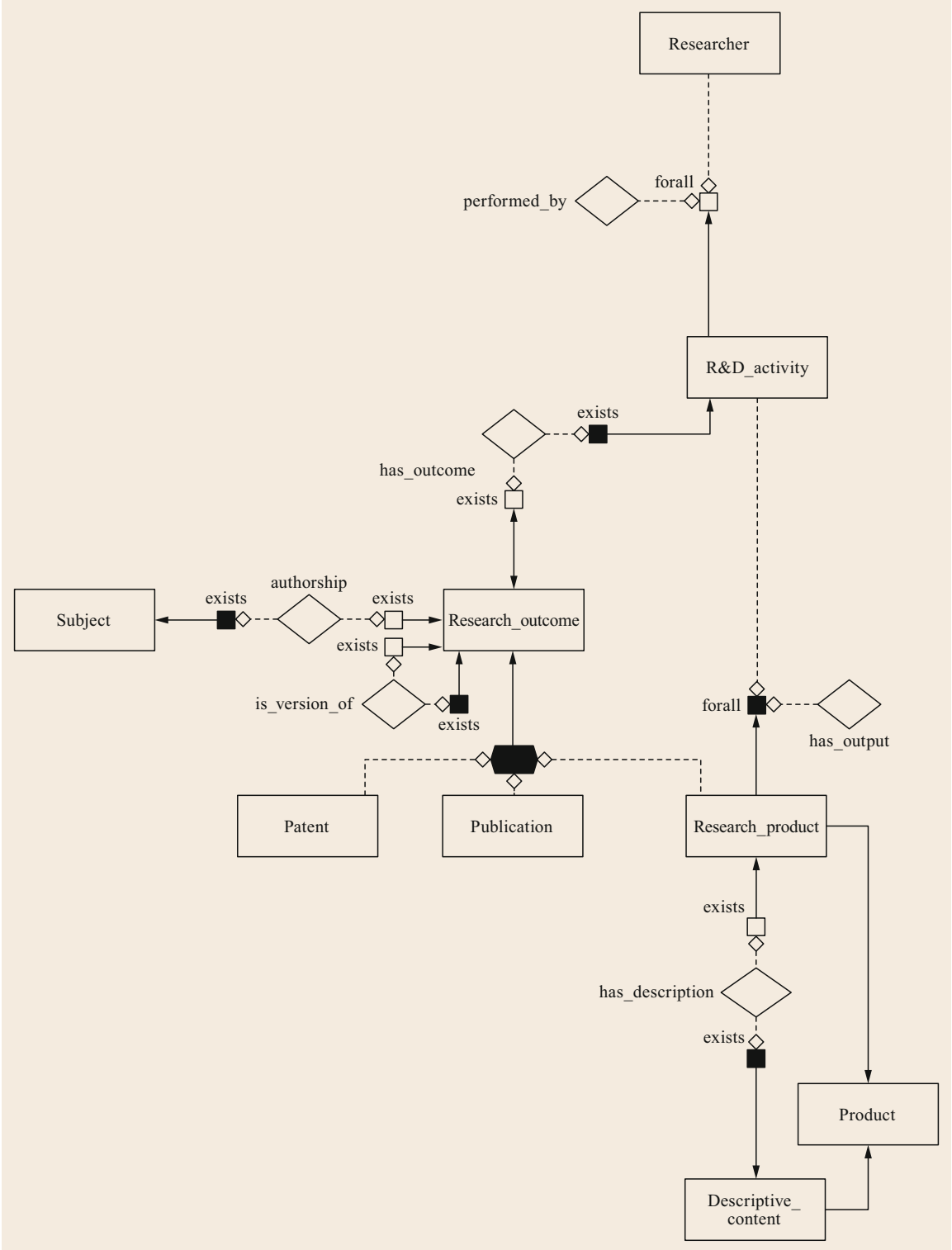



**Fig. 15.9** An illustration of *Sapientia* from Module 3 R&D. Part I: from researcher to descriptive content

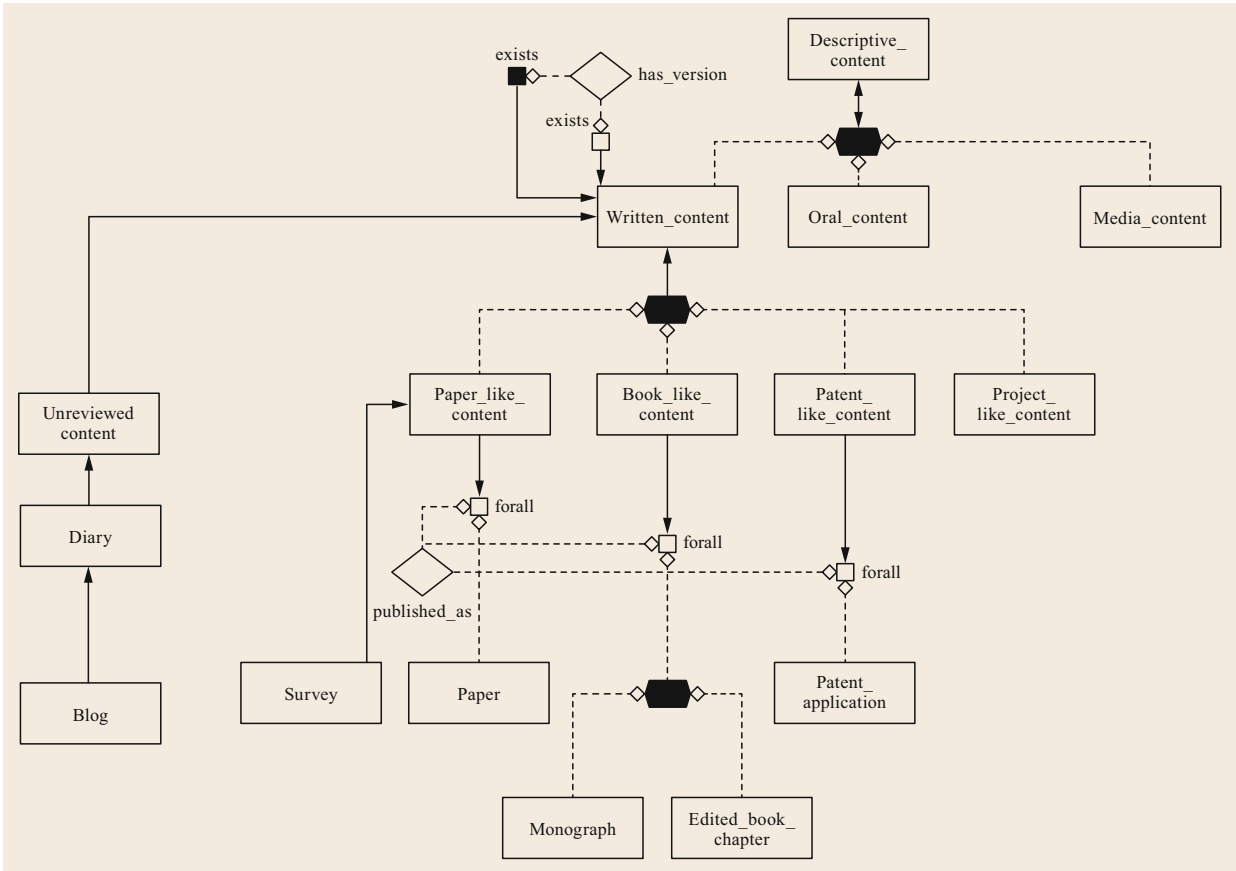



**Fig. 15.10** An illustration of *Sapientia* from Module 3 R&D. Part II: from descriptive content to publications

with ER techniques through *modularization*. In addition, the ontological model offers the advantage of analyzing authors and affiliations together by means of a *conceptual* flexibility through the Module 11 Representations.

So far, we have implemented blocking mechanisms, mainly based on the Inverted Index. As with pairs of comparisons, blocking mechanisms are based on a common interface and specific implementations for single data sources. They act in a *spiral model* that will allow us to refine ER algorithms in subsequent iterations. In this respect, the ontology can be supportive to identify the strengths and weaknesses of the different ER algorithms that will be tested and implemented to our domain.

*De Giacomo* et al. [15.14] in the current challenges for OBDI mention "the problem of devising methodologies and tools for developing mappings for OBDI" and state that it "is largely unexplored."

### 15.5.3 Reasoning over the Data and Indicators

Other possible reasoning tasks are those on data sources and on the indicators that may be built over the data. *Daraio*, *Lenzerini*, et al. [15.33] showed the usefulness of an OBDI system for R&I integration from a data quality perspective. *Daraio*, *Lenzerini,* et al. [15.42] showed that an OBDI approach allows for an unambiguous specification of indicators according to its four main dimensions: ontological, logical, functional, and qualitative (Table 15.2).

Reasoning over *Sapientia* permits the characterization of each indicator along its four dimensions listed above. This is extremely useful for studying the specification of indicators in the context of an *ontological approach* to the evaluation of research. The specification of the indicators within an OBDI framework aims to protect the analyst from the *risk of reductive conclu-*

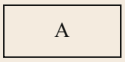
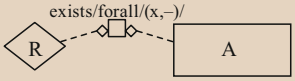
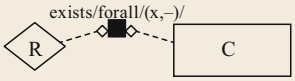
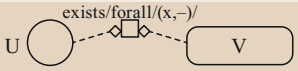
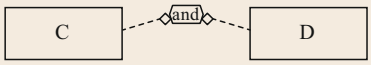
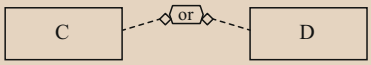
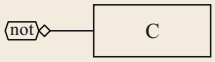
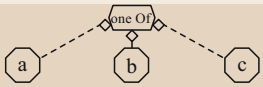
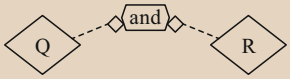
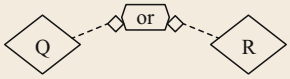
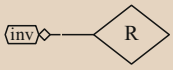
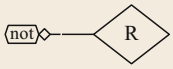
| Expression | DL-syntax | Graphol syntax |
|---|---|---|
| Atomic concept | A |   |
| Domain restriction on role | ∃R.C ∀R.C<br>≥ xR.C ≤ yR.C |   |
| Range restriction on role | ∃R⁻.C ∀R⁻.C<br>≥ xR⁻.C ≤ yR⁻.C |   |
| Domain restriction on attribute | ∃U.V ∀U.V<br>≥ xU.V ≤ yU.V |   |
| Concept intersection | C ⊓ D |   |
| Concept union | C ⊔ D |   |
| Concept complement | ¬C |   |
| Concept one-of | {a, b, c} |   |
| Atomic role | P |   |
| Role intersection | Q ⊓ R |   |
| Role union | Q ⊔ R |   |
| Role inverse | R⁻ |   |
| Role complement | ¬R |   |
| Role chain | Q ∘ R |   |

**Fig. 15.11** Legend of the symbols illustrated in Figs. 15.9 and 15.10 (after [15.38, 39])

*sions*, which would focus on the logical and functional aspects of the specification, ignoring the ontological and quality assessment parts of the process.

In a broader perspective, *Daraio* and *Bonaccorsi* [15.43] identify two trends in indicator development:

- Trend towards *granularity* of indicators ("new indicators are explicitly requested to allow various kinds of aggregation and disaggregation, preserving desirable statistical properties, in order to address new policy needs"), detailing granularity in territorial, institutional and disciplinary areas.
- Trend towards *cross-referencing* ("the ability of indicators to be combined in meaningful ways, preserving their statistical properties").

*Sapientia* and its OBDI system may be a suitable infrastructure to develop indicators that satisfy the policy requirements of granularity and cross-referencing in a coherent and consistent way.

**Table 15.1** Main concepts and roles illustrated in Figs. 15.9 and 15.10

| Term | Type | Definition |
|---|---|---|
| Researcher | Concept | A researcher is an agent that carries out research activities allowing the scientific community to advance the state of the art of knowledge |
| R&D activity | Concept | A research and experimental development (R&D) activity, according to the Frascati manual [15.40, p. 28] "comprise creative and systematic work undertaken in order to increase the stock of knowledge—including knowledge of humankind, culture and society—and to devise new applications of available knowledge." |
| Research_product | Concept | A research product is a product that is an output of a research activity. In *Sapientia*, we included all the research products of the *Research Excellence Framework* [15.41]. |
| Research_outcome | Concept | A research outcome of a research activity R is the output of an activity (not necessarily a research activity) participating a value chain where R has an enabling role (without R that output would not be generated). |
| Subject | Concept | A subject is any entity that can act as an agent and, playing such role, performs some activities. There are two types of subjects: natural persons and organizations. |
| Patent | Concept | A patent is a right that may be owned in an ownership, and shall confer to its owner the following exclusive rights: where the subject matter of a patent is a product, to prevent third parties not having the owner's consent from the acts of: making, using, offering for sale, selling, or importing for these purposes that product where the subject matter of a patent is a process, to prevent third parties not having the owner's consent from the act of using the process, and from the acts of using, offering for sale, selling, or importing for these purposes at least the product obtained directly by that process. Patent owners shall also have the right to assign or transfer by succession the patent and to conclude licensing contracts (Article 28 of the Trade-Related Intellectual Property Rights (TRIPS) Agreement administered by the World Trade Organization (WTO) that sets down minimum standards for many forms of intellectual property). Patents grant their owner a set of rights of exclusivity over an invention (a product or process that is new, involves an inventive step and is susceptible of industrial application). The legal protection conferred by a patent gives its owner the right to exclude others from making, using, selling, offering for sale or importing the patented invention for the term of the patent, which is usually 20 years from the filing date, and in the country or countries concerned by the protection. This set of rights provides the owner with a competitive advantage. Patents can also be licensed or used to help create or finance a spin-off company. It is therefore possible to derive value from them even if their owner does not have its own manufacturing capability (e. g., universities). |
| Publication | Concept | A publication is a particular kind of product consisting of an atomic or complex media including some content. There are three kinds of publications: atomic, collections and series. |
| Product | Concept | A product is an output of an activity, an entity (that might satisfy a want or need expressed by someone—or something—different from the agent who carried out the activity) which appears in the domain as consequence of an activity. Notice that the activity does not need to be finished at the time one of its products appears. |
| Descriptive content | Concept | A descriptive content is an interpretable object from which a human or an artificial intelligence can capture a meaning. It can use linguistic expressions and or media content. |
| Written content | Concept | A written content is a set of resources suited to be included in a single publication. These can be texts, technical drawings, diagrams, photographs and so on. These resources come together with their organizations (chapters, paragraphs, index). The linguistic parts of the resources are written in one or more natural languages (has_language). |
| Oral content | Concept | A descriptive content is oral if is the content of a speech. |
| Media content | Concept | A descriptive content is a media content whether it represents the way one or more events stimulate human sight and/or hearing. |
| Paper-like content | Concept | A paper like content is a written content suitable to become a paper, with respect to its internal organization, its length, its illustrations (technical drawings and diagrams, for example) and so on. Every paper has its content but not every paper-like content is published as a paper (for example it exists before the publication). |
| Book-like content | Concept | A book like content is a written content suitable to be published as (a part of) a book, with respect to its internal organization, its length, its illustrations (technical drawings and diagrams, for example) and so on. Every book has its content but not every book-like content is published as a book (for example it exists before the publication). |

**Table 15.1** (continued)

| Term | Type | Definition |
|---|---|---|
| Patent-like content | Concept | A patent like content is the a written content suitable to a patent application, with respect to its internal organization, its length, its illustrations (technical drawings and diagrams, for example), and so on. Every application has its content but not every patent-like content is published as an application (for example, if it exists before the application). |
| Project-like content | Concept | A project-like content is written content suitable to apply to answer a call with respect to its internal organization, its length, its illustrations (technical drawings and diagrams, for example), and so on. A project like content is not, in general, an object of publication. |
| Survey | Concept | A paper-like content is a survey if it is an attempt to summarize the current state of understanding on a topic or a knowledge area. |
| Paper | Concept | A paper is an atomic publication. It contains original research results or reviews existing results. Before publication, the content of the paper has undergone a process of *peer review* by one or more referees (who are experts of the same field) who have checked that the content of the paper is suitable for publication in the journal. Such content may undergo a series of reviews, revisions, and re-submissions before finally being accepted or rejected for publication. |
| Monograph | Concept | A monograph is an atomic publication. Although a monograph has, in general, a structure, all its parts share the same group of authors. A monograph can be physically distributed in more volumes. |
| Edited_book_chapter | Concept | An edited book chapter is an atomic publication that is a part of an edited book with specific authors. |
| Unreviewed_content | Concept | It is a written content that has not been reviewed. |
| Diary | Concept | A diary is a record (possibly in handwritten format) with discrete entries arranged by date reporting (typically) on what has happened over the course of a day or a period. |
| Blog | Concept | A blog (a truncation of the expression weblog) is a discussion or informational site published on the World Wide Web and consisting of discrete entries (*posts*) typically displayed in reverse chronological order (the most recent post appears first). Blogs often cover a single subject. A blog is a diary, since the posts are arranged by date. |
| Published_as | Relation | Binds a written content to the publication which disseminates it (if any) |
| Has_version | Relation | Binds a written contents to its new versions (if any) |
| Has_description | Relation | Binds a research activity with descriptive content that descriptive it (if the activity has descriptions); notice that the descriptive content of a research activity may not be one of its outputs, and an output of a research activity may not be a description of the activity itself |
| Has_output | Relation | Binds any activity to its outputs; an activity may produce its outputs at any time when it is operative. |
| Authorship | Relation | Binds a research outcome to the subject which is responsible for it |
| Has_outcome | Relation | Binds a research activity to any output of any activity (not necessarily a research activity) participating in a value chain where the research activity has an enabling role (without the research activity that output would not be generated). The figure shows chains schemes that justify the outcome of the research activity shown in the left. The arrows represent the role has_output. |
| Performed_by | Relation | Binds an activity to the agent who performs it |

**Table 15.2** Main indicator dimensions and their role in the process of indicator development

| Dimension | Description | Role |
|---|---|---|
| Ontological | Conceptual characterization (knowledge representation) of the domain of the indicator | Conceptual definition (*meaning*) of the indicator (a benchmark for the qualitative dimension) |
| Logical | Logical specification of the query(-ies) needed to retrieve all the information (data) needed to calculate the indicator | *Data* definition: selection of the relevant information through the query |
| Functional | Mathematical expression of the indicator (to be applied to the results of the queries) | *Mathematical* definition: related to the selected method of calculation of the indicator (most relevant for the user: the user is interested in the value of the indicator!) Note that the method is outside the ontological domain |
| Qualitative | Ontological questions related to the meaningfulness of the indicator. | Definition of the criteria for the *assessment* of the obtained result (degree of meaningfulness of the indicator) |

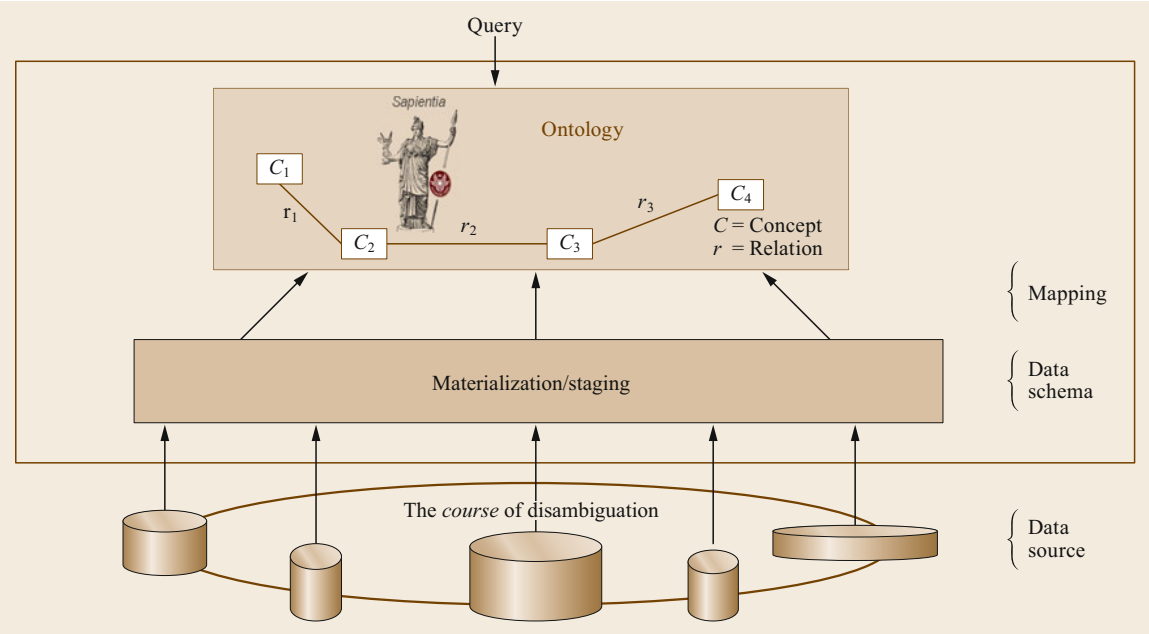



**Fig. 15.12** Illustration of the materialization phase in an OBDI system

## 15.6 Conclusions

In this chapter, we introduced the main challenges in data integration for R&I. We discussed the two main existing approaches to data integration, namely procedural and declarative. We followed the latter approach and focused the subsequent analysis on the OBDI approach. The key idea of OBDI is to resort to a three-level architecture, constituted by the ontology, the sources, and the mapping between the two.

*Daraio*, *Lenzerini*, et al. [15.42] introduce the OBDI 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. They outline the main advantages of OBDI with respect to the traditional silos-based approach to data integration, namely: *conceptual access to the data*, *re-usability*, *documentation and standardization*, *flexibility*, *extensibility*, and *opening of the system* (Table 15.3).

The three main advantages of OBDI for research and innovation analysis [15.33], which encompass and further expand those listed in Table 15.3 are: *openness*, interoperability, and data quality.

An OBDI approach may be an adequate platform/infrastructure to embrace and coordinate in an effective way (i. e., ensuring interoperability and high level of data quality standard), the many initiatives that are going on in research and innovation data collections.

**Table 15.3** Main advantages of an OBDI approach over a traditional *silos*-based approach (after [15.42])

| Advantage | Short description |
|---|---|
| Conceptual access to the data re-usability | Users can access the data by using the elements of the ontology. |
| Documentation and standardization | The mapping layer explicitly specifies the relationships between the domain concepts and the data sources. It is useful for documentation and standardization purposes. |
| Flexibility of the system | You do not have to merge and integrate all the data sources at once which could be extremely costly. |
| Extensibility of the system | You can incrementally add new data sources or new elements (ability to follow the incremental understanding of the domain) when they become available. |
| Opening of the system | Provide a conceptual framework that can be used as a common language by the community. |

Figure 15.13 shows an outline of the main component of an open OBDI infrastructure.

An open STI data platform may encourage and support new research developments in the generation of new indicators carried out by scientists, which exploit the accessibility and transparency of data. In this way, there may be opportunities for the creation of new indicators beyond the short-term needs of policy-makers. The open-data framework, offers the possibility of full documentation on data and explicit articulation of logical linkages. It makes the traditional training and accreditation approach obsolete, in which users of indicators were dependent on the training provided by the owners of the data. Communities of users can, in fact, contribute to the improvement of the documentation and identify pitfalls and shortcomings of indicators. *Sapientia* and OBDI are two technologies to operationalize the consideration of data as infrastructural resources, as they are "shared means to many ends" that satisfy the three criteria of infrastructure resources [15.44]: a) non-rivalrous goods; 2) capital goods; and c) general-purpose inputs. *Sapientia* and OBDM could, indeed, be two *enabling technologies* to improve the exploitation of data for supporting growth and well-being, as proposed by OECD [15.45], and pushing towards the realization of an open science [15.46].

As we have showed in this chapter, the application of OBDI for the integration of data in R&I can be an interesting technology to further explore and exploit. However, it is important to point out that OBDM is not a *panacea* able to solve all the main challenges in data integration for R&I. Nevertheless, it can be a useful tool for *reasoning* over the assessment of research and innovation for different purposes and may lead to a more aware and careful specification of data and indicators useful in the evaluation process. *Sapientia* and its OBDI system may be at the heart of an open and collaborative platform around which to build a knowledge infrastructure for the assessment of research and its impacts.

Besides, the field of OBDI in computer science is far from being stable and consolidated. It is a dynamic and evolving field. *De Giacomo* et al. [15.14] discuss the main challenges related to OBDI that currently deserve investigation, namely querying rewriting optimization, meta-modeling, and meta-querying, non-relational data sources, OBDI methodology and tools, OBDI evolution, and going beyond data access. This is to say that the field of OBDI introduced in this chapter is a relatively new discipline to apply to interesting and complex issues, such as the evaluation of research and innovation, to corroborate the existing methodologies, and develop new and appropriate tools and solutions for data integration in these fields.

We conclude this chapter, recalling the final observation of *Daraio* and *Glänzel* [15.9]:

> One of the main Grand Challenges that remains to address is the exploitation of data availability, Information Technology and current state of the

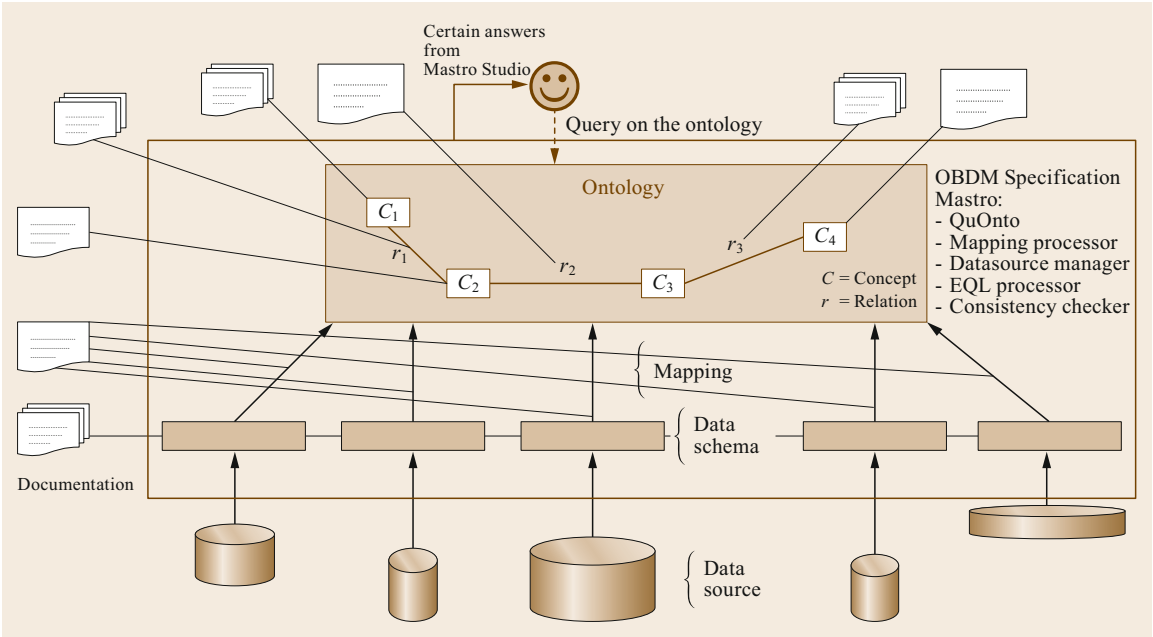



**Fig. 15.13** Illustration of an open OBDI information system (after [15.33])

> art in science and technology for the dynamical setting of standards in a data integration framework in use for multiple purposes.... Within this framework, to deal with this Grand Challenge, the interaction with stakeholders for ensuring an efficient and effective sustainable model is crucial. It depends also on the ability to successfully address, in a systematic way, the other problems highlighted above.

What would be needed to deal with this Grand Challenge and interact with stakeholders in order to develop and implement an OBDI approach for the integration of research information systems and the building of indicators upon them is a *long term investment* in the technology behind the OBDI and its related research. As recalled above, it is a relatively new discipline, which is worth further exploration and exploitation to address the existing challenges in data integration for R&I.

*Acknowledgments.* Financial support from the Project Sapienza Awards 2015 n. C26H15XNFS and the Project FILAS RU 2014-1186 is gratefully acknowledged.

## References


15.1 J. Chen, Y. Chen, X. Du, C. Li, J. Lu, S. Zhao, X. Zhou: Big data challenge: a data management perspective, Front. Comput. Sci. **7**(2), 157–164 (2013)

15.2 H. Ekbia, M. Mattioli, I. Kouper, G. Arave, A. Ghazinejad, T. Bowman, V. Ratandeep Suri, A. Tsou, S. Weingart, C.R. Sugimoto: Big data, bigger dilemmas: A critical review, J. Assoc. Inf. Sci. Technol. **66**(8), 1523–1545 (2015)

15.3 C.L. Borgman: *Big Data, Little Data, No Data: Scholarship in the Networked World* (MIT Press, Cambridge 2015)

15.4 Z. Majkić: *Big Data Integration Theory, Texts in Computer Science* (Springer, Switzerland 2014)

15.5 X.L. Dong, D. Srivastava: Big data integration, Synth. Lect. Data Manag. **7**(1), 1–198 (2015)

15.6 M. Lenzerini: Data integration: A theoretical perspective. In: *Proc. 21st ACM-SIGMOD-SIGART Symp. Princ. Database Syst. PODS2002* (2002) pp. 233–246

15.7 C. Parent, S. Spaccapietra: Database integration: the key to data interoperability. In: *Advances in Object-Oriented Data Modeling*, ed. by M.P. Papazoglou, Z. Zari (MIT Press, Cambridge 2000) pp. 221–253

15.8 C. Daraio: A framework for the assessment of research and its impacts, J. Data Inf. Sci. **2**(4), 7–42 (2017)

15.9 C. Daraio, W. Glänzel: Grand challenges in data integration—state of the art and future perspectives: An introduction, Scientometrics **108**(1), 391–400 (2016)

15.10 OECD: *Quality Framework and Guidelines for OECD Statistical Activities* (OECD, Paris 2011)

15.11 W. Glänzel, S. Katz, H. Moed, U. Schoepflin: Preface, Scientometrics **35**(2), 165–166 (1996)

15.12 W. Glänzel, H. Willems: Towards standardisation, harmonisation and integration of data from heterogeneous sources for funding and evaluation purposes, Scientometrics **106**(2), 821–823 (2016)

15.13 W. Glänzel: The need for standards in bibliometric research and technology, Scientometrics **35**(2), 167–176 (1996)

15.14 G. De Giacomo, D. Lembo, M. Lenzerini, A. Poggi, R. Rosati: Using ontologies for semantic data integration. In: *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years, Studies in Big Data*, Vol. 31, ed. by S. Flesca, S. Greco, E. Masciari, D. Saccà (Springer, Cham 2018)

15.15 C. Daraio, M. Lenzerini, C. Leporelli, P. Naggar, E. Fusco, A. Balducci: Sapientia (the ontology of multidimensional research assessment) and OBDM (ontology based data management) as two key enabling technologies for the development of integrated data platforms for science, technology and innovation (STI). In: *OECD Blue Sky 2016, Ghent* (2016)

15.16 J.D. Ullman: Information integration using logical views. In: *Proc. Int. Conf. Database Theor., ICDT'97*, LNCS, Vol. 1186 (Springer, Berlin, Heidelberg 1997) pp. 19–40

15.17 A.Y. Levy, A.O. Mendelzon, Y. Sagiv, D. Srivastava: Answering queries using views. In: *Proc. 14th ACM-SIGMOD-SIGART Symp. Princ. Database Syst., PODS'95* (1995) pp. 95–104

15.18 A.Y. Halevy, A. Rajaraman, J. Ordille: Data integration: The teenage years. In: *Proc. 32nd Int. Conf. Very Large Data Bases, VLDB 2006* (2006) pp. 9–16

15.19 N.F. Noy, A. Doan, A.Y. Halevy: Semantic integration (editorial), AI Magazine **26**(1), 7 (2005)

15.20 D. Calvanese, G. De Giacomo, M. Lenzerini, R. Rosati, G. Vetere: DL-Lite: Practical reasoning for rich DLs. In: *Proc. Int. Workshop Descr. Log., DL2004*, CEUR, Vol. 104 (2004), http://ceur-ws.org

15.21 D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, R. Rosati: Tractable reasoning and efficient query answering in description logics: The DL-Lite family, J. Autom. Reason. **39**(3), 385–429 (2007)

15.22 A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, R. Rosati: Linking data to ontologies. In: *J. Data Semant*, Vol. 4900 (Springer, Berlin, Heidelberg 2008) pp. 133–173

15.23 M. Lenzerini: Ontology-based data management. In: *Proc. 20th ACM Int. Conf. Inf. Knowl. Manag., CIKM'11* (2011) pp. 5–6

15.24 C. Daraio, M. Lenzerini, C. Leporelli, H.F. Moed, P. Naggar, A. Bonaccorsi, A. Bartolucci: Sapientia: the ontology of multi-dimensional research

assessment. In: *Proc. 15th Int. Soc. Scientometr. Informetr. Conf., Istanbul*, ed. by A.A. Salah, Y. Tonta, A.A. Akdag Salah, C. Sugimoto, U. Al (Bogaziçi Univ. Printhouse, Turkey 2015) pp. 965–977
15.25 F. Baader, D. Calvanese, D. McGuinness, D. Nardi, P.F. Patel-Schneider (Eds.): *The Description Logic Handbook: Theory, Implementation and Applications*, 2nd edn. (Cambridge Univ. Press, Cambridge 2007)
15.26 T. Imielinski, W. Lipski Jr.: Incomplete information in relational databases, J. ACM **31**(4), 761–791 (1984)
15.27 S. Ceri, G. Gottlob, L. Tanca: *Logic Programming and Databases* (Springer, Berlin 1990)
15.28 R. Fagin, G.P. Kolaitis, R.J. Miller, L. Popa: Data exchange: Semantics and query answering, Theor. Comput. Sci. **336**(1), 89–124 (2005)
15.29 P.N. Edwards, S.J. Jackson, M.K. Chalmers, G.C. Bowker, C.L. Borgman, D. Ribes, M. Burton, S. Calvert: *Knowledge Infrastructures: Intellectual frameworks and research challenges* (Deep Blue, Ann Arbor 2013), http://hdl.net/2027.42/97552
15.30 N. Georgescu-Roegen: The economics of production, Am. Econ. Rev. **60**(2), 1–9 (1970)
15.31 N. Georgescu-Roegen: Process analysis and the neoclassical theory of production, Am. J. Agric. Econ. **54**(2), 279–294 (1972)
15.32 N. Georgescu-Roegen: Methods in economic science, J. Econ. Issues **13**(2), 317–328 (1979)
15.33 C. Daraio, M. Lenzerini, C. Leporelli, P. Naggar, A. Bonaccorsi, A. Bartolucci: The advantages of an ontology-based data management approach: Openness, interoperability and data quality, Scientometrics **108**(1), 441–455 (2016)
15.34 X. Li, J.D. Johnson: Evaluate IT investment opportunities using real options theory, Inf. Resour. Manag. J. **15**(3), 32–47 (2002)
15.35 C.Y. Baldwin, K. Clark: *Design Rules – The Power of Modularity* (MIT Press, Cambridge 2000)
15.36 D.L. Parnas: On the criteria to be used in decomposing systems into modules, Commun. ACM **15**(12), 1053–1058 (1972)
15.37 H.A. Simon: The architecture of complexity, Proc. Am. Philos. Soc. **106**, 467–482 (1962)
15.38 D. Lembo, D. Pantaleone, V. Santarelli, D.F. Savo: Easy OWL drawing with the graphol visual ontology language. In: *Proc. 15th Int. Conf. Princ. Knowl. Represent. Reason., KR2016* (2016) pp. 573–576
15.39 D. Lembo, D. Pantaleone, V. Santarelli, D.F. Savo: Eddy: A graphical editor for OWL 2 ontologies. In: *Proc. 25th Int. Jt. Conf. Artif. Intell., IJCAI* (2016) pp. 4252–4253
15.40 OECD: Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development. In: *The Measurement of Scientific, Technological and Innovation Activities* (OECD, Paris 2015), https://doi.org/10.1787/9789264239012-en
15.41 REF Research Excellence Framework: Panel Criteria and Working Methods. https://www.ref.ac.uk/2014/media/ref/content/pub/panelcriteriaandworkingmethods/01_12_1.pdf (2012)
15.42 C. Daraio, M. Lenzerini, C. Leporelli, F.H. Moed, P. Naggar, A. Bonaccorsi, A. Bartolucci: Data integration for research and innovation policy: An ontology-based data management approach, Scientometrics **106**(2), 857–871 (2016)
15.43 C. Daraio, A. Bonaccorsi: Beyond university rankings? Generating new indicators on universities by linking data in open platforms, J. Assoc. Inf. Sci. Technol. **68**, 508–529 (2016)
15.44 B.M. Frischmann: *Infrastructure: The Social Value of Shared Resources* (Oxford Univ. Press, New York 2012)
15.45 OECD: *Data-Driven Innovation Big Data for Growth and Well-Being* (OECD, Paris 2015)
15.46 OECD: Making Open Science a Reality. In: *OECD Science, Technology and Industry Policy Papers*, Vol. 25 (OECD, Paris 2015), https://doi.org/10.1787/5jrs2f963zs1-en



**Maurizio Lenzerini**

Department of Computer, Control, and Management Engineering
Sapienza University of Rome
Rome, Italy
*lenzerini@diag.uniroma1.it*

Maurizio Lenzerini is a Full Professor of Data Management and Semantic Technologies at the University of Rome La Sapienza. He has provided fundamental contributions to data management, knowledge representation and automated reasoning, and ontology-based data access and integration. He is the author of more than 350 publications. He is the recipient of two IBM Faculty Awards, a Fellow of EurAI, Fellow of the AAAI, and Fellow of the ACM.

**Cinzia Daraio**

Department of Computer, Control, and Management Engineering
Sapienza University of Rome
Rome, Italy
*daraio@diag.uniroma1.it*

Cinzia Daraio is an Associate Professor of Management Engineering at the University of Rome "La Sapienza", where she teaches Productivity and Efficiency Analysis, Quantitative Models for Economic Analysis and Management and Economics and Business Organization. She is specialized in science and technology indicators, higher education microdata and methodological and empirical studies in productivity and efficiency analysis. She has authored more than 150 publications.