Data Integration for Research and Innovation (R&I) Policy: challenges, *Sapientia* and state of the art of the KIMAR Project

SAPIENZA
UNIVERSITÀ DI ROMA

Cinzia Daraio (cinzia.daraio@uniroma1.it)

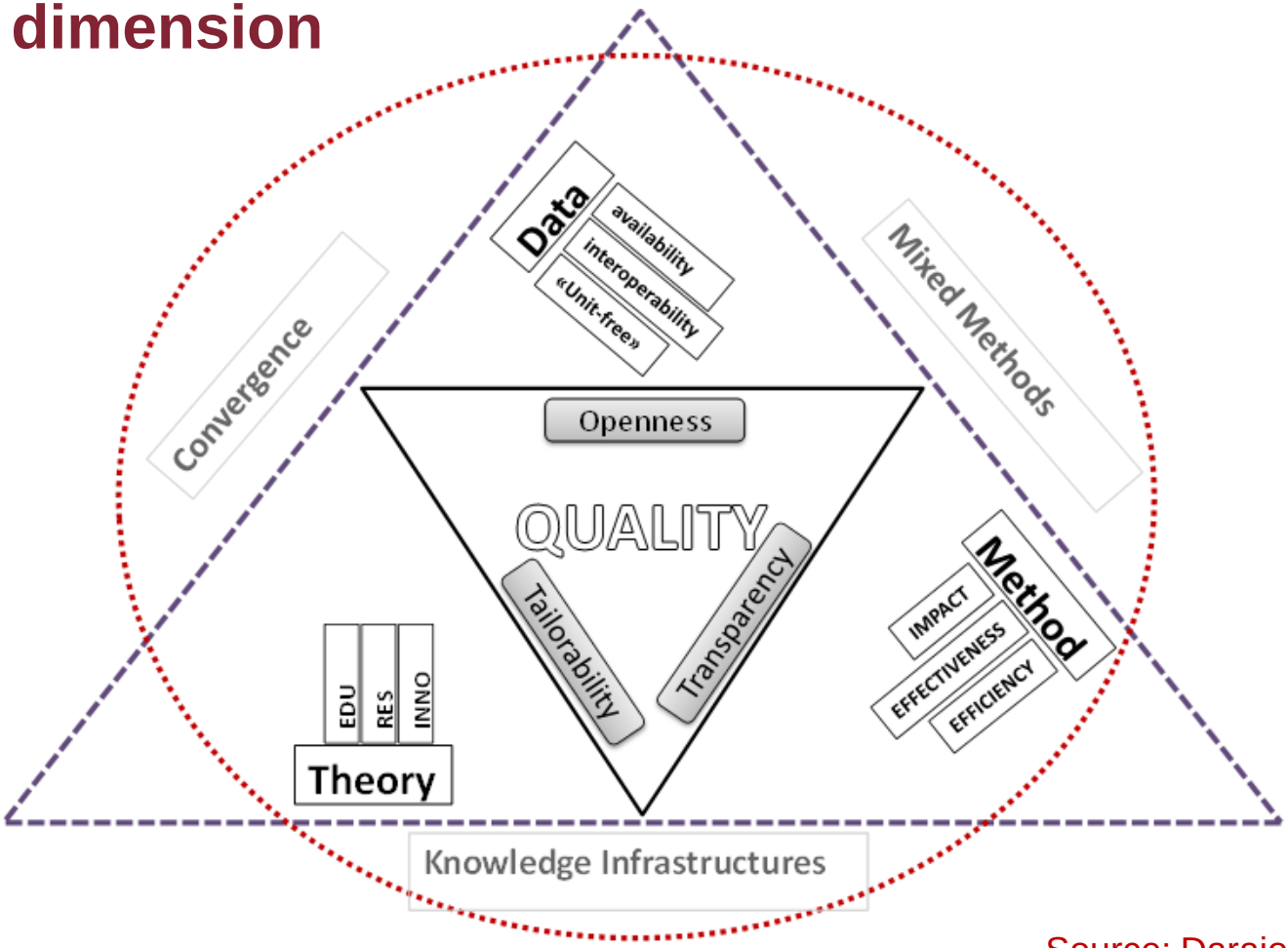
University of Rome La Sapienza

Joint EU-OECD Semantic Workshop Bruxelles, 19 June 2017

Introduction and outline

- Recent trends in research assessment, computerization of bibliometrics, altmetrics, complexity of research assessment, granularity, increasingly demanding policy needs.
- Crucial role of data!
- Our proposal: an Ontology-Based Data Management Approach (OBDM)
- *Sapientia*: the Ontology of Multi-Dimensional Research Assessment (includes also altmetrics!)
- State of the art and works in progress
- An OBDM approach to design and implement a new generation of Science, Technology and Innovation (STI) indicators

A new holistic framework for the assessment of Research and Innovation which includes «Data» as a dimension



Source: Daraio (2017).

Grand Challenges in Data Integration for Research and Innovation (R&I) Policy.

The Grand Challenges identified (Daraio and Glanzel, 2016) were:

- Handling Big Data,
- Coping with Quality Issues,
- Anticipating New Policy Needs.

Framed in four areas of intervention:

1. data collection/project initiatives,
2. open data, linked data and platforms for Science, Technology and Innovation (STI),
3. Monitoring performance evaluation
4. stakeholders, actions, options, costs and sustainability.

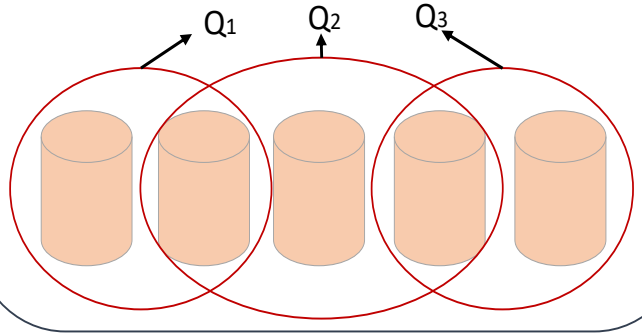
Grand Challenges in Data Integration for R&I.

The identified (not exhaustive) critical issues:

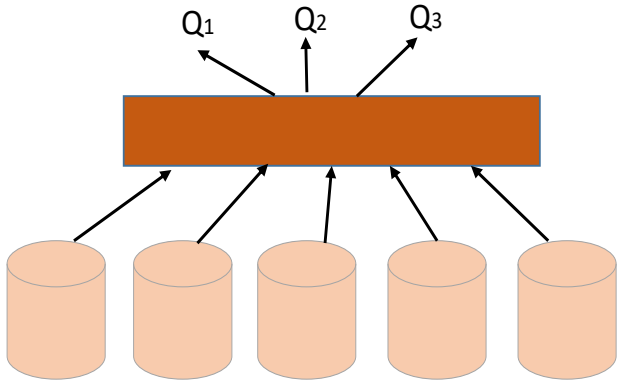
- data quality (considered as “fitness for use” with respect to user needs, see OECD, 2011) issues - completeness, validity, accuracy, consistency, availability and timeliness;
- comparability problems related to heterogeneous definitions of the variables, data collection practices and databases;
- lack 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.

Approaches to data integration for R&I

Procedural or bottom-up
(called in gergo silos approach):
For every «indicator need», figure out which data you need and how they can be accessed, and design/realize a corresponding service



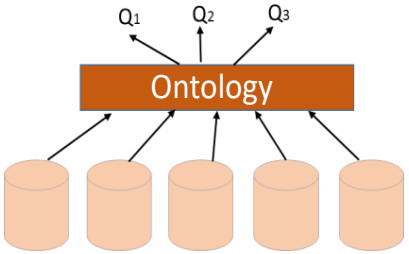
Declarative or top-down:
Define a global structure which is valid for all source data, link this structure to the data, use this structure to specify the «indicator needs» and automatically extract the right data from the sources



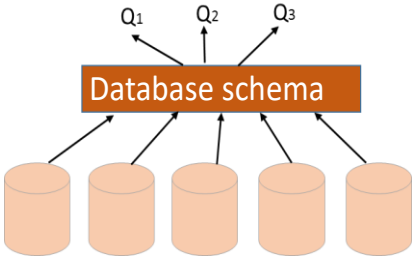
OBDM (Ontology-Based Data Management):
A new declarative paradigm for STI data integration and governance

- Use knowledge Representation and Reasoning principles and techniques for managing data.
- Leave the data where they are.
- Build a conceptual specification of the domain.
- Map such knowledge structure to concrete data sources
- Express all the indicators over the abstract representation
- Automatically translate conceptual indicators to data

OBDM

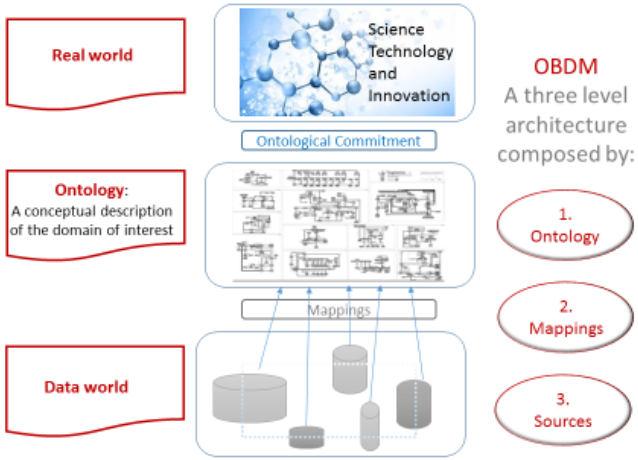


Declarative database integration



The main purpose of an OBDM System

- is to allow information consumers to query the data using the elements in the ontology as predicates.
- it can be seen as **a form of information integration**, where the usual global schema is replaced by the conceptual model of the application domain, formulated as an ontology expressed in a logic-based language.



The OBDM Approach (Calvanese et al. 2010; Lenzerini, 2011; Poggi et al. 2008)

- Key idea: a **three-level architecture**, constituted by:
- **The ontology**: is a conceptual, formal description of the domain of interest (expressed in terms of relevant concepts, attributes of concepts, relationships between concepts, and logical assertions characterizing the domain knowledge).
- **The sources**: are the repositories accessible by the organization where data concerning the domain are stored. In the general case, such repositories are numerous, heterogeneous, each one managed and maintained independently from the others.
- **The mapping**: is a precise specification of the correspondence between the data contained in the data sources and the elements of the ontology.

The main advantages of an OBDM Approach

1. Users can **access** the data by using the elements of the ontology.
2. By making the representation of the domain explicit, we gain **re-usability** of the acquired knowledge.
3. The mapping layer explicitly specify the relationships between the domain concepts and the data sources. It is useful also for **documentation and standardization** purposes.
4. **Flexibility** of the system: you do not have to merge and integrate all the data sources at once which could be extremely costly!

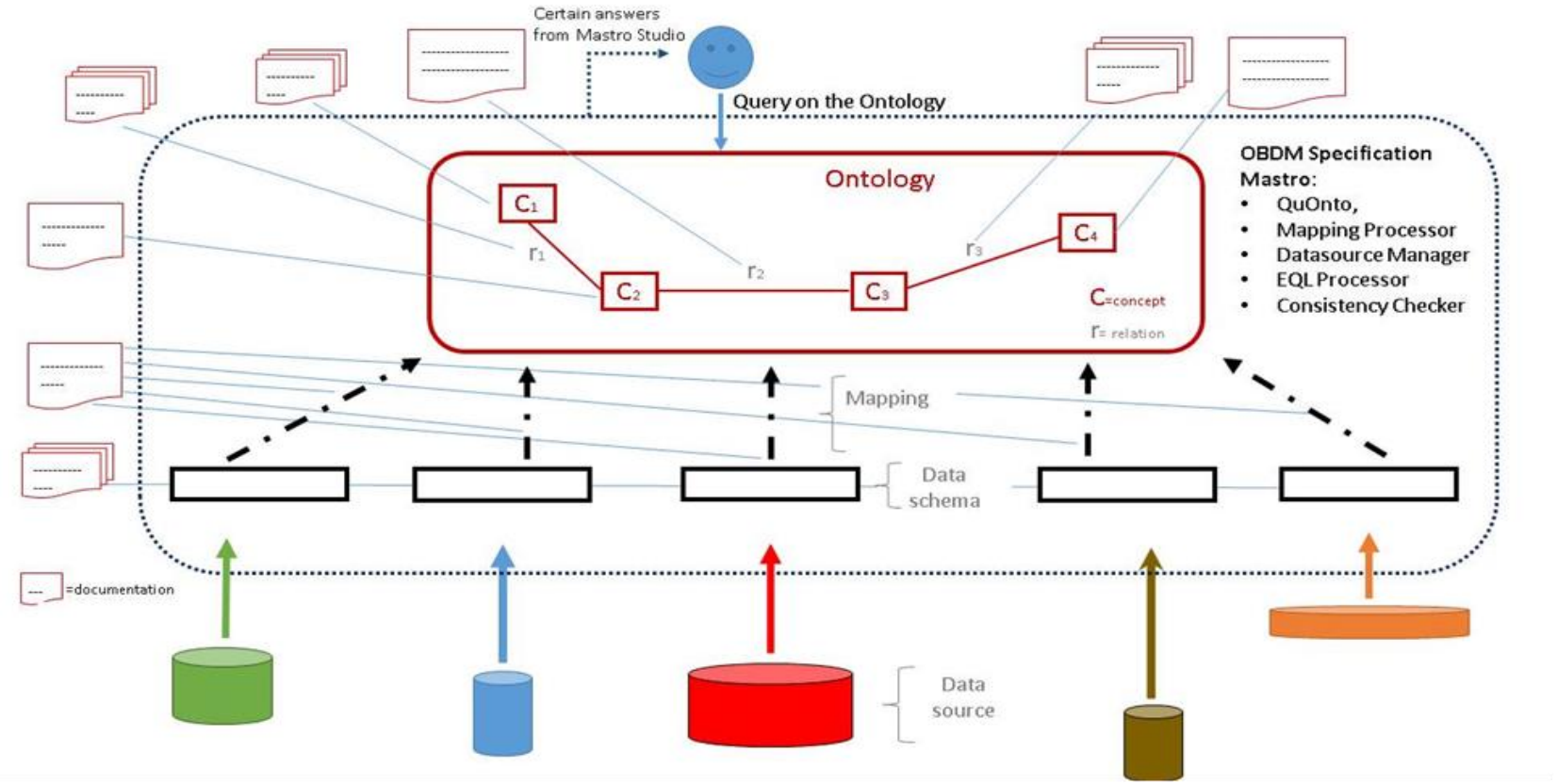
The main advantages of an OBDM Approach (cont-)

5. 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!

6. Opening of the system: provide a conceptual framework which can be used as a common language by the community.

7. A step towards an open science system!

The main advantages of an OBDM Approach: interoperability, openness, data quality



Scope of *Sapientia*

- The main objective of *Sapientia* is to model 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.
- This is a difficult task that needs to be addressed with a systemic view accounting for all the interactions of research with education and innovation

Sapientia's principles

1. We started with a *top-down* modelling 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.
2. 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.
3. 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.

Sapientia's principles

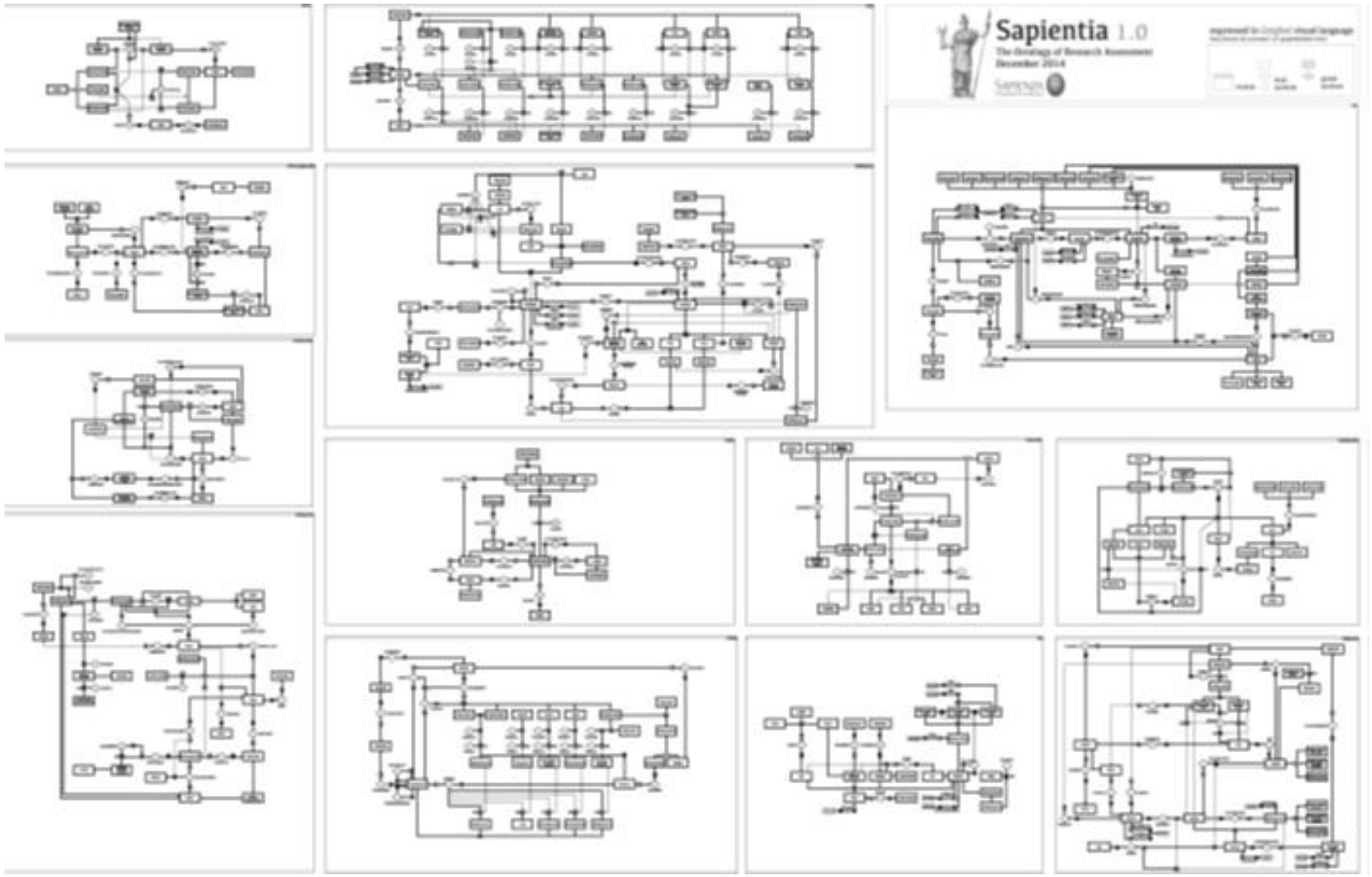
4. We pursued a modelling approach based on *processes*, which are conceived as collections of activities. A process is composed by inputs and outputs.

6. Individuals and *activities* are the main pillars of the ontology.

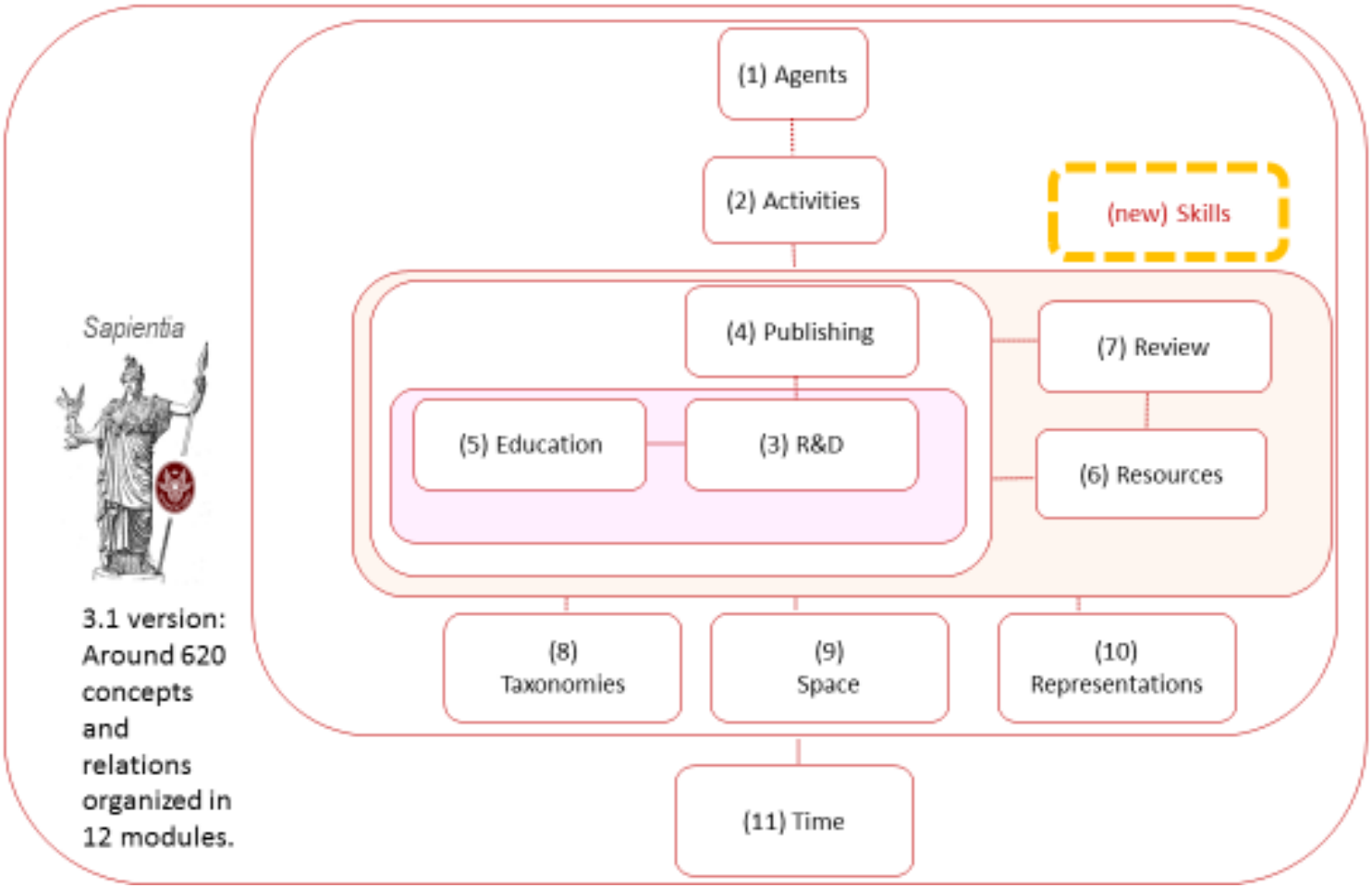
7. We followed a modelling approach based on a *modularization* of the system. Our ontology is organized in modules. As we shall see later, we have two kind of modules: *functional* modules and *structural* modules. By functional modules we mean modules which model the main agents and activities of our domain (namely Agents, Activities, R&D, Publishing, Education, Resources and Review). *By structural modules we mean those modules which represent the constituent elements of the ontology to ensure its long lasting and general-purpose functionality (namely, Taxonomies, Space, Representations and Time).*

THERE ARE HUNDREDS/THOUSANDS OF ONTOLOGIES WORLDWIDE AND IT IS DIFFICULT TO COMPARE THEM: WE DID A PILOT EXERCISE!

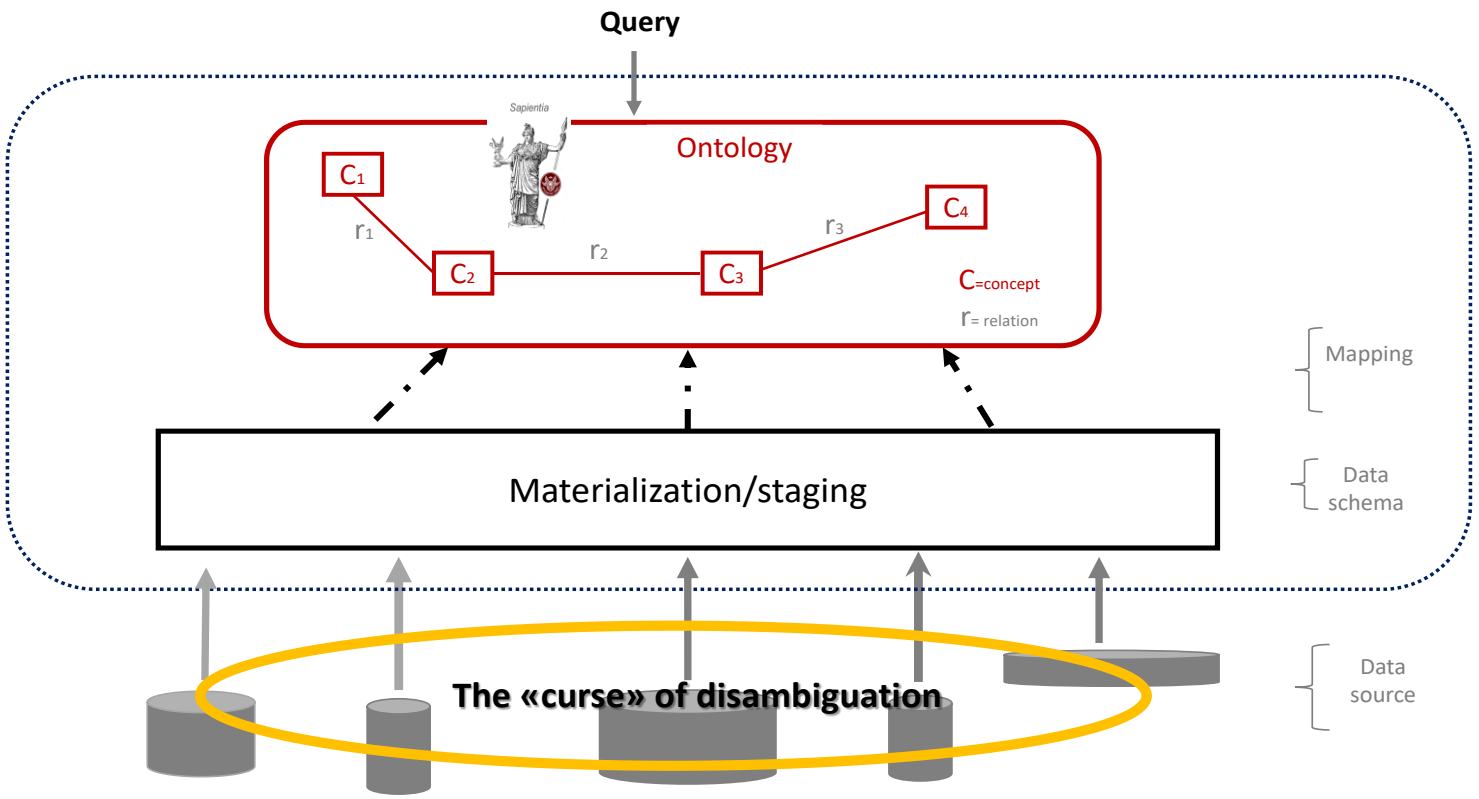
First version of *Sapientia: 1.0*



Current version of *Sapientia: 3.1*



Crucial role of Mappings: Illustration of the materialization phase in an OBDM system



Sapientia and KIMAR works in progress

- Consolidation and mappings (importance of research on mappings!)
- Mappings with ETER data/test of data quality within RISIS
- WoS Data (KOL Project)
- Scopus Data
- DBLP
- Open data from the EU
- Patents
- Inclusion of the Skills module (matching competences and key enabling technologies- Project for Lazio Region)
- We would be available/interested to make a joint cooperation with the REITER PROJECT to implement policies (defined over concepts maybe easier than defined over data)

Problems in STI indicator development and application

- Concepts not defined clearly (e.g. “publication”)
- Informal definitions based on everyday language
- One concept name may refer to different concepts
- Ad hoc definitions of indicators based on available datasets or specific user needs
- Indicators non re-usable in future contexts
- Database content is not fully transparent
- Aggregate indicators cannot be decomposed into smaller units
- Increase in data sources, assessment criteria and actual use asks for an overarching structure

Advanced specification and calculation of STI indicators

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)

Benefits of the OBDM Approach

- The formal specification of the indicators is made *independently* of the data
- OBDM offers the opportunity to compute “comparable” indicator values at different level of aggregation
- It offers a reference system to *check the comparability* level among the heterogeneous data sources
- OBDM approach permits an *unambiguous* way to define and compute the indicators
- Knowledge on the indicator system (concepts and data sources) is embedded in a formal framework
- This knowledge can be transferred more easily to new generations of producers and users

Is the current «indicator factory» fitting with a new Quadruple helix interactive innovation model?

Bonaccorsi, Catalano, Daraio, Moed (Blue Sky, 2016)

- Quadruple helix: Add two new dimensions to the University-Government-Industry triple:
- Citizens and their organizations
- Practice-based, bottom-up learning models
- This requires multidisciplinary knowledge, involve public-private interaction, need radically new business models and public governance models.
- The established factory in not adequate then.
- The new indicators must be created ex-novo and must be designed and produced interactively!

Generating new indicators from an open platform

- The development of new technologies in database design, data analysis, data quality, data integration, data visualization and related allow a conceptual jump in the search for a solution.
- The first step is conceptualizing research and innovation systems as complex systems formed by more elementary *entities*.
- The second step is to develop *concepts* associated to entities. The development of concepts may be helped by the users requirements.
- The third step is to look for lists and standards, or more generally in the database language, Authority Files.
- The fourth step is to establish links between concepts.
- Once these steps are carried out what we obtain is a graph of actors, each with associated a graph of concepts (direct and indirect link).
- We then are ready to implement a principle for the integration of research and innovation information systems. The informal principle we follow is this: ***put your information in the region of the graph in which the potential for generation of other information is greater***. By using this approach a number of new indicators can be generated.

Generating new indicators from an open platform

- We aim at introducing a change in the way indicators are designed. We suggest a departure from the traditional approach:
- start top down with a thorough conceptual analysis of entities in the research and innovation system.
- draw all linkages between concepts, particularly the sharing of concepts between entities.
- examine the linkages down to the lowest possible level of granularity, in order to ensure that the sharing is complete
- generate new indicators through the combination of concepts shared across entities.
- implement the new indicators with existing data, and to invest into the areas that may produce the largest increase in connectivity among entities.
- Following this approach, once indicators are created the feasibility is immediately visible. If there is a hole in the data, then it is clear where to invest in production of data.

Some Selected references

- Daraio C. (2017), A framework for the assessment of Research and its impacts, DIAG Technical Report n. 4, 2017, Sapienza University of Rome.
- Daraio C., Bonaccorsi A., (2017), Beyond university rankings? Generating new indicators on universities by linking data in open platforms, Journal of the Association for Information Science and Technology.
- Daraio C., Lenzerini M., Leporelli C., Moed H.F., Naggar P., Bonaccorsi A., Bartolucci A. (2015). Sapientia: the Ontology of Multi-Dimensional Research Assessment, in Salah, A.A., Y. Tonta, A.A. Akdag Salah, C. Sugimoto, U. Al (Eds.), Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference, Istanbul, Turkey, 29 June to 3 July, 2015, Bogaziçi University Printhouse, pp. 965-977.
- Daraio, C., Glänzel, W. (2016). Grand challenges in data integration - state of the art and future perspectives: an introduction. Scientometrics, 108 (1), 391-400.
- Daraio, C., Lenzerini M., Leporelli C., Naggar P., Bonaccorsi A. & Bartolucci, A. (2016b). The advantages of an Ontology-based Data Management Approach: openness, interoperability and data quality. Scientometrics, 108 (1), 441-455.
- Daraio, C., Lenzerini, M., Leporelli, C., Moed, F. H., Naggar, P., Bonaccorsi, A. & Bartolucci, A. (2016a). Data integration for research and innovation policy: An Ontology-Based Data Management approach. Scientometrics, 106 (2), 857-871.