Introduction: the challenge of Data Quality for Research and Higher Education

SAPIENZA
UNIVERSITÀ DI ROMA

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Cinzia Daraio (E-mail: cinzia.daraio@uniroma1.it)
DIAG Dipartimento di Ingegneria Informatica, Automatica e Gestionale Antonio Ruberti

Rome, 15-17 September 2020, RISIS Methodological Course “Data Quality for Research and Higher Education Studies”

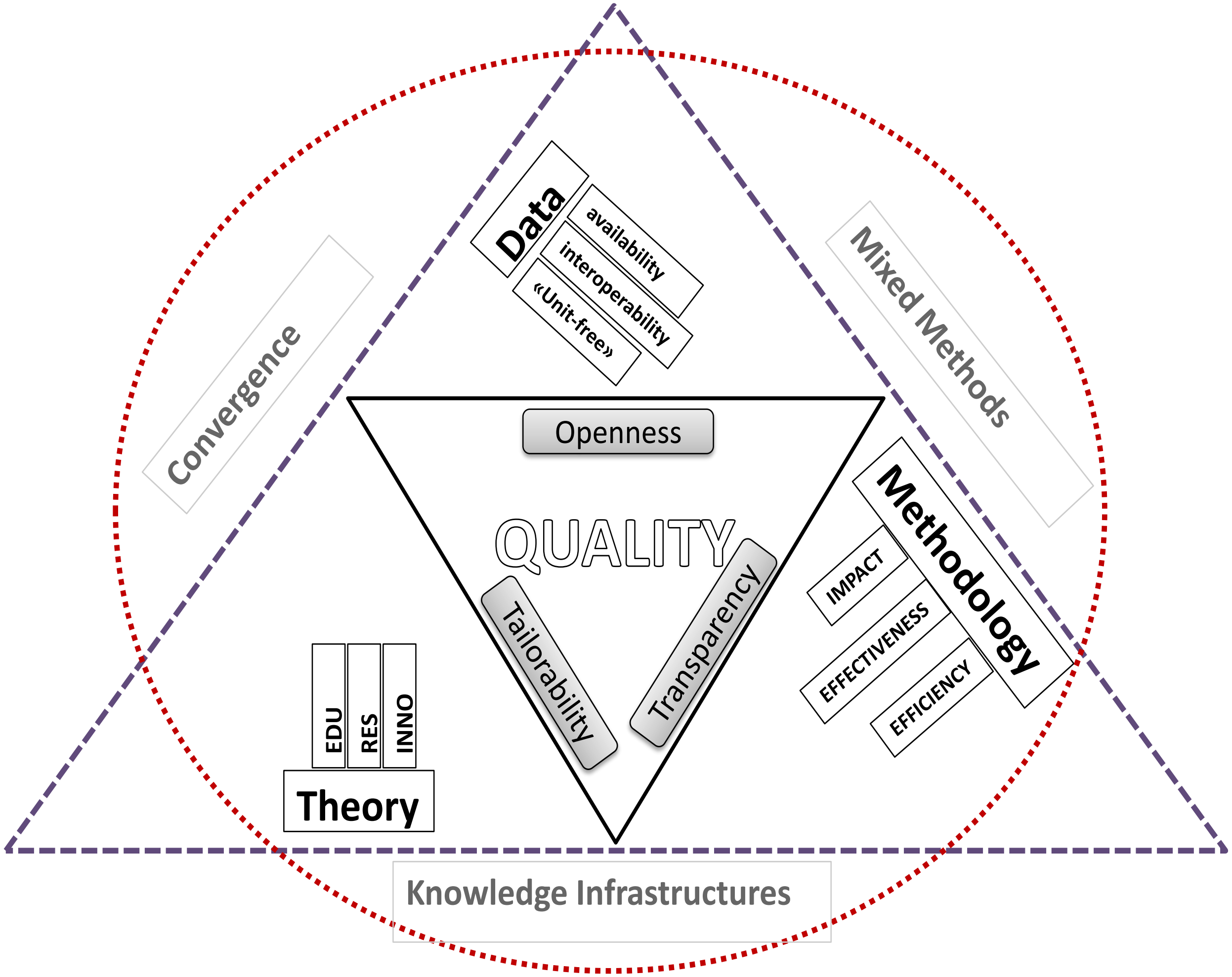
Outline

- Introduction: the challenge of Data Quality for Research and Higher Education
- What the Course is about
    - Objectives
    - Content
    - Structure
    - Attendance, Questionnaire and Certificate of Participation
- *Tour* of participants

Introduction

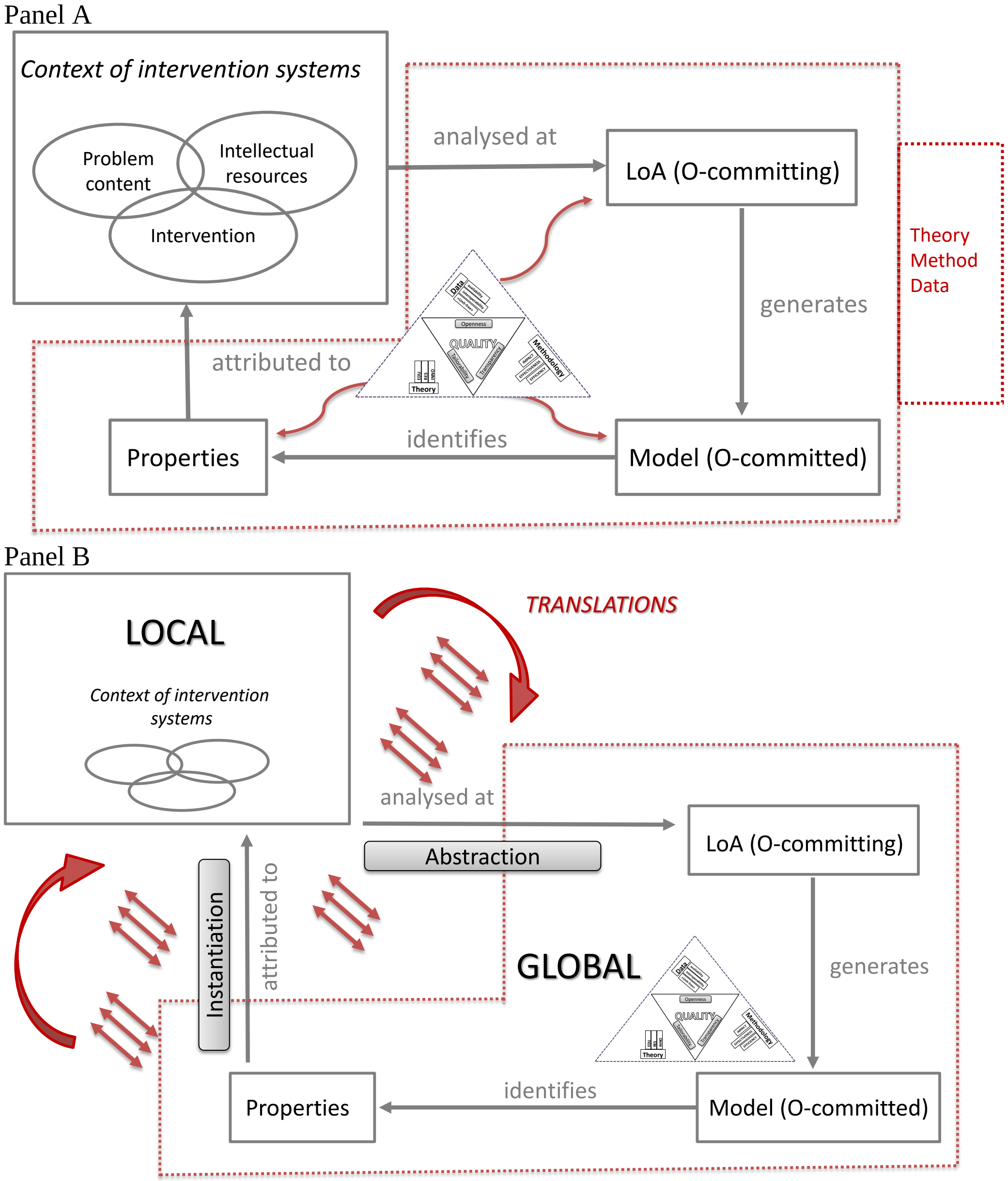
- The quality of data and of related information is *crucial* to add value and improve the awareness and better exploitation of the available data, enhancing data *quality-aware empirical investigations* for studies in Research, Education and Innovation.
- We need a broader framework in which Quality is included and considered along all the dimensions and their interconnections.
- A framework is required to develop models of metrics/indicators and assess their robustness.

Introduction: Relevance of Quality in a broader framework



Source: Daraio (2017a).

Need of an overall framework: *difficulties* due to the implementation problem



Panel A

Context of intervention systems

Problem content

Intellectual resources

Intervention

analysed at

LoA (O-committing)

generates

Model (O-committed)

identifies

Properties

attributed to

Theory
Method
Data

Panel B

LOCAL

Context of intervention systems

TRANSLATIONS

analysed at

Abstraction

LoA (O-committing)

generates

GLOBAL

Model (O-committed)

identifies

Properties

attributed to

Instantiation

Source: Daraio (2017b).

DATA AND INTEROPERABILITY: a crucial component

The quality of data is *context-dependent* and an appropriate quality of a single dataset, for a specific purpose, is not enough.

The linkages between different datasets are relevant as well. The compatibility, interchangeability and the connectability of a given dataset with other related data are fundamental aspects which need to be taken into account (Daraio and Glanzel, 2016).

DATA AND INTEROPERABILITY: a crucial component

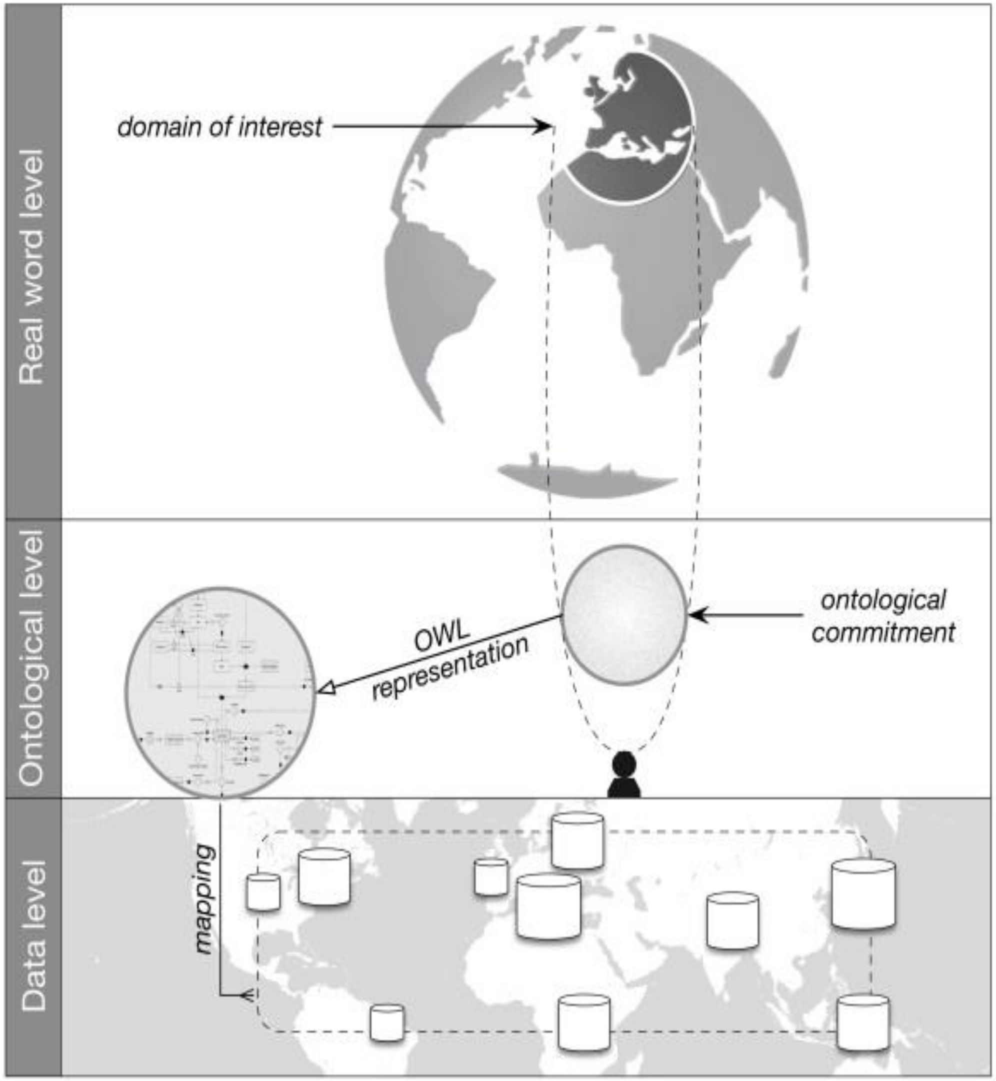
Data integration is the problem of combining data residing at different sources, and providing the user with unified view of these data (Lenzerini, 2002).
According to Parent and Spaccapietra (2000), ***interoperability*** is the way in which heterogeneous systems talk to each other and exchange information in a meaningful way.
They identified three levels of interoperability:
- lowest level (no integration),
- intermediary level (the system does not guarantee consistency across database boundaries)
- higher level that has the goal of developing a global system on top of existing system, to provide the desired level of integration of the data sources.

DATA AND INTEROPERABILITY: a crucial component

Several levels of conceptual interoperability have been identified in the specialized literature. For instance, Tolk and Muguira (2003) propose the following 5 levels of conceptual interoperability:

Level 0: System specific data (isolated systems);
Level 1: Documented data (documentation of data and interfaces);
Level 2: Aligned static data through Meta Data Management (use of common reference models/common ontology);
Level 3: Aligned dynamical data and "Implemented processes" (common system approach/open source code);
Level 4: Harmonized data and processes, conceptual model, intend of use (common conceptual model/semantic consistency).
The formal and precise means to achieve level 4 of interoperability (harmonized data and processes) is a *logic-oriented ontology language*.
This is exactly what the OBDM approach allows for: see next slides

An illustration of the OBDM Approach (Source: Daraio et al, 2016 a,b)



The OBDM Approach (Source: Daraio et al, 2016 a,b, Lenzerini and Daraio, 2019)

- 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 complexity and trade-offs of Data Quality

- Data quality is a very complex topic, in which the theory and practice often differ.
- In practice, data quality does play an important role in the design of data architecture.
- All the data quality efforts must start from a solid understanding of high-priority use cases, and use that insight to navigate various trade-offs to optimize the quality of the final output.
- The followings are trade-offs related to data quality:
- Should we select data for cleaning based on the cost of cleaning effort or based on how frequently the data is used or based on its relative importance within the data models consuming it? or a combination of those factors? What sort of combination?
- Is it a good idea to improve data accuracy by getting rid of incomplete or erroneous data? While removing some data, how do we ensure that we do not introduce distortions or bias?

References

- Daraio C. (2017a), A framework for the assessment of Research and its Impacts, *Journal of Data and Information Science*, Vol. 2 No. 4, 2017 pp 7–42.
- Daraio C. (2017b), Assessing research and its impacts: The generalized implementation problem and a doubly-conditional performance evaluation model, *ISSI 2017 - 16th International Conference on Scientometrics and Informetrics, Conference Proceedings*, pp. 1546-1557.
- Daraio C., Bruni R., Catalano G., Matteucci G., Daraio A., Scannapieco M, Wagner-Schuster D., Lepori B. (2020), A tailor-made Data Quality Approach for Higher Educational Data, *Journal of Data and Information Science*, 5(3), 129–160.
- Daraio C., Glänzel W. (2016), Grand Challenges in Data Integration. State of the Art and Future Perspectives: An Introduction, *Scientometrics*, 108 (1), 391-400.
- Daraio C., Lenzerini M., Leporelli C., Moed H.F., Naggar P., Bonaccorsi A., Bartolucci A. (2016), Data Integration for Research and Innovation Policy: An Ontology-based Data Management Approach, *Scientometrics*, 106 (2), 857-871.
- Daraio, C., Lenzerini, M., Leporelli, C., Naggar, P., Bonaccorsi, A., Bartolucci, A. (2016). The advantages of an Ontology-Based Data Management approach: openness, interoperability and data quality. *Scientometrics*, 108(1), 441-455.
- Lenzerini M. and Daraio C. (2019), Challenges, Approaches and Solutions in Data Integration for Research and Innovation, in *Springer Handbook of Science and Technology Indicators* edited by Glänzel W., Moed H.F., Schmoch H. and Thelwall M., 397-420, ISBN 978-3-030-02511-3.

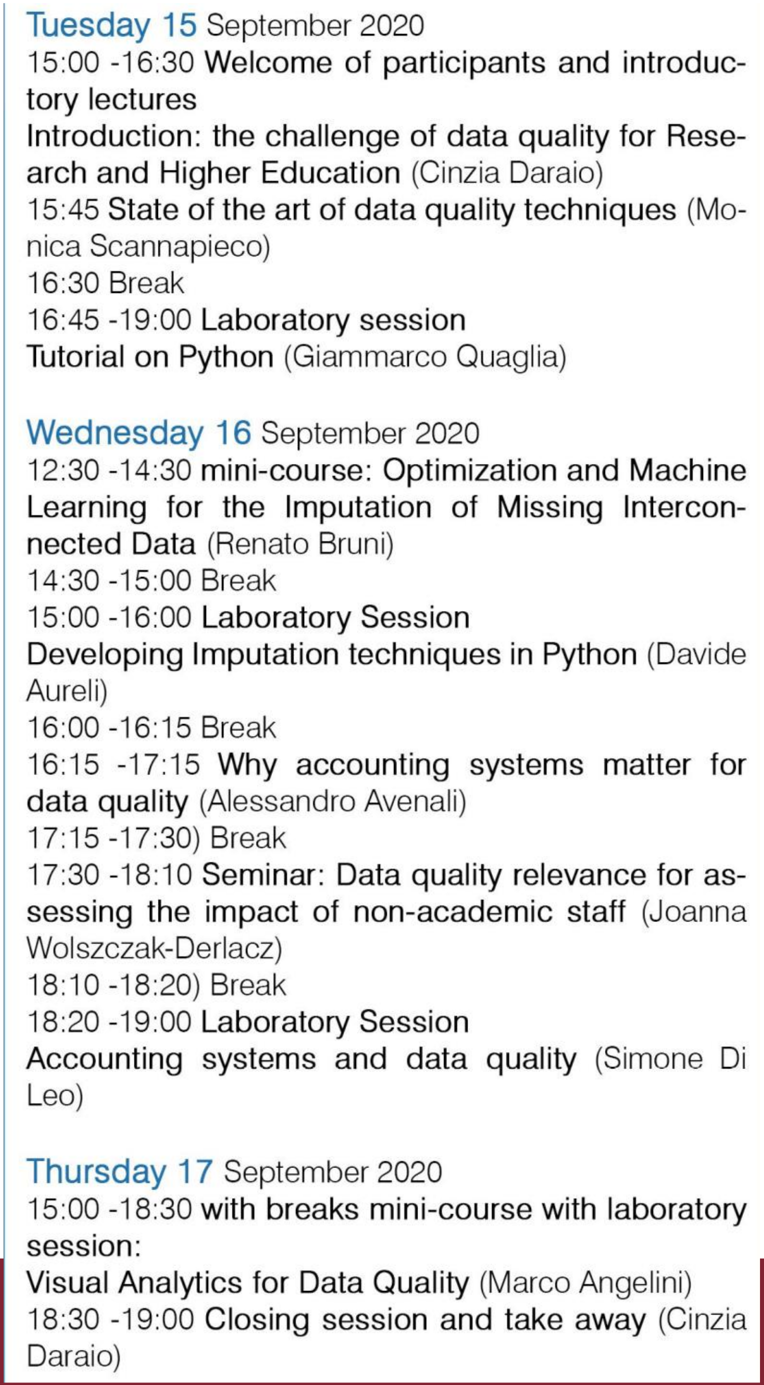
What the Course is about: Course Objectives

- Introduce the participants to the importance of Quality issues for Data on research and higher education studies;
- Provide the participants with the basic knowledge for understanding and use data quality techniques in their context;
- Propose tools for implementing data quality techniques in their context of application;
- Offer tutorials on the main software that will be used during the course;
- Encourage the participants to explore the proposed tool with their own datasets;
- Offer the possibility to interact with the Course’s lecturers to have advice on their own specific needs;
- Expose participants to seminars on advanced topics related to data quality issues and methodological approaches.

What the Course is about: Contents

- Data quality for research and higher education studies: where we stand and where we are going
- Laboratory sessions: introduction to the software
- Accounting for data quality by analyzing accounting systems
- Data quality tools: introduction and practical exercises
- Imputation techniques: presentation and implementation
- State of the art of advanced methods and challenges ahead.

What the Course is about: Program



Tuesday 15 September 2020
15:00 -16:30 Welcome of participants and introductory lectures
Introduction: the challenge of data quality for Research and Higher Education (Cinzia Daraio)
15:45 State of the art of data quality techniques (Monica Scannapieco)
16:30 Break
16:45 -19:00 Laboratory session
Tutorial on Python (Giammarco Quaglia)

Wednesday 16 September 2020
12:30 -14:30 mini-course: Optimization and Machine Learning for the Imputation of Missing Interconnected Data (Renato Bruni)
14:30 -15:00 Break
15:00 -16:00 Laboratory Session
Developing Imputation techniques in Python (Davide Aureli)
16:00 -16:15 Break
16:15 -17:15 Why accounting systems matter for data quality (Alessandro Avenali)
17:15 -17:30) Break
17:30 -18:10 Seminar: Data quality relevance for assessing the impact of non-academic staff (Joanna Wolszczak-Derlacz)
18:10 -18:20) Break
18:20 -19:00 Laboratory Session
Accounting systems and data quality (Simone Di Leo)

Thursday 17 September 2020
15:00 -18:30 with breaks mini-course with laboratory session:
Visual Analytics for Data Quality (Marco Angelini)
18:30 -19:00 Closing session and take away (Cinzia Daraio)

What the Course is about: Attendance, Questionnaire and Certificate of Participation

- Attendance of all the sessions/activities of the Course is compulsory
- At the end of the Course there will be a Questionnaire to ask for your feedback and suggestions for improving the next courses/activities
- Certificate of Participation will be given to all participants who attended all the sessions/activities and filled in the Questionnaire