Data Driven Digital Libraries: The Case of Data Citation

Gianmaria Silvello
@giansilv

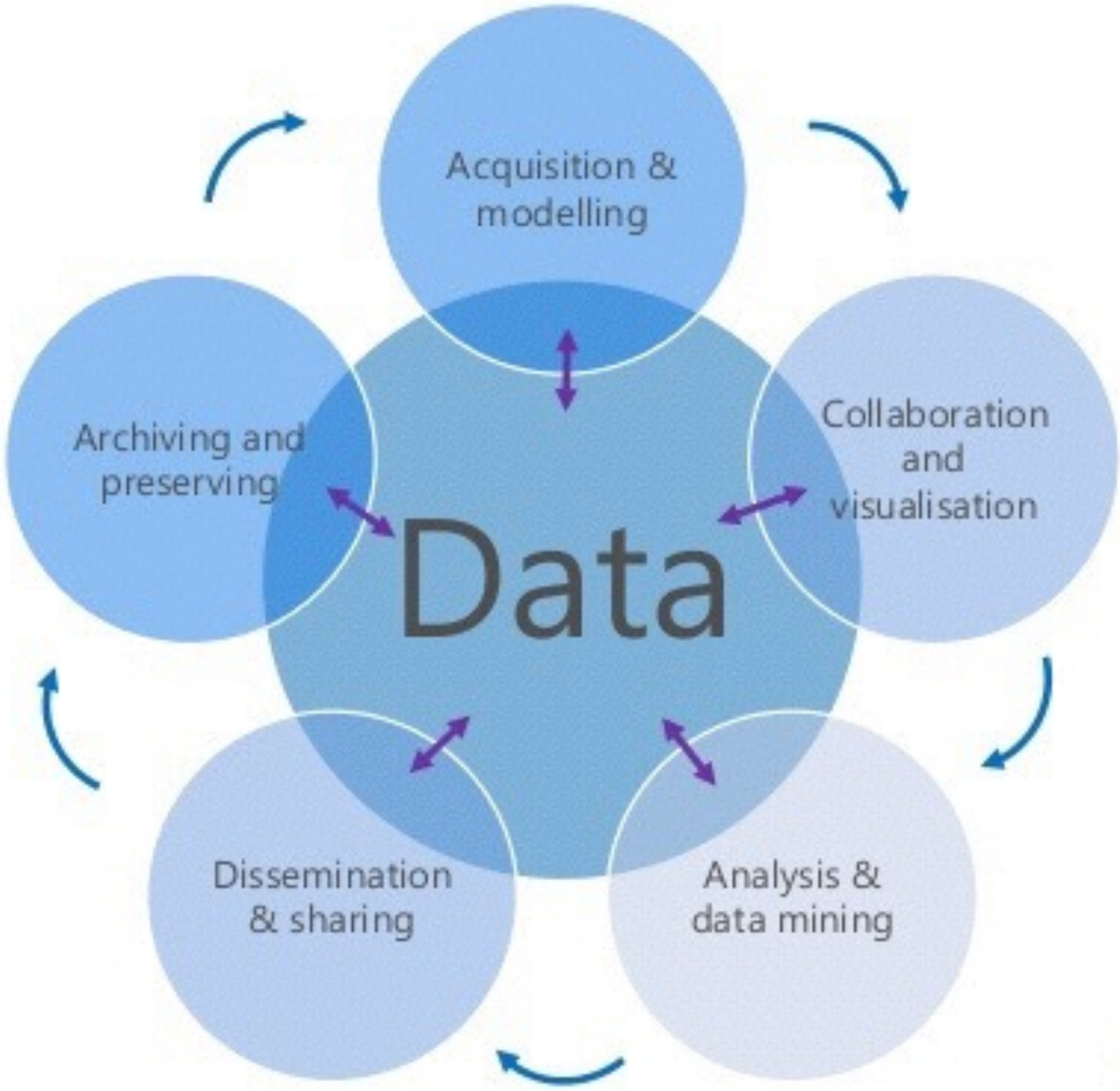
Information Management Systems Research Group
Department of Information Engineering
University of Padua, Italy

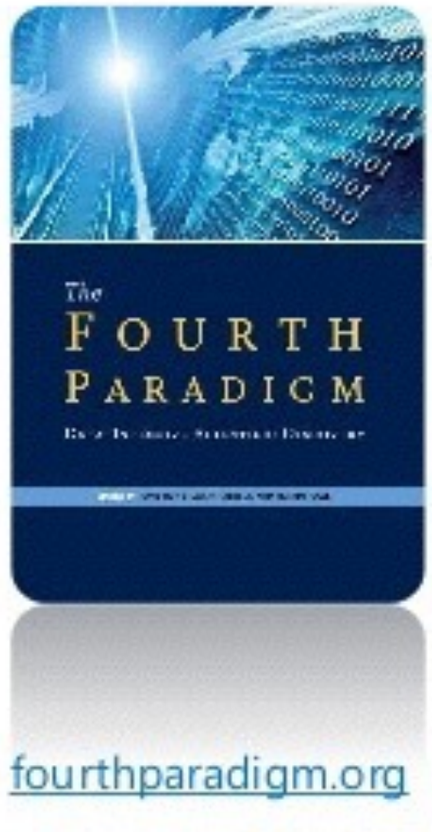
Digital Humanities, Digital Libraries and Information Science: what relation?

Florence, Italy, 24 October 2016

Data-Driven Science Discovery

Data-intensive Research





Science@Microsoft
THE FOURTH PARADIGM IN PRACTICE

Data-Driven Science Discovery

- Data is increasingly central for scientific progress

- Also text and traditional references now can be seen as data

- How do digital libraries have to evolve for handling the transition to data centric scientific discovery?

New DL Data Centric Services

- Data Provenance
- Data Citation
- Data Quality
- Data Discovery

Data Citation

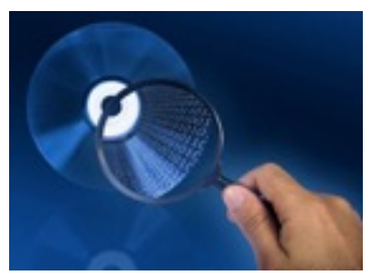
Let’s focus on data citation as a case study

Why Data Citation is Important?

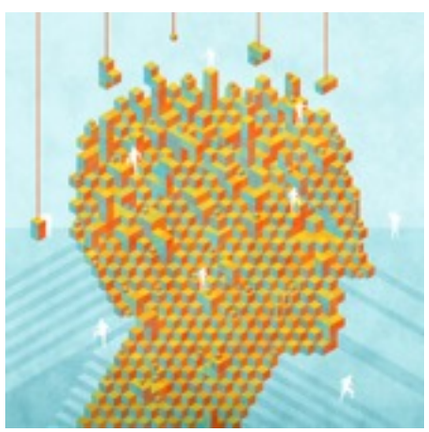
- Give credit to data creators and curators (and institutions)

- Repeatability, reproducibility and generalizability of research

- Referencing data in order to identify, discover and retrieve them



- Building and propagating knowledge



~~Why~~ Data Citation is Important ~~?~~

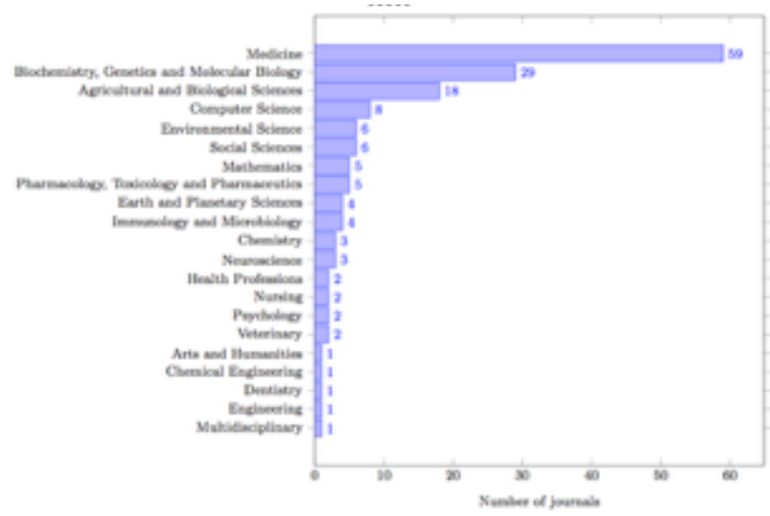
A lot of work has been done…

- Principles of data citation

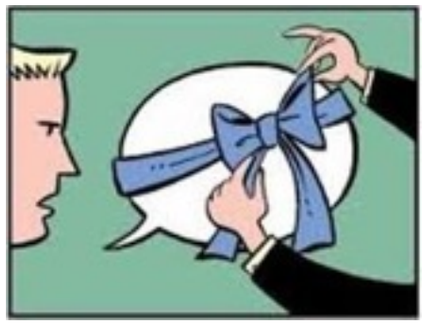
- Recommendations for data citation systems

- Data publishing infrastructures and data journals

- Indexes and dataset impact

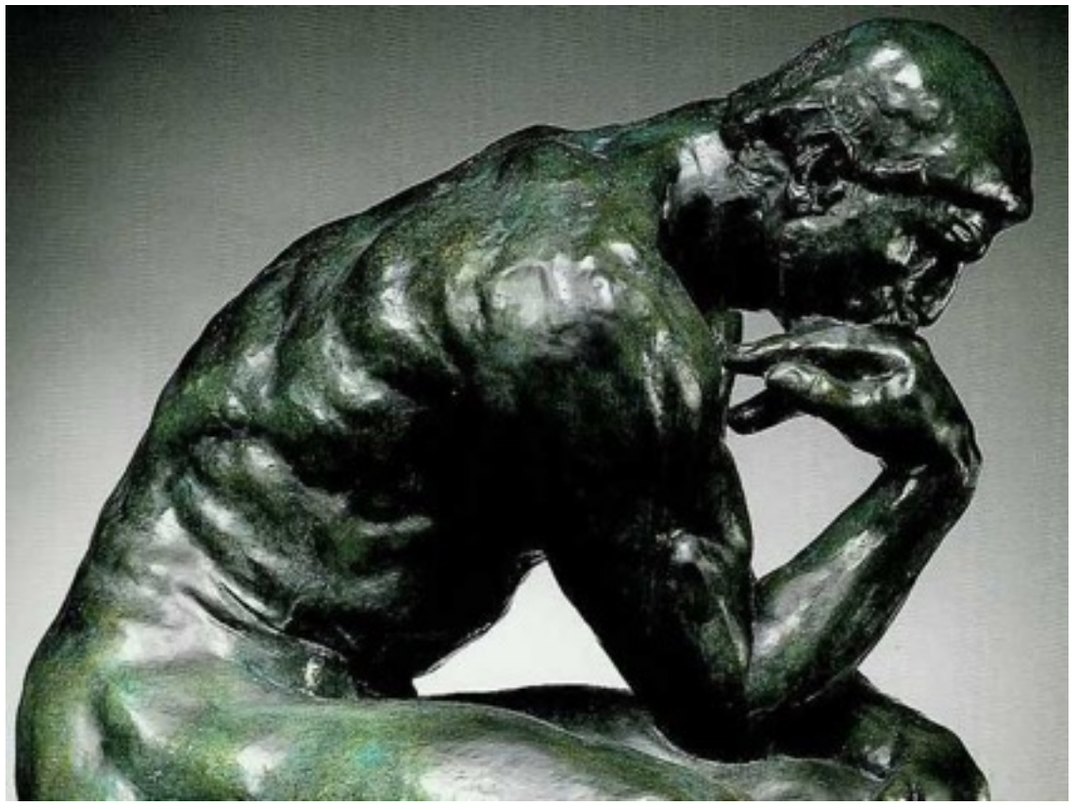


~~Why~~ Data Citation is Important ~~?~~



(euphemism)

The practice of citing data is still not pervasive in scientific publishing



Is data citation known in the digital humanities context?

Is data citation known in the digital libraries context?

What is data citation?

User perspective

- The generation of human- and machine-readable citations should be automatic

- Cited data should be uniquely identified: e.g., DOI

- Citing data should be easy: click, generate, copy and paste

- Setting up and maintaining a citation system should require low (no) effort to data creators/curators

What is data citation?

Computer scientist perspective

- Data is not (always) fixed, it changes

- Persistent identifiers are (only) part of the solution

- Variable granularity (deep citations)

- Automatic generation of citations (yes, but how?)

- Different data types and formats

We started to build automatic techniques

- XML: Rule-based system [Buneman&Silvello, 2010]
- XML: View-based system [Buneman et al., 2016]
- XML: Learning to cite framework [Silvello, 2016]
- Relational DB: View-based model [Davidson et al., 2017]
- Relational DB: Queries as proxies for data [Rauber et al., 2016]
- RDF: Named graphs based method (again views) [Silvello, 2015]
- RDF: View-based (new) method (more general) [2017]

Questions for the audience

- What data do we use in the digital humanities context?

- How do we refer these data?

- Are there connections on data usage in the digital humanities context and in the scientific (e.g., pharmaceutical) context?