Data curation glossary: a survey on

terminology and interdisciplinary perspectives

Anna Maria Tammaro
Professor

University of Parma (Italy)
annamaria.tammaro@unipr.it

Vittore Casarosa
Researcher

ISTI-CNR Pisa (Italy)
casarosa@isti.cnr.it

Abstract

Research is increasingly undertaken by interdisciplinary collaborative teams, which bring expertise in different research areas and in different aspects of the research process. The research workflow is evolving towards a global cyber-infrastructure providing preservation and access to research data and research results, and providing discipline-specific tools. It has to be noted that data infrastructures are stimulated by open science, data science and data sharing policies promoted by the European Commission. In this context, the role and tasks of information professionals and other professionals supporting the research process and the data management are to be re-examined.

The paper explores the first results of a survey investigating, from literature and documentary review, data curators in Europe about the state of art of the job, collecting the main concepts and terms in this domain, as well as collecting interdisciplinary perspectives.

These essential roles have been developed in an ad hoc way, and new roles have been defined, such as:

- *Data Creator: researchers with domain expertise*
- *Data Scientist: information professionals working with data creators, engaging in creative inquiry and analysis*
- *Data Librarian: information professionals specializing in digital stewardship, including the curation, preservation and archiving of data*
- *Data Manager: computer scientists or information technologists responsible for computing facilities, storage, continuing access, and preservation of data*

Where are the actual information professionals? Memory Institutions (traditionally Libraries, Archives and Museums – the LAMs) provide tools and expertise that support research and scholarship and have the institutional structure and many of the resources needed to advance and sustain scholarly workflows. However, can the role of LAMs evolve from one of holders and providers of knowledge resources to one of being an active partner in the research workflow, for what concerns the data creation, curation, access and preservation ?

Keywords: data scientist, open data, data sharing.

Introduction

The research workflow nowadays is evolving towards a global cyber-infrastructure providing preservation and access to research data and research results, and providing discipline-specific tools. Data infrastructures are stimulated in Europe by open science, data science and data sharing policies promoted by the European Commission and more organizations are establishing data science centers of excellence. With these ongoing programs, organizations are fostering standardization, reuse, collaboration, governance and automation within and across data science initiatives. Data infrastructure requires that researchers, professors and students receive adequate support by new professionals skilled in computing and networking, as well as in handling, analysing and storing large amounts of digital content. The authors interest is focused on professional recognition of these new professionals and on the development of appropriate curricula, developing skills which are crucial to ensure effective services to institution staff and students. Training opportunities should be available at all levels and for all communities potentially engaged in research and innovation related activities. However, formal education for the new emerging professions of data scientist, data archivist or "data librarians" hardly exists today.

The reason the authors undertook this survey for “data scientist” is that there is a pressing need for interdisciplinary professionals who understand research data curation and management.

Methodology

The paper explores the results of a literature and documentary review investigating data scientists in Europe, about the state of art of the job and collecting the main concepts and terms in this domain, as well as collecting interdisciplinary perspectives and finally trying to arrange the results in a taxonomic classification.

The first objective is to clarify the competencies and role of these professionals: they have their specialties but also need generalist skills that let them blend the wide range of interdisciplinary methods needed to manage research data.

The second objective to be investigated is: where are the actual information professionals? Memory Institutions (traditionally Libraries, Archives and Museums – the LAMs) provide tools and expertise that support research and scholarship and have the institutional infrastructure and many of the resources needed to sustain scholarly workflows.

The research questions are:

- What competencies make a data scientist?
- How to leverage traditional LAM skill for a career in research data management?

The literature search has been done for job titles like these:

- data modeler
- data curation specialist
- metadata librarian
- data mining specialist

- data visualization specialist

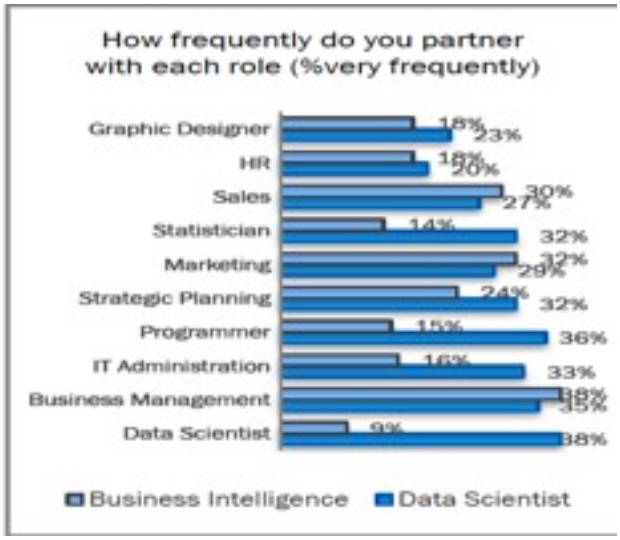
- digital curator
- data manager

- data services librarian

Literature review

Data science is a growing field. The absence of clear boundaries defining data science, and the many job titles defining the new professionals, are evidencing the present confusion and the interdisciplinary function of data management.

Starting in 2008 Jeff Hammerbacher and D.J. Patil have coined the term 'data scientist" building the analytics respectively of Facebook and Linkedin. In 2009, Google Chief Economist Hal Varian predicted that the statisticians will be the next sexy job and according to the Harvard Business Review data scientist job openings are increased 15,000 percent in 2012 over 2011.



The EMC Data Science Community Survey interviewed 497 data scientists and business professionals from around the world, including deliberate intelligence samples in the United States, India, China, the United Kingdom, Germany, and France. They tried to distinguish between business intelligence and data scientist roles. Their results are evidenced in Table 1.

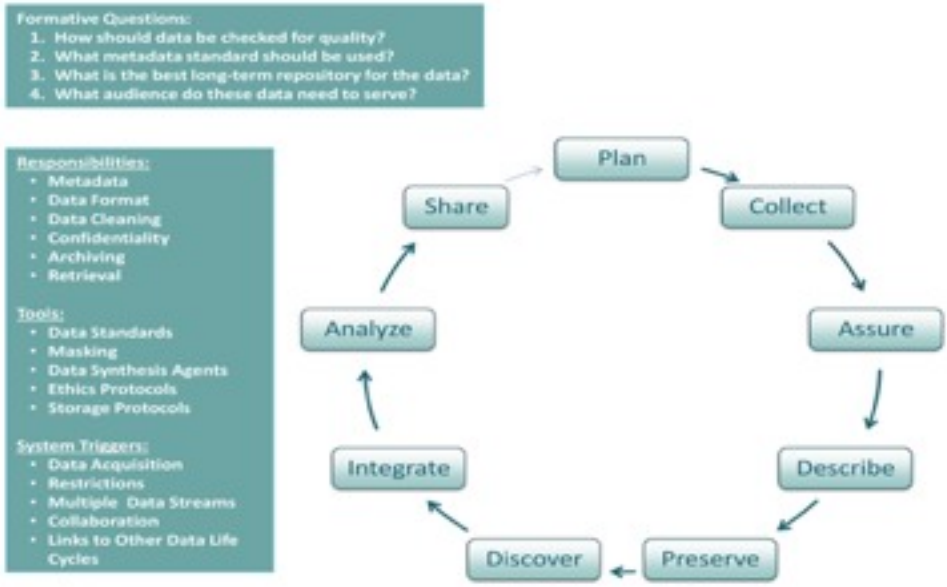
Data Scientists at Work is a collection of interviews with sixteen of the world's most influential and innovative data scientists from across the spectrum of this new profession. Gutierrez asked to his interviewees' earliest data projects how they became data scientists, their discoveries and surprises in working with data, their thoughts on the past, present and future of the profession, their experiences of team collaboration within the organizations to which they contribute, and the deep insights they have developed organizing raw data into objects of commercial, scientific, and educational value for their organizations and clients.

In cultural institutions, the question is: can the role of LAMs evolve from one of holders and providers of knowledge resources to one of being an active partner in the research workflow, for what concerns the data creation, curation, access and preservation? Librarians have always been great at information management and organization. This is a core skill in data science; it

manifests most strongly in the data curation component of the big data problem. The preservation of culture and modern knowledge is becoming more dire every day as electronic records are lost. Digital preservation is a brand new field where stakes (losing our cultural heritage) are high. (Read about DigCCurr for more info.)

The research done by PLOS (2013) has evidenced that successful data science often requires working with teams, including specialists who are able to divide labor and collaborate on the final outcome. As more corporations begin making decisions based on the analysis of data, they need creative curious data scientists who can work collaboratively with programmers, graphic designers, and statisticians to run data experiments (Fig. 1).

Figure 1. The life cycle of data: the steps needed to responsibly collect, record, store, and steward data.



Hartter J, Ryan SJ, MacKenzie CA, Parker JN, et al. (2013) Spatially Explicit Data: Stewardship and Ethical Challenges in Science. PLoS Biol 11(9): e1001634. doi:10.1371/journal.pbio.1001634
http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001634

PLOS | BIOLOGY

Findings

These essential roles have been developed in an ad hoc way, and new roles have been defined, such as:

- Data Creator: researchers with domain expertise
- Data Scientist: information professionals working with data creators, engaging in creative inquiry and analysis
- Data Librarian: information professionals specializing in digital stewardship, including

the curation, preservation and archiving of data

- Data Manager: computer scientists or information technologists responsible for computing facilities, storage, continuing access, and preservation of data

Data are best understood as representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship (Borgman). Inside the black box of data is a plethora of research, technology, and policy issues.

Data practices are local, varying from field to field, individual to individual, and country to country. In research data management, they are embedded in the rapidly changing landscape of scholarly work in the sciences, social sciences, and the humanities. In 2012, Allard et al. have described (Fig. 2) data practices involved in the research flow and in the framework of the Fourth Paradigm of science.

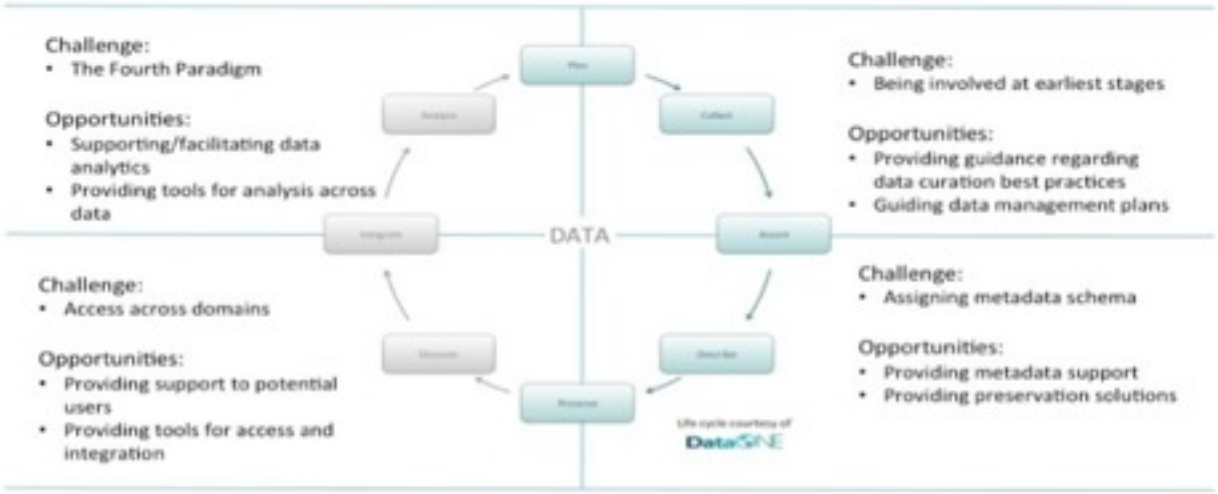


Fig. 2 Data Life Cycle and Information Professionals

Conclusion

Hal Varian's quote of data scientist is:
The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it.

The authors have evidenced that the data scientist has to be defined by its context and the practices of the domain commercial, educational or scientific. In this context, the role and tasks of information professionals and other professionals supporting the research process and the data management are to be re-examined. Doing so will require an additional educational commitment,

and quite possibly less attention to certain traditional topics.

References

Allard, S. (2012) Th Data Life Cycle and Information professionals. Third Annual ASIST Research Data Access and Preservation Summit. New Orleans

Borgman, C. L. (2014). Data Scholarship in the Humanities. Presented at the New Trends in eHumanities, Meertens Institute, eHumanities Group, Amsterdam.

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

Lesk, Michael. Data Curation: Just in Time or Just in Case? New Brunswick, NJ: Rutgers School of Communication and Information, 2011. http://youtu.be/DopPBlOkZ3c.

Varian H. (2009) “Data scientist is the sexiest job in the 21st century" Harvard Business Review