Who is the data curator? Defining a vocabulary

Anna Maria Tammaro[1], Vittore Casarosa[2]

[1] University of Parma
[2] ISTI-CNR

Abstract. In 2016, the IFLA Section Library Theory and Research has (partially) funded the research project “Data curator who is s/he?” to clarify the profile of data curator. The main goal of the project was to define characteristics of roles and responsibilities of data curators in the international and interdisciplinary contexts.

The research questions of the Project were:
R1: How is data curation defined by practitioners / professional working in the field? - R2: What terms are used to describe the roles for professionals in data curation area? - R3: What are primary roles and responsibilities of data curators? - R4: What are educational qualifications and competencies required of data curators?
In this paper we present briefly some of the results related to research questions R1 and R2, namely what terms are used to describe the roles for professionals in data curation area.
Keywords: Data curator, Data curation, Vocabulary.

1 Introduction

Data curation has been defined by the University of Illinois as “the active and ongoing management of data through its life cycle of interest and usefulness to scholarship, science, and education. Data curation activities enable data discovery and retrieval, maintain its quality, add value, and provide for reuse over time, and this new field includes authentication, archiving, management, preservation, retrieval, and representation.”

After the introduction of the term "curation" by the DCC centre, an interdisciplinary professional community has investigated the role and responsibilities of a new or renewed profile called “data curator”. Data curators are specialists who do an important role, whose definition is still in progress and not agreed. The activities of the new profile are also unclear and there is no professional association that has analysed and prepared a competence list.

In 2016, the IFLA Section Library Theory and Research has (partially) funded the research project “Data curator who is s/he?” to clarify the profile of data curator. The main goal of the project was to define characteristics of roles and responsibilities of data curators in the international and interdisciplinary contexts.

The methodology used in the IFLA Project has been based on a mixed-methodology, with content analysis of job announcements for data curators and for

librarians to be involved in Research Data Management activities on one side, and semi-structured interviews with professionals working as data librarians, data curators, or research data managers on the other.

The research questions of the Project were:

- R1: How is data curation defined by practitioners / professional working in the field?
- R2: What terms are used to describe the roles for professionals in data curation area?
- R3: What are primary roles and responsibilities of data curators?
- R4: What are educational qualifications and competencies required of data curators?

The project, funded by the IFLA, was articulated in three main phases:

- Phase I - Literature review and vocabulary.
- Phase II – Content analysis of job announcements offering positions with data curation responsibilities in libraries, archives, and research centers. The job offers were collected from:
  - Job postings from the American Library Association (http://joblist.ala.org/)
  - The community driven site of Code4lib (http://jobs.code4lib.org/jobs/data-curation/)
  - The IASSIST Jobs Repository (http://www.iassistdata.org/resources/jobs/all)
- Phase III – Document analysis of interviews with data curators and questionnaires distributed to data curators.

In this paper we present briefly some of the results related to research questions R1 and R2, namely what terms are used to describe the roles for professionals in data curation area.

2 Term extraction from corpora related to Digital Curation/Digital Curator

One of the main intent of the project was to identify a set of terms (a vocabulary) and possibly an ontology related to Digital Curation, by analysing relevant textual data in the field. Six different corpora were collected, identified with the following nicknames.

- Bibliography Old. Text extracted from abstracts and keywords of papers related to Digital Curation and published up to 2015.
- Bibliography New. Text extracted from abstracts and keywords of papers related to Digital Curation and published in 2016 and 2017.
- Positions/Job offers. Text extracted from job offers and positions, searching for “digital curators”, mostly from academia.
- Questionnaire. Text extracted from a set of questionnaires distributed to professionals already working as data librarians, data curators, or research data managers.

- Interviews. Text extracted from the transcript of interviews with selected respondents of the questionnaires.
- Edison project. Text extracted from deliverables of the Edison project, a European project aiming at defining the skills and the roles of the new profession of Data Scientist.

The system used to extract relevant terms from the corpora, more generally called "key phrases" is the Keyphrase Digger (KD, see http://dh.fbk.eu/technologies/kd), developed at the Fondazione Bruno Kessler (FBK, see http://ict.fbk.eu/). KD scans a given corpus, and computes the "scores" of candidate key phrases, based on term frequency measures and linguistic syntactic information (Part of Speech patterns). It then return the key phrases in descending order of their score. The Keyphrase Digger has three main parameters:

- n, the number of key phrases to be returned, which was set to 50 for all corpora except the Interviews. In this case, given the highly unstructured and colloquial text, n was set to 80 in order to try and capture more relevant key phrases.
- p, which gives a boost to more specific key-concepts (ie. multi-token expressions). Depending on the value of p there will be more or less multi-token expressions in the result. It can have values NO, WEAK, MEDIUM, STRONG.
- m, which indicated the maximum number of words that can be used in the multi-token expressions.

In order to try and extract the maximum amount of information, KD was run on each corpus several times, with different values of the p and m parameters. A first series of five runs was done, with p=WEAK and m (the maximum number of words in a key phrase) going from 1 to 5. Then the results of the five runs were merged into a single list eliminating duplicates and terms clearly not related to Digital Curation (based on subjective judgement). It has to be noted here that KD does not have any "semantic" knowledge, and the key phrases identified are based only on frequencies and syntactic information. Therefore the returned list may contain "key phrases" not related at all with Digital Curation, even if the system has a stop word list to eliminate the most common words.

The same process of five runs was applied to the same corpus with p=STRONG, obtaining a second list of relevant key phrases. The two lists were finally merged into a single one, eliminating duplicate terms and obtaining the set of terms related to Digital Curation for that corpus.

After having extracted the set of key phrases form the six different corpora, it was apparent that there was a minimal overlap between any two pairs of corpora. To make this observation more precise, we generated the intersection of the 15 possible pairs of corpora, obtaining the key phrases in common between any two sets. In Appendix A there is the table with all the overlapping terms. The results confirmed the initial observation, as the "intersection sets" go from a minimum of 3 elements for the pair Interviews/Edison to a maximum of 18 elements for the pair Bibliography New/Bibliography Old.

This result may not be not surprising, if we consider that the texts in the corpora are coming from different communities, and we assume that each community may have its own terminology. What remains to be understood (possibly in a continuation of the project) is whether the differences are just a matter of "terminology", i.e. different communities use different terms to indicate (more or less) the same set of concepts, or is rather a difference in the set of concepts related to Digital Curation, assuming that the "relevance" of a concept (and therefore its appearance in the final results of KD) is actually depending on the community using it.

3 Data curation concepts

To start trying to understand if and to what extent the overlapping in terminology between the different corpora, we have matched some of the common terms extracted from the corpora with the definitions found in Wikidata. We have also started building a table of "related terms" in order to arrive to a more complete and articulated taxonomy in the field of Data Curation. The initial preliminary results are shown below.

- Data curation (Q15088675)
  work performed to ensure meaningful and enduring access to data
- Data management (Q1149776)
  all disciplines related to managing data as a valuable resource
- Digital curation (Q5276060)
  selection, preservation, maintenance, collection and archiving of digital assets
- Digital Preservation (Q632897)
  formal endeavor to ensure that digital information of continuing value remains accessible and usable
- Preservation (Q1479406)
  maintenance of objects as closely as possible to their original condition, also called conservation
- Research Data (Q15809982)
  collection of facts produced through systematic inquiry
- Research Data Management RDM (Q30089794)
  activities around the life cycle of research-related data

Term	Definition	Related Term	Code
Research Data Management (RDM)	Activities around the life cycle of research-related data	Research Data: collection of facts produced through systematic inquiry (Q15809982)	(Q30089794)
Data curation	Work performed to ensure meaningful and enduring access to data	Digital curation: selection, preservation, maintenance, collection and archiving of digital assets (Q5276060)	(Q15088675)
Data management	All disciplines related to managing data as a valuable resource	Data Management Plan (Q17085509)	(Q1149776)
Digital Preservation	Formal endeavor to ensure that digital information of continuing value remains accessible and usable	Preservation: maintenance of objects as closely as possible to their original condition also called conservation (Q1479406)	(Q632897)
Data Science	Interdisciplinary field about processes and systems to extract knowledge or insights from data	Data Scientist: a person studying and working with data (Q29169143)	(Q2374463)

Appendix A

	Bybliography Old	Bibliography New	Positions/Job Offers	Questionnaire	Interviews	Edison
Bibliogr. Old (60 key phrases)		curation//data curation//data curation education//data curation profiles//data management//data quality//digital curation//digital preservation//escience professionals//large scale information management problems//metadata//plurality of curation roles//preservation//repositories//research data//research data curation//research data management	curation//data curation//data management//digital curation//digital preservation//preservation//research data//research data curation//research data management	curation//data curation//data management//data quality//data sharing//data stewardship//digital curation//digital preservation//preservation//repositories//research data//research data management	area of data//curation//data curation//data management//data management plan//digital curation//metadata//research data//research data management	curation//data curation//data management//data quality
Bibliogr. New (69 key phrases)	curation//data curation//data curation education//data curation profiles//data management//data quality//digital cura-		curation//data curation//data management//data management planning//digital curation//digital preserva-	curation//data curation//data management//data management planning//data quality//digital curation//digital preserva-	curation//curators//data curation//data management//digital curation//metadata//research data//research data management	curation//data curation//data management//data

	Bybliography Old	Bibliography New	Positions/Job Offers	Questionnaire	Interviews	Edison
	tion//digital preservation//escience professionals//large scale information management problems//metadata//plurality of curation roles//preservation//repositories//research data//research data curation//research data management		tion//preservation//research data//research data curation//research data management	tion//information management//preservation//repositories//research data//research data management//research data services		quality//data science
Positions (40 key phrases)	curation//data curation//data management//digital curation//digital preservation//preservation//research data//research data curation//research data management	curation//data curation//data management//data management planning//digital curation//digital preservation//preservation//research data//research data curation//research data management		curation//data curation//data management//data management planning//digital curation//digital preservation//preservation//research data//research data curation//research data management	curation//data curation//data management//digital curation//research data//research data management//scholarly communication	curation//data curation//data management
Question. (56 key phrases)	curation//data curation//data management//data quality//data sharing//data stewardship//digital	curation//data curation//data management//data management planning//data quality//digital cura-	curation//data curation//data management//data management planning//digital curation//digital pre-		curation//data curation//data management//digital curation//research data//research data mana-	curation//data curation//data manage-

	Bybliography Old	Bibliography New	Positions/Job Offers	Questionnaire	Interviews	Edison
	curation//digital preservation//preservation//repositories//research data//research data management	tion//digital preservation//information management//preservation//repositories//research data//research data management//research data services	servation//preservation//research data//research data management		gement	ment//data quality
Interviews (34 key phrases)	area of data//curation//data curation//data management//data management plan//digital curation//metadata//research data//research data management	curation//curators//data curation//data management//digital curation//metadata//research data//research data management	curation//data curation//data management//digital curation//research data//research data management//scholarly communication	curation//data curation//data management//digital curation//research data//research data management		curation//data curation//data management
Edison (25 key phrases)	curation//data curation//data management//data quality	curation//data curation//data management//data quality//data science	curation//data curation//data management	curation//data curation//data management//data quality	curation//data curation//data management	