Data Librarian, the Steward

Nicolaie Constantinescu
Information Architect
Kosson Community (www.kosson.ro)
E-mail kosson@gmail.ro

Background. *By the 2020 mark, The European Union should have a functional Digital Single Market. One of the policies sustaining the efforts in such endeavour is building a European Data Economy (European Commission 2017a). The aim is to build a* common European data space *a transformative space even for the information science specialists.*
Objectives. *The study is investigative in nature aiming to extract the traits out of the trends leading to a possible specialisation having a broad scope in searching possible career paths for librarians in data science.*
Results. *A graphical explanatory framework indicating requirements, skills and competencies needed by the librarians to achieve new cross-disciplinary engagements with patrons and the research environment.*
Methods. *Because this study is put in the context of Open Science, and particularly addressing the European Union space, the late developments reflected in the policies and initiatives were taken under scrutiny coupled with relevant European Commission project deliverables. Other studies concentrating on the new roles of the librarians pertaining to* data librarian *were consulted, and also the body of information gathered around the Open Science Cloud. To couple existing studies' conclusions with the demand in the librarianship, a body of job listings were investigated to find common traits* data librarian *job descriptions are exposing. To complete a picture, some training facilities were taken into the mix.*
Debate. *Out of many possible roles envisaged for the librarians, there are a few which trigger some focus on the future skills needed to be acquired. These possible scenarios involve data management planning guidance, data stewardship and curation, and data visualisation. Some or all of them might lead to a deep transformation of the librarianship as a craft we were used to up to the Open Science rising tide. The article invites all the information specialists to look into what are the needs to shape up or even to rehash the careers of library and information science specialists.*

Keywords: *data librarians; data stewardship; skills set; FAIR data; Open Science*

1. Introduction - speaking data

Everyday interaction with computers ensue acts of data input into a system, no matter what is the nature of interaction. Libraries and their staff are part of the new data ecosystem due to the continuous manipulation of catalogue data, involvement in digitisation projects, curation of digital assets or guiding the users through information products, and data silos provisioned by the big publishers´ services. Metadata is the object of manipulation in the world of libraries, data nonetheless, and the process of acquiring the necessary skills goes back in time since the beginning of informatisation of library services. In the meantime the growth of data volumes and importance it gained in evaluation, re-evaluation and reproducibility lead in a silent upgrading for all the actors, libraries´ staff included. As soon as the interest in data exploitation rose, the libraries needed to find stable paths leading to better articulation of responses endowing its staff with new skills.

For the purpose of setting the stage right from the start, we do need to put some meaningful milestones along the path of our inquiries. Because we will explore quite a few documents, initiatives, policy documents and job adds, it is wise to retain some definitions beforehand (Merriam-Webster Dictionary 2019):

Revista Română de Biblioteconomie şi Ştiinţa Informării = Romanian Journal of Library and Information Science
ISSN 2559-5490, ISSN-L 1841-1940 · Volume 14 Issue 4 2018 pp. 113-121
https://doi.org/10.26660/rrbsi.2018.14.4.113

- *Curation* - "organizing and maintaining collections";
- *Stewardship* - "conducting, supervising, or managing of something"; "In recent years, the long-established <<management>> sense of *stewardship* has evolved a positive meaning, <<careful and responsible management>>;
- *Management* - "conducting or supervising of something".

Out of the definitions and considering all practical aspects involved with data, it might be possible to infer that the state of data could be regarded under stewardship when it is being curated. We have another relevant definition to abide to (Council on Library and Information Resources 2019):

"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".

2. Setting up the stage

2.1 Historical milestones

The traits of the data librarian is closely linked to the data archives beginnings, a late arrival of the sixth decade of the last century. In 1967, Great Britain decides to establish UK Data Archive focusing on managing data sets arriving from social sciences (UK Data Archive 2017).

Because on the European level the numbers rose, in 1976 the Consortium of European Social Science Data Archives (CESSDA ERIC) was set. The function of this body was to establish a common management point for a federated structure of services in the field of social sciences (Consortium of European Social Science Data Archives 2018). Research and training of the actors in this newly established field of data management is coupled with the skills enabling complex qualifications demands entailed by the labours they are exposed to (Consortium of European Social Science Data Archives 2018). This is one of the multiple concurrent streams leading to the data multiverse of today. We are to recognize as the biggest catalyst, the appearance thirty years back of the World Wide Web as means to sharing research data. Soon enough research institutions explored and envisaged ways to exploit better the Massive Data (National Research Council 1997) already at hand in the advent of what is today Big Data.

At the beginning of the first decade of the new millennium, the interest on data shifts gears caused by the volume and importance, and also the potential benefits coming from exploitation. In 2007, *OECD Principles and Guidelines for Access to Research Data from Public Funding* based on the objectives and principles of the Paris Declaration in 2004, amended in 2006, set a path fulfilled later on by the outcomes of FAIR data principles (Organisation for Economic Co-Operation and Economic Development 2007).

In 2009, the European Commission embarked on a policy concerning Open Access and preservation of the research output. It started out immediately after Recession *Preparing Europe for a New Renaissance* stating a *paradigm shift* coupled with the need to *Riding the Wave* of data (European Commission 2010), and later on concluded with the ascension of Open Science on the firmament having libraries right from the start as main actors. *Riding the Wave* set the stage for new specialisations in libraries exposing an important question: *How can we foster the training of more data scientists and data librarians, as important professions in their own right?*

In 2012, The European Commission issued a *Recommendation on access to and preservation of scientific information* (2013a), with a reversion in 2018 (2018b). The community reacted to the new directions looking into the matters at hand (European Commission 2013b). An important moment shaped the policies concerning Open Access and data management in respect - *Joint Declaration of Data Citation Principles* from FORCE11 (Martone 2014). In 2014, the FAIR principles start to shape up arriving to the final form in march 2016 (Wilkinson et al. 2016). FAIR principles may be considered an invitation to the scholarly community to reconsider the role of the scientific data in the search for re-use.

2.2 A demand for skills in continuous dynamics

Since the arrival of networked communication, the librarians got involved with data community because by themselves were producers and managers of data, let that be from the academic corridors or cultural heritage digital representations. The community of practice got into contact with data since the times of electronic catalogues and remote search through Z39.50 information retrieval protocol for bibliographic data protocol. The body of experience gained working already in the field of data (Ford 2013), led to gradually accruing skills day by day or on a project bases, or sometimes via specialized training programmes for librarians.

Little by little, besides continuing traditional services, the libraries become the electronic access to the Internet, and doing so, to the specialized research databases or services actioned via Application Programming Interfaces (APIs). Actually, the library becomes a mining ground for resources, from the personal physical space carved from a soaked informational context, to digital resources to which the library became a transactional actor and a gatekeeper. The key term to be observed is *the resource*, which by the laws governing the communication protocols, is a representation identifiable through a Uniform Resource Identifier (Tim Berners-Lee et al. 2004). Librarians silently became managers of resources in a constant need to be better organized, cared for, and kept safe.

With the advent of World Wide Web, in Europe, the information science professionals started to look into the present and future roles and career paths - see, for example, *European Curriculum Reflections on Library and Information Science Education* (Kajberg & Lørring 2005). At that time, data for most of the librarians, besides the catalogue and bibliographic standards, were actually coming from digitisation projects as digital assets and their attached metadata. New skills were needed and a good picture of those times may be seen in *The skills, role and career structure of data scientists and curators: an assessment of current practice and future needs* (Swan & Brown 2008).

The rhythm picked up by the time a reality needed to be solved; Big Data with its peaks in research data. The European Commission embarked on a path of changing how research should be conducted in the new European Research Area starting with investigations on scientific publication markets, and arriving to the launch of the European Open Science Cloud. There is a deep interest in bringing to fruition the exploits of the Big Data through Machine Learning, and ultimately with Artificial Intelligence.

3. Traits out of needs

On European and international level, managing research data or cultural heritage sets, is a continuous business closely related to those who are the practitioners. The business of data is evolving threefold: producers (research and cultural heritage), managers (private bodies, consortia, governments), and consumers (researchers, business, media, general public). The managers needs tooling, resources and qualified personnel.

There is an increasing pressure to reveal instruments and data able researchers and librarians to cope with the increase in numbers of scientific articles accompanied in a growing trend by its respective data sets. Researchers are not able any more to cover the entirety of the research outputs arriving from the global scale. In this context opening FAIR data is a must following Open Access, an already established publishing model if not a scientific conduct in itself enveloped in what is called Open Science.

3.1 Data librarian - the steward

Definitions are handy for getting an upper view on the state of the art, and *data librarians* is no exception.

One of the most complete definitions is the one used by the Research Data Alliance through the Term Definition Tool: *"data librarians are professional library staff engaged in managing research data, using research data as a resource, or supporting researchers in these activities"* (Data Foundation and Terminology 2017). The CASRAI Dictionary (2015) takes this definition, but it emphasizes the expertise in working with the data: *"data experts who have a librarian background"*. CASRAI marks another issue in the effort of setting the profile of a would be data librarian. That is the continuous *"overlap between data librarians, data managers, and data stewards"*.

We should be aware at all times that the data librarian is savvy with the technical parts of data management, being at all time involved with data repositories management and development. This permanent association leads to a technical skill set needed from the start that would be in a continuous updating with the latest as the daily tasks demands it.

Most of the new skills required like data citation (an organic elongation of bibliographic services), data description (metadata and descriptor handling), or data identification (Digital Object Identifiers), are in fact abilities already attached to every librarian's tool belt. On these the technological layer needs to be updated and raised to new levels.

3.2 Roles for librarians

Seeking the paths leading to an established profile for the data librarian produced a number of research articles already. One of the recent ones, *Defining data librarianship: a survey of competencies, skills, and training* (Federer 2018) gathered data about librarians involved in data curation already. One of the important implications is that *"many libraries have never had a specialized data librarian job"* and therefore *"have unclear expectations of the skills and knowledge that the successful job candidate should have"*. Often, the librarians are *"in the position of implementing new and previously untested services"*. Although these conclusions relate to an environment close to the information technologies, a data librarian could start using the existing tools developed by the data communities, tackling the challenges on a moderate level. For example, in the case of working with data, one may employ tools like ISA framework (http://isa-tools.org).

Nonetheless, the technical skill set needs to be updated and infused with new abilities and knowledge aimed on data curation. A short overview at the Digital Curation Centre *Curation Lifecycle Model* (2019a) and the *Curation Reference Manual* (2019b) helps spotting needs for a data librarian in terms of skills needed.

Another important study (Burton & Lyon 2017) highlight a real skill gap: *"many librarians lack the technical skills to be effective in a data rich research environment"*. Interestingly enough, there is an opinion expressed by a librarian: "I learn new skills, but I still need to do my old job" (Burton & Lyon 2017) that goes in contrast with another present in Federer's (2018): "I feel less and less like a <<librarian>> and more and more like … something else". This indicates an emergent role, one that needs support if it is to become a career path.

Although an important part of the literature has defined the object of activities being research data management, looking back to the Curation Lifecycle Model, we should be able to see that curating the data is inferred.

To achieve a paradigm change through Open Science, the proper set of skills had to be looked into, and with this need in focus, some documents explored the tasks a future librarian might have in a data infused work context.

One important study (Schmidt & Shearer 2016) explored the competencies librarians have in their managerial role. These competencies are actually inscribed in the library's scope provided in three areas: access to data, awareness and support for managing data, and managing a data collection. The last one falls under the curation life-cycle model' steps. The core competencies imply a vast knowledge of the field of data, and also of the policies and project management. Most of these core competencies are technical in nature. A list of possible jobs is offered as well showing that some of them are more broad in data knowledge scope, some are more targeted.

3.3 Training and lifelong learning

The gap is in the scope of several international programmes like Data Scientist Training for Librarians (DST4L), or Library Carpentry. For a more comprehensive list, National Network of Libraries of Medicine provides valuable references (National Network of Libraries of Medicine 2019).

Librarians who will be engaged with the researchers active in European projects are expected to be able to give support and insight on all matters related to data curation. Complementary to the librarian skill set there should be the researcher's. A lot of training programs address the researcher in their declared mission to augment her/his data managing capabilities. Instrumental for understanding the effort would be a look into *Providing researchers with the skills and competencies they need to practice Open Science* (European Commission 2017b): *"particular attention needs to be paid to developing and growing the cohort of information professionals (which can include librarians, data scientists, data stewards and others)".*

The accent falls on the training of data stewards, and further down the path, the Report even names a new breed of librarians in the context of practising Open Science: *the Open Science librarians*. The interesting bit on the two is a profound acknowledgement for training, and for the resource allocation it demands.

3.4 A dynamic job market

Data librarians need to emerge as a fully fledged position within the libraries because universities and even public libraries are investing in computing infrastructures and access services for researchers and patrons alike.

For the benefit of this study, a set of job ads collected from the website of the International Association for Social Science Information Services and Technology (2018) was analyzed. Out of the many job adds provided, only those containing data librarian were hand-picked taking into consideration that title of the job is enough to extract the common traits for a possible career path. The sample under analysis might be considered biased due to its USA and UK coverage. For a global view, there is a need for an extended sample in the future.

From adds descriptions were isolated seven sets for the skills and abilities needed:

- in what is collaborating with others;
- what she/he organizes;
- in what is involved;
- for what is responsible of and what she/he elaborates;
- to what she/he contributes to;
- fuzzy (a set of skill dependant by the specificities of the institution);
- qualifications and specific requirements.

It is easy to see that the skill sets a librarian in charge also with data has, acts more like a middleware in the economy of data for the institutions. It has a liaison trait (a live human bridge with the staff members and researchers), it has to be productive in terms of services (from development, to documentation, to writing on the blog and social media, to assistance), it needs to be familiar with funding opportunities and how to apply, and more than that, to be technical savvy meaning at times mastering several programming languages.

The following information tends to become the norm for most of the post ads. Some of the original descriptions and attributes were truncated in the dataset attached for the economy of the text.

For answering to such demands, the information specialists need a brand new training framework. Realisation of such framework could be an extension of the existing ones like the one FOSTER provides (FOSTER 2019), or starting on national levels in accord with local specificities, but echoing the European level standards. One possible solution is accessing Digital Education Action Plan - Action 5 Open Science Skills (European Commission 2018a).

4. A possible profile

The first mandate of the data librarian is to ensure the safe keeping of data. In research environment, keeping data safe, means actually designing and implementing Data Management Plans (DMP). A data librarian should be able to understand how to gather the information, and being knowledgeable of the field specificities with regards to the types of data, and able to design a viable DMP. The community already designed tools in aid like *DMPTool* (California Digital Library - University of California Curation Center 2019) or *DMPonline*, another instance of DMPTool.

As much as it was gathered from the job ads, there is to be observed a need for a person that not only is a skilled technologist with data, but also a communicator, and most of all, an innovator. These requirements are the proper answer for future challenges.

Labouring through the documents there is a distinguishable profile that could lend itself to further scrutiny.

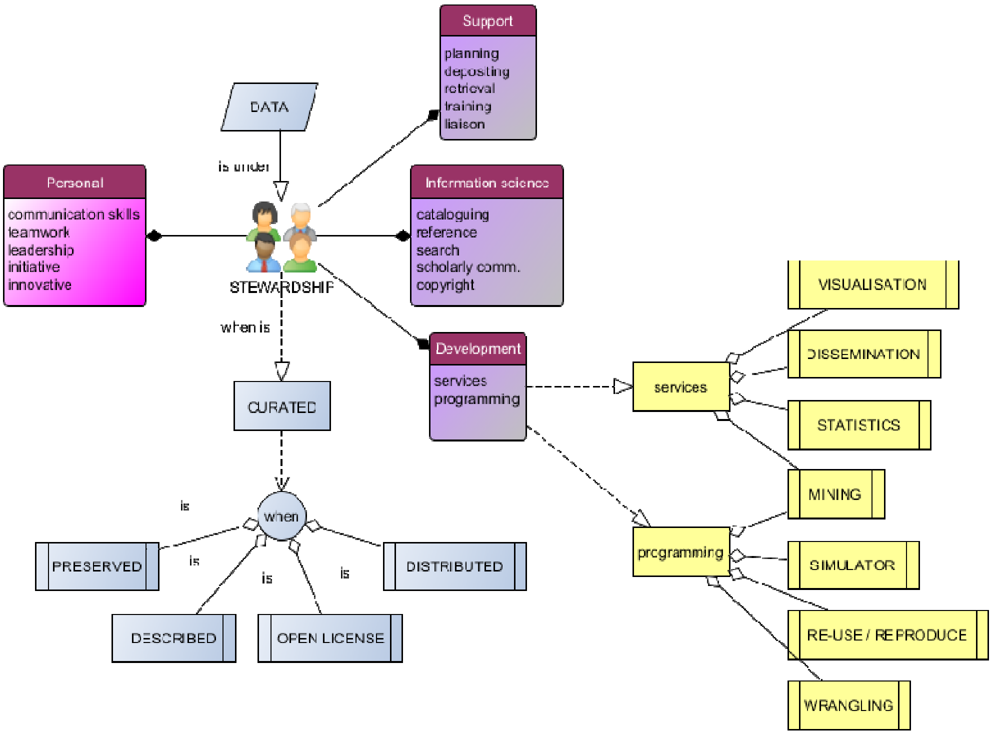

Figure 1. *A possible picture of skills, and provided services*

6. Debate

The diagram proposed is not impinging a fully hydrated model for answering any possible scenario, but it gives a possibility to assess the curation life-cycle of data as states on which stewardship is manifested.

Data is travelling the life cycle, and at every change of state, there is a need to have a suitable response from those tending to it. This response in in fact a set of skills, may that be soft or technology driven. Librarians are still the first choice in managing data that needs a good contextualisation. Libraries are the main actors in evolving the global knowledge graph through the renewed services they provide.

7. Conclusions

Librarians in general and academic librarians in particular need a new long life learning framework. Librarians need to re-skill in data curation as well in order to become a functional part of the data ecosystem (Dickersin 2019). Data curation as practice environment requires deep technical skills.

The data librarian emerging profile needs to step out as a real career path as it is not a new profile, but an evolved one having now a distinct contour. There is an observable overload in all the job adds analysed and that is the future candidate to cover all dimensions as match as possible, which in reality is rarely to find such embodiment. This state has the negative effect on scarring the professional who are actually doing the job already.

Open Science spurs the incentives for skill augmentation enough to see the rise of the data librarian, the steward of research and cultural heritage.

References

Berners-Lee, T. et al. (2004) *Architecture of the World Wide Web*, available: https://www.w3.org/TR/webarch/ [accessed 12 February 2019].

Burton, M. and Lyon, L. (2017) Data Science in Libraries, *Bulletin of the Association for Information Science and Technology*, 43(4) pp. 33-35, available: https://doi.org/10.1002/bul2.2017.1720430409 [accessed 10 March 2019].

California Digital Library - University of California Curation Center (2019) *DMPTool*, available: https://dmptool.org/ [accessed 14 March 2019].

CASRAI Dictionary (2015) *Data librarian*, available: https://dictionary.casrai.org/Data_librarian [accessed 14 February 2019].

Consortium of European Social Science Data Archives (CESSDA) (2017) *The Expert Tour Guide on Data Management*, available: http://bit.ly/rrbsi42018g [accessed 2 February 2019].

Consortium of European Social Science Data Archives (CESSDA) *About - CESSDA*, available: https://www.cessda.eu/About [accessed 23 February 2019].

Council on Library and Information Resources (CLIR) (2019) *Data Curation*, available: https://www.clir.org/initiatives-partnerships/data-curation [accessed 2 March 2019].

Data Foundation and Terminology (DFT) (2017) *Edit RDA: Data librarian*, available: http://bit.ly/rrbsi42018h [accessed 2 March 2019].

Dickersin, K. (2019) Lost Knowledge: Open Science is One Solution to Hidden Data, *Hopkins Bloomberg Public Health Magazine*, available: https://magazine.jhsph.edu/2019/lost-knowledge-open-science-one-solution-hidden-data [accessed 23 March 2019].

Digital Curation Centre (DCC) (2019a) *Curation Lifecycle Model*, available: http://www.dcc.ac.uk/resources/curation-lifecycle-model [accessed 23 February 2019].

Digital Curation Centre (DCC) (2019b) *Curation Reference Manual*, available: http://www.dcc.ac.uk/resources/curation-reference-manual [accessed 13 February 2019].

European Comission (2010) Digital Agenda to unlock the full value of scientific data: High-Level Group presents report, *Digital Single Market*, available: http://bit.ly/rrbsi42018i [accessed 6 January 2019].

European Comission (2013a) Commission Recommendation on access to and preservation of scientific information, *Digital Single Market*, available: http://bit.ly/rrbsi42018j [accessed 6 January 2019].

European Comission (2013b) EC Public consultation on open research data on 2 July in Brussels, *Digital Single Market*, available: http://bit.ly/rrbsi42018k [accessed 6 January 2019].

European Comission (2017a) Building a European data economy, *Digital Single Market*, available: http://bit.ly/rrbsi42018l [accessed 2 December 2018].

European Comission (2017b) *Providing researchers with the skills and competencies they need to practise Open Science: Open Science Skills Working Group Report*, available: http://bit.ly/rrbsi42018m [accessed 13 February 2019].

European Comission (2018a) Digital Education Action Plan - Action 5 Open Science Skills: teaching, learning and assessing open science skills, *Education and training*, available: http://bit.ly/rrbsi42018n [accessed 14 March 2019].

European Comission (2018b) Recommendation on access to and preservation of Scientific Information, *Digital Single Market*, available: http://bit.ly/rrbsi42018o [accessed 6 January 2019].

European Union (2010) *Riding the wave: How Europe can gain from the rising tide of scientific data available*, available: http://bit.ly/rrbsi42018p [accessed 21 December 2018].

Federer, L. (2018) Defining data librarianship: a survey of competencies, skills, and training, *Journal of the Medical Library Association*, 106(3) pp. 294-303, available: http://jmla.pitt.edu/ojs/jmla/article/view/306 [accessed 13 February 2019].

Ford, J. (2013) How to Beat Bibliographic Data into Submission, pt. 1, *Data Scientist Training for Librarians*, June 4, available: http://bit.ly/rrbsi42018r [accessed 15 February 2019].

FOSTER (2019) *Courses*, available: https://www.fosteropenscience.eu/courses [accessed 13 February 2019].

International Association for Social Science Information Services and Technology (2018) *Job postings*, available: https://iassistdata.org/resources/jobs/all [accessed 13 February 2019].

Kajberg, L. and Lørring, L. (eds.) (2005) *European Curriculum Reflections on Library and Information Science Education*, Copenhagen: The Royal School of Library and Information Science.

LIBRIS (s.a.) *Technical and format information*, available: http://librishelp.libris.kb.se/help/tech_eng.jsp?open=tech [accessed 13 February 2019].

Liscouski, J. (1997) 2. The Data Librarian: introducing the Data Librarian, *The Journal of automatic chemistry*, 19(6) pp. 199-204, available: https://doi.org/10.1155/s1463924697000242.

Martone, M. (ed.) (2014) *Data Citation Synthesis Group: Joint Declaration of Data Citation Principles*, San Diego CA: FORCE11, available: https://doi.org/10.25490/a97f-egyk [accessed 8 January 2019].

Merriam-Webster Dictionary (2019), available: https://www.merriam-webster.com/dictionary [accessed 24 January 2019].

National Network of Libraries of Medicine (2019) *Courses and Workshops*, available: https://nnlm.gov/data/courses-and-workshops [accessed 24 January 2019].

National Research Council (1997) *Bits of Power: Issues in Global Access to Scientific Data*,The National Academies Press: Washington, DC, available: https://doi.org/10.17226/5504.

Organisation for Economic Co-Operation and Economic Development (OECD) (2007) *OECD Principles and Guidelines for Access to Research Data from Public Funding*, OECD Publications: Paris, available: http://www.oecd.org/sti/inno/38500813.pdf [accessed 24 January 2019].

Schmidt, B. and Shearer, K. (2016) *Librarians' Competencies Profile for Research Data Management*, available: http://bit.ly/42018s [accessed 13 February 2019].

Swan, A. and Brown, S. (2008) *The Skills, Role and Career Structures of Data Scientists and Curators: An Assessment of Current Practice and Future Needs*, available: https://eprints.soton.ac.uk/266675/ [accessed 13 February 2019].

UK Data Archive (2017) *Across The Decades*, available: https://data-archive.ac.uk/about/archive/decades [accessed 22 November 2018].

Wilkinson, M.D. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data*, 3, 160018, available: https://doi.org/10.1038/sdata.2016.18 [accessed 22 November 2018].