Feature Article

On the importance of being a data-savvy librarian

Annarita Barbaro
Istituto Superiore di Sanità, Rome, Italy

Abstract
In the last years a sense of the importance of making all data related to scientific research open, shareable and therefore reusable, has spread in all the areas of scientific research. As this has generated a need to develop policies, infrastructures and services in order to assist researchers in creating, collecting, manipulating, analysing, storing and preserving datasets, this is a unique opportunity for libraries to play an active role in the research process. Librarians, along with their traditional skills, such as their ability to retrieve and organize information, consult and teach, should develop an understanding of the technical aspects of data management, developing new skills and capabilities.

Key words: librarians; data curation.

Introduction

Along with the open access movement, a sense of the importance of making all data related to scientific research open, shareable and therefore reusable, has spread in all the areas of scientific research. As with open access to journal articles, research funders are nowadays playing an important role in mandating the sharing of research data by introducing policies (or tightening existing ones) requiring the deposit and sharing of the data produced by the research they are funding. In order to ensure that research data are well-managed in the present, and prepared for preservation and usability in the future, some funders require also a Data Management Plan (DMP), a formal document outlining how the data resulting from the funded project will be handled both during the research, and after the completion of the project. The National Institutes of Health, for example, has mandated since 2003 that research data arising from projects they fund be made openly available, and the US National Science Foundation implemented a similar requirement in 2011. One of the biggest world funders, the European Union, has launched the Open Research Data Pilot – as part of EU Framework Programme for Research and Innovation Horizon 2020 (2014-2020) – to make the research data generated by selected Horizon 2020 projects accessible with as few restrictions as possible (1, 2). A policy of making data sharing a condition for publication has been enhanced also by some scholarly journals such as Nature (3) or all the journals published by the Public Library of Science (4).

While nowadays the benefits of making research data available are widely accepted, researchers are not always strongly motivated to share their own data because of concerns about intellectual property, confidentiality, or lack of academic credit, or fear that their research might be scooped or misinterpreted. Furthermore, preservation and planning for future use are often carried out in a localized and unsystematic fashion, usually limited to backup and storage for future reuse by the same research group. Moreover, no coordinated set of practices or even instructions exist to enable the majority of researchers to introduce an effective research data management into their workflow (5).

As these factors generate a need to develop policies, infrastructures and services in order to assist researchers in creating, collecting, manipulating, analysing, storing and preserving datasets, this can be a unique opportunity for libraries to play an active role in the research process. This role can be fulfilled in several ways: for example, libraries can provide consulting services related to research data

Address for correspondence: Annarita Barbaro, Istituto Superiore di Sanità, Viale Regina Elena, 299. 001621, Rome, Italy. E-mail: annarita.barbaro@iss.it

management and curation, can provide the infrastructure, or at least the front end, for data storage and curation, can build a bridge between administrative staff and researchers, and can "clean" data for analytic use (6).

The new role of the librarian in a data driven environment

Given the library's longstanding role in scholarly communication, librarians have a long history of facing changes in technology that reshape their work. The digital age has brought incredible changes in the way information and data are produced, consumed, adapted, and shared, requiring a transformation of resources and services. The NMC Horizon Report: 2015 Library Edition, which identifies and describes the important developments in technology likely to have a large impact over the coming five years in libraries around the globe, underlines the increasing focus on research data management, and that "[l]ibraries now have a responsibility to educate researchers about how to select the proper medium to fit their findings – and to ensure that what is submitted can be stored and published in their databases". This report states also that librarians can be genuinely useful for their technical expertise in areas including the optimization of taxonomies, citing data, and addressing any intellectual property issues (7).

The ACRL (Association of College & Research Libraries) Research Planning and Review Committee also identified library involvement in data curation, including collaboration with their research communities, as one of the 2012 top ten trends in academic libraries. According to the ACRL report, data curation offers opportunities for "finding new ways to communicate the value of the skills librarians already possess and in developing roles that were not previously associated with librarians" (8).

There is a substantial overlap between the areas where researchers face substantial challenges in developing data policies and the areas of expertise already associated with one of the core roles of the research librarian: collecting and arranging information in a way that guarantees that it will be retrievable and usable for a broad range of users in the future. Librarians have competencies that can be useful in supporting researchers during every stage of the research data lifecycle (plan, collect, manage, share, and publish datasets). In detail, they can help the researchers to write Data Management Plans, give them the appropriate information regarding licensing data, and direct them to the right repositories. Moreover, as the researchers prepare for the long-term curation of data by creating identifiers and depositing data in a trustworthy repository, librarians can help with their expertise in this field offering workshops or promoting metadata best practices. The aspect of assigning the appropriate metadata to data sets is an obvious concept for librarians but most of researchers are unaware of their value for an effective data sharing, citation and reuse (5, 9).

As a step in this direction, libraries at research institutions or universities have started with a variety of data awareness services on their websites. The MIT libraries have created a page on data management (http://libraries.mit.edu/data-management/) with guidance on planning, storing and sharing of data, addressing a broad range of technical, administrative and confidentiality issues. The University of Edinburgh data library, funded by Jisc, developed an online course, MANTRA (http://datalib.edina.ac.uk/mantra/), covering the essentials of research data management for doctoral students and other researchers. The course provides also a DIY training kit for librarians (http://datalib.edina.ac.uk/mantra/libtraining.html).

Old and new skills

Broadly speaking, the mentioned skills already fall within the traditional competencies related to the work of a librarian, but in practice working directly with research data also involves a number of practical issues which may be unfamiliar. To be of effective help, librarians need increasingly to become data-savvy themselves and to have a deeper understanding of the research data lifecycle in order to enhance the services they offer.

Data curation is still an emerging field in librarianship so there is still some disagreement on the kinds of skills librarians need to perform this role effectively. One perspective is that the main requirement is a basic familiarity with how various software tools can transform data. For example, a librarian does not need to also be competent as a statistician or a graphic designer but every data

librarian should have a clear understanding of how basic tests of numeric data can be used and to recognize the features of an effective data presentation.

An alternative and more flexible approach is for data librarians to learn how to code (10). This also has the potential to make librarians more effective in other parts of their work. In a blog post titled "Why would a librarian learn to code?", Tom Sykes, a Deputy Librarian at a Cambridge college library, sums up four reasons why librarians should learn how to code: to optimize workflows by automating repetitive tasks involving messy data, to improve usability of library services, to better communicate with IT and software vendors and, last but not least, improve their creativity, which is indispensable for problem solving (11). In a sign of more widespread interest in coding in libraries Library Technology Reports (journal published by the American Libraries Associations) published an issue entitled "Coding for Librarians: Learning by Example", which reports more than fifty interviews with librarians who have written code in the course of their work (12).

Although librarians are broadening their expertise to adapt to this new environment, the scale of the challenge in terms of infrastructure, skills and culture change requires concerted action by a range of stakeholders, and librarians need to collaborate with IT services and researchers along with other key players such as research support offices. Acquiring more technical skills is useful in itself but it is also valuable for improving collaboration with IT staff who have a complementary role in managing databases. Not all librarians are interested in becoming data scientists but they can make a difference by collaborating with experts bringing their competence and knowledge in handling metadata and applying it in the context of data. This collaborative approach to service, which connects library expertise with different stakeholders, can create opportunities to build networks within and beyond the library, to integrate library support into the research process, and to support open access to research data.

Currently there is still no coordinated effort in providing targeted opportunities for professional staff development in this area and most library staff are building their knowledge and skills in research data management on the job, through self-training, or participating to conferences or workshops. However, several initiatives have started to prepare librarians for more open and data-intensive scientific research. Among them, the Harvard-Smithsonian Center for Astrophysics John G. Wolbach Library and the Harvard Library, have developed, two years ago, an experimental course to train librarians to respond to the growing data needs of their communities: the Data Scientist Training for Librarians (DST4L) (13). In this free, hands-on course, librarians were taught about the research data lifecycle and they started to learn the basics of some of the latest tools for extracting, analyzing, storing, and visualizing data. In 2015 the course was hosted in Europe by the Technical University of Denmark Library of Lyngby, Copenhagen (http://www.altbibl.io/dtu/), and I was one of the 40 librarians who were selected to participate.

In this three day course we learned how to use several such tools through hands-on practice: OpenRefine, an open source desktop application for data cleanup and transformation to other formats, GitHub, a web-based collaborative platform for code and content management and review, Python, a programming language. The course also included an introduction to the basics of data visualization.

One of the great things about DST4L is its hands-on approach to data-driven projects. The course is useful also for librarians who don't intend to work directly on these kinds of projects because after working directly with messy, unavailable or difficult-to-access data it is possible to have a more complete vision of the different issues the researchers have to face when working with data. Much of the DST4L course material is currently accessible through a WordPress site along with blog entries, written by participants, with accompanying notes, code, data, and anything else used in the sessions (http://altbibl.io/dst4l/).

Conclusion

Data management is still in its emergent phase and funding agency mandates are relatively new so only a few libraries have already developed strategies to assist their researchers with their data and in creating Data Management Plans. It is already clear that librarians, along with their traditional skills, such as their ability to retrieve and organize

information, consult and teach, should develop an understanding of the technical aspects of data management, developing new skills and capabilities. Experiencing the research data lifecycle firsthand and upgrading to data savvy skills could help librarians improve outreach and services to scientists but could also, at the same time, help them to explore new ideas to improve the workflow at their library, making them more efficient in their existing roles.

Received on 5 February 2016.
Accepted on 15 February 2016.

REFERENCES

1. European Commission. Directorate-General for Research & Innovation. Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020. October 2015. Available from: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf
2. European Commission. Directorate-General for Research & Innovation. Guidelines on Data Management in Horizon 2020. October 2015. Available from: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
3. Nature Publishing Group. Availability of data, material and methods. Available from: http://www.nature.com/authors/policies/availability.html
4. Public Library of Science. Data Availability. Available from: http://journals.plos.org/plosone/s/data-availability
5. MacMillan D. Data Sharing and Discovery: What Librarians Need to Know. J Acad Librariansh . 2014; 40 (5): 541-549
6. Tenopir C, Birch B, Allard S. Academic librarians and research data services: an ACRL White Paper. 2012. Available from: http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf
7. Johnson L, Adams Becker S, Estrada V, Freeman A. NMC Horizon Report: 2015 Library Edition. Austin, Texas: The New Media Consortium. Available from: http://cdn.nmc.org/media/2015-nmc-horizon-report-library-EN.pdf
8. ACRL Research Planning and Review Committee. 2012 Top Ten Trends in Academic Libraries. Coll Res Libr News June 2012; 73 (6): 311–320. Available from: http://crln.acrl.org/content/73/6/311.full.pdf+html
9. Auckland M. Re-skilling for research: An investigation into the role and skills of subject and liaison librarians required to effectively support the evolving information needs of researchers. 2012. London: Research Libraries UK. Available from: http://www.rluk.ac.uk/wp-content/uploads/2014/02/RLUK-Re-skilling.pdf
10. Erdmann C. Teaching Librarians to be Data Scientists. Inform Outlook. 2014; 18 (3): 21-24
11. Sykes T. Why would a librarian need to code. [blog post]. 2016 Jan 24. In: Code and the librarian. Available from: http://codeandthelibrarian.wordpress.com/2016/01/24/why-would-a-librarian-learn-to-code/
12. Coding for Librarians: Learning by Example, Libr Technol Rep. 2015; 51 (3): 1-30
13. Erdmann C. Data Scientist Training for Librarians, Library and Information Services in Astronomy VII: Open Science at the Frontiers of Librarianship ASP Conference Series. Vol. 492: 31-37