LIBER QUARTERLY
The Journal of the Association of European Research Libraries
Vol. 27, no. 1 (2017) 89–96 | e-ISSN: 2213-056X

Book review

Robin Rice and John Southall. *The Data Librarian's Handbook*. London: Facet Publishing, 2016, 192 pp.

Brunella Longo

Information Management Adviser, UK
bl@brunellalongo.co.uk

This is a book correctly defined by the publisher as "an insider's guide to data librarianship". Better then to warn the reader before we start: the following is an outsider's book review. As such it may cause some cognitive discomfort.

Revolving around very well-known aspects of data management in universities contexts, the handbook is not for those who may seek or hope to find a bit of librarians' wisdom while dealing with the craziness or absence of requirements for big data and Internet of Things (IoT) projects. Neither it is likely to satisfy the appetite for new ideas among librarians and researchers or other practitioners who have been dealing with metadata management for electronic resources or with linked data for a while and still do not really know which general theory or practical semantic principles apply to make information visible (Flichy, 2013) in massive, global big data projects. Little is said that could be relevant for data scientists, managers, librarians or data engineers working in corporate information and documentation centres, international or non-governmental organisation or social enterprises (from where comes a huge proportion of research data available in the public domain).

On the contrary, in a certain sense, the book seems written to suggest to these experienced audiences that they themselves could make an effort to answer innumerable theoretical and practical questions on what should be the core focus of data librarianship science as a discipline and as a practice. Who could

Uopen Journals | http://liberquarterly.eu/ | DOI: 10.18352/lq.10206

better answer the question of what is the relationship between data librarianship and the current holistic omnipresent and jelly-like concept of data science? Or, on the contrary, between data librarianship and the invisible results of data engineering in public discourse?

On the other hand, it is fair to say that this "insider's way" to data librarianship is aligned with current general academic goals and plans in this field, in the UK and internationally. In fact, it offers an excellent summa, very well organised, to make librarianship students, volunteers and new professionals aware of what to expect when they that start working in data literacy across the faculties.

The authors, Robin Rice (Edina, University of Edinburgh) and John Southall (Bodleian Libraries, University of Oxford) introduce their arguments with a certainty that will please the youngest and reassure everybody, predicting that the "future of data librarianship lies with academic libraries": these are seen as the secure place supporting learning and knowledge management in academia. It is hard to disagree but for noticing that the strength of the tie with their own organisations and their disciplinary boundaries constitutes also the main weakness of Rice and Southall's approach to the subject.

In fact, the best practices the authors have considered are soundly rooted in experiences known among specialists' networks in the social sciences (ESRC, IASSIST, CESSDA) but they do not even mention the existence of data management approaches, cases, problems or trends in other disciplines or sectors. It is unlikely that such compartmentalised vision of research data will survive in the long term as we already see the best results and more funded initiatives looking for innovations in opposite directions, with multidisciplinary, interdisciplinary and participatory projects having a big policy influence (Von Hippel, 2007).

The first (*Data librarianship: responding to research innovation*) and the second chapters (*What is different about data?*) explain such a backdrop through a concise and convincing excursus into the history of the concept of data collection, considered as the evolution of those treatments of machine readable formats that North American academic libraries have pioneered since the 1960s.

Recognising the "accidental nature of the profession" helps managers and librarians to see a rationale or to identify best practices among the diverse

organisational solutions proposed by different academic institutions. Those solutions are intertwined with organisational and funding models (more special data centres supporting the core research programmes of their audiences in the USA versus more centralised data centres sponsored by national research councils in Europe). There are no firm and standardised solutions yet in respect of topical issues like copyright and intellectual property rights or about the crucial relationship between data and metadata creation and processing. These are seen as "emerging areas of interest" in which the characteristics (volume velocity and variety) of new software applications and datasets production determine not only technical developments but also new conceptual and methodological challenges like processing the "long tail" of data as open sources delivered over the internet, the need of data citation, the call for policies and procedures for data curation. And yet, research libraries have to find solutions and answers acceptable as standard within their target disciplines. The "key take-away points" and "reflective questions" that close the second chapter, asking the reader for instance to identify the typical presumptions about research data in their discipline, say all about the lack of consolidation and maturity of data librarianship as a universal practice, offering at the same time a practical checklist useful as a starting point to write or to discuss a new business brief or a project proposal. It might have been out of scope but perhaps more adherent to the reality of the market if these two initial chapters had included references to practical data mining problems that are more logical and related to design rather than to statistical or computational aspects—such as lack of metadata standards, missing data, mis-categorisation, lack of standards for data pre-processing and other domain-independent properties that assure algorithms' reliability, particularly in heterogeneous applications (ACM DL, 2012).

At this point I put down a few more additional questions for myself: think about those increasingly debated investments or experiments into artificial intelligence and machine learning we are bombarded with in daily news and talk shows. Are they leading to more demand of data curation, quality or librarianship? How does the advice offered by Rice and Southall to the "situational data librarian" stand in respect of the guidelines on management of research data by the Research Council UK (2015) or other academic bodies and working groups around the world?[1] The first rule to solve a wicked problem is to get rid of the weirdness that makes it unmanageable and to redefine it so that it becomes structured and treatable. Savvy enough, also the book stops here with the theoretical reflections on data librarianship and departs

from what could have been an excessively normative or controversial presentation of a discipline still at its conception stage.

From Chapter 3 on, the Data Librarian's Handbook enters the fields of what is currently going on in many academic libraries concentrating on building data collections (Chapter 4) and working across the institution to define research data management services and policies (Chapter 5).

Chapter 3 (*Supporting data literacy*) contains a short paragraph that in less than two pages offers a quite timid proposition on what I would like to see at the centre of the discussions and reflections on data librarianship and data librarians' identity, the title of which is simply "categories of data".

Here the authors mention distinctions that are crucial in the data universe such as those between published statistics and survey data, micro and macro data, panel and cohort data in longitudinal studies, time series, geospatial data. Unfortunately these distinctions—so fundamental in computational social sciences—are seen as functional, framed within the narrative of information and data literacy discourse and are not further elaborated *per se* into what would be a much needed new taxonomy of datasets. Questions that could reopen the theoretical and policy debate are carefully avoided—I mean questions such as at what level of granularity data librarians should and could decide to apply what type of metadata to their data collections, manually or through algorithms? what terminology or content analysis should determine assurance criteria useful to responsibly share and reuse data among different contexts and disciplines or through the open web? Nothing is said on how a general data theory could help with the design of innumerable data applications and algorithms that rely on clustering, for instance, or about scheme and records matching, for data cleaning or for the similarity based ranking criteria that are so crucial not only for research data in biology or medicine but also for all sorts of data managed by search engines, e-commerce platforms, social media.

The authors refer to known training initiatives (like the University of Edinburgh's MANTRA) that have shaped the data librarian training curriculum in recent years as an extension of the general library instruction or information literacy programme in support of any discipline, helping with operational matters like defining a data storage and security policy, organising access and data sharing, writing a data management plan, improving data

handling skills and even navigating the semantic web. It seems that the basics of data librarianship should be in all that … knitting of symbols, commons, dots and apostrophes typical of the linked data syntax.

And yet, thrown out from the main door of the job description's convenience or operational priorities, the classificationist's approach to data librarianship may come back from the window inevitably left opened on categorisation, indexing and quality issues.

It may capture the reader's attention the fact that, for instance, at the end of the excursus on how to support data literacy we have plenty of practical suggestions on data citation, reference tips, creation of "fit-for-purpose teaching datasets" but nothing is said on how to handle an overwhelming mass production of datasets pouring on students, librarians and teachers. Predictable, simple ways to deal with the data streams across different disciplines, projects and application profiles seem a long way off. Shouldn't be part of the mission of data librarianship to give research datasets more chances of being retrieved with precision and pertinence in data searching and data mining applications? Is it or is it not a prerogative of any library collection to describe items in such a way that relevant attributes and formats (for instance, Character Encoding) can be used consistently to identify, access, exchange and potentially reuse data on a global basis and through machine-to-machine transactions?

Chapter 4 (*Building a data collection*) and 5 (*Research data management service and policy: working across your institution*) take the long view and indirectly answer these concerns, looking at different roles and functions in the traditional organisation of academic libraries, all converging towards new data goals: policies for acquisitions and licensing, vendor trials, changes in management of institutional support and plans, collaboration with others, storage space, roles and responsibilities and so on and so forth, including toolkits for the evaluation of data curation profiles and audit or assessment or maturity guide in respect of the whole research data lifecycle.

The book offers at this point concise and effective references to the best theoretical proposals elaborated within UK universities in the last five years including workflows in place at Oxford, Edinburgh and Imperial College and eventually leans again towards the existential initial question: "What is the library's role?".

Chapter 6 (*Data management plans as a calling card*) presents further case studies of data management plans prepared by eight librarians: students and volunteers will love it as they are an invitation to… cut and paste. Excuse my malice, but I believe it would be realistically wiser to avoid the publication of such recipes not only because they may propagate wrong or not optimal procedures but also because they demotivate a more creative and investigative attitude towards innovation among librarians that is absolutely critical for advances in data and research data management by their end-users, including teachers and researchers.

Chapter 7 (*Essentials of data repositories*) offers basic definitions and references to repositories platforms. It is only at this point of the book that the authors introduce the issues of "choosing a metadata scheme" and "converging on a standard", in relation to repository requirements. Again, the frustration of the classificationist could not be greater while considering that the only reason for choosing a metadata scheme seems to be to allow research data to be queried by search engines and application programming interfaces (APIs) so they can drive more traffic to the Universities' websites. However, this chapter is a good starting point for readers not familiar with data repositories.

Chapters 8 to 10 (*Dealing with sensitive data*, *Data sharing in the disciplines* and *Supporting open scholarship and open science*) presents the current debate on privacy, collaboration, open access policies and can surely give the reader confidence to join or start the conversation on these topical issues. The quality or assurance aspect of any artificial intelligence or machine learning project is the elephant in the room: who is going to take responsibility of an ethically driven development of big data applications using a massive amount of research data streams?

The array of applications, methods and literature available on research data management outside the library space and even outside the academia is so huge that it is hardly believable that the authors are not aware that a sort of data librarianship culture is developing outside the physical walls and the disciplinary boundaries of the university library concept. The idea of data librarianship as an ancillary field helping with access and storage of data produced by academic researchers and as an extension of traditional library instruction and information or media literacies activities has undoubtedly many organisational, political and tactical advantages: it does not scare anybody, it is easy to share in the whole academic hemisphere, provides valid arguments

in support of cross-department collaborations and can reduce friction among faculties on new big data projects, enhances the quality and transferability of the practical skills offered to students and young researchers in any discipline and does not interfere with priorities, methods and sense making processes that fall into the remit of single faculties.

However, sooner than once imagined, the same patrons may require librarians to have a more proactive and constructive vision of their own role, to fill what has been called the "data trust deficit" and to make sense of the unmanned, though efficient, lack of predictable quality in data intelligence automation. The authors avoid putting forward any suggestion in this direction but it must be said that only very recently ethically driven policy developments have been seen in the engineering, statistics and computer science communities.[2]

In conclusion, the book does what it says on the tin: introduces to a sort of common research data management idiom still very immature that in the effort to be universally accepted and institutionally neutral and never controversial, as an Esperanto, risks remaining anchored to a list of potential unanswered questions about data librarianship—but perhaps I should also concede that a list of potential unanswered questions is all what data ultimately consists of.

References

ACM DL. (2012). Mining big data. Special issue of *ACM SIGKDD Explorations Newsletter 14*(2), 1–81.

Flichy, P. (2013). Rendre visible l'information. Une analyse socio technique du traitement des données. *Réseaux 178–179*, 55–89. Retrieved April 27, 2017, from https://www.cairn.info/revue-reseaux-2013-2-page-55.htm. English translation by Liz Libbrecht: Making information visible. Retrieved April 27, 2017, from http://www.cairn-int.info/article-E_RES_178_0055--making-information-visible.htm.

Research Council UK. (2015). *Guidance on best practice in the management of research data*. Retrieved April 27, 2017, from http://www.rcuk.ac.uk/documents/documents/rcukcommonprinciplesondatapolicy-pdf/.

Tenopir, C., Talja, S., Horstmann, W., Late, E., Hughes, D., Pollock, D.,…, & Allard, S. (2017). Research data services in European academic research libraries. *LIBER Quarterly, 27*(1), 23–44. Retrieved April 27, 2017, from https://www.liberquarterly.eu/articles/10.18352/lq.10180/, https://doi.org/10.18352/lq.10180.

Verhaar, P., Schoots, F., Sesink, L., & Frederiks, F. (2017). Fostering effective data management practices at Leiden University. *LIBER Quarterly, 27*(1), 1–22. Retrieved April 27, 2017, from https://www.liberquarterly.eu/articles/10.18352/lq.10185/, https://doi.org/10.18352/lq.10185.

Von Hippel, E.A. (2007). Horizontal innovation networks—By and for users. *Industrial and Corporate Change 16*(2), 293–315. Retrieved April 27, 2017, from http://ssrn.com/abstract=1411719, https://doi.org/10.1093/icc/dtm005.

Notes

[1] See for instance projects described and references in recent articles published in this journal: Verhaar, Schoots, Sesink, & Frederiks (2017) and Tenopir et al. (2017).

[2] For instance, in December 2016, the IEEE has launched a *Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems*. The consultation initial position paper is available at http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.