THE ROLE OF LIBRARIANS IN TRANSFORMING THE WORLD THROUGH OPEN DATA AND OPEN SCIENCE

Ina Smith
Project Manager
African Open Science Platform
Academy of Science of South Africa,
Pretoria, South Africa

and
Susan Veldsman
Director
Scholarly Publishing Programme
Academy of Science of South Africa
Pretoria, South Africa

Abstract

More and more – both in the world of business and research – data is referred to as becoming the new oil. Artificial Intelligence – informed by data (incl. Big Data) and data instruments - is widely used to inform decisions, resolve problems, and predict trends. Social media such as facebook and Twitter are built around data only, and are billion dollar industries. Scientists internationally generate huge quantities of data on a daily basis, so that in turn it can be used to address challenges faced, including the challenges referred to in the 2030 Agenda for Sustainable Development. When data is FAIR – findable, accessible, interoperable, and reliable – it can be re-used, accelerating innovation and growth towards sustainable solutions. This paper will focus on the role of research librarians in making sure that data can be used for the unforeseeable future. The roles of the different role players in libraries will be addressed, as well as the knowledge, skills and services required to remain competent and relevant in an increasingly data driven world.

Keywords: *open science, open data, data repositories, research data management*

Introduction

In an increasingly data-driven world, a number of key stakeholders – including librarians – need to align and understand how they can contribute towards achieving the UN sustainable development goals. It cannot simply be business as usual, and digital technology developing faster than ever requires librarians to be fully digital competent, with the ability to use a wide spectrum of tools with great ease –also as far as data concerns, for example to analyse data, clean data, and visualise data. Librarians always had the role of making sure citizens and

researchers have access to much needed quality information to advance research, and it is no different with data – another information source that can be utilised by researchers. We are on a daily basis confronted with fake news, fake information, and also – more and more in future – might be fake data. The data underlying all forms of research output has to be openly accessible so that it can be re-used, and the outcomes of studies be verified where in question.

As custodians of research in research intensive institutions, librarians need to embrace the opportunity to remain relevant, contributing through the specialised skills they have, through:

- Distinguishing between good quality data managed in a responsible way, versus the opposite, and
- Managing data generated by researchers in a responsible and ethical way.

Other stakeholders in making sure that quality data remains open and accessible include:

- Governments (responsible for open science policies);
- Institutions (responsible for strategically adhering to policy);
- Researchers (collecting data in an ethical and trusted way so that it can be re-used);
- Statisticians (processing, analysing and visualising data);
- System engineers (to maintain a network and allow for data to be digitally transmitted), and
- Libraries (managing and organizing the data, and making sure it is digitally preserved for the unforeseeable future).

As far as the Open Access movement worldwide concerns, librarians have played a huge role in implementing institutional research repositories, archiving institutional research output such as theses, dissertations and copies of published research papers, for the unforeseeable future, making it openly accessible for all to benefit. In addition to repositories, advocating for publishing in open access scholarly journals have also received much attention, with some libraries offering extending this role to publishing their own journals using journal management systems.
In addition to the above, libraries can also play an extended role in data management, which will be addressed through this paper.

African Open Science Platform

The African Open Science Platform (AOSP) – an outcome of the international accord on Open Data in a Big Data World - now in its second year - is managed by the Academy of Science of South Africa (ASSAf) with input from the

International Council for Science Regional Office for Africa (ICSU ROA) being hosted by ASSAf. It is funded through the SA National Research Foundation, and in collaboration with the International Council for Science (ICSU) Committee on Data for Science and Technology (CODATA).

The project aims to develop an open science and innovation dialogue platform in order to increase awareness, accessibility and visibility of African science and data, at the same time reflecting on progress made on the African continent in terms of the following areas:

- Open science/data policy and strategy;
- Open science/data information technology (IT) infrastructure;
- Capacity building/training to support open science/data; and
- Incentives for sharing science output and specifically the underlying data sets, in an open and transparent way.

As part of this project, the team has so far engaged with numerous stakeholders, presented numerous workshops, and upcoming is a National Dialogue on an Open Science policy with Ugandan high level key stakeholders. Policy is often seen as a barrier, but it is much needed to understand how a country manages and protects its data from being exploited, for example where researchers would come and conduct research, leave the country, and then the data – of which the nationals and its very assets were the objects – never benefit in the end, and the data cannot be used to achieve the challenges mentioned in the sustainable development goals.

Importance of open research data

Research data is often referred to as the new "oil", with unlimited potential to inform and predict trends, develop new applications (apps), teach or learn machines (machine learning), and more. Africa itself has many great examples where data has changed lives, but more needs to be done for the data to be exploited to its full potential. To be fully exploited for all to benefit, data needs to be FAIR – findable, accessible, interoperable and re-usable. Librarians can play an important role in fulfilling the role of data stewards, managing copyright and licensing, assigning metadata, and preserving data for the long term through data repositories.

According to Gurin (2015) the primary purpose of open data initiatives worldwide is to help governments, businesses and civil society organizations utilize the already available digital data more effectively to drive sustainable development. Many Open Data initiatives involve taking data that is already publicly available and putting it into more usable formats, making it a powerful resource for private sector development, jobs creation, economic growth, and more effective governance and citizen engagement. Through the African Open

Science Platform project we are trying to establish what exactly is happening on the continent in terms of African research data.

The African Open Data Impact Map provides a very conservative view, and many African data initiatives are not registered on this map. Through advocacy – also by librarians – researchers and institutions can play a much bigger role in creating an awareness for the importance to share data, the existence of the data to avoid duplication, and encourage all to register initiatives on platforms such as re3data.org, the Open Data Barometer and the Open Data Impact Map.

Figure 1. Open Data Impact Map - Africa

Benefits of open data

Open data benefits society in general, but to highlight a few:

- It helps predict trends and allow for informed decisions to be made;
- It drives development and improves the livelihoods of citizens of the country;
- More and more entrepreneurs are using data in innovative ways, creating more jobs which is much needed on our continent;
- It helps improve service delivery;
- It provides evidence for research conducted;
- Data potentially has far more outcomes when open, with a higher impact; and
- Only if research and data are open and democratized so that all can have equal access, it would be possible to work towards achieving the 2030 Sustainable Development Goals.

Fears researchers experience

Regardless of the benefits, researchers still often have their doubts, and are hesitant to share their data openly because of the following reasons, which is understandable. The African Open Science Platform hopes to come up with an incentives framework to guide governments, institutions, funders and more to how this issue should be dealt with. More and more funders make it a requirement that the underlying data to an article or research project/theses/dissertation be made openly accessible, e.g. NRF (South Africa) and Horizon 202 (European Commission).

Researchers are hesitant to share their data because they are afraid of:

- Getting scooped by other researchers;
- They feel they have invested resources and effort, and why should others have it "easy";
- Fear of someone else finding a path-breaking application of the data that the original researcher hasn't considered;
- Fear of problems/errors in the measurement process being exposed;
- Confidentiality/privacy of respondents might be a problem, although all proper and trusted research should be approved and subjected to ethics clearance; and
- Intellectual Property Rights are often signed away, with little understanding of copyright, licensing and intellectual property rights of the individual or organisation.

For all the above there are solutions, and the AOSP Incentives Framework – once published – will be addressing these.

Examples of data achieving the sustainable development goals

Great success has been achieved through making data openly accessible, and a few examples are shared below. It is important that these stories be collected, so that the impact can be monitored and researchers who are hesitant can be encouraged to make the right choice.

The struggle to eradicate malaria continues

Malaria remains to be a challenge, and is difficult to accurately measure because it shares symptoms with many other diseases. According to [this paper] there is however a way to accurately determining the quantity of malaria in any given area, providing valuable data to decision makers, researchers, funders and more. The study conducted had to rely on sources mostly hidden in old government archives or curated by the World Health Organisation. Most of the records were either poorly stored, burnt or were missing. In some countries like Kenya, Senegal, Tanzania, South Africa, Botswana, Namibia and Burkina Faso the surveys dated back 1950s. Conversely, recent surveys have been easier to locate through more modern web based searches.

To obtain village or school level data published in most journals or reports, scientists and government officials provided the raw data. This is a testament to a new era of data sharing where over 800 people have contributed finer resolution data.

The final report covers over 50,000 surveys dating back 115 years. This is the largest repository containing information on over 7.8 million blood tests for malaria. The study suggests that the prevalence of malaria infection in sub-Saharan Africa today is at the lowest point since 1900. But more needs to be done, and through making data openly available in a properly managed way, great progress can be made.

The Open Data Institute (ODI) published a report - *Supporting sustainable development with open data* - in which there are numerous examples of how data from Africa can contribute to achieving the sustainable development goals.

Protecting banana farmers' livelihoods in Uganda

Data was provided to the Uganda government on the banana bacterial wilt with real-time information on the spread of the disease. They were able to quickly identify the most affected areas and direct the limited treatments for the disease to prevent further advances. At the same time, they could disseminate information directly to the public via SMS on treatment options and how to protect their crops. Within five days of the first messages being sent out, 190,000 Ugandans had learned about the disease and knew how to save bananas on their farms.

A particularly beneficial use of this data would be to build a global subnational map of the prevalence of underweight children that could be used by governments and aid groups to target nutrition interventions to where they are needed most.
Other case studies from the ODI which is worth looking at, to demonstrate how data can benefit the people, include:

- Using maps to increase access to education in Kenya
- Monitoring child malnutrition in Uganda

Role of librarians

From the examples mentioned, it is clear that librarians – as data stewards – can position themselves strategically towards playing a role in achieving the SDGs through managing data. We hope that it receives prominence in institutional and library strategic plans, and that libraries in Africa are actively contributing to accomplish making valuable data openly valuable, managing copyright, licenses, file formats, metadata, migration to readable formats, call for proper citation of data, and more.

Burnett (2013) refers to ten recommendations – the result of numerous workshops - for libraries to get started with research data management:

1. Offer research data management support, including data management plans for grant applications, intellectual property rights advice and information materials. Assist faculty with data management plans and the integration of data management into the curriculum.
2. Engage in the development of metadata and data standards and provide metadata services for research data.
3. Create Data Librarian posts and develop professional staff skills for data librarianship.
4. Actively participate in institutional research data policy development, including resource plans. Encourage and adopt open data policies where appropriate in the research data life cycle.
5. Liaise and partner with researchers, research groups, data archives and data centers to foster an interoperable infrastructure for data access, discovery and data sharing.
6. Support the lifecycle for research data by providing services for storage, discovery and permanent access.
7. Promote research data citation by applying persistent identifiers to research data.
8. Provide an institutional Data Catalogue or Data Repository, depending on available infrastructure.
9. Get involved in subject specific data management practice.
10. Offer or mediate secure storage for dynamic and static research data in co-operation with institutional IT units and/or seek exploitation of appropriate cloud services.

From the above it is clear "why librarians have become natural partners in the research data management (RDM) process" (Burnett 2013). They have highly relevant information standards and organizational skills, including expertise in setting up file structures, knowledge of workflows and collection management, describing data in accordance with established metadata schemes and controlled vocabulary, collection curation/ preservation and service provision in the form of helpdesks, training, availability of subject specialists, etc.

Required skills librarians should have

It is crucial that library schools constantly adapt themselves, and train prospective librarians in the field of managing data. Existing librarians need to upskill themselves, learn from the literature and implementations by libraries worldwide, and attend courses (also free and online) in order to upskill themselves as part of Continuous Professional Development.

Both the *Author Carpentry* and *Library Carpentry* online workshops provide valuable direction as to the new skills to be acquired to better support researchers, which can be summarised as follows:

- Introduction to the terminology of data and computing, and the use of regular expressions to search and update text;
- Unix-style command line interface, allowing librarians to efficiently work with directories and files, and find and manipulate data;
- Cleaning and enhancing data in OpenRefine and spreadsheets;
- Introduction to the Git version control system and the GitHub collaboration tool;
- Relational database management system using SQLite;
- Web scraping and extracting data from websites;
- Using Python as a general purpose programming language that supports rapid development of scripts and applications;
- Scientific writing in useful, powerful, and open mark-up languages such as LaTeX, XML, and Markdown;
- Formulating and managing citation data, publication lists, and bibliographies in open formats such as BiBTeX, JSON, XML and using open source reference management tools such as JabRef and Zotero;
- Transforming metadata documenting research outputs into open plain text formats for easy reuse in research information systems in support of funder compliance mandates and institutional reporting;
- Establishing scholarly identity with ORCiD and managing reputation with ORCiD-enabled scholarly sharing platforms such as ScienceOpen;
- Crediting authorship, contributorship, and copyright ownership in collaborative research projects;
- Demonstrating best practices in attribution, acknowledgement, and citation, particularly for non-traditional research outputs (software, datasets);
- Identifying reputable Open Access publications and Open Institutional/Open Data repositories;
- Contributing to, and gaining value from, scholarly annotation and open peer review;
- Investigating and managing copyright status of a work, and evaluating conditions for Fair Use;
- Open sharing of research using Creative Commons licenses, waivers, and public domain marks;
- Provide data management services;
- Provide sustainable and trusted data repositories where researchers can upload their raw datasets, and preserve for the unforeseeable future; and
- Assign metadata and clean metadata.

The above not exhaustive, and to be continuously monitored.

Conclusion

Libraries – where data has not been embraced yet – can no longer afford to sit on the side-line and watch. They are important stakeholders in making sure the SDGs are accomplished, and should become data custodians and stewards. Existing practises need to be revisited and prioritised, clearly distinguishing between what researchers more and more can do for themselves, and where libraries can serve a bigger purpose in terms of data. To summarise in three points, the following are required from librarians:

- Upskill themselves and learn about the role they can play;
- Implement as part of the library strategy;
- Advocate for data to always remain open; and
- Make sure data is FAIR through data repositories.

If all stakeholders playing a role in the research and data lifecycle actively participate, we can all together achieve the sustainable development goals as far as data to providing solutions and predicting trends concern.

References

Author Carpentry. (n.d.). Available at: https://doi.org/10.7907/Z96H4FFZ . [Accessed on 6 Jan. 2018].

Burnett, P. (2013). *What is the role of a librarian in research data management?* Available at: http://blog.inasp.info/research-data-management-role-librarians/. [Accessed on 6 Jan. 2018].

Gurin, J. (2015). *How open data can drive sustainable development.* Available at: http://blogs.worldbank.org/ic4d/new-discussion-paper-how-open-data-can-drive-sustainable-development . [Accessed on 6 Jan. 2018].

Library Carpentry. (n.d.). Available at: https://librarycarpentry.github.io/ . [Accessed on 6 Jan. 2018].

Open Data in a Big Data World. (2015). Available at: http://www.science-international.org/sites/default/files/reports/open-data-in-big-data-world_long_en.pdf . [Accessed on 6 Jan. 2018].

Open Data Institute. (2015). *Supporting sustainable development with open data.* Available at: https://theodi.org/news/new-report-reveals-how-open-data-is-fuelling-problemsolving-in-the-developing-world-from-mapping-ebola-to-protecting-banana-crops . [Accessed on 6 Jan. 2018].

Snow, B. (2017). *What 115 years of data tells us about Africa's battle with malaria past and present.* In: The Conversation. Available at: https://theconversation.com/what-115-years-of-data-tells-us-about-africas-battle-with-malaria-past-and-present-85482 . [Accessed on 6 Jan. 2018].

Supporting sustainable development with open data. Available at: https://theodi.org/news/new-report-reveals-how-open-data-is-fuelling-problemsolving-in-the-developing-world-from-mapping-ebola-to-protecting-banana-crops . [Accessed on 6 Jan. 2018].

Transforming our world: the 2030 Agenda for Sustainable Development. (2015). Available at: http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E. [Accessed on 6 Jan. 2018].