IJoL International Journal of Librarianship

ISSN: 2474-3542 Journal homepage: http://journal.calaijol.org

Lives in Data: Some Prominent Data Librarians, Archivists and Educators Share Their Thoughts

Kristi Thompson and Guoying Liu

Abstract:

We asked several data librarians, archivists and educators who have had prominent and interesting careers if they would be willing to let us profile them and share some of their thoughts on the field. Six graciously agreed to be interviewed via email. Many of our respondents played key roles in developing data services and infrastructure in their respective countries, while others are involved in building the future of the field through education, advancing standards, and advocacy.

Our virtual panel includes Tuomas J. Alaterä, Finland; Ann Green and Jian Qin, United States; Guangjing Li, China; Wendy Watkins, Canada; and Lynn Woolfrey, South Africa.

To cite this article:

Thompson, K. & Liu, G. (2017). Lives in data: Some prominent data librarians, archivists and educators share their thoughts. *International Journal of Librarianship*, *2*(1), 66-72. https://doi.org/10.23974/ijol.2017.vol2.1.35

INTERNATIONAL JOURNAL OF LIBRARIANSHIP, 2(1), 66-72
ISSN:2474-3542

Lives in Data: Some Prominent Data Librarians, Archivists and Educators Share Their Thoughts

We wanted to gain a wider-ranging international perspective on the evolving field of academic data librarianship. We asked several data librarians, archivists and educators who have had prominent and interesting careers if they would be willing to let us profile them and share some of their thoughts on the field. Six graciously agreed to be interviewed via email. Many of our respondents played key roles in developing data services and infrastructure in their respective countries, while others are involved in building the future of the field through education, advancing standards, and advocacy.

Our virtual panel includes **Tuomas J. Alaterä**, Finland; **Ann Green** and **Jian Qin**, United States; **Guangjing Li**, China; **Wendy Watkins**, Canada; and **Lynn Woolfrey**, South Africa. Responses were compiled by Kristi Thompson and have been excerpted and edited for brevity.

Tuomas J. Alaterä is an IT Services Specialist at the Finnish Social Science Data Archive, and current president of the International Association of Social Science Information Services and Technology (IASSIST), a prominent international organization for professionals who support research and teaching with data.

Tuomas on how he became involved in working with data:

By a phone call! Back in year 2000, and before the reign of Google, curated subject specific web resources were a thing. As a trainee at the Department of Information Studies, with a background in political science and web design, I was hired to work on a project on political science resources. Rather soon we set to build an online learning environment for teaching and learning quantitative methods. The simple key to success was to make real research data available, and walk students through the analysis process in SPSS and provide information on how to interpret the results.

The FSD has its roots in traditional European social science data archive model. It grew out of a need to archive research data for reuse and provide information services related to that purpose. Now our spectrum of services has widened a lot. We offer a data portal for researchers and students to browse, search and download datasets. Data cleaning and producing descriptive metadata about the datasets is a big part of the daily work, but guidance and instruction on research data management, data protection and privacy is gaining more and more visibility.

So one could say that at first I was mostly concerned with accessibility and data literacy. I then started thinking about data services on a broader scale. I've always found building networks to be the way to the profession, because then, and even now, there is fairly little formal training on how one become a data librarian or data support professional.

IASSIST on the other hand strives to foster the development of data professionals. Though our annual conference, interest groups and mailing lists we provide our members a network that is there for collegial support, exchange of knowledge and learning. And even for safety, in case someone finds her or himself thrown in at the deep end of data support duties. In future we hope to run a series of webinars as well.

On how data services have evolved over the course of his career:

The demand for data services has grown immensely. I would argue that we have moved from the sidelines to centre field. Nowadays it is technically easier to share, access, analyse and visualise data online, and data sharing has become ideologically more desirable. Not to say that it didn't happen before, but the culture is changing. Requirements from the funders and journals on making the data available have further accelerated this development. The business model of scientific publishing is changing because of Open Science. Services for making also data and methods available in order to secure the reproducibility of research are in demand right now.

Many modern research data infrastructures are now being built on and around the structures that social sciences have had in place for decades. For sure, the scales are different, the pace is different and the ways of using the data are different. But this makes me proud of the practitioners of "soft sciences". It shows that we correctly identified the needs and benefits that data sharing can have.

Advice for someone getting started in data services:

Collaborate. Seek networking options, check webinars, and convince your superior to allow you to attend conferences or workshops. Like IASSIST, for example! Most academic institutions need local data support services and repositories. However, the services requested are fairly similar in most. This is not the time or place to start reinventing the wheel, but to figure out how to get the wheels turning and keep them rolling in the best possible way for your institution.

One key element is to learn to know your customers, and the policies within which you will be operating. Don't start from scratch but reach out, interview other service providers and find ways to collaborate. If there is no data policy in place, then advocate that your institution should start on defining one.

Services are also becoming more standardised. Best practices like the FAIR Data Principles, or criteria for trustworthiness like the Data Seal of Approval, can act as a yardstick for emerging services.

Ann Green is currently an independent research consultant. She was previously the Data Archivist at the Cornell Institute for Social and Economic Research (CISER) and later director of the Yale University Social Science Statistical Laboratory. She recalls:
As the first Data Archivist at the Cornell Institute for Social and Economic Research (CISER), I was charged with helping to establish a social science data archive and statistical consulting service to support research and teaching needs at Cornell. It was an exciting challenge at a time (about 30 years ago) when other data archives and data services were in development across the world.

At a time when a data archivist was a rare professional title, I was fortunate to be "drafted" to take on the data archivist position at Cornell, having been a reference librarian at Cornell's graduate library. It meant building a data collection from scratch, taking programming courses, visiting data libraries across the country, and diving in to set up new services for faculty and graduate students in the social sciences. The university library worked closely with CISER to address the new challenges of building a digital archive and providing a range of support services.

On how data services have evolved over the course of her career:

Before data analysts could access data directly via FTP and the internet, local social science data archives provided carefully managed collections of data that their research and teaching communities needed. Local data collections (on tape and then disk) helped reduce duplication and loss of data, provided a home for consultation and training, and gave a voice for user requirements and program development. In the '80's and 90's, data archives were housed in libraries, research centers, government agencies, and in academic departments…

With the advent of internet based dissemination of data, some of these local collections and services were reorganized or absorbed into libraries. Most recently, mandates and incentives to share data, along with an increasing emphasis upon replication and reproducibility, have seen a resurgence of interest in making research data available along with the necessary information to make it usable for the long term. Data archives and repositories with the mandate to guarantee access and understandability over time should build upon these decades of data archiving expertise and the data management best practices that have been developed and refined throughout the history of data archives.

Advice for someone starting out:

My advice would be to connect with other professionals in the data archive world. For example, IASSIST is an organization of individuals (working primarily in the social sciences, but increasingly in other fields) who meet annually and are connected by an active email support network. IASSIST conferences (and others) provide an excellent opportunity to present ideas, track down solutions, find mentors, and build collaborations. Connections can also be made through workshops sponsored by ICPSR and other major archives, and by taking part in data user groups and research data management organizations.

It is important to continually build expertise: learn about methods, applications, and data analysis in the fields being supported by data services, take programming classes and write code, and be familiar with data collection techniques. Cross boundaries between data science initiatives, data librarians, and digital archivists. Become active in a solution to a data challenge – contribute to the development of a metadata standard, participate in an assessment standard like Data Seal of Approval, focus upon specific tool development, or take part in data rescue efforts. Advocate for and take on the commitment to long term accessibility, understandability, and reuse of critical data resources. Explore the dynamics of digital preservation including metadata standards, format migration, replicated storage, and usability over time.

Guangjian Li is a professor in the Department of Information Management, Peking University as well as a PhD student supervisor and the dean of the department. He holds positions on numerous national societies and boards which are helping shape the development of data service and infrastructure in China, and also sits on the editorial boards of four journals in information science. His responses were translated from the Chinese by Xiaoai Ren, one of the editors of the IJoL.

Guangjian's initial involvement in the data field came about through his work in digital infrastructure:

I participated in the construction of the network and automation system in the new National Library of Science, Chinese Academy of Sciences (CAS) and the construction of the CAS National Digital Library. I also participated in the research and development of the library portal, federated search, and intelligent search.

Later I worked for the National Science Library, Chinese Academy of Sciences (NSLC). The NSLC is the research library service system of CAS as well as the National Library of Sciences in Chinese National Science and Technology Libraries (NSTL) system. It functions as a key library for collecting information resources and providing information services in natural sciences, inter-disciplinary fields, and high tech fields, for the researchers and students of CAS and for the researchers around the country. It also provides services in information analysis, research information management, digital library development, scientific publishing, and promotion of sciences. My current employer is in higher education.

On how data services have evolved over his career:

Data services at the early stage were mainly about data collection development and sharing for academic and research purposes. The service model was a passive service in which we provided access to what we had. As big data and machine learning develops at rapid pace, recent data services are focusing more on providing customized data based on users' need. The new service model is one of active service which we provide what the user needs.

Advice for someone starting out:

Be curious! Think critically and be willing to give it a try and make a difference. First and foremost, continuous learning is the best way to quickly merge into this field and keep growing; second, critical thinking ensures one's creativity and motivation; third, practice leads to innovation; lastly, do not fear change. It's the best way to move forward.

Jian Qin is a professor at Syracuse University iSchool, where she teaches courses on information organization, metadata, and fundamentals of digital data. Her research areas include metadata, schema representation of information, ontological modeling, research data management, impact assessment, and collaboration networks. She co-authored the book *Metadata* with Marcia Lei Zeng.

On how her initial research into data literacy snowballed:

About 10 years ago I obtained my first NSF grant to develop a scientific data literacy course. During the project life, I had the opportunity to work with faculty in other disciplines and study

their data management needs, which snowballed into a few other funded projects related to research data management. From the conversations and interactions with scientists, I gained in-depth understanding of the issues, practices, and requirements for research data management.

I have been working on an NSF-funded big metadata analytics project since 2013. This project takes the metadata from GenBank (an international data repository for genetic sequences) as the source to study the dynamics and structures of collaboration networks and the impact of cyberinfrastructure on data-intensive science' productivity and knowledge transfer and diffusion. The data collection is very large and has a high demand for computational power. The university academic computing worked with us to meet our needs for data storage, computing resources, and technical support.

On how data librarianship has changed during her career:

As an educator having worked in library and information science for the last 30 years, the biggest change in data services is the knowledge and skills required of librarians to perform their duties. The continuing increase in data science and services prompted the demand for data science knowledge and skills. A strong research method training and graduate level of disciplinary knowledge is becoming the new desired qualification for librarians who will engage in data science and services. Another change is the proactive, agile response to data service trends in libraries, that is, starting new data services from small and adjusting frequently and rapidly before a full-scale deployment. This approach allows libraries to make a quick response to service programs rising from the change of bigger environment while minimize the risk of doing (or not doing) it.

Advice for people (and institutions) starting out:

Because data services are still new compared to traditional libraries services, even a small-scale experiment requires the support from administration at university and library levels, which means committing the funding and resources needed by the new data services. To make such a support long-term and part of library operations, institutionalization of data services is key. This means that a data service project needs to promote awareness among constituencies (faculty, administrators, etc.) and build a community of practice early on. Policies about data services are also part of the process.

Wendy Watkins is the retired head of the Library Data Centre at Carleton University in Ottawa, Canada, and is currently an active member of the Portage Training Expert Group and a founding donor of Evidence for Democracy. She was a major force in creating Canada's Data Liberation Initiative, a project that greatly increased the availability of Canadian government data to academics.

She reflected on some highlights of her career:

I came to be a Data Librarian by a very circuitous path. I have never been to library school. Instead, I came to Carleton's library from the Sociology Department where I reported to the Dean of Social Sciences. My first professional job was as a hockey sociologist for the Canadian Hockey Review, a judicial enquiry. From there I held a number of research positions that involved data analyses in

a variety of contexts. In 1981 was offered the position as Data Archivist in the Faculty of Social Science at Carleton where I both provided a data service as well as active research services for 10 years.

I was then offered a 2-year sabbatical to work with Ernie Boyko at Statistics Canada, where I wrote "Liberating the Data", a paper that became the basis of Canada's Data Liberation Initiative (DLI), now in its 20th year. On returning to Carleton, my position moved to the library where my research duties remained the same. I was appointed as Data Librarian in 2005 and promoted to Librarian IV in 2010. In addition, I have held many positions in local, national and international organizations.

On how data librarianship has changed during her career:

When I began you required an IT background as the job involved knowing how to run jobs on a mainframe computer with SPSS/Fortran coding. There was no internet but since there were so few institutions we formed a cohesive group and shared what we could via BitNet and ArpaNet. At that time, it was also imperative that you have a background in research methods and statistics at an advanced level. This was needed to 'debug' programs and advise researchers on appropriate statistical methods for the data they were planning to analyse.

Through the efforts of Ernie Boyko and others, Data Liberation (DLI) came on stream and things changed drastically. Most of our early colleagues had been in social science research centres; the new members were primarily in libraries and had no research background or statistical expertise. Canada's data libraries expanded from nine to more than 50 virtually overnight and now the DLI boasts nearly 80 members.

Advice for a new Data Librarian:

My first piece of advice to a would-be data librarian is to choose an undergraduate degree that is rich in research methods and statistics. That way, the data end of the job is covered. Trying to pick up these advanced skills on the job as a Librarian is very difficult; and without a library degree after, it is nearly impossible to obtain a faculty position within a library.

My second would be to attend professional meetings and follow listservs with groups like IASSIST, the Research Data Alliance, etc. as well as taking part in as many hands-on workshops as possible. These tend to marry the fields of research data and librarianship.

And finally, finding a mentor within the field can be invaluable. The international data librarianship community is replete with colleagues who are more than willing to lend a hand to their newest members. You will be amazed by their generosity.

Lynn Woolfrey manages the research data infrastructure and services at DataFirst, a research data service based at the University of Cape Town, South Africa. Her career provides a glimpse into how data and politics intersect.

She shares:

I worked as an Information officer/librarian for a research unit at the University of Cape Town, the Southern Africa Labour and Development Research Unit. We were tasked with undertaking the first inclusive household survey in South Africa (previous household surveys either left out "black" households, or enumerated them with a separate questionnaire). The data from this survey was used to inform the Post-Apartheid government's socioeconomic policies. I was tasked with curating the data from the survey, and liaison with data users. In 2001 we obtained funding to establish a research data service to curate the growing collection of government survey data. We expanded our services to include census and survey data from other sources, such as the private sector or donor organisations.

On how evolving tools and standards have changed how she works:

When I started curating data for research, there were few tools for my work. Now there are a proliferation of useful and often free tools. For example, we used to pay to use the Nesstar server software to publish data. We now use free software, the National Data Archive (NADA) dissemination tool, available from the International Household Survey Network, an initiative of the World Bank. Since then, too, data curation standards have been established by our community of practice. We adhere to these in our data preparation, metadata creation and data publishing activities. For example, we use the Dublin Core Standard to describe the documents we share with the data, and the Data Documentation Initiative standard for describing our datasets.

Also, data sharing requirements. In the 1990s we had to persuade projects and government departments to share their data, and we had to provide training to build quantitative skills, as these were in short supply among African academics. These days, we are approached by data owners to curate and publish their data, as this is often a funding requirement. Governments also now realize the legitimacy of their data depends on them sharing the data with the academic community.

Advice for someone starting their career in data, or anyone developing a new service:

Use standards and best practice. We have built up our reputation as a trusted repository, and garnered awards (we are the only African data repository to be certified with the Data Seal of Approval).

Use community resources. Make use of the expertise that exists in our community of practice, and ask for help from organisations like IASSIST, the UK Data Service, and the ICPSR.

Build healthy relationships with data owners. I suggest working with institutions within government and academia who collect data, to build a relationship of trust. This encourages data owners to deposit data with the service.

Many thanks to our panelists for taking the time to share their insights with our readers.

Edited by:
Kristi Thompson, Guest Editor, *IJoL*
Guoying Liu, Editor-in-Chief, *IJoL*
With thanks for their assistance to editors **Keven Liu**, **Xiaoai Ren** and **Yongming Wang**