The Data Librarian's Handbook

Every purchase of a Facet book helps to fund CILIP's advocacy, awareness and accreditation programmes for information professionals.

The Data Librarian's Handbook

Robin Rice and John Southall

facet publishing

Published by Facet Publishing
7 Ridgmount Street, London WC1E 7AE
www.facetpublishing.co.uk

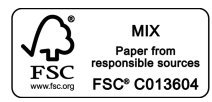
Facet Publishing is wholly owned by CILIP: the Chartered Institute of Library and Information Professionals.

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library.

ISBN 978-1-78330-047-1 (paperback)
ISBN 978-1-78330-098-3 (hardback)
ISBN 978-1-78330-183-6 (e-book)

First published 2016

Text printed on FSC accredited material.



Typeset from authors' files in 10/13 pt Palatino Linotype and Open Sans by Facet Publishing Production.
Printed and made in Great Britain by CPI Group (UK) Ltd, Croydon, CR0 4YY.

Contents

Acknowledgements

We would first like to thank our spouses and our bosses for offering us their support and especially patience as we wrote this book without our work and private lives slowing down. Helen Carley, our publisher, always showed faith in us, even when we worried that the field of data librarianship was changing faster than we could even fix our knowledge onto the page. Laine Ruus read and critiqued our early drafts, debated with us about some of our assumptions, and added a fresh perspective. Members of the International Association for Social Science Information Service and Technology (IASSIST), our main professional society, have helped us crystalize our knowledge about data librarianship throughout our careers, and provided a supportive and fun community allowing us to thrive in our work.

Robin Rice and John Southall

Preface

This is not the first book written about data librarianship, and hopefully it will not be the last, but it is one of very few, all written within the past few years, that reflects the growing interest in research data support. Academic data librarians help staff and students with all aspects of this peculiar class of digital information – its use, preservation and curation, and how to support researchers' production and consumption of it in ever greater volumes, to create new knowledge.

Our aim is to offer an insider's view of data librarianship as it is today, with plenty of practical examples and advice. At times we try to link this to wider academic research agendas and scholarly communication trends past, present and future, while grounding these thoughts back in the everyday work of data librarians and other information professionals.

We would like to tell you a little bit about ourselves as the authors, but first a word about you. We have two primary groups of readers in mind for this book: library and iSchool students and their teachers, and working professionals (especially librarians) learning to deal with data. We would be honoured to have this book used as an educational resource in library and information graduate programmes, because we believe the future of data librarianship (regardless of its origins, examined in Chapter 1) lies with academic libraries, and for that to become a stronger reality it needs to be studied as a professional and academic subject. To aid the use of this book as a text for study we have provided 'key take-away points' and 'reflective questions' at the end of each chapter. These can be used by teachers for individual or group assignments, or by individuals to self-assess and reinforce what they may have learned from reading each chapter.

Equally important, we empathetically address the librarian, academic, or other working expert who feels their working life is pulling them towards

data support or that area of academic activity known as research data management (RDM). We appreciate that this subset of readers will bring many pre-existing abilities and knowledge to this area, so we attempt to fill in the missing portions as pragmatically as we can, while linking daily tasks to broader goals and progressive initiatives, some of which you will be well familiar with and others less so, depending on your area of expertise. As will become apparent in Chapter 1, virtually everyone working as data librarians today received no special training beyond learning on the job, professional development opportunities and, if we were lucky, some personal mentoring.

We hope that by foregrounding these groups we have closed a significant gap in this nascent body of literature. We have also considered the requirements of other potential readers, be they library managers hoping to create new data librarian posts, policy-makers in libraries and academia developing strategies for research data, or academic librarians and other support professionals compelled to add data support to an existing workload who could use a primer on the subject.

Although between us we have over 20 years of experience as data librarians we still find it tricky to describe our work (at the proverbial cocktail party). In that sense, writing this book has been a welcome opportunity to explore our own professional activities and proclivities, to compare and contrast with each other and with other data librarians and data professionals, and to draw out what is consistent, lasting and of most value in what we offer to the research communities we serve. Typically for data librarians, as we shall see, one of us comes from a library background, the other from research (sociology), and while we are both UK-based, one of us began our data librarian career in the USA (at the University of Wisconsin-Madison), so we aim for a cross-Atlantic view. Although we aim to provide a single voice to the book it has certainly been the case, given the variety of approaches to data support by both institutions and individual data librarians, that 'two heads are better than one' for this endeavour.

A few of our conventions are worth mentioning here. It is our intention to always use the word data in the plural form. Some uses in the singular may slip through, as it does in general culture, but we find that you can get used to using the term 'properly' if you try. Figure 0.1 sums up the situation well.

Figure 0.1 *Data: singular or plural?*

© XKCD Comics, 'Data'. Used in accordance with https://xkcd.com/license.html.

Where we cite literature, and especially the more seminal literature that has grown up in our field, we provide complete references at the end of the book in the time-honoured manner. However, as our working world is very much one that is always online, web-based resources are sprinkled throughout the book, not in separate footnotes, but embedded in the text, so that you may have a look and a play as you are reading. The fact that some of these URLs are bound to disappear over time is one we regretfully accept, but we hope there is enough context given for the reader to find either the resource discussed or a newer, equivalent tool for the job.

Some of the terminology we use may be unfamiliar to you. We find that much of it is authoritatively explained in the community resource called Open Research Glossary (www.righttoresearch.org/resources/OpenResearchGlossary), which we encourage you to use as a companion to the book. A note on referring to library patrons, which in itself can be revealing of different traditions and presumptions: some institutions use established terms such as 'reader'; some libraries or archival services developed specifically to support work with data refer to their main audience as 'users'. This will be discussed further in Chapter 1, but in our opinion both reader and user are acceptable terms, since one refers to the relationship of the researcher to a library-based support service and the other to their relationship with the data.

A final point, we are grateful that Facet Publishing have their own reasons for believing in a book on this topic at this time, and we very much welcome your interest as readers in data librarianship – a term we embrace that seems to encompass both the very new and the traditional in libraries – and hope that you find at least the beginnings of what you are seeking, and wish you well on your data journey or career.

Robin Rice and John Southall

CHAPTER 1

Data librarianship: responding to research innovation

The rise of data librarians

A university has been defined as 'just a group of buildings gathered around a library' (https://en.wikiquote.org/wiki/Shelby_Foote); in any case, the role of the library in academic life is a central one. Those working within libraries make a valuable contribution to supporting research and teaching as well as shaping the character and intellectual life of individual institutions. Whether a university focuses on the humanities, physical sciences, classics or any other number of disciplines, the librarian ultimately works to support learning and the spread of knowledge. This may take many established forms but increasingly there is a need to support new forms of information. Digital data is one particular new form. In the case of data collections and research data creation this has also led to the rise of a new kind of library professional: the data librarian. But to what extent is this in fact a new role and in what ways does it differ from traditional librarianship?

For example, one role of the librarian is to deal with what may be called the lifecycle of information resources. These are the varied tasks to do with evaluation, selection, purchasing and promotion, and preservation of materials within the library. This relies on having a good working knowledge of what readers in a particular area need for their work. It also draws on a familiarity with what is being made available by publishers and other suppliers of information resources. The terms employed to describe a researcher also indicate the orientation or origin of research support services. Some may prefer traditional terms such as patron or – as favoured at the University of Oxford – reader, since this gives continuity to existing provision. The medium or methodologies being applied to the data are unimportant. On the other hand those working on support services created specifically to deal with digital data may feel older terms are inappropriate or anachronistic. Since digital information is often used in conjunction with software it is no

longer 'human-readable' at all and its value lies in the fact it can be easily supplied to researchers. Their role is to manipulate, interpret, analyse, watch, listen to, or more generally 'use' the data. For this reason data centres or repositories often refer to 'users' of data. Finally, a sense of what characterizes a particular library is also important in how these different elements relate to each other. This often forms the basis of collections development policy.

Traditional – that is to say established – library activity also covers developing procedures and materials that help make collection items discoverable and accessible. Cataloguing and organizing of materials is an ongoing area of work that forms a foundation of much of librarianship. Preservation and curation is another key responsibility – especially when access has to be maintained for material that is harder to find or no longer in print. Reference and user services are a common feature in most libraries that build on maintaining collections. Consultancy and training workshops that seek to support readers in analysing problems, framing research questions and working with information resources in a meaningful way are as well. Librarianship then begins to be understood not simply as something that supports discovery of and access to published titles or information resources but also as something that engages with the *conduct* of research and academic enquiry.

These are ways that librarians are responding to the needs of the university as well as to the specific intellectual needs of readers. However, new areas of activity are emerging that reflect changes in the research environment or expectations of the kind of support library professionals should offer. These are not necessarily new in themselves – and may have been undertaken by other sections of an institution's infrastructure to some extent. Issues to do with licensing of research materials are a common example. Another is giving advice on sources of funding and completing funding applications, reference management software, statistical analysis software and Computer Assisted Qualitative Data Analysis Software packages. Working with readers to access, manipulate or share research data is a way to demonstrate libraries' responsiveness to academic needs.

In the past these areas of activity have often been seen as administrative or technical stages of research that need to be dealt with but that are unrelated to traditional librarianship. In a seminal article on cyberinfrastructure, data and libraries in *D-Lib Magazine* Anna Gold has characterized these areas that fall outside the usual comfort zone of academic librarians as 'working upstream' in the research process, before the point of publication (Gold, 2007). Working upstream means not only working with information that has not yet been published, but understanding the processes by which various types of data are used to generate information. The rest of this chapter will show how research data have become an archived resource over the last 40 years or so, and how this is becoming normalized as just another information resource.

It will also show how research data are no longer the specialist reserve of IT departments but are indeed now part of the remit of academic libraries.

Addressing early demand for data services in the social sciences

The origins of data libraries and data archives in the 1960s and 1970s owe as much to the way that the social sciences were developing as an empirical research domain as they do to the rise of centralized computing in research.

The social science disciplines (politics, economics, psychology, sociology, anthropology, etc.) are generally thought of as softer than the physical or hard sciences, in part because of their methodologies – which at times may be more like an art than a science – and in part because their subject matter – humans and their behaviours, individually or collectively – are so hard to pin down or predict. The rigorous application of the scientific method towards the social sciences resulted in the rise of quantitative methods – statistics applied to samples of populations in order to describe, explain and predict behaviours. New computer processing techniques combined with quantitative methods gave social scientists much power to view social phenomena in an objective or scientific light, using measures such as psychological experiments, social surveys and economic indicators.

With the power of hindsight it is easy to see why, in the second half of the 20th century, there was a backlash of sorts by many social researchers against 'positivism', or a tendency to explain or reduce all human behaviour to statistical trends (Williams, Hodgkinson and Payne, 2004). In the UK the social sciences are still recovering from the effects of this backlash against quantitative methods, to the point that the primary funder of social science research, the Economic and Social Research Council (ESRC), has declared there is a dearth of quantitative skills and has been investing in a number of programmes to beef up the statistical literacy and numeric skills of researchers and students (Jones and Goldring, 2015). As with librarians, many of the students entering the social sciences do not think of themselves as 'numbers people' and gravitate more naturally towards qualitative methods (such as interviewing), the findings of which may be quite rich but involve sample selections and sizes that can seldom be generalized to a population.

Happily, these days the social sciences have moved beyond the 'quantitative–qualitative divide' of the last century and most social scientists believe that both methodological approaches are valid and even symbiotic (Brannen, 2005). For example a mixed methods approach generally turns to quantitative methods for discovering *how* people behave, and a qualitative approach is embraced for uncovering reasons *why*, especially when that is less well known. Or, if one is starting with collecting qualitative data, one might wish to compare the characteristics of the selected subjects against a

reference population, using quantitative sources as benchmarks for contextualizing qualitative studies.

Then as now, social scientists are likely to turn towards secondary sources when they require quantitative datasets (consisting of one or more numeric data files, which may be encoded within software such as a spreadsheet, along with descriptive documentation about its contents). The use of data sources by the UK Data Archive (UKDA) roughly follows Pareto's 80/20 rule: 20% of resources are used by 80% of users, with 80% of datasets rarely if ever consulted, according to a previous director of the UKDA. (The Pareto distribution will be revisited in Chapter 4.) More often than not the highly used datasets are the well known national surveys and population census collections. Such resources have the advantage of being rich in variables (e.g. survey questions asked) so that new research questions can be interrogated, as well as having large, nationally representative samples. This is important for being able to study a sub-population (such as an ethnic minority group) and still have the statistical power to obtain useful, generalizable results. With a few exceptions (such as the British Election Study), it is government agencies which have the funding to carry out such extensive surveys and censuses, and not academic departments, let alone individual researchers. Another reason to share data about human subjects is to lower the response burden of individuals. Targeted commercial marketing has helped to create an atmosphere in which telephone and postal surveys are held in contempt by the intended subjects, lowering response rates across the board.

In this sense the social sciences, as the poor cousin of the more well endowed physical sciences, were the first disciplinary group to embrace data sharing and the re-use of data for reasons of economy or efficiency. However data about human subjects, like other observational data such as weather conditions, can never be replicated in the same circumstances – another strong rationale for sharing.

The earliest data libraries and archives – an American chronology

No one is better placed to recall the origins of academic data libraries in North America than Judith Rowe, retired Senior Data Services Specialist of Princeton University. In her entertaining speech at IASSIST's 25th anniversary conference banquet in Toronto on 20 May 1999, 'The Decades of My Life', she reminisced about the chronological developments in social science data collection and methods that she had witnessed (Rowe, 1999).

Rowe began her story as far back as the 1930s with the computing ideas of Alan Turing and Vannevar Bush and the existence of punch-card machines, but also the beginning of large-scale sampling, the Gallup Polls and the Brookings Institution. In addition to the massive global upheaval taking place

in the 1940s, the World Bank, International Monetary Fund and the United Nations were founded; scaling and multivariate analytic techniques were developed; and 'Immediately after the [US] election in which the pollsters [incorrectly] chose Dewey over Truman the [US] Social Science Research Council appointed a committee . . . to find out why' (Rowe, 1999).

In the 1950s IBM produced its first 'real' computer, programming languages COBOL (common business-oriented language) and Fortran were invented, the Institute of Social Research at Michigan and the Bureau of Applied Social Research at Columbia was producing survey data galore, and the Roper Center had archived over 3000 surveys from 70 countries. But sampling and survey data were not yet mainstream: 'Walter Cronkite used UNIVAC 2 to predict the 1952 election. Unable to believe the computer report of such a complete Eisenhower sweep, he failed to report it' (Rowe, 1999). There was an echo of this in the 2012 US presidential elections, in which pundits derided statistician Nate Silver's probability-based predictions before he correctly predicted the presidential race outcomes in every state plus the District of Columbia – and thus the re-election of President Obama. This was characterized on the internet as a triumph for big data. (See, for example, Triumph of the Nerds: Nate Silver wins in 50 States, http://mashable.com/2012/11/07/nate-silver-wins.)

By the 1960s batch processing with punch cards was common, statistical packages SPSS (a social sciences package), SAS (a statistical analysis system) and plenty of others were being invented and developed at universities, and 'local data services were in place at Princeton, Northwestern, at the Universities of British Columbia and North Carolina as well as at Wisconsin and Yale' (Rowe, 1999). The Inter-university Consortium for Social Research at Michigan (later to have a P for Political added to become ICPSR) was founded as a consortium of eight institutions; the Library paid the membership fee on behalf of Princeton. The Council of Social Science Data Archives was funded by the National Science Foundation, and cross-Atlantic meetings about data archiving began. The ASCII standard was established and the Unix operating system was invented (Rowe, 1999).

A data archiving profession was in place by the 1970s and the pace was clearly accelerating. IASSIST was organized at a meeting sponsored by the World Congress of Sociology in Toronto in 1974 and met in London, Edinburgh, Cocoa Beach, Toronto, Itasca, Uppsala and Ottawa. 'The U.S. Census released off-the-shelf data products, both aggregate and [microdata] sample data, and significantly there was a growing involvement of traditional libraries in providing data services. The American Library Association constituted a subcommittee to recommend rules for cataloguing machine-readable data files' (Rowe, 1999).

Canada and the Data Liberation Initiative

Research institutions in Canada, such as the Universities of British Columbia, York University, Western Ontario and Carlton University, were also fostering data libraries and data support services in the social sciences in the same decades. However Canada has long suffered from a lack of national data archiving infrastructure. No central data archive, whether government-funded or consortium-based, has ever been established.

Statistics Canada, the national statistical service, for many years had a liberal policy of making aggregate data, microdata and other types of administrative data (postal codes, geocoded files, etc.) readily available and affordable. However, with a Conservative government in the mid-1980s that policy changed radically and all types of data became very expensive, and in the case of microdata more and more restricted from the mid-1990s.

Nevertheless the Canadian presence in IASSIST has always been a strong one (apparently sometimes known as Can-IASSIST), and its members eventually fostered the Data Liberation Initiative (DLI) beginning in the mid-1990s to pressure Statistics Canada to increase the amount of data released as public use microdata files and to reduce steep pricing regimes for 'special' data files or microdata (Boyko and Watkins, 2011). The movement was fairly successful; there is now a DLI section on the Statistics Canada website listing DLI products and members (www.statcan.gc.ca/eng/dli/dli). DLI is defined as 'a partnership between post secondary institutions and Statistics Canada for improving access to Canadian data resources'. DLI's early lobbying for more affordable pricing and release to the academic sector of all standard products was achieved, but critically microdata have largely fallen out of the standard products categories, and are now mainly available in secure data facilities only.

European data archives take shape

Meanwhile, in Europe there was a different trend than in North America: this was for centralized rather than localized social science data archives to be established as the result of encouragement and investment by funding bodies. In the social sciences there was an interest in preserving the data that were being generated by research into voting behaviour, for example (a former director of ICPSR singles out the efforts of two European political scientists – Stein Rokkan and Erwin K. Scheuch (Rockwell, 2001). There was also an appreciation of the slow but steady accumulation of government-sponsored survey data.

In 1976, the Council of European Social Science Data Archives (CESSDA) was founded – then a loosely tied federation of national data archives, which co-operated with each other in resource discovery through a shared catalogue,