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

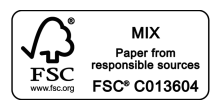
Facet Publishing is wholly owned by CILIP: the Chartered Institute of Library and Information Professionals.

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.

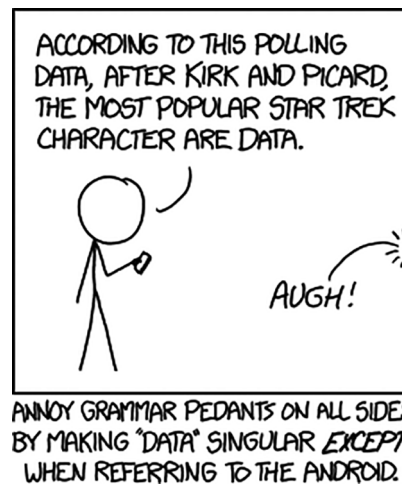


Figure 0.1 *Data: singular or plural?*

© 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 3

Supporting data literacy

Information literacy with data awareness

Academic librarians have excelled in promoting the information literacy agenda and developing bibliographic instruction as part of learning research skills. Not only is it quite common for library sessions to be a key part of new student inductions, but examples abound of successful partnerships with university instructors for getting library-based training into the classroom and coursework, and sometimes even assessed. Virtual learning environments, online courses and distance learning all offer new ways for librarians to interact with learners and teachers as well. A little bit of 'data awareness' can go a long way in extending traditional information literacy and bibliographic instruction programmes.

In some ways existing forms of library instruction lend themselves easily to the addition of concepts of data management and re-use. For example, in teaching about doing a literature search in a given discipline, librarians may give instruction in using standalone or online reference management tools, such as EndNote, Reference Manager, Zotero or Mendeley. For some disciplines in which the data used are mostly in textual form (e.g. law, history), such tools may even be the best method of conducting data management throughout a research project. For disciplines using other data types, further organizing options may need to be explored, such as those described in the Research Data MANTRA training course Organising Data (http://datalib.edina.ac.uk/mantra/organisingdata).

Bibliographic search methods may also be useful for teaching skills in data discovery. Students will not find all of the references needed for a literature search solely by consulting Google or Wikipedia, nor will they find all the supporting non-textual data for their research that way. Data librarians can teach the use of specialized data portals, disciplinary data centres, government statistical websites and repository registries alongside the use of

specialized publication databases to aid discovery of published literature and data. Besides, in the case of many online databases, licensed products, or datasets requiring permission to access, content cannot be indexed by Google, and so familiarity with potential sources is essential. Training in data sources may be offered to particular classes or as part of information skills programmes. Alternatively web-based resources may be offered, for example the Bodleian Data Library web pages (www.bodleian.ox.ac.uk/data).

Students can also be taught to look for citation trails to key datasets through publication lists of seminal articles. For example, where a publication has been written that is based on an original dataset, the author should provide instructions for obtaining the dataset for re-use, if not a complete citation to it. If nothing else the author's contact details may be used to track down an incomplete reference to a dataset (or the data librarian can help by contacting the author on behalf of the student). Some data repositories (such as ICPSR and the UKDA) offer lists of published articles and papers associated with datasets in their collection. These are especially useful for determining whether a line of enquiry has yet been pursued on a given dataset, for example to help determine an original research topic.

Promoting data citation

One area in which data librarians and others can easily incorporate data-aware information literacy in their instruction is by instilling the importance of data citation. At its essence, data citation is simple, and analogous with textual publication citation: in order to stand one's own work up to scrutiny a proper reference list must be provided, with enough detail for the reader to track down the relevant content of the referenced object. Just as one would not dream of putting a reference such as 'Smith, 2013' in a citation list or bibliography, it is surely unacceptable to print a table with the caption 'OECD, 2013' with no further citation appearing in the reference list. Yet this has been common and accepted practice until quite recently.

What has changed? As shared datasets, images, video clips and other non-textual digital objects become more valued in exchanges of scholarly communication in their own right, the provenance of these objects gradually becomes as important to the scholarly record as the peer-reviewed, published papers which describe and analyse them. This is not only important for the reader who wishes to track down a copy of the original object; it is equally important for the object's creator who wishes to receive career rewards based on the academic value of their work, as measured through citation counts and other impact measures that show the data have been recognized, consulted, downloaded or cited in other studies (including replication studies).

So data librarians wishing to support best practice in information literacy

can advocate 'proper' data citation along with well established bibliographic citation practices, stressing not only that it should be done but also advising on how it can be done. Some but not all major style guides provide assistance in how to cite data, so in many cases applied judgement is required. An interest group of the IASSIST has created a short, practical Quick Guide to Data Citation (www.iassistdata.org/community/data-citation-ig/data-citation-resources), which provides examples based on applying the general principles of three major style guides to an agreed set of minimum citation elements.

Data citation can be trickier than bibliographic citation as the data can vary so much: for example, citing small fractions of large datasets, citing a constantly changing database or map, citing underlying data elements going into a compiled dataset or data visualization, or perhaps citing an unpublished data object that has been shared informally. Some of these problems are being tackled by the research community, such as various working groups of the Research Data Alliance. Others can be solved through good examples. Statistics Canada has made a useful attempt to provide citation guidance for a number of different forms of data: tables, maps, graphs, published and unpublished datasets, e.g. microdata and custom tabulations (Statistics Canada, 2009). Good practice is encouraged by the fact use of their data for published analysis is conditional on proper citation. The guidance is also useful for building a citation for data of various types from any provider (Statistics Canada, 2009).

Most of what has been said above applies to human-readable citations. However it is just as important to be able to have machine-readable or computer-actionable citations. A website address or URL is an example of the latter, because it provides the means to access the reference (through a hyperlink or by pasting the string into a web browser). However the problem with URLs, as we have all experienced, is that they change or disappear – they do not tend to persist. The Hiberlink project carried out by Los Alamos National Laboratory Research Library and University of Edinburgh in 2013–15 investigated the extent of 'reference rot' – defined as a combination of 'link rot' or '404 Not Found' errors for URLs – and content drift, in which the referenced web page has changed its content or underlying structure since the citation was given. The project found that overall, 'One in five articles suffer[s] from reference rot' (Burnhill, Mewissen and Wincewicz, 2015, 56). One existing solution used by many publishers is to instead use persistent identifiers (or IDs), such as handles or DOIs, which provide a mapping from the URL to a persistent ID maintained by a central facility. This improves the chance that a resource will be found as well as fixed in an unchanged state, but only if the content continues to be hosted somewhere and the mappings are updated as websites change or publishers merge or go out of business.

The Data Citation Index, by Thomson Reuters, is an example of a newer

service which tracks data citations, though its impact on the academic community is not as well established as its sister service for publication references, Web of Science. A few alternatives also exist. Google Scholar has the merit of being free and more widely accessible. Also, it does not discriminate between textual publications and datasets, as long as it recognizes the source of the dataset as a publisher (such as a repository). Altmetrics and related services may also help to make the impact of data creation and citation more visible.

Finally, remind students to cite the data they create or collect themselves! This is 'Rule 5' from 'Ten Simple Rules for the Care and Feeding of Scientific Data': 'Link Your Data to Your Publications as Often as Possible. Whether your "data" include tables, spreadsheets, images, graphs, databases, and/or code, you should make as much of it as possible available with any paper that presents it. Your data can even be cited before (or without) its inclusion in a paper' (Goodman et al., 2014, 3).

Data discovery

Established researchers often know what datasets are available in their field of study, or at least the main sources and providers from which to seek them. They are immersed in the literature and knowledgeable about the activities of their peers, they know the institutes and principal investigators that are producing research relevant to their work. This is usually not the case with postgraduate students, even less so with undergraduate students, and nor is it the case with researchers exploring the boundaries of their disciplines or doing cross-disciplinary research. Fortunately for data librarians, it is possible to become knowledgeable about sourcing datasets in a given field without being an expert in that field. A simple prompt or series of suggestions of possible avenues to explore enables them to be in a position to come to the rescue of researchers needing to source existing data.

Clearly data librarians will provide online guidance for data discovery as well as offer face to face support. Existing data libraries offer wonderful examples of creative online outreach material of all sorts. Just to give two examples from the authors' own workplaces, there is a library guide 'Data and Statistics for the Social Sciences' from the University of Oxford (http://ox.libguides.com), and the portal Finding Data from Edinburgh University Data Library (www.ed.ac.uk/is/finding-data). As both of these lean towards social science data, here is another example by Brian Westra, University of Oregon Libraries: 'Chemistry – Handbooks, tables and reference data' (http://researchguides.uoregon.edu/c.php?g=366305&p=2475534).

Staff and students may be reluctant to schedule a face to face appointment asking for support for various reasons, so making them aware you are

available for consultations in a variety of ways is useful. Scheduling office-hours, drop-in sessions or surgeries is one approach. Building relationships with teachers and supervisors of research students is another: helping teachers to construct a data-related assignment, introducing a special resource in the classroom, or putting a reference list in a syllabus are all ways to make your service known and contact details readily available to inquiring students. By offering support to teachers you may also be encouraging them to raise their game in teaching by providing live data sources, in turn encouraging students to create their own research questions and to learn how to answer them.

One of the authors was involved with a large UK study on barriers to the use of data in the classroom in social, geographic and health-related disciplines, those thought to be the biggest users of secondary data, from February 2000 to June 2001. In institutions with no dedicated data support service, the situation is unlikely to have changed very much:

> The survey uncovered a number of barriers experienced by teachers in the use of these services, namely a lack of awareness of relevant materials, lack of sufficient time for preparation, complex registration procedures, and problems with the delivery and format of the datasets available. These problems were elaborated in open-ended comments by respondents and in the case studies of current teaching practice. . . A compounding problem is the lack of local support for teachers who would like to incorporate data analysis into substantive courses. A majority of the survey respondents said that the level of support for data use in their own institutions was ad-hoc. Peer support was more common than support from librarians and computing service staff, and over one-third received no support whatever. The top three forms of local support needed were data discovery/locating sources, helping students use data, and expert consultation for statistics and methods (for staff).
>
> (Rice and Fairgrieve, 2003, 19)

The good news for librarians is that here is a demand that can be readily addressed in the form of support for data discovery. Albeit, the demand may have to be stirred up where academics do not expect data support to come from the library. In Chapter 5 we explore further how libraries can begin to establish credible research data services where they have had none before.

Data reference interviews

Once the enquiries are coming in, whether from undergraduates, postgraduates or academic staff, how do data librarians match users to the data required, to make liberal use of one of Ranganathan's Rules: 'Every

reader his book' (https://en.wikipedia.org/wiki/Five_laws_of_library_science). The answer is largely already known to the trained librarian – through a reference consultation or interview.

The trick to giving reference interviews well is to realize that the question or data query being asked by a user may not be exactly what is required. Through a process of active listening and prompting for more information using open and closed questions, the librarian is able to translate the user's initial query into a question that can be looked up using available sources. By avoiding shortcuts, jumping to conclusions or making assumptions, the librarian spends a little extra time getting the query right, and saves the time of the user by finding the answer to the right question instead of an irrelevant one. An example is a UK-based researcher who asks how to log on to get data files from ICPSR, when what she really wants is access to the Canadian Election Surveys microdata files, which are available freely on the web as well as from the ICPSR and a number of other data providers. Even this example is simpler than most, because the researcher already knew the title of the dataset she required. Data-related queries tend to be more difficult, more time consuming, and sometimes even bewildering to the non-subject expert trying to provide assistance to the experienced researcher. The following advice was given to organizational representatives of the UKDA in 2000 (primarily librarians and IT support staff), with respect to conducting data reference interviews:

> There are a few angles that data supporters can take in leading a reference interview. First, be sure to determine the level of enquiry the user is committed to undertake (tactfully, of course). Is the work for a complex analysis or a class paper? Will it be part of a planned research project lasting several months, or background statistics for a quick and dirty proposal with an encroaching deadline? Is the user experienced with using secondary data or a novice? Is s/he well-grounded in the subject or methodology, or traversing new terrain? What statistical packages is s/he prepared to use? You are likely to encounter some users again (and again) as they progress through the stages of their analysis; others may have just a one-off request. Finding out this information at the outset will help you anticipate future needs, and will prevent wasted efforts, such as referring users to published sources when they want the raw data, or ordering data files when all they want is a printed table from a book.
>
> (Rice, 2000, 8)

In order to help a user find an appropriate dataset to answer a research question successfully, you may need to prompt the researcher for further information about their requirements, such as:

- general subject or research question
- unit of analysis (e.g. country or other geographical area, household, individual, company)
- key variables required
- required format of the values (e.g. is age by year essential or do age groupings suffice?)
- dates or years required, whether very recent data are required or not.

In practice, if the researcher is starting with an original research question rather than a known study or dataset, it is likely they will be forced to make trade-offs about all of the details they require: for example, settling for an older date of collection than wished, or a larger geographic area than desired, or a proxy variable because data for the preferred variable do not exist. Factors that affect the need to settle for such trade-offs include amount of time the user has allocated to complete the work, their level of data analysis proficiency, willingness to struggle with a raw dataset, and perhaps whether they have the qualifications to apply for use of more sensitive datasets with special conditions.

For example, to apply for Eurostat microdata, which requires completion of an application form and a confidentiality agreement before access, the user must also present proof they are a member of staff of a recognized research entity. After submitting the application they must wait a number of weeks for each country's statistical agency to sign off their approval of the research project before they are given a copy of the data. In many cases the researcher is unaware of these hurdles at the beginning of the process, and you may find yourself needing to manage expectations or locate a less ideal but more accessible source.

Categories of data

It is enormously helpful to narrow down the type of data required at an early stage, especially if microdata – data about individuals – or macrodata – data aggregated at a higher level, such as country or region, are expected. A user seeking 'labour statistics' or 'data on fertility', for example, may require either. Further discussion is needed.

If the research question requires survey data (responses from questionnaires), could a single, one-off survey or poll answer the question or does it involve change over time? If the latter, the options may be limited to known survey series that repeatedly sample the population at different points in time – (known as cross-sectional surveys, such as *British Social Attitudes*).

However, a study of social mobility, for example, may require data points at different times in individuals' lives – such as their parents' occupation at

birth, their grades in school, whether they completed higher education, and their salary or occupation at a given age. Perhaps only longitudinal studies, where the sample is made up of a constant set of individuals who are traced over time, can address these sorts of research questions. Examples include both panel studies such as the UK's long-running household panel study, *Understanding Society*, and 'cohort' studies, which follow a sample of a particular generation over time, often from birth, such as *Growing up in Scotland*.

Another category of data, often consulted by economists and business researchers, is known as time series: as the term implies, these data usually consist of a single variable over several points in time, often arranged in columns, whereas comparisons (e.g. between companies, equities, countries, regions, etc.) are listed in rows. Data queries requiring time series often begin, 'I need data for X (variable) going back 20 years/quarters/months, etc.'

Geospatial data, on the other hand, focus on comparing one or more variables for different but comparable regions, government districts, zones or countries, often using a standardized code, such as the European NUTS 1, 2, 3 (where the numbers represent varying geographic sizes of regions), Census codes, or postal code. This will also often come in a row and column format, with geographic names or codes in the rows, and variables – often including an X and Y geographic co-ordinate – in columns. Often the researcher will wish to use the geographic co-ordinates to map the data, for example in a GIS analysis package, often by adding a new layer to a base or topographical map that shows some basic features. The researcher may build their map either with points (for example, to represent the centre of the region) or boundaries, in which case the extent of a numeric value can be displayed in shades of colour (this is called a choropleth map). There may or may not be a time series component, but if there is, this would normally result in more than one map.

Top tips for the reference interview

Since every research question is unique (unlike every classroom assignment), it is easy to feel daunted when confronted by a new one. These are some helpful hints to guide you through a difficult reference interview:

- Buy yourself time by asking more questions before trying to come up with a source; avoid making assumptions about the user's requirements, prior knowledge or viewpoint.
- Find out if the user is basing their query on a published article; ask for the citation or a copy to help you with the context. If a student, ask who is their teacher or supervisor.

- Ask the user to explain acronyms and jargon they use in their language; you do not have to pretend to be an expert in their area of study to help them.
- Take notes and write down key phrases as the user speaks (if meeting in person or on the telephone).
- Even if you are unable to find the perfect source for your user, you can probably give them some useful starting points for their search, based on your knowledge of data sources, or that of your peers.
- Do not be afraid to take time to think, search and consult others; always take the user's e-mail address for future contact or to follow up.
- If you remain stumped, resort to asking others: immediate colleagues, peers at other institutions, government statistical agencies, data providers and publishers. Once you have done your homework and ruled out obvious contenders you can also post to a library mailing list or the member list of IASSIST.
- If the query is about using a dataset rather than finding one, take time to read the documentation, try out the interface yourself or reproduce the problem before turning to others for help.
- If the user does not voluntarily let you know their query has been satisfied, follow up in a reasonable amount of time to see if you can offer further assistance.

When doing data reference work there are often ample opportunities when helping a user discover or work with data one on one gently to improve their understanding and skills in data literacy. While you are looking up data resources, for example, speak aloud about why you are going first to this source (government agencies often have authoritative data on this subject) or that source (the UKDA has very good metadata for searching at the variable level), so the student learns how to evaluate data from different sources.

Many datasets and published databases are useful to researchers and students because of the statistical content they contain. In the case of microdata (individual-level data with cases displayed as rows and variables displayed as columns), summary or descriptive statistics must be created from the microdata through statistical analysis or spreadsheet software. In the case of data taking the form of pre-compiled descriptive statistics – aggregate data – such as some government statistical serials, the descriptive statistics have been produced for public consumption. This may be in the form of tables and charts. Regardless, the user needs a basic level of statistical literacy in order to make best use of the information received. A higher level of statistical competency might be required if the published tables are not up to the task of answering the research question and so the original dataset needs to be queried directly, for example to produce new tables from a single year of age

instead of age groups, or with a cross-tabulation of two variables that do not appear in the pre-compiled version.

What has statistical literacy got to do with it?

How is statistical literacy related to data literacy? There can be no single answer to this question in these fast-changing times, but the IASSIST social science data library and archives community has long held the notion that to be data literate one must be both numerate and possess some level of statistical literacy. Numeracy has various definitions but is sometimes treated in common with functional literacy – the ability to do basic arithmetic, or perhaps the absence of numbers anxiety. In addition to being able to manage one's household budget and other common tasks associated with adult numeracy, news media commonly convey numeric information about rates, ratios and percentages – so a non-expert level of numeracy is essential for responsible citizenship. For students and researchers to learn to be comfortable using numeric data sources they must in the first instance be numerate or 'comfortable with numbers'.

Statistical literacy goes beyond basic numeracy to being able to understand and evaluate statistical results as they appear in research literature. It is not necessarily as deep a knowledge as statistical competency or proficiency. In part it means simply being able to understand information displayed in tables and graphs. It also involves understanding other statistical concepts well enough to judge the claims being made in research literature. Examples include understanding how response rates, sample types and sizes contribute to the accuracy of statistical claims, and measures of statistical 'significance', or the likelihood that finding a given result is not due to chance, as commonly expressed in P values. Furthermore, to critically understand statistical claims that others make (which could be one definition of statistical literacy), it is necessary to understand how a sample was selected or chosen, and therefore whether or not it is representative of a larger population. (Note that this is the kind of additional documentation that the original data creator or claim maker needs to provide to make the data understandable and the analysis comprehensible.)

Unfortunately many statistical concepts – to do with odds, chance, randomness and probability – must be learned, and are simply not intuitive. As David Spiegelhalter, a statistician at the University of Cambridge who specializes in conveying concepts of probability to the public, has wryly written, 'Why do so many people find probability theory so unintuitive and difficult? After years of careful study, I have finally found it's because probability is unintuitive and difficult' (Spiegelhalter, 2011).

Milo Schield, who established and edits the website Statistical Literacy

(www.statlit.org) aimed at improving the statistical education of undergraduates, claims that what information literacy, statistical literacy and data literacy have in common is the evaluation of information: 'All librarians are interested in information literacy; archivists and data librarians are interested in data literacy. Both should consider teaching statistical literacy as a service to students who need to critically evaluate information in arguments' (Schield, 2004, 6).

In studying the behaviour of his own business students at a liberal arts college in Minnesota, Schield has made a number of interesting observations about the tendencies of statistically illiterate readers. For example, when reading a research article, some skip over tables and charts, preferring to take in the information in narrative form. Unfortunately, much of the critical thinking involved in judging the importance of tables and graphs comes from studying all of the elements going into the visual display of the numbers.

Bear in mind it should be easier to teach statistical literacy, skills to evaluate statistics used as evidence in scientific and social arguments, than full statistical competency. To be statistically competent one must know how to apply statistical techniques correctly to research problems, which requires more knowledge and experience. Combined with a healthy dose of critical thinking, teaching statistical literacy can even be fun. What nonsense can occur from reading row percentages when only column percentages make sense? What fun can be made of looking at completely unrelated variables that have 'statistically significant' correlations simply by coincidence? Statistical fallacies are commonly found in news media articles over-simplifying scientific results, such as equating correlations with cause and effect. Similarly, 'bad' charts can be critiqued for common tricks such as not starting the Y axis at zero, or otherwise exaggerating small differences.

There is no shortage of material for this kind of training activity. Full Fact (fullfact.org) is one dedicated organization, based in the UK, which focuses on statistical and other kinds of fact-checking to correct misleading claims by journalists and politicians about current affairs. The home page has a range of recent news stories corrected or confirmed. Academic blogs by critical thinkers, such as David Colquhoun's Improbable Science (www.dcscience.net), can be useful sources as well. Ben Goldacre, a physician and blogger on Bad Science (www.badscience.net), has been so successful at this sort of critical writing he first became a *Guardian* newspaper columnist and then a best-selling author.

Data journalism and data visualization

Data journalism sites such as the Guardian's datablog (www.theguardian.com/data) offer good examples for making statistics visually compelling, and

telling stories with data. Data visualization sites like Flowing Data by Nathan Yau (flowingdata.com) and Information is Beautiful by David McCandless (www.informationisbeautiful.net) are equally inspiring. Indeed, students' attention may be captured more easily with workshops on data visualization or data journalism than data or statistical literacy per se. Duke University Libraries Data and Visualization Service (http://library.duke.edu/data), as the name indicates, foregrounds data visualization tools and workshops along with topics in data analysis and data management. Studying and creating infographics may be another way to capture the imagination of students who are reluctant to learn about statistics.

Topics in research data management

Until research funders began promoting the importance of RDM – and sharing, as a way to increase efficiency of scarce funds for a growing scientific agenda in the last decade (see Chapter 7) – the support that did take place occurred in a very individualistic way, and was hardly taught by anyone, let alone librarians. So it is not surprising that data librarians may need to work to create demand for such training. However basic data management principles and practices are not difficult to teach, and can easily be incorporated in an information literacy training programme. The overlap with furthering open access to publications and the desirability of getting datasets into the preserved scholarly record should be enough to justify the inclusion of RDM in the information literacy training programme, especially for postgraduates, who are likely to be creating new datasets in some form. As with open access, librarians are in a position to affect academic culture change, which unlike rapid technological developments can be notoriously slow. Presenting the benefits of good practice in data management and sharing to early career researchers and students before less desirable habits are formed is one way to speed it up.

Fortunately, there are many benefits of undertaking data management, especially for researchers who have never considered it in a conscious way before – not least, avoiding the disaster scenario of losing all of one's research data owing to lack of a back-up plan. If this seems unlikely just go do a Twitter search on 'lost USB stick', and you will no doubt find similar messages to the one shown in Figure 3.1, which appeared in an Edinburgh neighbourhood newspaper, the Broughton *Spurtle*, on 16 August 2015.

Of course the number of topics to address in RDM training partly depends on the length of the training, the audience and indeed the trainers; librarians are comfortable with slightly different subjects from those taught by IT professionals or statisticians, for example. The University of Edinburgh conducted a training programme in RDM for its academic service librarians

in 2012–13 covering topics that it was agreed were within the sphere of the librarians' professional interests and, potentially, areas in which they could provide support for researchers or training for students. These were:

- data management planning
- organizing and documenting data
- data storage and security
- ethics and copyright
- data sharing.

Figure 3.1 *Lost USB stick*
Source: www.broughtonspurtle.org.uk/news/please-help-find-lost-usb-stick.

Data management planning is perhaps the most important topic on which to focus. (There is more about support for data management planning in Chapter 6.) The other topics focus more on the practice of managing data throughout the project and what is involved with sharing data – presumably at the end of their project, after a thesis has been written, or a paper has been published. Of course, data management plans are not always required for postgraduate research; it is much more common to be required by a funded grant. So unless you are able to work with the graduate schools of departments to require a data management plan (DMP) of some kind as part of a thesis proposal, training on writing a plan aimed at students may be little more than awareness raising. However, by covering the other practical topics – many of which students are unlikely to be exposed to in their postgraduate training – they will be more likely to pick up the immediate value of what they are being taught.

The closer the training can come to their actual experience of working with data, the better. If students want to learn how to organize a database efficiently, they will not be interested in the finer points of using a lab notebook. If they are compiling information from textual references for a paper about changes in policy or legislation over time, they will not wish to learn about how to transform image data from one format to another. This is why subject librarians can be invaluable partners in designing and delivering appropriately focused data management training.

While designing domain or discipline-specific data management training can be time consuming as well as challenging, many lessons in data management really can be presented in a more generic fashion. This is because there are commonalities across domains for working with data as evidence or

as proxy for ephemeral content (a dance performance) or analogue content (rock samples from a geology field trip). Part of what is common is the digital nature of the data. As disciplines have gradually 'gone digital' in their practice, academic lecturers naturally continue to emphasize the subject content of doing research rather than the practicalities of working with similar data in digital form, or with new tools or software. It is therefore sensible for data librarians or IT professionals to focus on the digital nature of the data and the skills needed to work with digital material – both data and code.

What does a well rounded data management training programme look like? The University of Edinburgh's open, online RDM training course, or MANTRA, is well used by those wanting to brush up their data management skills in a range of disciplines, and is employed by institutional trainers in many countries to supplement face to face and other bespoke training.

MANTRA consists of eight main topics, each designed to take about an hour to step through, and including interactive quizzes, video clips of researchers describing how they have dealt with data management issues, and further reading. An additional section on data handling extends the skills into working within four particular software environments. On the 'About' page of the MANTRA website (http://datalib.edina.ac.uk/mantra/about.html) the primary benefits for research students of taking the course are listed:

1. Understand the nature of research data in a variety of disciplinary settings
2. Create a DMP and apply it from the start to the finish of your research project
3. Name, organise, and version your data files effectively
4. Gain familiarity with different kinds of data formats and know how and when to transform your data
5. Document your data well for yourself and others, learn about metadata standards and cite data properly
6. Know how to store and transport your data safely and securely (back-up and encryption)
7. Understand legal and ethical requirements for managing data about human subjects; manage intellectual property rights
8. Recognise the importance of good RDM practice in your own context
9. Understand the benefits of sharing, preserving and licensing data for re-use
10. Improve your data handling skills in one of four software environments: R, SPSS, NVivo, or ArcGIS.

Although there is a nominal ordering of topics, MANTRA is designed for mixing and matching, letting readers dip in and out of topics of interest – one strategy for coping with learners from a wide variety of backgrounds.

The avatars on the home page guide users at varying career points towards suggestions of topics to look at first. The inexperienced research student who

may not have worked with data before is pointed towards 'Research data explained', which introduces the wide variety of research data in existence, with a smattering of disciplinary approaches to using data. The career or post-doctoral researcher is already well immersed in data use, but may not have written a DMP before, so is pointed to 'Data management plans'. The senior academic, who is experienced in all stages of a funded research project and supervises postgraduate students, is pointed towards 'Sharing, preservation and licensing', because they may be less familiar with publishing data than papers, and towards the 'Data handling tutorials' in case they can use them in a practical workshop with their students. A fourth avatar is for the information professional, who may have the most to learn in the unit on data protection, rights and access, but may also be interested in organizing a few colleagues to study in a small, supportive group by using the 'DIY Training Kit for Librarians'.

MANTRA was one of the early arrivals of RDM training, launched in 2011 as the deliverable of a funded project in the first Managing Research Data Programme of Jisc, a UK provider of higher education services, and regularly updated by the Data Library team at EDINA, Information Services, University of Edinburgh. Training materials in RDM are now abundant, and some excellent training materials can be found on the RDM pages of several institutions. We list here a few highly recommended training resources or portals; readers are asked to observe copyright and licensing conditions when re-using materials:

- DCC's Resources for Digital Curators at www.dcc.ac.uk/resources.
- New England Collaborative Data Management Curriculum at http://library.umassmed.edu/necdmc/index.
- DataONE (Data Observation Network for Earth) at www.dataone.org/education-modules.
- FOSTER (Facilitate Open Science Training for European Research) training portal at www.fosteropenscience.eu.
- IASSIST Resources: Data Management and Curation at http://iassistdata.org/resources/category/data-management-and-curation.
- companion page for the UK Data Service's book *Managing and Sharing Research Data*, with related presentations and group exercises at www.ukdataservice.ac.uk/manage-data/handbook (Corti et al., 2014).
- Essentials 4 Data Support, from Research Data Netherlands (in English or Dutch) at http://datasupport.researchdata.nl/en/.
- RDM and Sharing MOOC (massive open online course) at www.coursera.org/learn/research-data-management-and-sharing (a free Coursera MOOC developed and delivered by the University of North Carolina-Chapel Hill and the University of Edinburgh for researchers and librarians. A certificate of completion is available for a small fee).

Training in data handling

Research students may agree to study topics in data management but what they really want to be doing, and therefore learning, is how to analyse their data. The reason the ninth module of MANTRA focuses on data handling skills in particular software environments is that the authors saw data handling as an intermediate skill between generic data management skills (which overlap with basic computer literacy) and data analysis – which is more domain-specific and normally taught within a curriculum.

Data handling, or data manipulation, is at the heart of pre-analysis data preparation – such as data cleaning, recoding and creating documentation, merging files based on common variables or cases, or exporting from one type of software and importing into another without losing data content (format migration). It is equally at the heart of post-analysis processing – preparing a 'golden copy' master of the data on which results are based for long-term preservation, documenting changes made to data files, and depositing in a data repository in an appropriate format.

Data librarians who provide help with research datasets in particular software packages tend to have or acquire excellent data handling skills. They are in a good position to provide training in these skills, the very skills which tend to get left out of curriculum-based data analysis courses and IT-based 'introduction to X software package' workshops.

Depending on one's background, there are a number of attractive data-related training courses that data librarians can and do offer their local communities. By partnering with others in the library or IT centre, academics or even outside the university, such as local or central government, one can combine knowledge to provide extremely attractive data-related skills-based courses.

Just a few imagined examples include Preparing a Successful Data Management Plan, Working with [the latest] Census Data, Data Visualization Basics, Mapping Your Data, Data Journalism 101, Data Preparation and Analysis in [insert your favourite software package here], Preparing your Data for Publication, Sharing Your Data for Impact, Making Your Research Reproducible and Keeping Your Data Safe.

Reskill yourself

There are many opportunities to not only create training but also receive it yourself. School of Data (http://schoolofdata.org) is an open educational resource that focuses on the skills needed to use government-linked open data, and 'works to empower civil society organizations, journalists and citizens with the skills they need to use data effectively – evidence is power!' Data Carpentry 'develops and provides data skills training to researchers'

based on the successful model of Software Carpentry. Data analysis and visualization in R, OpenRefine for cleaning datasets, database management with SQL (Structured Query Language), and coding in Python are hot topics currently covered by Software and Data Carpentry. Various summer schools focus on honing data-related skills as well. ICPSR and the UKDA often host summer schools in data analysis and data support in Ann Arbor, Michigan, USA, and Essex, England. The Research Data Alliance (RDA) and CODATA (Committee on Data for Science & Technology) have begun hosting a week-long School of Research Data Science in Europe, as well. Sources such as Lynda.com offer multiple online training tools in various kinds of software. Conferences such as IASSIST, iPRES (International Conference on Digital Preservation) and the International Digital Curation Conference offer pre-conference workshops on data and curation-related topics.

Supporting data literacy is an ongoing objective of the data librarian. It will also involve the use of skills that raise your profile with researchers and deepen your involvement with their work in many cases. Inevitably the skills that stead you well now will go out of date or need augmenting. Keeping up your own data literacy involves not only going to conferences, reading the literature and monitoring trends on social media, but actually learning new skills. Data skills are bound up with software skills. Software is not static, but changes with each new edition of the program, introducing new capabilities, and dropping others deemed no longer to be necessary, and you must keep up with these changes. Also, software at its essence is code; there's no avoiding learning a bit of coding to handle data well. If you insist that you are 'not a programmer' you will not only limit yourself forever to the limitations of graphical user interfaces, but you will also miss some wonderful opportunities to interact with young people who are coming together to share skills in open source communities. Although some of these initiatives may not be coming from the perspective of a researcher, they all share enthusiasm for the data-driven world we are inhabiting, and a desire to be in control rather than controlled by it. As a data librarian, learning on the job is a necessity – but also a privilege.

Key take-away points

- Basic data literacy can be incorporated in mainstream library instruction and information literacy training.
- Librarians and data librarians are in a position to promote good data citation practice, even where examples in style guides are lacking.
- A relatively easy way for data librarians to help researchers is in sourcing existing datasets.
- Incorporating live data into a teaching setting is time consuming; data

librarians may be able to assist those in teaching roles by creating fit-for-purpose teaching datasets.

- As with reference enquiries, data enquiries need not be taken at face value; dig for additional context before trying to answer the question. Follow up to find out whether the information offered was useful or more help is required.
- By being aware of different data types (published statistics, microdata, macrodata, survey data, longitudinal data, geospatial data), the data librarian is better equipped to match a researcher with an appropriate data source.
- Numeracy is the bedrock on which statistical literacy can be learned. Statistical competence goes beyond being able to critically understand statistical claims, to being able to make them.
- Data management planning helps keep a research project on track from start to finish – avoiding disasters such as lost data.
- Data librarians can cover a range of data management topics, appropriate to the level and discipline of the group – from planning, to file management, to documenting and controlling versions of data, to storing and transporting data securely, to legal and ethical requirements of data collection, to applying metadata for sharing and preserving data. Free, vetted training resources on these topics abound.
- Challenge yourself to learn new data-related tools and software. Take time to play with data.

Reflective questions

1. Why is it becoming more important for researchers to cite the data they use in their work?
2. How might you approach a teacher to offer your support to their students in the use of datasets?
3. What do you consider is an appropriate amount of time to devote to a data enquiry?
4. Which do you think is more important for being statistically literate: numeracy or critical thinking skills?
5. Which data management topics would you be comfortable teaching yourself, and which would you call in an expert or colleague? To which groups at your institution?
6. How can you fit learning new data-related skills into your schedule?