1

Open data

Introduction

'Open data' is not the result of a single, idealistic movement that has resulted in the opening up of vast amounts of data from a host of different organizations, instead it is the result of numerous individuals and organizations interested in data being made publicly available for both selfish and selfless reasons. There are large international organizations that recognize the publishing of data provides an opportunity for them to benefit commercially from the skills of developers and experts outside their organization. There are governments who are not only responding to calls for increased transparency from members of the public, but also recognize the economic potential of the data they own, as well as the possibility for savings in public departments. There is the scientific community, where open science offers the opportunity for reducing the duplication of work and enabling faster and cheaper progress for scientific discovery. Also, there are significant sources of data that can be opened up by libraries to provide their users with better information services, although this is an area that has hardly begun.

This chapter looks at the breadth and context of the open data that is being made available by academia, government, industry and the library community. The scientific community has a long tradition of publicly publishing findings within academic journals, although this is a very small proportion of the information and knowledge that is actually collected. The commercial sector has been one of the leaders in making data available online, and trying to tap into the potential of the wisdom of the crowd that has been demonstrated so ably in the open-source community. Over the past two years governments around the world have been making huge efforts to make public data available, and not only is it a trend that is set to continue, but also some governments are set to be leaders in the open-data field in comparison to the commercial sector. The library community also has a well

established tradition of sharing information, and although this has principally been restricted to those within the library community, there are increasing attempts to share the data more widely. The scientific, commercial and government sectors do not work in isolation, but rather interact with one another as well as the wider public in the innovation process (Etzkowitz, 2008), and it is important that data is opened up and shared amongst all sectors of society if it is to reach its potential.

The publication of all this data has no point unless there are people who can make use of it. In the same way as the library has traditionally facilitated user access to documents, bringing together a collection few individuals could afford or had the requisite skills to locate on their own, now the librarian can facilitate access to the data that users could not otherwise make use of. Facilitating access to the web of data, where the data will not necessarily fit into the neat document formats that users are used to, can be seen as just as important as facilitating access to the web of documents, if not more so.

Open science

Science helps people understand the world through the systematic collection and analysis of data. Open science has been defined as 'making methodologies, data and results available on the internet, through transparent working practices' (Lyon, 2009); using the potential of the web to move beyond the limitations of traditional technologies. Progress in science is achieved through the sharing of findings with one another, and technology has always played a pivotal role in making science more open; as Bernal noted (1939), 'The growth of modern science coincided with a definite rejection of the idea of secrecy.' The scientific journal has not only been at the centre of the spread of scientific ideas, but also the principal means by which scientists have received recognition. A scientist's publication list not only provides a quantifiable measure of their output, but through citation analysis a quantifiable measure of their impact in their field. Although books and journals have been a remarkably successful format for the distribution of research findings they have a number of limitations: they require a work to be complete before it is published; have limited scope for feedback and later amendments; whilst publishing and printing costs require a significant print-run to be worthwhile.

In many ways 'open science' is a misleading term, suggesting the idea that science is either open or closed. In fact, as we have seen throughout the

history of science, openness is a continuum. At its most closed, science can be a secret world, more akin to alchemy than our understanding of modern science, where not only are the results held secret, but even the existence of the experiments may be unknown. At its most open the whole scientific process may take place in the public arena or under public scrutiny, at least as far as technological limitations allow (e.g., people are not fully aware of how or why they do the things they do, and we have no way of recording their thought processes). Between these two points there are many steps, from publishing the results years later, to having an ongoing discussion throughout the research process.

Despite a large increase in the number of journal titles throughout the 20th century as publishers responded to the growth in the number of scientists and increasing specialization, at the start of the 21st century the suitability of the traditional publishing process has increasingly been questioned. Both because of the economics of the journal publication process, and the large amount of data that traditional publishing processes fail to capture.

Open access

Developments in information technology at the end of the 20th century coupled with rising journal costs have led to questions about the appropriateness of the existing journal model. For many people the traditional publishing system seems inherently unfair: public taxes pay for much of public science, scientists write research papers and hand over the copyright to publishers for free, other scientists peer-review the research for free, after which scientists have to pay (at least via their institutional subscriptions) to get access to the work. When the price of the published work is beyond the means of many libraries in developed countries, let alone those in developing countries, or the members of the public whose taxes have paid for much of this work, questions start to be asked about whether there is a better way.

In the words of John Willinsky (2006, xii): 'A commitment to the value and quality of research carries with it a responsibility to extend the circulation of such work as far as possible and ideally to all who are interested in it and all who might profit by it.' Such openness underpins Merton's (1973) essential principles of science: communism, the idea that scientific results are the property of the entire scientific community; universalism, the idea that all scientists can contribute to scientific debate; disinterestedness, the idea that scientists should separate personal belief from science; and organized

scepticism, the idea that claims should be critically appraised.

It is clear in our increasingly educated and connected world that those who are interested in and might profit by research extends far beyond a limited group of Western scientists, and that there is a need for research findings to be circulated beyond the research library walls. The open-access publication of science has not been without some critics. Some have been concerned about how the public may misinterpret or misuse findings, by taking results out of context or by exaggerating the implications or conclusions. Others have been concerned about the effect of open access on the publishing process, possibly undermining the peer-review process and the established journals of record (Worlock, 2004). Most recognize, however, the potential for faster scientific discovery, and the wider public good, when we reduce the barriers for information exchange.

For many the argument about open access to the traditional journal article is now over, with many research-funders insisting that funded research results are freely available online, at least within a pre-print format. The question now is what the best way to make them freely available is. Whilst there is an increasing need for access to research, it is necessary to find a sustainable economic model, that does not risk the quality of the research. Before the invention of the web the additional unit costs of each copy restricted the circulation to those who could afford to pay for it. Online, where each additional copy can be provided for free, or as close to free as to be negligible, there is only a fixed cost to be covered. The developments in web publishing have enabled the adoption of the various open-access models, although there is not necessarily a single best solution to extending the circulation of work, as attempts are made to keep costs down and the quality of work up. The two main variants of open access that have emerged are the so-called green road to open access and the gold road to open access (Harnard et al., 2004). Green open access is where journals allow self-archiving in institutional repositories, or on personal websites. Gold open access is provided by publishing in an open-access journal. There are also hybrid open-access journals which only provide gold open access to those individual articles where the author has paid a publishing fee.

There are significant variations in different fields, both in terms of the proportion of published papers that are freely available online and the routes that are taken to make these papers available, although overall the proportion of papers freely available online continues to be a minority. Björk et al. (2010) found that for a sample of research papers only 8.5% were freely available on

the publisher's website (i.e., the gold), with a further 11.9% freely available after doing a Google search. At the same time as recognizing the potential from open research data, it is important to realise the difficulty in getting even a widely accepted change for the better put into practice.

Importantly, the move to the web not only enables open access and the recognition of public rights to access, but also the provision of more innovative services.

E-science and e-humanities repositories

Whilst the idea of a public right to access data is indelibly associated with the rights and advantages of open access, providing access to the raw data as well as the journal article is by no means a new idea. Following the establishment of data banks in the United States, Germany and Holland, the UK's first data archive was established in 1967. The creation of a data archive not only recognized the need for the preservation and curation of high-quality data for secondary analysis, but was also a reaction to a perceived loss of information to the USA who bought the information (UK Data Archive, 2007). Unsurprisingly the technologies have changed considerably since the Social Science Research Council Data Bank, as the UKDA (UK Data Archive) was then known, was established, most noticeably in that the data sets that were once deposited and issued on punch-cards are now available on the website. In many ways, however, the archive continues in its traditional role: it is a curated archive, aimed at the traditional researcher, and deals with post-research data sets.

Professional curation of digital archives is designed to increase the accessibility, usability and reliability of data sets. Supplemented with extensive and standardized metadata, data is made available in the most appropriate format, with the legitimacy of the data source verified before it is included within the collection. Such curation is, understandably, very resource intensive, and is not necessarily the most appropriate way to meet the increase in the quantity of data available to be deposited at the start of the 21st century any more than a widely useful web archive could be created by manually selecting websites for submission (although a number of such examples do exist). It is not surprising to find, therefore, that the traditional curated collections have been joined by an increasing number of self-archiving collections, for example, the UKDA-store at http://store.data-archive.ac.uk was launched in 2008 for the self-archiving of data by ESRC

(Economic and Social Research Council) funded researchers. As well as self-archiving collections potentially containing less consistent metadata, they are not necessarily providing the data in the most appropriate format (too often in proprietary formats thus limiting access). Nonetheless they are able to scale to meet users' needs.

Whilst the web means that there are fewer barriers of access to the data, the UK Data Archive continues to be aimed primarily at the researcher with specific research in mind rather than the casual user. Not only is it necessary to register to gain access to the data, but it is necessary to register for each new use of data, providing a usage title and a brief description of usage of at least 30 words. Whilst this provides researchers with useful case studies of how others are making use of the data, and may not be thought to be overly arduous, it does not encourage a sandbox area for the serendipitous discovery and combination of data. Often people will not know exactly how they are going to use the data, or even if they are definitely going to use the data, until they start playing with it.

Scientific data archiving at the start of the 21st century is an increasingly complex and dynamic landscape with numerous players with various interests in the research process. As well as research-funding agencies financing archives, there are also archives associated with journal publishers (e.g., Dryad at http://datadryad.org) and with particular fields (e.g., www.pangaea.de for geoscientific and environmental data), data archived within institutional repositories (e.g., Edinburgh DataShare at http://datashare.is.ed.ac.uk), whilst there are also commercial data stores not specifically aimed at the academic community (e.g., Many Eyes at http://manyeyes.alphaworks.ibm.com/manyeyes). As with the outsourcing of any service, there is always the risk of a service disappearing, and this was emphasized in 2008 with the depreciation of the Arts and Humanities Data Service (www.ahds.ac.uk), which is no longer funded to provide a national service. However as is seen in many cases, where research groups have been involved in the publishing of digital resources online, the resources have often disappeared when the funding has.

The increased number of archives not only reflects increasing recognition of the importance of making data available, but also the fact that data archiving, along with open access, is increasingly being incorporated into the requirements of research funding (Sherpa Juliet, 2009), with the USA's National Science Foundation now requiring researchers to submit data-management plans with their grant proposals (Parry, 2011). Equally

important is the fact that in the UK the public has the right to request certain information under the Freedom of Information Act. In July 2010, the *Times Higher Education* reported that the Information Commissioner had ruled that an academic who had been analysing the rings in Irish trees for more than 30 years must release the data to someone who had asked for the data under a freedom of information request, despite the fact some of the findings had yet to be published by the academic (Fearn, 2010). Whilst such a ruling will potentially have widespread ramifications amongst much of the academic community, there are already a number of scientists looking for ways to make science more open and accountable, most noticeably in the realm of open data. The Panton Principles (http://pantonprinciples.org), a set of principles for open data, not only recognize the importance of publishing data, but also making it available with the necessary waiving of rights, so that others may reuse and build upon the data.

Open data is an area that has become increasingly important since the recent email controversy regarding the Climatic Research Unit at the University of East Anglia in the UK, where hacked email accounts suggested evidence of the manipulation of data by climate scientists (Revkin, 2009), as well as problems with the computer analysis software (Merali, 2010). Although the scientists were later exonerated, there has nonetheless been damage to the public perception of climate science.

Open source science and Open Notebook Science

Open source science and Open Notebook Science are two overlapping approaches to open science that are built around the idea that carrying out science in the open enables less duplication, greater robustness and greater progress.

Open source science is the application of the open-source approach, which has been so successful in the creation of software, to the realm of science. Open software is seen, along with mass-contributory projects such as Wikipedia, as one of the more successful examples of collaboratively created content, with great swathes of our computer activities now being accomplished with open source software that is not only free, but in many cases is also considered better than the proprietary alternatives (see Table 1.1). Open source software development works by making the software source code freely available so that it can be manipulated by outside developers (often volunteers), and any suggested developments reincorporated within

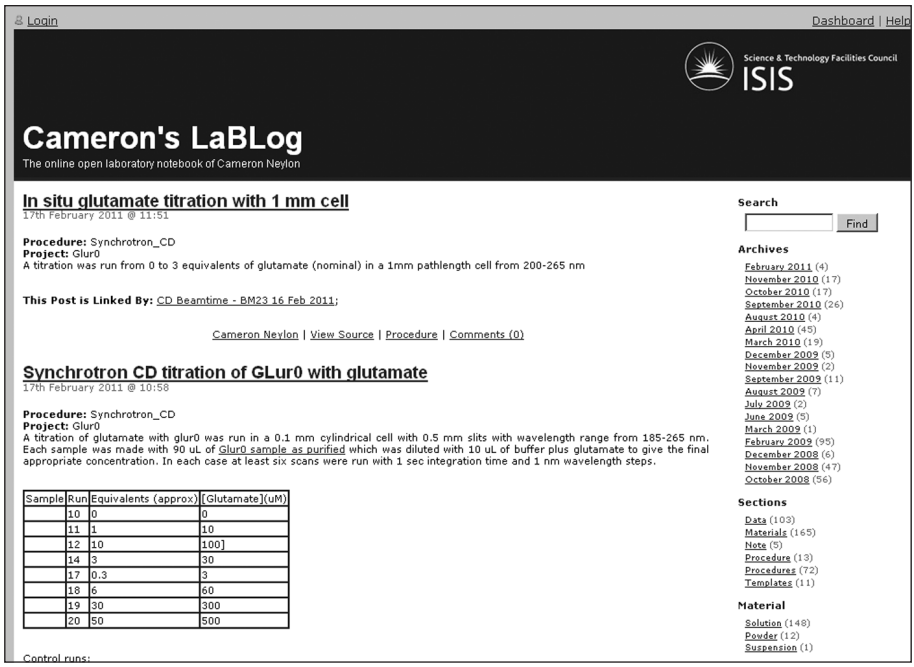
Table 1.1 *Open source software examples*

Proprietary Software	Open-Source Alternative
Microsoft Office: Word, Excel, PowerPoint, Access	Open Office: Writer, Calc, Impress, Base
Adobe Photoshop	GIMP (GNU Image Manipulation Program)
Internet Explorer	Mozilla Firefox
Windows Operating System	Linux
Windows Server	Apache
Talis ILS	Koha; Evergreen

the official software release if deemed an improvement by the community. Around many of these open software projects, vibrant communities have emerged sharing ideas and helping to solve one another's problems. Such communities have been encouraged by so-called 'copyleft' licences that are often employed by open-source projects. These licences require anyone modifying and distributing a work to include the same rights as the original work. For example, the code is available for all the variations of the Linux operating system as the original version was published under the GNU General Public Licence (Benkler, 2006).

Open source science is designed to encourage equally vibrant communities around areas of scientific interest, and is seen as an especially appropriate method for dealing with those areas that do not provide the return on investment that is a prerequisite for a lot of commercial investment. The Synaptic Leap (www.thesynapticleap.org) is one such example, bringing together biomedical science researchers to investigate diseases where profit-driven research is failing (Kepler et al., 2006).

Open Notebook Science (ONS) is another approach to carrying out science in the open: 'organising the scientific production based on the public disclosure of achievements and failures, and their related data and procedures, so that they are analysed and discussed openly to further advance science by solving and addressing specific problems' (Vera, 2009). Whereas open source science is problem-centric, Open Notebook Science is more person-centric, focusing on the ongoing publishing of a scientist's research findings, for example, Cameron Neylon's lab log at http://biolab.isis.rl.ac.uk/camerons_labblog publishes the results of various ongoing experiments (see Figure 1.1). Lab logs not only provide access to positive data, but also negative results, the so-called dark data of failed experiments


Login

Dashboard | Help

Science & Technology Facilities Council
ISIS

Cameron's LaBLog

The online open laboratory notebook of Cameron Neylon

In situ glutamate titration with 1 mm cell

17th February 2011 @ 11:51

Procedure: Synchrotron_CD
Project: Glur0
A titration was run from 0 to 3 equivalents of glutamate (nominal) in a 1mm pathlength cell from 200-265 nm

This Post is Linked By: CD Beamtime - BM23 16 Feb 2011;

Cameron Neylon | View Source | Procedure | Comments (0)

Synchrotron CD titration of GLur0 with glutamate

17th February 2011 @ 10:58

Procedure: Synchrotron_CD
Project: Glur0
A titration of glutamate with glur0 was run in a 0.1 mm cylindrical cell with 0.5 mm slits with wavelength range from 185-265 nm. Each sample was made with 90 uL of Glur0 sample as purified which was diluted with 10 uL of buffer plus glutamate to give the final appropriate concentration. In each case at least six scans were run with 1 sec integration time and 1 nm wavelength steps.

Sample	Run	Equivalents (approx)	[Glutamate](uM)
	10	0	0
	11	1	10
	12	10	100]
	14	3	30
	17	0.3	3
	18	6	60
	19	30	300
	20	50	500

Control runs:

Search

Find

Archives

February 2011 (4)
November 2010 (17)
October 2010 (17)
September 2010 (26)
August 2010 (4)
April 2010 (45)
March 2010 (19)
December 2009 (5)
November 2009 (2)
September 2009 (11)
August 2009 (7)
July 2009 (2)
June 2009 (5)
March 2009 (1)
February 2009 (95)
December 2008 (6)
November 2008 (47)
October 2008 (56)

Sections

Data (103)
Materials (165)
Note (5)
Procedure (13)
Procedurez (72)
Templates (11)

Material

Solution (148)
Powder (12)
Suspension (1)

Figure 1.1 *Cameron Neylon's Open Notebook Science lab log*

(Goetz, 2007). This information is imperative to the progress of science, not only to prevent duplication of research, but the positive bias of results. In the same way as the International Committee of Medical Journal Editors has a requirement to register clinical trials before the trials take place to prevent bias in their journals (ICMJE, 2009), such a requirement for the publication of all data would be equally useful for demonstrating the validity of the interpretation of results.

Open source science and Open Notebook Science not only potentially provide access to far more scientific data than ever before, but a far greater variety of data. This data will not necessarily be deposited within the established official data archives, but is likely to be scattered across the web, making use of various web and web 2.0 technologies. Whilst the library and information professionals within the academic sector have an important role in highlighting the benefits of open research data to the academic community, and helping them with the publishing of data, there is also a role for library and information professionals in other sectors to encourage the opening up of academic research data through freedom of information requests.

Commercial sector

At the same time as the scientific community has been making data available to aid in the generation of new knowledge, many commercial organizations have also been making increasing quantities of data available for more commercial reasons. Web-based organizations and more traditional organizations have released data sets both as part of competitions to find solutions to specific problems as well as for seemingly more noble purposes. Data is not only being made available by many leading-edge web-based companies, such as eBay and Twitter, but also many traditional organizations such as the supermarket chain Tesco, who have built web services to enable external organizations and services to interact with their product information.

Whilst it could be argued that the scientific community has focused primarily on making its data available to the rest of the scientific community, with the wider public of secondary consideration, commercial organizations have generally been far less discerning about who they make their data available to. Their primary concern is generally opening up their data to contributions from as large a number of people as possible, recognizing that it does not matter where the solutions come from as long as they get the solutions they are looking for. One method of engaging with the public is simply to state a problem, make the data set available and ask the public for the solution.

One of the best known and earliest examples of this is that of the Goldcorp mining company, covered in detail in Tapscott and Williams' (2006) bestseller *Wikinomics: How mass collaboration changes everything*. With Goldcorp employees failing to find sufficient deposits in their Red Lake mine, the CEO (Chief Executive Officer) took the innovative step of making all the company data about the mine publicly available and launching their Goldcorp challenge in March 2000 with $575,000 worth of prize money to those who could identify where the gold was located. Whilst some submissions came from geologists, others also came from other fields, applying solutions to the problem that would not have been thought of otherwise, and resulting in the discovery of significant quantities of gold.

In 2006 Netflix, the online video rental service, took a similar step. They provided access to a set of anonymized rating data and set a competition with a $1,000,000 prize to the individual or group who developed a movie recommendation algorithm that improved on Netflix's own system by more than 10%. The prize was won in September 2009, but whereas there had been plans for a follow-up competition, it was discovered soon after that the

anonymized data could actually be traced back to the original users (Lohr, 2010). Whereas most of the user information would have been publicly available on the site anyway, albeit in a format that was not so easily mined for information, the fact that supposedly anonymized data could be traced would understandably worry users. Especially when seen within the context of AOL's earlier search data scandal.

In 2006 AOL Research released search log data for a random selection of 658,000 users, primarily as a data set for the research community. Such information is of use to a wide range of researchers in various disciplines, from psychologists and social scientists trying to understand the way people think and behave to computer and information scientists trying to improve user interfaces and information retrieval systems. Whilst the data was 'anonymized' through the use of identification numbers rather than usernames or IP addresses, privacy advocates highlighted that many users could be simply traced through their history of searches, and a number of the users have been since. For a user's supposedly private searches to be made publicly available was a significant breach of trust and the data was, understandably, quickly taken down soon afterwards. However, there were people who both recognized the value of the data and the likelihood that it would be quickly taken down once AOL realized their error, and a number of copies of the data set were made and placed on other sites online.

The continuing online presence of the AOL log data raises a number of ethical questions for those researchers for whom it potentially offers a valuable source of information. Should such sources never be used as the basis of research? Or is it acceptable as long as it is used in the manner for which it was originally released, and no attempts are made to identify individuals? The AOL log data is by no means the only controversial data set freely available online. In September 2010 the personal details of thousands of broadband customers who were alleged to have illegally shared music, films and pornography, were obtained from ACS:Law and published on the message board at www.4chan.org, allegedly in retaliation to the company's bullying tactics in pursuing file-sharers. The publication of such controversial and sensitive information is by no means isolated. WikiLeaks' (http://wikileaks.ch) sole purpose is to publish and comment on leaked documents. Examples of data sets published on the WikiLeaks site include the leaked British National Party membership and contacts list, the 570,000 intercepts of pager messages on the day of the September 11 attacks, the hundreds of thousands of classified documents about the war in Iraq and the US embassy cables.

In addition to complete data sets, many websites also offer access to their services via application programming interfaces (APIs), establishing a relatively simple method for external actors to engage with an organization's web services. APIs provide the opportunity for an organization to spread its services beyond its own website and to tap into the wisdom of the crowd in the innovation process. Providing developers with a means of accessing a service's data allows them to display the data in new ways or combine it with data from other sites to create mashups. For example, www.22books.com uses the Amazon API so that users can build and share lists of books that they like, whereas www.auctionsnearyou.com combines eBay search results with Google Maps. Such services also provide a wealth of potential information to the research community.

The advantage of online retailers such as Amazon and eBay enabling developers to engage with their services away from their primary site is obvious. It makes little difference to Amazon or eBay whether the items they sell are through their site or via a partner site. From their perspective what is important is that as many items are sold as possible. Whilst Amazon and eBay are very successful in selling items in their own right, it would be naïve to believe that there is a single, ideal shop-front best suited to all customers.

With large online players such as Amazon and Google having strong interest in the publishing industry, it is not surprising to find that they offer a number of book-related APIs. A number of libraries have used these feeds to add additional functionality to their OPACs. For instance, Bath University library have added Google's 'Preview' feature to their OPAC, so that where previews are available users can see what the work contains without having to go to the shelves first (see Figure 1.2). Bath University has also added to each page a link to Blackwell's bookshop, so that they can quickly buy the book if they wish. Such linking to commercial retailers may not only be seen as an added service, but also a potential source of income for a library, albeit probably a small one.

Not all commercial APIs are primarily designed for such explicitly commercial ends; instead they can provide a means of driving users to use a service, both through the adoption of additional services and through the distribution of a service onto multiple platforms. The microblogging service Twitter provides a good example of a web service that has grown to a large extent through the capabilities of its API. Through the Twitter API, multiple different interfaces have been built for different platforms, both desktop (e.g., Windows, Linux, Apple OS) and mobile (e.g., Symbian, Android). There are

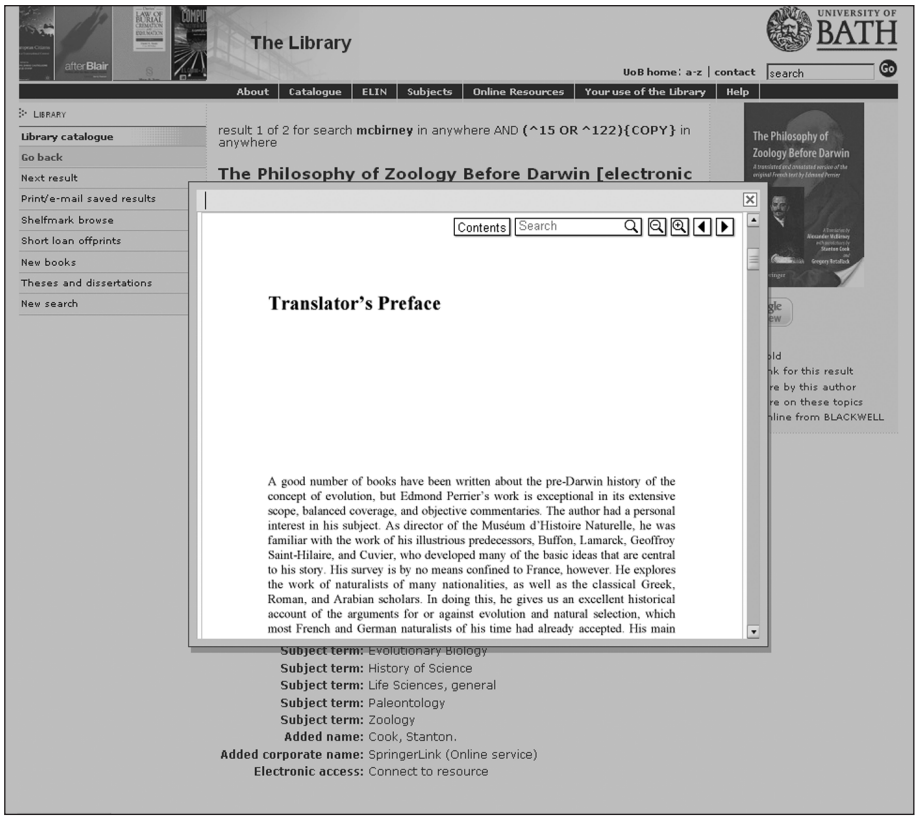


Figure 1.2 *Google Preview feature embedded within University of Bath OPAC*

also numerous additional services that have been established to provide rich content to the basic 140 character Twitter service, for example, photos (e.g., www.twitpic.com) and videos (e.g., www.twiddeo.com), as well as applications to integrate the Twitter service into existing social network services. In August 2010 Twitter turned off its basic authentication API method and replaced it with the more complicated OAuth method; whilst OAuth is more secure, the change is significant in that it shows Twitter's changing attitude to external services. It has also begun to incorporate an increasing number of features that were initially only provided by external services, for example, it now allows users to add location information to updates and organize the people they are following into lists. Now that Twitter has sufficiently established itself it seems as though it is less concerned about enabling as many applications as possible, but rather is interested in more secure applications.

The importance of data and lightweight programming models were recognized as key features of web 2.0 websites in Tim O'Reilly's seminal paper on the topic (2005), and it is unsurprising to find them as a key part of many online services, from the internet giant Google to local library catalogues. The Programmable Web website at www.programmableweb.com provides details on many of the APIs available, as well as providing a directory of some of the services that make use of them. Whilst they are often designed to be incorporated into other online services, they can also provide a valuable source of information for researchers from a wide range of disciplines. For it to be useful, however, it is necessary for researchers to be aware of the type of information that is available to them, and for them to have the necessary skills to collect the data.

The wide variety of data sets available online are often of use to individuals and organizations unthought-of by the publisher of the data, and unlike much of the scientific data, it will not be available from a small number of relatively well known archives and websites. It will also often lack the necessary metadata and documentation to aid a researcher in discovering and making use of the data. However, whilst there have been noticeable attempts by commercial organizations to open up some of their non-personal data sets, the move is by no means universal. In a survey of British businesses in April 2010, after the launch of the UK government's data store, 68% said they would not be prepared to open up access to their own data despite recognizing the potential commercial benefits of sharing that data (Arthur, 2010b). There is still a long way to go before the potential of open commercial data is realized.

Government data

Governments have always been amongst the largest collectors and producers of information, not only taking Francis Bacon's oft-quoted dictum 'knowledge is power' to heart, but using its knowledge to dictate the discourse with its citizens. However at a time when trust in politicians has fallen to an all time low (Campbell, 2009), there have been increased calls for transparency in government both by the media, e.g., *The Guardian* Datablog in the UK, and non-profit organizations, e.g., Sunlight Foundation in the USA. Providing people with the relevant information will allow them to make more informed decisions about the issues that affect their lives, regarding health, education and, importantly, government itself. Providing

access to the data rather than a government's interpretation of the data allows for a rebalancing of the discourse between citizen and state. The importance of transparency has led to the suggestion that it should be used for the benchmarking of the provision of e-government services (e.g., Osimo, 2008; Freed, 2010), as traditional benchmarking indicators, such as the provision of certain specific services, quickly become saturated and meaningless for differentiating between different countries' offerings.

Coupled with the recognition of the economic potential of much of the data to the private sector, as well as savings in the public sector, the last two years has seen an increasing number of governments around the world making their data publicly available through the use of new one-stop data portals. For example, in the USA (http://data.gov), in the UK (http://data.gov.uk), in Australia (http://data.australia.gov.au) and in Greece (http://geodata.gov.gr/geodata). Some of these one-stop data portals offer little more than directories pointing to data previously published elsewhere on government websites; some focus on one particular type of data (e.g., Greece tries to bring together all the government's geospatial data), whereas others have been coupled together with legislation requiring the publication of more data (e.g., the USA's data.gov), or republish the data in a more useful format (e.g., the UK's data.gov.uk). It is an idea that is not restricted solely to developed countries of the West, for example, Kenya also has an open data website (www.opendata.go.ke).

There is also a lot of economic potential in the data that governments and government agencies collect about their citizens, potential that is not being realized while the information is not made public, as well as the potential for governments to save money. As has been shown in the commercial sector, opening up the data to outside users provides access to ideas that either would not have been thought of or not given the time to put into action. In the same way that commercial web services such as Twitter have used external developers to distribute services, so can governments. A plethora of government departments spent thousands of pounds in 2009 and 2010 to launch applications for the iPhone, despite the iPhone only being used by a small percentage of the population. By releasing the raw data, communities of developers for the different platforms could have quickly provided applications that could be used by far more users.

Looking at any of the government data archives quickly gives an idea of the depth and breadth of the information that governments around the world collect and are increasingly publishing. Examples of data sets from

data.gov.uk include: the costs of, traffic on and satisfaction with government websites; public servants earning over £150,000 per annum; perceptions of crime; and children travelling to school – mode of transport used. Unsurprisingly there are differences in the data that has been made available within various countries, this is not only due to different data having been collected in different countries, but also historical precedent regarding the public right to data from government agencies. For example, Ordnance Survey data has traditionally been sold to generate revenue, however certain data sets were released in April 2010 (Thorpe and Rogers, 2010), albeit not as much as some would have liked.

Of primary importance in the release of government data is that the data is published, whatever the format, although as will become clear throughout the course of this book, the format makes a big difference to the ease with which the data can be used by people. At one end of the spectrum data may be an image published as a PDF, where the information is to all intents and purposes stuck within the document as it is only comprehensible to the human reader. At the other end, data may be published as Linked Data, not only structured in a semantic format, but also allowing the data to be linked to/from other data on the web (see Chapter 4 'The semantic web: the RDF vision' for more details). Unlike many other countries, the republishing of certain data sets as Linked Data is a key part of the UK's data strategy, and as a major publisher of authoritative data it may be considered to be playing an important role in the move towards the vision of a more semantic web.

The potential value in the information can only be realized if communities emerge around the data. As such, a number of governments have encouraged users to engage with the data by running competitions, for example, *Show Us A Better Way* in the UK (www.showusabetterway.co.uk) and Apps4Finland (www.verkkodemokratia.fi/apps4finland/en). The World Bank has also run a competition for applications related to the Millennium Development Goals (www.worldbank.org/appsfordevelopment). Whilst such competitions can gather interest in the short term, eventually there will be many more data sets than there are interested developers, and developers will only ever come up with a small proportion of the ideas about how the data may be used. It is important to open the data up to the wider community who may have ideas, but not the necessary development skills.

Whilst commercial data sets may be of global interest, certain government data sets, especially those at a local level, are likely to be of interest to far fewer users. It is therefore important that librarians play a part in the

promotion of these data sets and make sure that they are used as much as possible. As such, it is important that librarians keep themselves acquainted with the huge range of information that is being released, not only responding to user requests.

Library data

As well as libraries having a pivotal role to play in facilitating access to external data, and helping the wider organization in which they are situated make data publicly available, libraries also possess a large amount of data that may be of use to external individuals and organizations; not only data sets from within their collections, but also data contained within their catalogues, and data about how their library resources are being used. The history of libraries sharing information pre-dates the web (Taylor, 2003), with the widely adopted Z39.50 protocol allowing for the searching of remote databases. However, the Z39.50 protocol has been primarily used by the library community rather than the wider public, and there are now increased efforts to lower the barriers of entry so that the same information can be made more easily available over the web in formats that are more widely used. This information has been made more accessible in a number of different ways. SRU (Search and Retrieve via URL [Uniform Resource Locator]) may be seen as the next generation of Z39.50, updated for the web and returning the files in an XML (eXtensible Markup Language) format, a format that most developers will have experience of. Some libraries have also incorporated lightweight APIs; whilst these may have reduced functionality they offer a simple way of collecting information. Others have made the whole catalogue available for download as a single file.

There are now a wide range of libraries around the world that are making their catalogue information available in more accessible forms, from national libraries such as The Library of Congress (www.loc.gov/standards/sru/simple.html) and the Bibliothèque nationale de France (www.bnf.fr/fr/professionnels/catalogage_indexation.html), to smaller university libraries such as Cambridge University Library (www.lib.cam.ac.uk/api) and North Carolina State University (www.lib.ncsu.edu/dli/projects/catalogws). It is not only catalogue information that has been made available, the Library of Congress has also made their widely used Subject Headings available as Linked Data (http://id.loc.gov/authorities), a data set that is not only of value in its own right, but, as will be returned to in Chapter 4, by being published

in a Linked Data format has the potential to add value to information published elsewhere on the web. These more accessible formats allow the data to be manipulated by those beyond the traditional library community.

Libraries, as much as other organizations, increasingly recognize the need to reach out beyond the traditional library website accessed via the desktop computer. There are now many different platforms with which to engage with users, for example, library services may be provided through social network sites, through mobile phones or even in virtual worlds. It would, however, be prohibitively expensive for each library to create an application for every possible platform. Enabling the interaction with the data allows the user community to create the applications where they are wanted and needed, rather than the library making the selection based on the headline-grabbing technology of the day (e.g., iPhone and iPad). In the same way as some governments and commercial organizations have attempted to incentivize the user community to engage with the data they have released, some libraries have as well. For example, Finland Libraries made their full catalogue available in a MARCXML format (http://data.kirjastot.fi/), with Apps4Finland incentivizing the creation of applications with a prize of 500 euros.

Whilst there has been a lot of focus on data within library catalogues, libraries in fact have access to a lot of other useful information that could potentially be released but has not been made widely available as yet: information about which books are being borrowed, the electronic resources that are being used and the OPAC searches that are being made. Whilst there are privacy issues to be overcome, there are without doubt advantages to be had from this information. The successful combining of catalogue data with usage data at the University of Huddersfield has managed to increase the number of books that are being borrowed at a time when throughout the academic sector the number of items loaned per user has actually been going down (JISC Mosaic, 2010). Such information would not only be of great interest to information science researchers, and those interested in the development of improved information retrieval systems, but could also form the basis of many new services as yet unthought-of by the information community. However, whilst the opening up of such information would enable insights from the wider community, it would also open library work up to wider criticism, as users question the allocation of funds and the gaps in the resources available.

Conclusion

The huge increase in data that is being made available is likely to be of interest to a wide range of users in every type of library. There will be members of public libraries who are interested in how the government is spending their money, both nationally and locally, as well as the performance of schools, hospitals and the emergency services. There will be users of academic libraries, both staff and students in all academic disciplines, interested in the vast research potential of data that is released, whilst people in industry are likely to be more focused on realizing some of the commercial potential in the data that is publicly available. There are also the libraries themselves, who can make use of much of the data available to provide improved traditional services.

For all this to happen, however, library users need to be served by a community of library and information professionals who are engaged with the web of data, able to promote the data that is available and have the skills necessary to help library users both access and make use of it. Whilst the format may be different, facilitating access to this data can be seen as a natural progression for the librarian from their traditional role of facilitating access to documents, although it does necessitate a willingness to engage with new technologies.