LIBER QUARTERLY 20 (1), September 2010 – ISSN: 1435-5205. P94–104
http://liber.library.uu.nl/
Igitur, Utrecht Publishing & Archiving Services

British Library Dataset Programme: Supporting Research in the Library of the 21st Century

J. Max Wilkinson, Tom Pollard, Adam Farquhar

The British Library, 96 Euston Road London NW1 2DB, United Kingdom,
max.wilkinson@bl.uk; tom.pollard@bl.uk; adam.farquhar@bl.uk

Abstract

Advances in computational science and its application are reshaping the social landscape and the practice of research. Researchers are increasingly exploiting technology for collaborative, experimental and observational research in all disciplines. Digital data and datasets are the fuel that drives these trends; increasingly datasets are being recognised as a national asset that requires preservation, attribution and access in much the same way as text-based communication.

The British Library is in a unique position to enhance UK and international research by extending its presence from the physical collection to the digital dataset domain. To meet this challenge and be a responsible steward of the scholarly record, the Library has defined a programme of activity to support the datasets that underlie modern research and promote them as a national asset. We are designing a mixed model of activity where specific service-level projects with clear goals will provide support for collaborative work aimed at revealing and clarifying requirements related to datasets.

For example, there is a clear community need for stable, scalable and agreed data citation mechanisms. In response the British Library became a founding member of DataCite, the International Data Citation Initiative which, as a member of the International DOI foundation, assigns Digital Object Identifiers (DOIs) to datasets. We are leveraging the services built for DataCite to actively partner with a number of UK data centres and data publishers to add value to their collections and facilitate the rejoining to the scholarly record by linking the published record with the datasets that underlie it. We are also implementing a similar strategy to promote dataset discovery services through the Library's catalogues and streamlining access to national external collections.

The British Library datasets programme will guide activities across the Library and provide a focus for stakeholder communities to address the challenge of integrating

datasets into their researcher services. This work will ensure that the integrity of the scholarly record remain intact, useable and vital for future generations.

Key Words: data citation; scholarly record; datasets; Digital Object Identifiers (DOI)

Introduction

Background

Advances in computational science and its application are reshaping the social landscape and the practice of research. Researchers are increasingly exploiting computational capacity and capability for collaborative, experimental and observational research in all disciplines. Digital datasets fuel these trends in cultural and scholarly endeavours and are increasingly being recognised as a national asset. For a number of reasons these data assets are not being exploited to the extent that traditional publications have been and there is a groundswell of community interest in bridging the widening gulf between publications and the datasets that underlie them.

In the UK, research councils and funding bodies have responded by mandating data management plans in requests for funding (Serco Consulting, 2008). Publishers have responded by accepting supplementary data alongside articles or requiring submission of data into recognised repositories. Specific disciplines, organisations and institutions have created data management centres for explicit data types irrespective of publication status.[1] Research libraries have been largely absent from the discussion and there is growing awareness of a widening fissure in the scholarly record — the gap between published research and the datasets that underlie it (Buneman *et al.*, 2000, 2004; Gray *et al.*, 2002; Hey *et al.*, 2003; Lord & Macdonald, 2003; Borgman, 2007; Lyon, 2007; RIN, 2008; Serco Consulting, 2008). While there are well-established services for published research, there is only an incomplete patchwork of services for datasets. For example, there are no agreed ways to identify, search, cite, or catalogue datasets. The British Library is in a unique position to enhance UK research and strengthen its profile as a critical component of modern research by extending its presence from the physical collection to the digital dataset domain.

The Library is one of the largest research libraries in the world. It has a statutory responsibility to acquire, preserve and make accessible the UK national

published archive. It holds over 150 million items ranging from historic manuscripts to modern electronic journals, digital music files and patents and is leading international collaborations to find solutions to ensure this rich and varied collection is sustained far into the future. Strategic priorities for the Library (2008–2011) are to build its digital infrastructure, capture digital publications and support research with innovative and integrated processes. The Datasets Programme is one of several programmes of work that will help reach these strategic goals.

Motivation

Awareness of the impact of the digital age on research is growing. The British Library Chief Executive, Dame Lynne Brindley DBE, observed that the biggest challenge facing the British Library is presented by 'the data deluge and the increasing need to integrate datasets that underlie published research with the more traditional formats and preserve these digital formats into the long term' (IWR Horizons, 2008). This view is supported by three studies conducted by the Library in 2007 and 2008 looking specifically at options and approaches to dealing with datasets in Science Technology & Medicine, Arts & Humanities and Social Sciences (Key Perspectives, 2007; Education for Change, 2008; Rothenberg & Hoorens, 2008).

In order to meet this challenge, be a responsible steward of the scientific record and improve its relevance to researchers, the British Library is undertaking a programme of activity to support the data that underlie modern research and promote them as a national, intellectual asset.

Scope

A dataset can be regarded as an organised collection of data. Our programme focuses on digital formats that exist as content within or external to the Library, including images, audio and numerical data. We consider datasets to be collections of data that are generated or consumed by users in the process of their research. The Library already contains many digital objects that can be considered datasets. Digitised maps, books and even the Library catalogues may be considered datasets based on specific requests or operational needs. In addition, mass digitisation programmes like the Library's newspaper digitisation projects generate large datasets.[2] External to the Library data centres provide for the preservation and persistence of data generated by UK-funded

research, government departments and national programmes, e.g., the UK Data Archive, the census and other national data collections.

As ever more data are generated, modern researchers are confronted with new challenges, such as how to persist their data and how to reference them as required. These challenges are all the more relevant as most UK research funding organisations are implementing data sharing in response to the agreed notion that the results of publicly funded research should return the maximum benefit possible to the research community and the wider public. As these data sharing policies and requirements are implemented, researchers find themselves with significant data management needs.

Data sharing is not new in the research paradigm. Many researchers and discipline networks share data as part of their common practices. Despite this, primary research data are often referenced via network locations on a plethora obsolete or near-obsolete media. As we explored possible contributions that the British Library could make, two clear requirements came to the forefront: data require preservation and data require persistent identification.

Persistent identification is necessary for the stable referencing of data and the Datasets Programme sees a clear role here. Once datasets are cited in an agreed, stable and scalable way, they can be connected with associated research publications and may be shared and reused more easily. Clearly referenced and accessible datasets help to create an audit trail in the research process and provide a foundation for quantitative metrics for tracking the reuse and re-purposing of data.

We will work across disciplines and promote a joined-up infrastructure to help address the demands of the researcher to be able to locate and access data, regardless of physical location. This will involve establishing partnerships, principles, guidelines and agreements with external stakeholders parallel to establishing Library services that facilitate dataset citation, cataloguing and access.

The Library has made significant progress in establishing its presence in the domain. The Library is a founding member of the European Alliance for Permanent Access to the Records of Science, it chairs the World Wide Science Alliance, it is a signatory to the UK National Data Strategy for Economic and Social Data, and it is a member of the Steering Committee for the UK Research

Data Service. In addition, it was a founding member of the UK Digital Preservation Coalition and the co-ordinator for the FP6 funded PLANETS digital preservation project and subsequent Open PLANETS Foundation.

Strategy and Objectives

Traditionally libraries were built around printed and visual media, which now have agreed standards of citation and bibliographic formats. In contrast quantitative empirical data have not kept pace with their printed counterpart and are often fragmented and unobtainable, existing in a range of locations on various media in a variety of formats.

To determine what roles libraries should have in the datasets domain the Datasets Programme has undertaken a number of strategic and exploratory activities and has established a strategy that outlines how the Library can build on work already achieved. The strategy[3] takes a cross-department and cross-discipline approach to coordinate dataset activity across the Library and across external stakeholders.

We have established a multi-discipline team that directly engages with key stakeholders to reveal and gather requirements based on user need and user behaviour. Efforts are being made to establish the Library as a key stakeholder and partner in the datasets domain and to build communication channels to place the Library as a hub to focus the wider community. Modern social technologies will be used alongside more traditional printed and directed tools of communication.

By investing in this programme the Library will further expand its skill base in the digital world and provide a solid foundation for providing high-value services to its users based on their requirements.

Core Activity and Areas of Focus

Core Programme

The British Library Datasets Programme undertakes both cross-department activity within the Library and external project-focused activity. All activities attempt to address a clear community need and often require stakeholder

relationships to be established or developed. We aim to build on current infrastructure and achievements to deliver new and useful services to the library community.

DataCite

The British Library joins many other organisations that believe research data are an essential component of the scholarly record and that increasing access to these data has a number of well-recognized benefits. Just as scientific research is global, it seems appropriate to take a global approach to dealing with the challenges of increasing access to data. This principle underlies the formation of DataCite[4], which aims to remove one of the barriers in the fragmented data landscape by providing efficient, effective and persistent data citation facilities.

We have agreed to work together to promote global access to research data. The long-term vision is to support researchers by providing methods for them to locate, identify and cite research datasets with confidence. In order to achieve this long-term vision, the members established a not-for-profit agency that enables organizations to register research datasets and assign persistent identifiers to them. The agency will take global leadership for promoting the use of persistent identifiers for datasets to satisfy needs of scientists. It will, through its members, establish and promote common methods, best practices and guidance. National members work independently with data centres and other data publishers in their own domains, respecting and being familiar with different research funding frameworks and practices.

The founding members of DataCite were: The German National Library of Science and Technology; The British Library; Technical Information Centre of Denmark (DTIC); TU Delft Library of the Netherlands; Canada Institute for Scientific and Technical Information (CISTI); The California Digital Library (CDL) and Purdue University. Following the first general assembly in Paris on 5 February 2010, this membership was increased to include: The Australian National Data Service (ANDS); The Library of the ETH Zurich; The French Institute for Scientific and Technical Information (INIST); German National Library for Medicine (ZB MED); and the GESIS-Leibniz-Institute for the Social Sciences.

DataCite builds on the approach developed by the German National Library of Science and Technology and promotes the use of Digital Object Identifiers (DOIs) for datasets. The DataCite model of governance is composed of

national representation in a global presence, because researchers create, share, and access data globally. Most data centres and data publishers are embedded within their national funding structures and research frameworks. With the DataCite governance structure national representatives provide DataCite services to national-level parties and are familiar with the environment in which the data publishers operate.

Collaborative Projects

The Datasets Programme has joined with a number of key stakeholders to establish complementary programmes aimed at specific issues of data citation and management. In addition the Library has actively supported several projects that illustrate specific issues of data citation and management.

Data journals: This project is a partnership between the British Library, a national data centre and an established publisher to develop the mechanisms required to run an operational overlay data journal. We aim to provide incentives for scientific researchers to deposit their data in accredited data repositories, thereby improving the transparency and traceability of the data which underpins important scientific conclusions.

Data repositories: In partnership with UK Higher Education Institutions, a number of publishers and an international consortium, we are undertaking preparative work for the establishment of a repository for bioscience research datasets linked to the peer-reviewed articles they underpin. This work targets disciplines that are not well served by the current data centre infrastructures and will present us with unique and valuable experience in dataset construct and variability.

Data attribution: This collaborative project will develop and test a citation framework for complex network models of disease and associated data. Citations of network models will be embedded in two leading publications and will be based on an international biomedical project that is collecting and analysing massive coherent datasets from a number of contributing partners.

Outreach

Identifying, engaging and maintaining a stakeholder community is essential to the Library's datasets strategy. The Programme has built an initial stakeholder map and will review and build on this map at regular intervals to

keep it relevant. From this map a strong communication strategy has been formed that articulates clear key messages to each stakeholder community to ensure consistent and correct communication. Modern communication tools will be used alongside traditional techniques to reach stakeholders in convenient and useful ways. Collectively our outreach and communications activities will provide a focus for the datasets community to identify needs and solutions, and will help to promote the emerging services of the Library.

Areas of Focus

Science, Technology and Medicine

During 2007/08, the Science, Technology and Medicine (STM) team commissioned a study to help define the British Library's approach to scientific data.[5] This work provided an initial list of over 400 bioscience open data resources, a set of recommendations for potential data services that the Library could develop and guidelines for the selection of 'datasets as content'. Dedicated STM staff have been working to develop these recommendations into tangible solutions for the Library and its users.

The team are developing and testing selection criteria for high-value reference datasets to include in the British Library 'collection', initially concentrating on biosciences and environmental sciences. In undertaking this work the team have developed working relationships with data centres, data curators and stakeholders to aid discovery of and access to datasets.

The STM team is piloting dataset discovery services through the Library's search engine facility in order to better connect users to datasets and data resources. This pilot will help the Library better understand researcher behaviour and requirements and provide a first look at resource discovery of datasets alongside other materials.

Complementary to this activity, draft selection criteria for datasets have been developed and are being trialled. The Library has well-established criteria, principles and mechanisms for selecting and acquiring published content in both the print and digital environment. These are being augmented for datasets, which introduce new challenges around quality, stability and sustainability of resources — especially when connecting to them, rather than holding them directly.

Arts and Humanities
A traditional area for the Library, the Arts and Humanities discipline offers unique challenges for the Datasets Programme. We are working with the Archaeological Data Service to enrich the current web presence with the ability to use and reliably cite their collections. Internal to the Library, the team is piloting citation services to collection items, specifically music files in the sound archive, so that users of the sound collections can reference with confidence particular collection items in their research.

Social Science Collections Research
In the social sciences, dataset activity has focused on support of and access to existing, often under utilised resources. This activity is generating guides to social science resources for Library users, streamlining access to established external data publishers and data centres and exploiting national events to showcase the relevance of the Library's facilities.

The social science team are creating resource guides to help researchers use different types of datasets and find commonly used datasets. Initially the guides will cover areas including: food studies, management and business studies, health and wellbeing, and socio-legal studies. In the future these will be expanded into further subject disciplines.

Prior to the London 2012 Olympic Games the team will launch a website for research around aspects of the Olympic Games. A separate portal for management and business studies is also being developed. Dataset resource guides and links to relevant datasets will be included in the websites at launch. Further activities, including an exhibition, are being planned to mark the 2011 UK census.

The UK Data Archive (UKDA) provides access to major governmental and ESRC-funded datasets and is a key resource for social science researchers. Currently British Library readers are unable to access UKDA if they are not members of UK Higher Education Institutions, so the social science team is working with the UKDA to widen the access to this service. If successful, many more readers would gain access to this rich resource of social science and humanities datasets.

In addition to these activities the social science team is working on a pilot project to expand the Library's searchable catalogue to include social sciences datasets.

Summary

Advances in information technology have provided a means to generate, manipulate and share ever increasing volumes of data. These data form an essential part of the scholarly record and there is great risk that this foundation of knowledge will be lost.

Datasets are essential to the British Library's mission to advance the world's knowledge and are becoming ever more critical to researchers and policy makers. Our activities will place the British Library at the forefront of institutions that are connecting researchers to data and will help the datasets community to identify and tackle the challenges presented by our changing digital landscape.

References

Borgman, C.L. (2007): *Scholarship in the digital age: information, infrastructure, and the Internet*, Cambridge, Mass.: MIT Press.

Buneman, P., S. Khanna and W.C. Tan (2000): 'Data Provenance: Some Basic Issues', *Foundations of Software Technology and Theoretical Computer Science*, pp. 13–15.

Buneman, P., S. Khanna, K. Tajima and W.C. Tan (2004): 'Archiving Scientific Data', *ACM Transactions on Database Systems*, vol. 29, pp. 2–42.

Education for Change Ltd. (2008): *Research into type, scope, location and availability of qualitative and quantitative datasets in the arts and humanities and social sciences,* Feb.

Gray, J., A.S. Szalay, A.R. Thakar and C. Stoughton (2002): 'Online Scientific Data Curation, Publication, and Archiving', *Arxiv preprint cs.DL/0208012*.

Hey, T. and A. Trefethen (2003): 'The Data Deluge: An e-Science Perspective', *Grid Computing: Making the Global Infrastructure a Reality*, pp. 809–824.

IWR Horizons (2008).

Key Perspectives Ltd (2007): *Research into Reference Datasets of High Scientific Value*, Nov. 2007.

Lord, P. and A. Macdonald (2003): *e-Science Curation Report. Data Curation for e-Science in the UK: an Audit to Establish Requirements for Future Curation and Provision*, prepared for the JISC Committee for the Support of Research, http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf.

Lyon, L. (2007): *Dealing with Data: Roles, Rights, Responsibilities and Relationships — Consultancy Report*, June.

RIN (2008): *Mind the Skills Gap: Information-handling Training for Researchers*, Research Information Network (RIN).

Rothenberg, J. and S. Hoorens (2008): *Enabling Longterm Access to Science, Technology and Medical Data Collections*, Rand Europe.

Serco Consulting (2008): *UK Data Service Feasibility Study — Interim Report*, HEFCE.

Notes

[1] For example, the UK Data Archive for historical and social sciences, the EMBL European Bioinformatics Institute's nucleotide sequence database, and the Defra funded Master Chemical Mechanism database for tropospheric ozone formation.

[2] http://www.bl.uk/reshelp/findhelprestype/news/newspdigproj/index.html

[3] To be available shortly at www.bl.co.uk/datasets

[4] www.datacite.org

[5] www.rand.org/pubs/technical_reports/TR567/