A Legacy of Inspiration and an Enduring Smile

by Peter Burnhill[1]

Abstract

This is written in appreciation of the pioneering contribution made by Sue Dodd to what we would now call metadata standards for research data files. It describes two occasions when I had good cause to cite her work, the first when writing in 1984/5 about data libraries and how these might develop in the UK. The context is the early years of Edinburgh University Data Library and the visit by Sue Dodd to present at a seminar and workshop in London and Edinburgh. The second occasion for citation was almost 30 years later, when writing about digital preservation of scholarly statement. That gives opportunity to place her work in the context of the new forms of scholarly publication in which research data form an increasing part, with new need to ensure appropriate citation for web-based resources.

Keywords: Cataloguing, Metadata, Seriality, Web, Registries, History

Introduction

I have this sense of having met Sue Dodd for the first time on three separate occasions: through her writing in the *IASSIST Quarterly (IQ)*; when we spoke on the telephone; and finally when we met in person at the start of her visit to the UK in 1985. I recall those moments with a smile. Her writings, voice and warm sense of person have continued in my thoughts, her mix of charm, insight, dogged determination and encouragement. We all have access to her writing and those ideas and insights live on in our practice.

We surely all have mixed thoughts when we realise that some variant of the following Abstract could have been written yesterday:

> In the last two decades … agencies … have invested heavily in the collection of … data, contributing to the proliferation of … data. However, … the ability to produce data [has] progressed much more rapidly than our capacity to organize, classify, and reference its availability. … The purpose of this article is twofold: (1) to outline some of the information components associated with … data files, and (2) to provide guidelines, examples, and a uniform vocabulary for the creation of a bibliographic reference. (Dodd, 1979)

There is little doubt at the prescience of the advice that "Information stored in a computer-readable form will soon become a legitimate library resource available to those patrons who need it" (op cit). However, even with the arrival of the Web and the passage of time, research data is only now top of the agenda for libraries, and seemingly with a supply-side perspective, rather than having focus on the demand-side for the data needed for secondary analysis.

I first cited Sue's work in 1985; I found the need to do so again when writing an article for *Serials Review* almost 30 years later. The interest in making those two citations serve as temporal bookends for the two parts of this appreciation, labelled Parts A & B:

> Part A has its focus on the first article, "Towards the Development of Data Libraries in the UK" (Burnhill, 1985). Not surprisingly, when I began writing about data libraries I gave emphasis to the importance of cataloguing data – the term metadata then had other meaning – and I would cite the work of Sue Dodd.
>
> Part B has its focus on the other, "Tales from The Keepers Registry: Serial Issues About Archiving & the Web" (Burnhill, 2013), issued almost 30 years later when writing about digital preservation of scholarly statement.

I want to use this as opportunity to say something about the early years of Edinburgh University Data Library which has now been operating for some 30 years. I also wish to say something of the new forms of scholarly publication in which data form an increasing part. Perhaps what is persistent is the concern to ensure that researchers, students and their teachers can have access, both ease and continuity of access, to the resources that they need for their scholarship.

My first encounter with Sue Dodd

The very first time I met Sue was through her writing. It was 1984 and I had just been appointed to develop the Data Library at the University of Edinburgh. I had landed a very good job at a young age to lead a small team of two and a half full time equivalent staff, to take charge of the Data Library and advised that I would need to win external funding for its development.

I was reading the *IQ* collection that my predecessors had been collecting in order that I might understand the varied institutional settings in which data libraries were set. I began at the beginning, with volume 1 issue no. 1 of what was then called the *IASSIST Newsletter* (November 1976).[2]

What stood out was the importance of standards for cataloguing datasets and the key role being played by Sue who was listed as the US chairperson of the Classification Action Group. The report of activity stated that the Action Group in the US gave emphasis "on the library cataloguing of machine-readable data files in public multi-media catalogues," and noted:

> Sue Dodd has used the rules recommended by the American Library Association's Subcommittee on the Cataloguing of Machine-readable Data Files to prepare a draft version of a Working Manual for Cataloguing Machine-readable Data Files which will be tested by members of the US Action Group.

The other actions noted were a committee to investigate a national union catalogue of catalogued MRDF, use of MARC and a critical review of controlled vocabularies – the latter to interact with the European members of the Classification Action Group led by the data archives in Europe which had their focus on Study Descriptions.

My background in my new role as 'Principal Consultant (Data)' was that of a statistician and social scientist but I would go on to work with a number of forward thinking individuals in internationally well-regarded computing service organizations in Edinburgh. The largest of these computing organizations was Edinburgh Regional Computing Centre (ERCC) which operated the network and the mainframes for Universities of Glasgow, Strathclyde and many a research institute across Scotland as well as the large research and teaching base of the University of Edinburgh. The University's Computer Science Department and the ERCC had pioneered the development of multi-access computing, supporting a system known as EMAS that allowed its users to make use of commands in the English language (not IBM JCL) and to program within this operating system, including use of a form of hypertext in a system called View. This enabled us to escape much of the tyranny of magnetic tapes being experienced elsewhere. File transfer and remote log-on to computers hosted in national and regional computing centres were becoming routine for the initiated, as was email (and I still retain access to folders of email from that time). In the UK, SERCnet was being re-launched as JANET as the Internet backbone for UK research computing.

Responsibility for application software was with another group at Edinburgh, the Program Library Unit (PLU). This had been set up in 1969 with a national (and international) role for 'knowledge based software facilities (or DATA)' also converting and distributing IBM mainframe source code software to run under the operating systems used for the British manufactured ICL hardware. The founding director of PLU, Marjorie Barritt was clearly the far-sighted-genius, with commitment to 'data handling software'.

Just prior to my joining, PLU had merged with the ERCC Database Group to form a software house called the Centre Application Software Technology (CAST). CAST was a relatively short-lived organisation merging into ERCC in 1989 to become the Computing Service, but for those five years CAST provided the Data Library with a loving nursery.

There had already been positive activity to establish the operation of a University Data Library by Trevor Jones, a lecturer in Sociology, and by Audrey Stacey who was the computing expert (Jones and Stacey, 1984) with policy support from Deputy Librarian Peter Freshwater. Researchers had petitioned for centrally-managed university wide provision of access to large-scale datasets, typically the decennial population censuses for Scotland, the annual agricultural censuses for England & Wales and for Scotland, the General Household Surveys and a range of digitized boundaries being used in what were still path-breaking ways to do computerized mapping. Trevor left to work for CACI in the emerging and lucrative geo-demographic industry, creating the vacancy that I had applied to fill.[3]

Part A (1985). Towards the Development of Data Libraries in the UK

A visit by Geoffrey Hamilton from the British Library to Peter Freshwater, the Deputy Librarian at the University, led to an invitation to present a paper by a member of the UK Committee of Librarians and Statisticians. This was a joint standing consultation body of the Library Association and Royal Statistical Society that was responsible for publishing a series on statistical sources, such as 'A Union list of statistical serials in British libraries'[4]. Geoffrey Hamilton was leading an initiative on indexing the statistical tables published in government documents and he was intrigued at the discovery of activity to catalogue the datasets behind those tables.

I set about re-reading those early issues of the *IQ* in order to research the topic. The resultant paper, entitled "Towards the Development of Data Libraries in the UK" (Burnhill, 1985), was duly presented to the Committee. The opening page begins with a quote from Sue Dodd when offering a definition of 'data' to complement a media-based definition of 'library':

> Data has been described as "a general term used to denote any or all facts, numbers, letters and symbols which refer to or describe an object, idea, condition, situation or other factor" (S. Dodd 1982). Clearly this is quite wide and describes much that anyone would want to analyze. The word library is derived from the Latin word liber, originally the rind between the wood and the bark, the medium on which the information was recorded before the invention of paper. At one time the reader of a book had to know how to treat that particular medium, but after a

> while all that was needed were literacy and the right to use a library. Access software and analysis software now free the researcher from having to worry too much about the physical characteristics of machine-readable data held in a data library.

Re-reading that now, I would take issue with what was said, by Sue and by myself. However, perhaps that planted the seed for the view I took later to separate 'data' from the 'digital', regarding the former as only being so if it (they?) could be regarded as having evidential value for some enquiry, and the latter prompting the question 'what is different about the digital?' with focus on the malleability of the medium.

I made another reference to the work of Sue Dodd on page 9 in the section on 'Documentation' and then again when discussing the value of the Abstract, before placing her words centre stage when discussing cataloguing of machine-readable data files. This was an opportunity to combine my new found 'cataloguing' knowledge with some of the practices I had learnt from my time working as a survey statistician and researcher with the Scottish Education Data Archive. The stated purpose for my report to the UK Committee of Librarians and Statisticians was to highlight the existence of the data behind those statistical tables in government publications, and of the value of what I termed 'an online meta-database'. I also wanted to think aloud and see what was wanted of a 'data library' from the different perspectives of a data analyst and of a data producer.

> In this paper I look at data libraries from each of two directions: from the point of view of those who want to use the data, and from the point of view of those who generate the data; that is, from the point of view of data analysts and data producers. The paper also includes a rough historical sketch of the development of data libraries in the academic (mostly social scientific) sector; a discussion of the importance of bibliographic control and the provision of an on-line meta-database. ('data about data'), and highlights the trend towards access to the data that produce statistical tables.

Although not formally published that article is now, belatedly, in the University's institutional repository – scanned from a printed copy – and reportedly still being downloaded every month (Burnhill, 1985). In what now looks like a 'use case workflow', I wrote:

> When using a data library the data analyst may be motivated either by the need to provide information for managers and decision makers, or by the wish to contribute towards some longer term research enterprise. Either way, the data analyst asks something like the following series of questions:
>
> 1 Would the problem in hand benefit from empirical evidence?
> 2 Are there data available which could shed light on this problem?
> 3 Where is the database located?
> 4 How may I negotiate access?
> - Permissions; Mode of access; Payment or funding implications
> 5 What is the provenance, status and quality of the data?
> - Questionnaire; Target population; Sampling scheme; Non-response
> 6 Can I obtain codebooks and allied documentation?
> 7 How may I re-cast my problems so that these data can contribute?
> 8 What software is available for data retrieval, manipulation, analysis and presentation?
> 9 Could I use this software myself?
> 10 How may I obtain hard copy of the results from the analysis?
> 11 What would be the cost in time and money?

Regrettably, I look back on that paper as something of a 'failed manifesto' as the development of data libraries in the UK was much delayed – even now they exist in very few universities. However, the paper was influential at the time as evidence in the joint enquiry by the ESRC (UK) and NSF (US), alongside a contribution from Alice Robbin, a Past IASSIST President (1979-82) and then Director of the Data and Program Library Service, University of Wisconsin-Madison. The ESRC leadership was provided by Howard Newby, previously a Director of the Data Archive at Essex who would go on to be Chairman and Chief Executive of the Economic and Social Research Council (ESRC), and then CEO of the Higher Education Funding Council for England (HEFCE).

My second encounter with Sue Dodd

The second time I first met Sue Dodd was when I spoke to her in person on the 'phone. I had come to the conclusion that there was insufficient knowledge in the UK 'Anglo' part of AACR2 about the new Chapter 9. I decided that I should try to persuade Sue to come to visit the UK and that the best way to achieve that was to reach out to her by tracking down her number at Chapel Hill, North Carolina, which I then dialled. The voice at the end was slightly taken aback, as transatlantic calls were far from usual, for either of us. I established that she was interested in participating in the two seminars I then proposed, one to be held at the University in Edinburgh and one in London under the auspices of RSS/LA Committee of Librarians and Statisticians.

Sue was not at the IASSIST Conference in Amsterdam, May 1985, the first I attended. However, I did meet a number of the other names I had come across in those issues of the *IQ*. I also began to see some differences and divisions in the European approach being taken, with the practice of the national data archives in Europe, and that adopted in the US/Canada in which there were many university-based data libraries.

My third encounter with Sue Dodd

The third time I first met Sue was the delight of meeting her in person when she did indeed accept our invitation to travel to the UK. I recall that she noticed the jet lag but was determined to be positive and helpful. We took the opportunity to enjoy a travelling exhibition of the Terracotta Warriors that was visiting Edinburgh. I learnt later of her graduate studies about China.[5]

Advertisements for the two meetings had been distributed over the Summer of 1985, including this one:

> SEMINAR ON BIBLIOGRAPHIC CONTROL OF STATISTICAL DATA FILES
>
> As the number of machine-readable statistical data files increases it is becoming ever more difficult for data users to find out about all the data which may be relevant to their work. The need for a comprehensive register, or national bibliography, of data files is becoming apparent. How could this be prepared? Could it be compatible with bibliographies and library catalogues of printed material? How might it relate to output from the European Access Project with which the ESRC Data

> Archive is involved? What is the role of data libraries in making data accessible to the user community?"
>
> In order to provide an opportunity for discussion of these and related questions, the Committee of Librarians and Statisticians is organising a seminar at the City University, London on Monday 23 September 1985. The principal speaker will be Sue Dodd, a Data Librarian at the University of North Carolina, whose pioneering work in developing standards for cataloguing machine readable data files has earned her an international reputation. Other speakers include Marcia Taylor and Bridget Winstanley (ESRC Data Archive), Peter Burnhill (University of Edinburgh Data Library Services) and Geoffrey Hamilton (British Library).
>
> . . .
>
> While she is in the United Kingdom, Sue Dodd will also lead a workshop on "Computer-based catalogues for describing computer files and their documentation" on Friday 20 September 1985 at the University of Edinburgh, 18 Buccleuch Place, Edinburgh.

The title for the Edinburgh workshop centred on what I still think is still moot, namely whether 'data file and documentation' necessarily and collectively constitute a multi-part object – indeed, whether there is any simple object where data files are concerned.

Unlike many data libraries in North America all data files at Edinburgh were online and spinning on disc, not stored physically on tapes held in labelled tape racks. Moreover there was an online 'catalogue' of what was held in the Data Library. Just before Sue visited, Alison Bayley had joined the Data Library as a part-time programmer.[6] Alison was developing the online information service 'datalib' enabling users to navigate a form of hypertext in 'eview' (called simply View in EMAS) to find information on services, facilities, filenames, access restrictions, etc. This had many descriptive fields of our own making.

I recall that Sue's visit prompted an attempt to create a catalogue record for the small area statistics from the 1971 Population Census for Scotland in the University Library's (OPAC) catalogue. This led to interesting discussion with Peter Berwick, the Library's Head Cataloguer, when it was suggested that we change the title in order to improve the way in which the item would be filed. There was nothing of a title found in the 'item in hand': what had been received had no header file with a descriptive title. We were introduced to 'Toward Integration of Catalog Records on Social Science Machine-Readable Data Files Into Existing Bibliographic Utilities: A Commentary' (Dodd, 1982a).

The seminar at City University was interesting, attracting a wide variety from the library world as well as the Data Archive at Essex. I recall that Sarah Tyacke was there, then Deputy Map Librarian at the British Library. The next year she became Director of Special Collections in the Library and subsequently Keeper of Public Records and Chief Executive of The National Archives where she oversaw the development of new strategies for dealing with the preservation of born-digital records.

Through her visit contact was made with Ray Templeton of the Library Association who had been working on standards for cataloguing the recent phenomena of software for microcomputers, (Templeton and Witten, 1984). Ray and I were later to share the task of editing a Guide that resulted from the ESRC Computer Files Cataloguing Group (Burnhill and Templeton, 1989) which drew much from Sue's *Cataloging machine-readable data files: an interpretive manual* (Dodd, 1982b).

The knowledge derived from Sue's work had practical application as the Data Library participated in the ESRC Regional Research Laboratory (RRL) initiative as part of RRL Scotland (Burnhill, Carruthers and Messer, 1988; Burnhill and Ewington, 1992). The 'RRL initiative' provided an opportunity to engage with the developing field of geographic information systems. Particularly significant was a symposium sponsored by the UK Association for Geographic Information on 'metadata in the geosciences' in 1990. This brought together several disciplines having interest in 'metadata' and its relation to 'cataloguing information', especially as this might relate to spatially-referenced data.[7]

Metadata was characterised within the database community as the data dictionary that gave formal definition for the objects in the database. There was the beginning of understanding that additional metadata were required to support resource discovery. The term 'actionable metadata' was used to go beyond that needed for data discovery to include information that could be read and acted upon by software, not only metadata to identify relevant data for a user but also to retrieve the relevant data from a (remote) database and produce a predefined product such as a map or table (Burnhill, 1991; Medyckyj-Scott et al 1995). There was attempt to juxtapose these new metadata requirements with the cataloguing fields from AACR2 Chapter 9 that had their focus on 'identification and availability', 'subject and content', 'characteristics of the media' and 'access and management' (Burnhill, 1991).

Returning from the 1995 IASSIST Conference, hosted in Québec, Canada, I learnt that the University of Edinburgh had decided to respond to a national (UK) call for a third national datacentre (at that time there were BIDS, at Bath, and MIDAS, at Manchester) and wished to put forward the Data Library as the basis of that bid. Three weeks later the bid went in. Two months later we learnt that the University was successful, and we were given five months to be up and running and delivering online services. We launched EDINA, the poetic name for Edinburgh, on 25 January 1996, on Burns Night, starting with BIOSIS Previews, a bibliographic database.

That event might signal the date when my energies finally shifted away from the sharp focus on the social science data file. The prior contact with database experts and working with geospatial and mapping data had already prompted the beginnings of that shift. I have come to remember the strap-line for the 1990 IASSIST Conference as "Words, Numbers, Pictures, Sounds: All will be digital and accessed from afar." In fact, although I recall proposing the strap-line in the Programme Committee, it was actually "Numbers, Pictures, Words, and Sounds: Priorities for the 1990's." During the early 1990s, my management responsibilities broadened, to be required to deliver computing support given to the Library: staff in the Data Library began to learn more about text, and to carry out project work that led us to launch SALSER[8], 'probably the first Web-based national union catalogue of serials'.

EDINA continues today with a very broad range of services, <http://edina.ac.uk>, and with the mission to develop and deliver online services as part of the 'Jisc Family'[9] in order to enhance research and education in the UK, and beyond. The best way to appreciate the present spread of activity is to download the 'Community Report'; perhaps the best way to appreciate the

variety of activity over the years is to dip into the online archive of past issues of *''EDINA Newsline'*.

The Data Library continues to flourish and have purpose: it has its data catalogue[10] as well as a set of services geared at benefiting researchers, students and their teachers at the University of Edinburgh.[11] My colleagues in the Data Library, which together with EDINA form part of Information Services at the University, also contribute nationally and internationally. Examples include significant contribution to the University's focus on research data management (Rice et al, 2013) and MANTRA[12] , an online course designed for researchers or others planning to manage digital data as part of the research process. That includes a module on metadata and documentation in which three broad categories of metadata are described as part of training for future researchers:

- Descriptive - common fields such as title, author, abstract,
- keywords
- Administrative - preservation, rights management & technical metadata
- Structural - how components of a set of associated data relate to one another, such as a schema describing relations between tables in a database.

Active participation in IASSIST continues, including recent secondment of Stuart Macdonald to Cornell University and the temporary addition at Edinburgh of Laine Ruus, one of the famous names I read about in those early editions of the *IQ* alongside Sue Dodd, and whom I also cited in that first article, "Towards the Development of Data Libraries in the UK" (Burnhill, 1985):

> What is needed is a union catalogue of all known disseminators of MRDF, and some efficient means to access information on what new data files are being created. The movement by ICPSR and the Roper Centre towards on-line remote access to their inventories is a major step towards information retrieval." (L. G. M. Ruus, 1980).

Part B (2013). Tales from the Keepers Registry: Serial issues about archiving & the Web

Fast forward some thirty years and I look back to when there was again need to cite the work of Sue Dodd. During those thirty years the digital medium was no longer confined to those machine-readable data files that Sue had focused upon: the digital medium had become the norm for scholarly statement, as with much in everyday life. The privileged access to the Internet had given way to mass engagement with the Web as an arena of interaction.

Invitation to contribute an article for *Serials Review* had prompted me to renew my acquaintance with the writings of Sue Dodd. I was writing about the arrangements being made in order that we might know what e-journals were being kept safe and what remained at risk. I had been asked to report on progress being made to ensure continuity of access to scholarly literature given the shift from print to digital format for all types of continuing resources, particularly journals, and the need to archive not just serials but also ongoing 'integrating resources' such as databases and Web sites. My principal reason for citing Dodd (1982a, 1982b) was to place her work within the history of AACR2, in part also to alert today's librarians to the work of social science data librarians now that research data from all disciplines was being listed high on their agenda.

I would like to use this occasion to alert social science data librarians to some ideas being taken forward now that scholarly content is issued as online resources, either issued in parts or changing over time. The article (Burnhill, 1985) contains three stories which centre on the Keepers Registry which monitors the extent of e-journal archiving.

The First Tale: The Keepers Registry

The first story in the "Tales from the Keepers Registry" describes the problem of e-journal preservation, as noted in a number of reports over the past 10 to 15 years and the emergence of organizations willing to act as 'digital shelves'. It also described the role of Keepers Registry as a global monitor on who is looking after what (how and with what terms of access). The Registry has enabled the generation of statistics that indicate the extent of archiving for e-journals is cause for concern.

Today researchers in the social sciences – as in all disciplines ranging from physics to philosophy - rejoice in the good news that scholarly statement is made available in ways that can be accessed any-time, any-place, and increasingly by any person and for any purpose. That advance had been greatly assisted by the emergence of the Web, the principle arena for interaction across the Internet. Authors can make their content available very readily, via publishers or directly (with or without explicit licence). Consumers of that content can shorten the time and effort required to discover, locate, request and access what they require (according to the licence). That is true for the produce of scholarship and for the resources that scholarship requires.

The bad news is that so much of this scholarly content is not in the custody of research libraries. Academic and research libraries continue to play a part but their role as intermediaries has been challenged, not least in their role as stewards of scholarly content that exists in digital form. Libraries depend upon e-connections; they do not have their own e-collections.
The shift to journal content that is digital, online and held remotely has challenged the essential responsibility that libraries have in

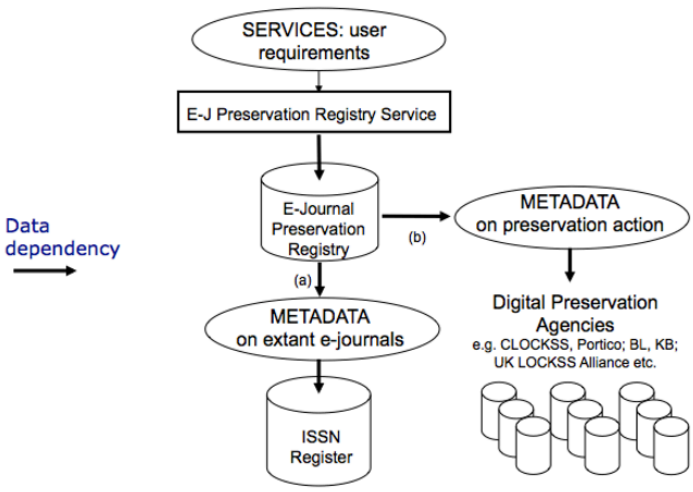

Figure 1 - Keepers Registry

ensuring continuity of access to scholarly content for their patrons. Following reports from studies and projects around the world, a small number of organizations stepped forward to act as long-term archives for e-journal content. Those reports noted the potential value of a resource that could address 'who was looking after what, how, and what are the terms of access?' The study commissioned by the JISC in 2007 (Sparks, Look, Muir and Bide, 2008) confirmed the feasibility and the perceived need for an e-journal preservation registry, indicating that such a registry could be built around the Serials Union Catalogue (SUNCAT), the national union catalogue in the UK developed at EDINA (Burnhill, Halliday, Rozenfeld & Kidd, 2004).

The Keepers Registry has now emerged as a global online facility[13], designed and built by EDINA at the University of Edinburgh in collaboration with the ISSN International Centre in Paris. The basics of the design are illustrated below, taken from Burnhill et al (2009), and show how the identifier for serials, the International Standard Serial Number (ISSN) and the ISSN Register is at the heart of the facility, against which the leading archiving organizations report on which serial titles (having ISSN) each is looking after: reporting metadata on how, to what extent and with what terms of access.

The real heroes in this first tale are those digital preservation agencies, the ten archiving organizations that are contributing to the Registry. As shown in the graphic above, the two main web-scale organizations of CLOCKSS and Portico were in from the start, as was the Global LOCKSS Alliance. The Library of Congress and HathiTrust are among those to have joined since - alongside the Archeological Data Service (UK), having a discipline-based archiving responsibility.

Considering the complexities of research data, it might be supposed that the preservation of e-journal content was easy and that the problem was solved. Unfortunately, that does not seem to be the case, as revealed by analyzing the archiving metadata that is aggregated in the Keepers Registry, as reported on the blog for the Keepers Registry.[14] Currently only about 22,000 e-serial titles of the 113,092 ISSN assigned to 'online serials in ISSN Register are reported as being 'kept safe' by the archiving organisations reporting into the Registry - and there are many 'missing volumes and issues'. In 2013, the simple coverage statistic is 19%, an increase from 2011 when it was 17% (being 16,558 / 97,563), and what is interesting is that the numerator and denominator are both increasing, as archiving organisations ingest more titles and as ISSN is assigned to an increasing number of 'points of issue', about which more later. Even if one narrows the focus to those serials that are considered important to libraries, the lists provided by Cornell, Columbia and Duke Universities, about 75% of e-serials (having ISSN) should be regarded as 'at risk'

The Second Tale: Metadata Matters

The second story in the "Tales from the Keepers Registry" is about the variety of metadata issues that had to be addressed during the PEPRS project, including a number that remain unresolved (Burnhill et al, 2009). Typically serials are 'well-published' with rich metadata made available to archiving organizations by publishers. However, there are challenges relating to identifiers; variants in publisher information (naming and identification, and reference to issuing bodies) and variability about 'holdings' information relating to issues, volumes, and other buckets of digital stuff.
The role of the ISSN has been key, the international standard identifier for a stream of content. The serial provides an entity which is 'economic' from an information management point of view, with discrete objects (typically as articles) made available in parts (typically as issues and volumes). Nevertheless, the article (file) remains the 'object of desire', being accorded its own identifier, the Digital Object Identifier (DOI). Attention is also given to the search for the universal holdings format in order to enumerate the extent of issued content, and thereby to check what is held and what may be missing.

The importance of another identifier is becoming plain. Until recently there was no universally accepted identification scheme for publishers, with name variants seen as part of the more general quest for authority files for personal and corporate names. A variety of name expressions is perhaps always to be expected, not just because of language differences. However, there is now a prospective solution with the emergence of the International Standard Name Identifier (ISNI), an ISO (International Organization for Standardization) Standard (ISO 27729), whose scope is identification for Public Identities[15]. The purpose of ISNI is to assist disambiguation of the public identities involved throughout the creation, production, management, and content distribution chain. That includes both organizations and persons (whether living or dead): there is a special allocation of ISNI numbers made available for assignment as ORCID.[16]

The first two Tales were presented in *Serials Review* in ways that were intended to engage serial librarians. The applicability of all of this for data librarians may not be self-evident but I would like to argue that there is much to be gained by considering how those matters might extend beyond such 'well-published' material as journals, especially to those social science data files that are generated from periodic enquiry and process. The emphasis is on identification rather than a full 'bibliographic record', and on the simplicity in a 'data registry' of knowing 'who is doing what'.

The central idea for a Registry such as an e-journal preservation registry, the model for which might be generalized and adapted for other purposes, is part of a four-point proposition:

1. Assign an identifier at the 'point of issue' for a stream of digital content
2. Ensure that (digital) content is archived routinely, and that arrangement is made to have others/peers do that for you too
3. Tell someone what you are doing and what you hold (and how)
4. Publish the terms of access for the archived content (now and when triggered as orphaned).

The Third Tale: Where data and journal content collide

The third story in the "Tales from the Keepers Registry" was also written with the serials librarian in mind. The intention was to look beyond the conventional journal to the new research objects that have now become to be recognized and to the implications of the dynamics of the Web. There is focus on the implications for citation, for notions of fixity, and for broader matters of digital preservation.

I wished to highlight for the serials librarian some of the consequences for scholarly statement now that the Web was becoming a principal arena for scholarly communication. Not merely a dominant means to access, the Web also enables rich aggregations of linked content into what have been termed 'Research Objects' having two classes: Archived Objects and Publication Objects that "are intended as a record of activity, and

should thus be immutable" and citable (Bechhofer, De Roure, Gamble, Goble and Buchan, 2010). This can be seen to have built upon an attempt "to distill some core characteristics of a future scholarly communication system" (Van de Sompel, Payette, Erickson, Lagoze & Warner, 2004) with both registration (and ultimately preservation) of a scholarly asset being central to its success within a workflow or pathway through various service hubs.

What are data librarians to make of these new scholarly objects that are growing in significance as part of the new information infrastructure for scholarship enabled by the Web? Thirty years ago it was important for so very many reasons to highlight the special case of social science data files as resources for scholarship and to contrast these with the apparent simplicity and fixity of what appeared as scholarly statement, as articles in journals and books on shelves. In the interim, scholarly statement has become digital and therefore malleable, with the characterization made above, it is now also extended to include data as intrinsic to that statement.

In that telling of the third tale I wanted to point out to serial librarians that the shift to a broader view of scholarly works in digital format should not necessarily be regarded as completely new and alien, noting that Sue Dodd had made important observation thirty years ago in the pre-Web era of the Internet. And we are reminded that she wrote that "There is no doubt that machine-readable data will play an even greater role in research and development programs of the future. More and more data needed for government and private research will appear in computerized form." (Dodd, 1982a, p352); "In the near future, libraries will have no choice but to become more involved with computerized files and programs." (op cit, p355). She was of course writing in the context of the publication in 1978 of AACR2 Chapter 9 on 'Machine-Readable Data Files,' renamed 'Computer Files' in the revision published in 1988.

On the other hand, this third tale could be interpreted and re-stated as a story that reflects upon the value of the concept of 'seriality' for data librarians and archivists. I have become convinced that this is a key concept for the structure of metadata for much that is issued on the Web and indeed for much of what we were and still are interested in for 'secondary data analysis of machine-readable data files'.

Complete revision of Chapter 9 saw it become 'Electronic Resources' in the 2001 amendments that were confirmed in AACR2 2002, which also saw Chapter 12 on 'Serials' renamed 'Continuing Resources,' driven by a wish to harmonize across AACR2 and other serials bodies, including ISSN. The motive was common belief in the usefulness of the concept of seriality for what was, following widespread adoption of the Web, being recognized as important points of issuance of content. The term 'integrating resources' was used to signify what was updated over time (differing from serials that are issued in separate discrete parts).

The manifesto noted above and described by Bechhofer et al (2010) is also reminiscent of work by Hunter and Choudhury (2006) and Hunter (2006) that focus, respectively, upon the preservation of composite digital objects using Semantic Web Services and the use of Scientific Publication Packages (SPPs) for linking the raw data, their associated contextual and metadata on provenance, as part of publishing and dissemination of scientific results and selective preservation of scientific data. There is determined focus upon a "unit of scholarly communication" that is not "journals and their contained articles." This evokes what are referred to as Compound units, "aggregations of distinct information units that, when combined, form a logical whole" and can be represented in a manner (OAI-ORE[17]) that enables them to be accessed and processed by machines and agents (Van de Sompel & Lagoze, 2007).

Seriality of issuance as such is not utilized in the argument put forward by Bechhofer et al (2010). However, now that the Web is recognized as an important point of issuance of scholarly content, both of scholarly product and of resource for scholarship, there is need for identification and 'minimally-sufficient' description of that stream, recognizing that some content is issued in separate discrete parts, and some changes (or is retrospectively updated/modified) over time.

What is particularly interesting about the article on Research Objects cited above was how it was made available; it was issued as a reviewed conference paper in *Nature Precedings*. At first sight, *Nature Precedings* resembles a journal, but it is not. Launched in 2007 and closed in 2012, it acted as an open access preprint repository for the Life Science community. It was an integrating resource and as such assigned an ISSN, 1756–0357. The ISSN assignment policy now is being extended to online repositories as first point of issue for an increasing number of scholarly works. It may yet extend to repositories, such as *figshare*[18], that exist to make research data and other forms of research output publically available. One wonders whether that ISSN assignment policy should and could extend to social science data archives.

This third tale mentioned a project being carried out jointly by the Research Library at Los Alamos National Laboratory and EDINA and the Language Technology Group at the University of Edinburgh.[19] That investigation into what is termed 'reference rot' is now underway (Sanderson, Van de Sompel, Burnhill and Grover, 2013). Reference rot describes when content referenced at the end of the link has evolved, has changed dramatically, or has disappeared completely; it is more than 'link rot'. An engaging overview is given in a talk by Van de Sompel (2011) about the use of the Memento tool to access prior versions of Web resources available from Web archives and content management systems by using their original URI and a constructed 'date-time stamp' for the desired version, a bit like 'Time Travel for the Web'. Preliminary work examining the survival of Web-based content cited in articles in two scholarly repositories noted that 28% of the resources referenced by the articles in an institutional repository had been lost, and 45% (66,096) of the URLs (in arXiv) that were found to still exist had not been archived (Sanderson, Phillips and Van de Sompel, 2011).

It may be fitting to end this appreciation on the topic of citation. The contrast with the fixity associated with earlier printed format for scholarly statement is obvious. That contrast with the past is less obvious for the dataset, despite the suggestion made by Dodd (1982a) to "conceptualize a singular MRDF to be an 'inert file' … that conceptually becomes the 'item in hand' to be described". That was clearly said with the librarian of the early 1980s in mind. However, today's data librarians and data archivists might be reassured to note, that Dodd (1982a) also drew attention to the "dynamic data base [as] one that is characterized by its fluid and constantly changing nature. It may be represented by economic time series, or bibliographic data bases, and may be corrected, revised retrospectively, updated, merged, partitioned, and blocked

into subfiles without changing its bibliographic identity." Although this latter observation predates the arrival of the Web it should underscore our recognition that the Web is dynamic. What may have existed, as indicated by citation, at the moment of reference can and does change. Once more we must pay renewed attention on how to cite the (web-based) data resources that are issued beyond the traditional journal literature.

References

Bechhofer, Sean, D. De Roure, M. Gamble, C. Goble, & I. Buchan. (2010, July 6). Research Objects: Towards Exchange and Reuse of Digital Knowledge. Nature Precedings. doi:10.1038/npre.2010.4626.1 <http://precedings.nature.com/documents/4626/version/1>

Burnhill, Peter. (1985) Towards the development of data libraries in the UK, Edinburgh: Centre for Application Software and Technology 1985. Available from: <https://www.era.lib.ed.ac.uk/handle/1842/2510>

Burnhill, Peter, Ann Carruthers, and Anne Messer. (1988) Bibliographic control of research data. Edinburgh: Regional Research Laboratory for Scotland.

Burnhill, Peter and Templeton, Ray. (Eds.) (1989) Cataloguing Computer Files in the UK: A Practical Guide to Standards. (Joint Report to the Computer Files Cataloguing Group, Economic and Social Research Council.) Colchester: ESRC Data Archive, University of Essex.

Burnhill, Peter. (1991) Metadata and cataloguing standards: one eye on the spatial, in Metadata in the Geosciences. Ian Newman, David Medyckyj-Scott, Clive Ruggles and David Walker (Eds). Loughborough: Group D.

Burnhill, Peter, and Ewington, Heather. (1992) Catalogue of digitized boundary files for England and Wales held by Edinburgh University Data Library. Working paper 33. Edinburgh: Regional Research Laboratory for Scotland.

Burnhill, Peter, Leah Halliday, Slavek Rozenfeld, & Tony Kidd. (2004) SUNCAT: a modern serials union catalogue for the UK. Serials 17 (1), 61-67. Available from: <http://uksg.metapress.com/content/c6y0ltjtxlhrgfr9/>

Burnhill, Peter, Francoise Pelle, Pierre Godefroy, Fred Guy, Morag Macgregor, Christine Rees and Adam Rusbridge. (2009) Piloting an e-journals preservation registry service (PEPRS). Serials 22 (1), 53-59. Available from: <http://uksg.metapress.com/link.asp?id=350487p5670h0v61>

Burnhill, Peter. (2013) Tales from The Keepers Registry: Serial Issues About Archiving & the Web. Serials Review 39 (1), 3–20. Available from: <http://www.sciencedirect.com/science/article/pii/S0098791313000178> and <https://www.era.lib.ed.ac.uk/handle/1842/6682>

Dodd, Sue A. (1979) Bibliographic references for numeric social science data files: Suggested guidelines. Journal of the American Society for Information Science 77–82.

Dodd, Sue A. (1982a) Toward Integration of Catalog Records on Social Science Machine-Readable Data Files Into Existing Bibliographic Utilities: A Commentary. Library Trends 30 (3). <http://hdl.handle.net/2142/7215>

Dodd, Sue A. (1982b) Cataloging machine-readable data files: an interpretive manual. Chicago: American Library Association.

Hunter, J. (2006). Scientific Publication Packages – A Selective Approach to the Communication and Archival of Scientific Output. International Journal of Digital Curation 1 (1) 33-52. doi:10.2218/ijdc.v1i1.4

Hunter, J. & Choudhury, S. (2006). PANIC – An Integrated Approach to the Preservation of Composite Digital Objects using Semantic Web Services. International Journal on Digital Libraries: Special Issue on Complex Digital Objects. 6 (2), 174-183. <http://www.springerlink.com/content/p480171432571j25/?MUD=MP>

Jones, Trevor and Stacey, Audrey. (1984) Data Library Service: introductory guide. Edinburgh: Centre for Applications Software and Technology, Edinburgh University.

Medyckyj-Scott, D. J., C. Monckton, P. Burnhill. (1995) Progress towards standards for spatial metadata, in Rowley, J. (Ed.), Standards supplement to the AGI '94 Conference, London: Association for Geographic Information.

Rice, Robin, Cuna Ekmekcioglu, Jeff Haywood, Sarah Jones, Stuart Lewis, Stuart Macdonald, and Tony Weir. (2013) Implementing the Research Data Management Policy: University of Edinburgh Roadmap. The International Journal of Digital Curation. 8 (2) 194-204 <http://dx.doi.org/10.2218/ijdc.v8i2.283>

Ruus, Laine G.M., (1980), User services in a data library. IASSIST newsletter 4 (2), 29-33

Sanderson, Robert, Herbert Van de Sompel, Peter Burnhill and Claire Grover. (2013) Hiberlink: towards time travel for the scholarly web. Proceedings of the 1st International Workshop on Digital Preservation of Research Methods and Artefacts (DPRMA '13). ACM. <http://www.deepdyve.com/lp/association-for-computing-machinery/hiberlink-towards-time-travel-for-the-scholarly-web-7bJrwFVcPA>

Sparks, Sue, Hugh Look, Mark Bide, & Adrienne Muir. (2010) A registry of archived electronic journals. Journal of Librarianship and Information Science, 42(2), pp. 1-11.

Templeton, Ray and Witten, Anita. (1984) Study of cataloguing computer software: applying AACR2 to microcomputer programs. London: British Library; Distributed by Publications Section, British Library Lending Division.

Van de Sompel, Herbert, S. Payette, J. Erickson, C. Lagoze, and S. Warner (2004). Rethinking Scholarly Communication: Building the System that Scholars Deserve, D-Lib Magazine. 10 (9). <http://www.dlib.org/dlib/september04/vandesompel/09vandesompel.html>

Van de Sompel, H., Lagoze, C. (2007, August). Interoperability for the Discovery, Use, and Re-Use of Units of Scholarly Communication. CTWatch Quarterly. 3 (3). <http://www.ctwatch.org/quarterly/articles/2007/08/interoperability-for-the-discovery-use-and-re-use-of-units-of-scholarly-communication/>

Van de Sompel, H., R. Sanderson, M.L. Nelson, L. Balakireva, S. Ainsworth, and H. Shankar. (2009, November 6). Memento: Time Travel for the Web. Arxiv preprint. <http://arxiv.org/abs/0911.1112>

Van de Sompel, H. (2011, December 2). Time Travel for the Scholarly Web. Talk given at STM Innovations Seminar 2011 (<http://www.stm-assoc.org/events/stm-innovations-seminar-2011/>). <http://river-valley.tv/time-travel-for-the-scholarly-web/>)

NOTES

1. Peter Burnhill is Director, EDINA & Data Library, Information Services, University of Edinburgh, Causewayside House, 160 Causewayside, Edinburgh EH9 1PR. <p.burnhill@ed.ac.uk>
2. The first issue of what was to become the IQ was based upon reports of an IASSIST meeting held as part of the International Political Science Association in August 1976. It had been hosted in Edinburgh which was also to host the IASSIST Conference on two later occasions, in 1993 and 2005.
3. Prior to that I had been working for almost five years as a survey statistician and researcher at the Centre for Educational Sociology in the University's Social Science Faculty, funded by the Scottish Education Department. With colleagues I was designing and conducting surveys of school leavers and helping with a collection of survey datasets known as the Scottish Education Data Archive - doing a lot of what we now call data curation.
4. A Union list of statistical serials in British libraries, Committee of Librarians and Statisticians. London, Library Association, 1972.
5. Typical of her generosity Sue insisted in buying me a figurine of one of those Chinese warriors that still has a pride of place on the mantelpiece at home.
6. Twenty years later Alison and I would do a joint presentation at IASSIST 2003, entitled 'Getting to Know the Score: Using the First 20 Years to Plan the Next', found at <http://datalib.library.ualberta.ca/conferences/2003/presentations/>
7. That was my first encounter with David Medyckyj-Scott who eventually joined EDINA and Data Library in 1995/6 in order to lead the development of Digimap and of metadata for geo-spatial systems more generally.
8. SALSER is the union catalogue of serials holdings for Scottish universities, the municipal research libraries of Edinburgh and Glasgow, numerous smaller Scottish research libraries and the National Library of Scotland. It was launched in 1994 and is available at <http://edina.ac.uk/salser/description.html>
9. Jisc (formerly the Joint Information Systems Committee, and still commonly referred to as JISC) is owned by the representative bodies of UK universities, colleges and skills organizations, <http://www.jisc.ac.uk/>
10. <http://datalib.edina.ac.uk/>
11. <http://www.ed.ac.uk/schools-departments/information-services/services/research-support/data-library>
12. <http://datalib.edina.ac.uk/mantra/>
13. <http://thekeepers.org>,
14. <http://thekeepers.blogs.edina.ac.uk/>
15. <http://www.isni. org/>
16. ORCID (Open Researcher and Contributor ID) is an alphanumeric code to uniquely identify scientific and other academic authors. <http://orcid.org/about/>; <http://www.isni.org/content/isni-other-identifiers>
17. Open Archives Initiative Object Reuse and Exchange (OAI-ORE) defines standards for the description and exchange of aggregations of Web resources. Sometimes called compound digital objects, these may combine distributed resources with multiple media types including text, images, data, and video. <http://www.openarchives.org/ore/>
18. <http://figshare.com/>
19. This project as funded by the Andrew Mellon Foundation was called 'Time Travel for the Scholarly Web' (TT4SW); the Hiberlink project website is at <http://hiberlink.org>

Keywords from Vol 37 N0. 1-4. Courtesy Tagxedo.com