Special Volume: Honoring the Work and Influence of a Pioneer Data Librarian

SPECIAL VOLUME
10 Papers

ON PAGE 5
Editor Karsten Boye Rasmussen

ON PAGE 74
Quick Guide to Data Citation
All word cloud / tag cloud graphics in this volume (p. 4, 27, 35, 44, 56) were prepared by Walter Piovesan, using the keywords for the articles in this volume plus the addition of the words “Sue Dodd” using software found at <tagxedo.com> and enhanced with Photoshop.

Online at: iassistdata.org/iq
In this issue

5 Editor’s notes
Karsten Boye Rasmussen

6 Guest Editor’s Notes
Margaret O’Neill Adams

8 Social Science Data Files and Bibliographic Control: Contributions of Sue A. Dodd
by Libbie Stephenson

15 Sue A. Dodd’s Lasting Influence: Libraries, Standards, and Professional Contributions
by Ann Gray

18 A Legacy of Inspiration and an Enduring Smile
by Peter Burnhill

28 Social Science Metadata and the Foundations of the DDI
by Karsten Boye Rasmussen

36 Building the DDI
by Ann Green and Chuck Humphrey

45 The DDI Matures: 1997 to the Present
by Mary Vardigan

51 DDI Timeline
by Mary Vardigan

57 Building on the Work of Colleagues: A Moment of Reflection
by Jonathan D. Crabtree

62 The Evolution of Data Citation: From Principles to Implementation
by Micah Altman and Mercè Crosas

71 A Practical Approach to Data Citation: The Special Interest Group on Data Citation and Development of the Quick Guide to Data Citation
by Hailey Mooney
Keywords from Vol 37 No. 1-4. Courtesy Tagxedo.com
Editor’s notes

Special issue: A pioneer data librarian

Welcome to the special volume of the IASSIST Quarterly (IQ (37):1-4, 2013). This special issue started as exchange of ideas between Libbie Stephenson and Margaret Adams to collect papers relating to the work of Sue A. Dodd. Margaret Adams (Peggy) acted as the guest editor and the background and content of this volume is described in her preface to this volume on the following page. As editor I want to especially thank Peggy and Libbie for pursuing and finalizing their excellent idea. I also want to thank all the authors that contributed to produce this volume. As one of the authors I can witness that Peggy did a great job.

Articles for the IASSIST Quarterly are always very welcome. They can be papers from IASSIST conferences or other conferences and workshops, from local presentations or papers especially written for the IQ. When you are preparing a presentation, give a thought to turning your one-time presentation into a lasting contribution to continuing development. As an author you are permitted “deep links” where you link directly to your paper published in the IQ. Chairing a conference session with the purpose of aggregating and integrating papers for a special issue IQ is also much appreciated as the information reaches many more people than the session participants, and will be readily available on the IASSIST website at http://www.iassistdata.org.

Authors are very welcome to take a look at the instructions and layout: http://iassistdata.org/iq/instructions-authors

Authors can also contact me via e-mail: kbr@sam.sdu.dk.

Should you be interested in compiling a special issue for the IQ as guest editor(s) I will also be delighted to hear from you.

Karsten Boye Rasmussen
April 2014
Editor
Guest Editor’s Notes

The Evolution of a Special Issue of the IQ in Honor of Sue A. Dodd: How remembering a pioneer data librarian became a tribute to the contributions of IASSIST over the decades

In the months before the 2012 annual IASSIST meeting, Libbie Stephenson contacted me to toss around ideas for some type of memorial we might plan for the coming meeting to honor our friend and colleague, Sue A. Dodd. Sue had passed away in October 2010. Libbie spoke of the important role Sue had played as a mentor in her data librarian career. She talked about how with the passage of time, newer members of our professional community might not know about the early days of IASSIST nor of the important contributions made then by Sue and others. I agreed and we decided to meet during the conference to discuss all of this further, allowing time for our ideas to percolate. When we met, quite informally, our discussion included others from Sue’s cohort: Tom Brown, Carolyn Geda, and Judith Rowe among them.

We agreed on the value of gathering original essays that would consider Dodd’s contributions as a conceptual starting point, especially regarding data description broadly conceived. We decided that in moving beyond this, the essays should address the history and development of related topics of special interest, reflecting the expertise of the authors. Subsequently, we discussed our ideas with the IQ editor, Karsten Boye Rasmussen, who was fully supportive. I volunteered to guest edit the collection and Libbie stated her interest in preparing a bibliographic essay on Sue’s publications. In subsequent months we contacted colleagues whom we thought might be interested in participating in the project and were pleased at the responses. As can be seen by a quick scan of the Table of Contents, they were so positive that our initial goal for “an IQ issue” grew to constitute the whole of IQ Volume 37 (2013)³.

Among the numerous themes that run through these essays, the collegiality within IASSIST is ever present. Perhaps the most interesting trait the essays have in common is the unique way each alludes to Dodd’s contributions. In explaining our concept of the project to potential authors, we did not expect that they would explicitly tie their work to hers, yet each essay does so in one way or another. Among the authors who knew Sue, some of the commentary reflects the personal as well as the professional. Two of the authors, Ann Gray and Jonathan Crabtree, share experiences of working at the Institute for Research in Social Science at the University of North Carolina, Sue’s professional home for more than 30 years. Peter Burnhill charms with memories of his first encounters with Sue while giving us a glimpse of the early days of the Edinburgh University Data Library and then EDINA, and his almost 30 years-later encounter with her writing as he embarked on new ventures related to digital preservation of what he terms “scholarly statement.”

Several authors note both implicitly and explicitly the manner in which ideas in the 1960s – 1980s were both prescient in terms of challenges still vexing the social science data community and foundational in terms of setting the course that has influenced the collaborative efforts of the international data community ever since. One of the more exciting results of the project is that it offered a venue for an exclusive: a thorough telling of the history of the Data Documentation Initiative (DDI). The comprehensive story of the DDI emerges for the first time in the essays contributed by Karsten Boye Rasmussen, Ann Green and Chuck Humphrey, and Mary Vardigan, and through the detailed and richly documented timeline of the DDI prepared by Mary Vardigan. Taking us from the early study description work at some of the European data archives and the contemporaneous documentation standards efforts in North America, their essays give us an overarching perspective on the context, motivation and requirements behind the design and development of metadata standards, the DDI, and the infrastructural and technological changes that contributed to the contemporary maturation of the DDI.

In addition to a glimpse of the legacy of Dodd at the University of North Carolina, Jon Crabtree’s essay explores some of the more contemporary collaborative ventures of the wider data community, especially the Data-PASS (Data Preservation Alliance for the Social Sciences) project. In the following essay Micah Altman and Mercé Crosas take us fully into the present and look to the future, discussing a range of work related to data citation in the context of contemporary frameworks for data management and research. Hailey Mooney brings the essays full circle with her discussion of IASSIST’s Special Interest Group on Data Citation and its development of the Quick Guide to Data Citation. Sue Dodd would be very pleased to see this guide, we think, as it seems to be exactly the kind of product that she expected of future generations of data professionals, as expressed in the quote by which Libbie Stephenson closes the introduction of her comprehensive bibliographic essay.

Editing this volume of the IQ was a stimulating and enjoyable experience; working with each of the authors introduced me to new facets of our shared experiences and interests in unexpected ways. My thanks to each of the authors for their enthusiasm, commitment, and
patience. Individually and as a group they personify what IASSIST has stood for all these years: collegial collaboration of social science data professionals across the world. That’s what made this endeavor so interesting. I discovered that in honoring the memory of one of our early members, we actually pay tribute to the many contributions of all of IASSIST. And as a bonus, the essays make for a good documentary read. Thank you, all.

Margaret O’Neill Adams (Peggy Adams)
April 2014
Guest Editor
peggyoadams@yahoo.com

Notes
1. Sue Anna Dodd, b. 6/18/1937 in Lexington, KY; d. 10/6/2010 in Pittsboro, NC.

2. There is an additional area of Sue Dodd’s professional contributions that is not covered in these essays because the work involved was outside their topical scope of consideration. However, in keeping with one of the goals for this project — to inform readers of Sue Dodd’s contributions — we add this note.

Building upon the work that her publications represent and her outreach to the professional library community, Dodd was among those who also contributed to the professional community of traditional archivists. She participated in training workshops for them, focusing especially on the description and documentation of archival machine-readable records. Most importantly and indicative of her civic commitment, she completed a multi-month consultancy during 1986-87, at the [U.S.] National Archives and Records Administration. She produced a substantial but unpublished report that covered her study of computer records and the Federal records management program, and case studies of the computer records in two Federal agencies: the Bureau of the Census and the Department of State. In the report she identified challenges, prioritized problems, and presented a set of recommendations, with cost estimates, to improve the management of computer records by the National Archives. (Dodd, S.A. “Computer Records and the National Archives: An Assessment With New Directions,” 1986-1987)
Abstract
Sue A. Dodd was active professionally in efforts to define and describe social science data files for library catalogs. Her two books on guidelines for cataloging data files and software were important contributions to our understanding of the core elements used for this purpose. Numerous reports and documents were produced through her work with IASSIST as part of the Classification Action Group. This overview discusses a selection of Sue Dodd’s published works and the following list of references are discussed individually. Also included are references for works of significance which relied heavily on Sue Dodd’s research.

Keywords: Cataloging, Bibliographic control, Metadata, Library catalogs

It is important to remember that until the work carried out by Dodd and others there were no widely known and systematically organized catalogs, inventories, or bibliographies of data files extant in the U.S. Some work had been discussed through the International Social Science Council (ISSC), largely due to the efforts of Stein Rokkan, and in the U.S. through the Council of Social Science Data Archives (Adams 2006). A publication calling for bibliographic conventions and standards was written by David Nasatir under contract to UNESCO to study “overcoming the barriers to realizing the fullest utilization of machine-readable social science data.” Nasatir called for not only the preparation of bibliographic details, but also for archives to enable variable-level searching across studies and across archives (Nasatir, D., 1973, p. 3 and p. 46).

The following items make interesting reading; many of the topics discussed are issues the data archive community is still concerned with today. A few details about these early days are illustrative. For example, there is mention of the “data explosion” to describe a situation not unlike the way today’s authors refer to the “data deluge.” And it should be noted that at that time, social science data primarily consisted of surveys, enumerations, public opinion polls, and some administrative records. Mainframe computers were the only electronic tools available to the researcher for carrying out statistical analysis. Today’s variety and size of file formats, including videos, images, simulations, games, etc., were not yet in the mainstream of what we now think of as research data. Still the volume of material even then was notable.

In addition to concerns about the amount of data available during the period when Sue Dodd was active, there was much discussion about the potential role for libraries in data management and calls for action. Even so this concept was not new and had been expressed since at least the late 1950’s when, as Nasatir mentions, “York Lucci and Stein Rokkan proposed a library centre of survey research data in a project sponsored by the School of Library Service at Columbia University in 1957” (Nasatir D., p 10). And library leaders such as Clifton Brock, whose 1967 work is quoted by Nasatir, state “Social Science Data Archives have developed entirely outside the scope of library systems,” and that “Scholars, organizations, and publications … [in the “data sector”] … are wholly outside the world of conventional librarianship” (Brock, 1967, p. 305). Brock’s
lament is echoed in a speech also described in Nasatir’s 1973 work in which Ralph Bisco in 1967 spoke about “Why should university libraries undertake data services when social science data archives are already providing them?” (Bisco, 1967). Further, the National Academy of Sciences, wrote Nasatir, conducted a study headed by Phillip E. Converse which concluded that “few research libraries are adequately staffed” … [and] “libraries appear overwhelmed by information revolutions on other fronts” (National Research Council, 1967, Chapter 3).

So it was within this environment that Sue Dodd began her work. The shift in thinking about roles and responsibilities for management of social science data files can be found in Sue’s writing in Drexel Library Quarterly, Journal of the American Society for Information Science, Journal of Library Automation, Library Trends, and Library Resources and Technical Services. These pieces are illustrative of increasing receptiveness of libraries to focus on the bibliographic aspects of describing and classifying social science data files. There were significant activities within the American Library Association to develop rules and guidelines for cataloging data files, and the Machine readable Catalog (MARC) format. But it would seem that the Report on the Conference on Cataloging and Information Services for Machine-Readable Data Files: March 28-31, 1978 had an impact that ensured the catalog rules would be implemented more widely in libraries. And Sue’s books, Cataloging machine-readable data files: A interpretive manual in 1982 and Cataloging microcomputer files: A manual of interpretation for AACR2 in 1985, aided even lone catalogers in small libraries and archives to create accurate records for their catalogs of holdings.

The following listing contains entries for publications, reports and other documents produced by Sue Dodd and her colleagues and cover the period between 1977 and 1990. Also included are pieces containing substantial reference to Sue Dodd’s work and her role within IASSIST. These works include Sue Dodd’s books on cataloging data files and cataloging software for what were then called micro-computers. There are reports she prepared for the IASSIST Newsletter (now known as the IASSIST Quarterly) in her role as a member of the IASSIST US Classification Action Group. Sue was active in professional organizations within the American Library Association and presented her work at meetings which were critical for their impact on and advancement of bibliographic identification of data files.

There have been a number of developments in both libraries and archives since Sue Dodd’s works were published. In the library cataloging realm Barbara Tillett’s Functional Requirement for Bibliographic Records (FRBR) entity-relationship model provided a new way to think about the links between bibliographic records for all types of materials and is concerned with entities, relationships and attributes (metadata) (Tillett, 2004). A new set of guidelines for cataloging will replace the Anglo-American Cataloging Rules; Resource Description and Access (RDA) will provide rules and instructions on recording data to reflect attributes and relationships associated with the entities defined in the FRBR. Further developments have led to the Resource Description Framework (RDF) model which extends the utility of entity-relationship models such as FRBR. And RDF links to metadata schema have been demonstrated. For example, Stefan Kramer et al. within the DDI Alliance, have produced a document on connecting RDF-Described datasets to other related resources … in the Semantic Web … more specifically, … to leverage the Data Documentation Initiative (DDI) to enable semantic linking of social science data to other data and related resources on the web” (Kramer, et al., 2012).

Work to ensure that datasets used in research are properly cited, to develop item-level, actionable metadata schema, and to link data to published content as well as to other resources has been driven by some of the same practitioners who were guided by Sue Dodd’s earlier efforts and by many newcomers to the field. One likes to think that efforts to consider the possibilities, build the theory and applications, and develop the tools available today have at least a kernel of kinship with the ideas and directions defined in Sue Dodd’s work. Perhaps it may be said without too much hyperbole that this was her expectation given that she wrote in the IASSIST Newsletter in 1978, “[s]uccess is invariably measured not by what you hope to achieve, but by what you are able to produce. At the same time, the sum total of the lessons learned and the refinements made in the initial stages of any new endeavor becomes the foundation for future successes. The work described here is still developmental but it is designed to be expanded and implemented by other parties.”(Dodd, 1978, p. 37).

**Selected bibliography of books, articles and citations**

*(Text extracted directly from works in quotes)*


Commentary: “Libraries are traditionally well equipped to handle … informational needs, in that they have standardized procedures for maintaining bibliographic control on a multimedia collection of materials. Coupled with a recent commitment to automated systems, library procedures offer an important way of dealing with the existing problems of organizing, classifying, and cataloging information on social science data files. This paper will focus on the feasibility and future implications of applying library procedures and standards to MRDF, starting with the most important step—cataloging” Dodd describes the process by which rules and guidelines for cataloging data files were developed through American Library Association Committees and Library of Congress. The IASSIST Classification Group work and projects are outlined. This article includes one of the first calls for national level efforts to “establish a national program of information services … through a shared and cooperative network.” One could easily repeat Dodd’s insistence in 1977 that “the time has come to focus our attention on our national and commonly held responsibilities.”


Commentary: “Agenda topics: discussion of useful areas of concern and future coordination of tasks; 2) Review of cataloging efforts to date … including discussion of the Working Manual for Cataloging Machine-readable Data files (MRDF) compiled by Sue Dodd; 3) Discussion of the ramifications of MRDF catalog records, such as the national union list of social science data; shared cataloging; cataloging-in-production; the MARC II record as a standard format for storing automated bibliographic records …; 4) Practical exercise in applying subject headings and descriptors for several large and uniquely held datasets, with a view towards compiling the beginning of the thesaurus or authority list of social science..."
terms for data files; 6) Discussion of the lack of adequate subject headings and sub-headings currently provided by the Library of Congress for social science data files, with a view toward providing constructive recommendations.”


Commentary: “The primary emphasis of the US Classification Action Group of IASSIST has been on establishing standards and on the study of library information systems as they may apply to social science data files. Some of the recent developments within the library system which the Classification Group is examining include: 1) the development of rules and guidelines for cataloguing machine-readable data files (WRDF); 2) the development toward the acceptance of the MARC (Machine-Readable Catalog) record format as a universal standard for the automated bibliographic record; 3) the development of networks and on-line information systems which allow for multiple input and immediate retrieval of information; 4) the development of thesauri and controlled vocabularies for social science terms; and, 5) the development towards future considerations of a national union list of available MRDF and their location.”


Commentary: “The primary focus of the CAG at the Canadian Working Conference centered on the first problem of how to cite properly a social science numerical data file in the published literature…Agenda topics: Current problems and associated tasks within the mandate of the Classification Action Group include developing (1) examples and guidelines for bibliographic references for social science numerical data files, (2) a ‘cataloguing-in-production’ scheme for major producers of social science data files, (3) a more ‘universally based’ classification scheme for social science data files; and, reviewing (4) the cataloging efforts to date and discussing any or all related problems and (5) existing printed thesauri in the social sciences in terms of their future applications to data files.”


Commentary: “Two recent attempts to compile ‘catalogs’ of social science data have encountered the lack of consistency among titles for the same data set. One attempt has been the recent cataloging efforts at the Universities of North Carolina, Wisconsin, Princeton, and Yale, whereby traditional library cataloging records are created for social science data generated by academic research. The other has been the efforts of the Association of Public Data Users (APDU) to compile a directory of publicly available data files which represent primarily government produced data. Both groups have experienced the same problem: variance of titles for the same data file. Yet, without some control over titles and some mutually agreed upon primary source of title information, there can be no bibliographic control of social science data and none of the related products such as a union list of machine-readable data files. This paper will attempt to offer some suggestions for remedying the situation, including guidelines for transcribing titles; for creating a “title page”; for compiling a bibliographic reference; and for establishing an “authority list” for titles.”


Commentary: “Social Science numerical and textual data files represent a vast amount of valuable and publicly available information. For example, they are widely used by students, faculty and policy makers engaged in research. Not only have such data files had an unprecedented growth in the last decade, but with the advance of small and relatively inexpensive computer terminals, data analysis and computer simulation models have moved into the classroom as legitimate instructional tools. Specialized files, often referred to as ‘educational data packages’ have been developed to teach students analytical skills, so as to better understand social and economic phenomena. According to Nesvold (1976): “Experience with machine-readable ‘laboratory’ materials should be as appropriate to the beginning social science student as is the laboratory for the beginning chemistry student.”

Unfortunately, many such data resources are not fully utilized because potential users are unaware of the existence and accessibility of social science data files. At the present time, information on usable MDRF is fragmented among varying government agencies, research institutions, and university computing and data centers. Among these various agencies, there is no common format for information on the existence of data files, nor is there any standardized structure that would facilitate retrieval of information from many different sources. Existing information on computerized files is available to some but not to all. What is needed is a central source of information within the public domain that would provide equal access to all interested users. What is needed is some sort of bibliographic control and national standards for social science files-- not unlike that which is available for printed materials.”


Commentary: “To coordinate the development and implementation of standards for controlling MRDF, a Conference on Cataloging and Information Services was organized by DUALabs… and supported by a grant from the National Science Foundation. The conference brought together at least 55 key persons having an active interest in establishing a framework within which a national program of cataloging and information services could be developed. The specific objectives of the conference were to identify key technical issues requiring resolution prior to implementing a coordinated cataloging effort, define the operational components of a centralized clearinghouse for MRDF cataloging… and… initiate a national program to catalog machine-readable data. Sue Dodd’s
presentation was focused on "the act of cataloging MRDF and the related information products."


Commentary: Reports on the National Conference on Cataloging and Information Services for Machine-Readable Data Files, held on March 29-31, 1978. The conference was an attempt at "establishing a framework within which a national program of cataloging and information services could be developed." The organizers might be forgiven for feeling these initial steps have yet to bear fruit as far as a national infrastructure. Even so, the conference resulted in a call to action for work to be carried out with major data producers in academic institutions and government agencies to use rules and guidelines for building a bibliographic record of data files created. Dodd's preliminary report was followed by a final Report on the Conference on Cataloging and Information Services for Machine-Readable Data Files, March 28-31, 1978 at Airlie House, Warrenton, Virginia. Produced by the MRDF Conference Secretariat, DUALabs. Arlington, Virginia. 1978.


Commentary: "In the last two decades, private research organizations, government agencies, and foundations have invested heavily in the collection of social science numeric data, contributing to the proliferation of machine-readable data. However, the development of information technology and the ability to produce data have progressed much more rapidly than our capacity to organize, classify, and reference its availability. There is an immediate need for some type of bibliographic control over MRDF, including guidelines on how to create a proper bibliographic reference. The purpose of this article is twofold: (1) to outline some of the information components associated with social science numeric data files, and (2) to provide guidelines, examples, and a uniform vocabulary for the creation of a bibliographic reference." Includes numerous illustrations and examples. Dodd hoped that these guidelines "will soon appear in the "authors' guide" section of social science journals and will eventually be included in such works as the Chicago A Manual of Style and Kate L. Turabian's A Manual for Writers. The ultimate goal would be to pave the way for social science data files to be included in printed bibliographies, end-of-work references, and indexing and abstracting works such as the Social Science Citation Index."


Commentary: "Explains how a multipurpose bibliographic/MARC data base of machine-readable data files (MRDF), created according to the MARC II record format, was conceived, and how much an information resource would benefit the general user and professional librarian." Dodd describes the experiences of the Social Science Data Library at the Institute for Research in Social Science at the University of North Carolina, Chapel Hill to build an online catalog of data files. The article provides an example of early efforts to use the facilities and expertise of academic main frame computing centers, MARC format for data files, and simple database design.


Commentary: "Models for the design of a database containing references to MRDF can be found in existing bibliographic information systems currently used for other types of research material - books, journal articles, technical reports, government documents, maps and audio-visual materials. The paper summarizes some recent developments in the area of bibliographic control of MRDF, outlines a model integrated system of data elements and indicates some products and services that could be derived from such a system."


Commentary (from the ALA press release for this book): "One of the effects of the information explosion is the proliferation of machine-readable data files (MRDF). In order to assure better bibliographic control over this medium, the Council on Library Resources awarded a grant to Sue A. Dodd to support her groundbreaking work on a manual for cataloging MRDF. The result of her efforts, CATALOGING MACHINE-READABLE DATA FILES will be published. . . by the American Library Association.

The first section of Dodd's manual is designed to demystify data files and computer programs and to make MRDF more comprehensible to those who must catalog and store them. Section two explicates rules for cataloging MRDF from the ninth chapter of AACR2, offering interpretations and examples and revealing the "how and why" of cataloging. Included is an extremely valuable outline of the steps for cataloging the microcomputer programs that are now found in many public and school libraries. The third section offers guidance on bringing bibliographic control to computerized files, including a bibliographic citation and a data abstract.

The only available work of its kind, CATALOGING MACHINE-READABLE DATA FILES provides the guidance which data producers, data archivists, and data librarians need to supply consistent bibliographic information for the MRDF they service. Newcomers to this rapidly developing field will appreciate the accessible presentation and the glossary of terms included in the manual."


Commentary: Comments on "significant steps that have contributed to current level of bibliographic control of social science machine-readable data files (MRDF). . . and . . . outlines some of the remaining problems to be considered before MRDF can be integrated into existing bibliographic utilities." Twenty references and MARC format, catalog entry, and data abstract for an MRDF are appended. The article summarizes and describes the process, groups and people who promoted the management and bibliographic control of data files, beginning in the late 1950's. Dodd includes details on the
role IASSIST played, including testing of a cataloging manual by several archives and reporting on the experience. There is a section describing how Dodd funded and wrote her first book *Cataloging Machine-Readable Data Files*. Development of a “multi-purpose automated cataloging system” at ICPSR is outlined as is work carried out to further test cataloging rules for data files. Writing in 1982, Dodd states “There is no doubt that machine-readable data will play an even greater role in research . . . more and more data will be needed for government and private research.” “[A] new dimension to the information explosion is now apparent; and with it an increasing demand for access to more and better documented data files.” “Communicating the availability of usable data is an inseparable part of research and an integral part of librarianship. In the near future, libraries will have no choice but to become more involved.”


Commentary: Dodd provides a detailed history and overview on public opinion polling, polling agencies, methods of data collection and problems in analyzing and interpreting poll data.


Commentary: Examines rules affected by revisions approved by American Library Association’s Joint Steering Committee for Revision of Anglo-American Cataloging Rules (AACR2) and those covered in the new “Guidelines for Using AACR2 Chapter 9 for Cataloging Microcomputer Software.” Changes still needed to provide adequate bibliographic control are suggested. Thirteen references are included.


Commentary: See discussion by Paden, 1986 (below)


Commentary: “A recent discussion among the participants of the E-Mail ‘Informal List for Official Representatives of ICPSR’ centered around citing computer files in references, footnotes, and bibliographies; whether to cite a codebook or file (providing you have both), and a discussion on citing primary or secondary sources. With respect to the last two concerns, there appeared to be adequate response indicating that it is better to cite the file as opposed to the codebook, and that one generally cites primary data sources. However, the first concern required more information and the ICPSR OR meeting was targeted as the next opportunity for such a discussion. Note: this paper was first presented at the ICPSR OR Annual Meeting in November 1989, but has been revised for the May-June 1990 IASSIST meeting in Poughkeepsie, N.Y. . . . The March 1981 issue of Social Forces was the first time that a major social science journal had provided instructions (in the ‘authors’ guide’ section) on how to cite a machine-readable data file (MRDF) — currently referred to as a ‘computer file’ . . . It is not possible in this discussion paper to provide anything but brief examples, but more detailed instructions on the components of a bibliographic citation are provided in the *JASSIS* article (Dodd, 1979) and in part three, chapter 9 of *Cataloging Machine-Readable Data Files* (Dodd, 1982).”


Commentary: Although this chapter was published in 1984, much of what Gray and Dodd say is still appropriate to ongoing debate about how libraries can take on the work of maintaining and providing access to data. The authors recognized that this would involve additional financial investment by libraries and they promote “cooperative arrangements which reduce the cost.” “Keeping in mind the goal of providing access to the data itself, the library should also solicit information from those facilities which might also provide services for accessing the data.” The authors suggest one option might be to share computing facilities and to provide statistical consulting. Discussion on acquisitions and collections, cataloging, and public service are well defined and could inform libraries still. The authors advise that libraries “should be prepared to evaluate new roles in providing access.”


Commentary: For many years, this manual was the de facto bible in the U.S. one could use to properly and fully document data files. Each of the guidelines is illustrated with a rationale and examples. The authors state that “bibliographic identity is provided by six kinds of information, . . . [that] which identifies . . . describes the content . . . classifies . . . in a set of descriptors or keywords . . . [and provides] information required to access, analyze, . . . [and] archive the MRDF.” There is a comprehensive section on the kind of information required to describe each variable in a dataset: wording of questions, variable names, variable labels, explanatory text, code values, category labels, frequency count, and universe definition. These elements were to form the basis of a data dictionary and/or users’ guide we think of as today’s codebook. A chapter includes a detailed checklist. This document played a significant role in standardizing the kind of information required to ensure usability of datasets and impacted data management practices of researchers and archivists alike.

**Works citing or making use of Sue Dodd’s research.**


Commentary: This paper describes the significant characteristics of archival materials and of archival methods of description and arrangement. Key sections of Archives, Personal Papers, and Manuscripts are explicated with particular reference to the ways in which the archival approach to descriptive cataloging reflects the nature of contemporary archival records and practice.
while remaining compatible with the style and structure of bibliographically oriented cataloging. The relationship of catalog records to other forms of archival finding aids is explained.


Commentary: An increasing proportion of government information is being disseminated in machine-readable and electronic form. An important subset of this information—information distributed on diskette or optical disk or available online—is accessible using microcomputer technology. This article examines the role that libraries can play in helping their users to locate and use this microcomputer-accessible government information, and the potential of such a role in helping libraries to fulfill their responsibility to provide broad access to government information.


Commentary: “Bibliographic records for microcomputer software in the OCLC Online Union Catalog are evaluated primarily for the purpose of focusing catalogers’ attention on selected areas in need of more consistent treatment. The degree of cataloging inconsistency evident in these records is examined with respect to the application of rules and prescriptions embodied in AACR2 Chapter 9, (and) the ALA Guidelines for cataloging microcomputer software… A secondary purpose of this quantitative/qualitative study is to provide a general assessment of the overall composition of microcomputer software cataloging.”


Commentary: “Standard catalog entries … constitute primary records by which computer-readable data files should be controlled and accessed. It is appropriate that academic and research institutions would want to record and provide access to files of data in machine-readable form in the public catalog of their libraries where entries already appear for other media… It is high time that the feasibility and desirability of incorporating records of machine-readable data files… in the public catalog… should be explored.”


Commentary: “Fortified with AACR1, AACR2, the Final Report of the Catalog Code Revision Committee, the Working Manual for Cataloging Machine-Readable Data Files and documentation from the Data Library at UCLA’s Institute for Social Science Research (ISSR), a test run of cataloging MRDF was undertaken. This article chronicles the difficulties encountered during that practical application of the developing cataloguing principles.”


Commentary: “In 1976, the International Association for Social Science Information Service and Technology (IASSIST) Classification Action group participated in a project to test the feasibility of cataloging and classifying MRDF. After applying the rules for descriptive cataloging presented in Sue Dodd’s Working Manual, based on the recommendations drawn up by the ALA Catalog Code Revision Committee’s Sub-Committee on Rules for Cataloging MRDF… The project resulted in three recommendations. “The first is that [Library of Congress Subject Headings] LCSH be used for catalog… subject description of MRDF… The second recommendation… is for people… to follow Dodd’s guidelines and to provide as many descriptive terms as are applicable to the study… The third… is for a group representing substantive academic disciplines, government agencies, and catalogers to draw up useful terms at all levels of the hierarchy… to evaluate terms… to submit suggested changes to LCSH… to combine… terms into interdisciplinary thesauri… [and] to coordinate the consistent use of standard terms…”


Commentary: This article “argues for the inclusion of a format for machine-readable data files in the existing bibliographic networks” given that a MArchine-Readable Cataloging (MARC) format has been developed. The author notes that “Large academic, research, and special libraries are requesting the capability of … [using] the US/MARC formats in combination with AACR2 (to) define the content of the data elements in a MARC record. Some of the elements … are a data file description which shows existence and source of data; a detailed abstract which includes the genesis and history of the file so as to link modified files; a keyword structure, physical characteristics of tape, file and software needed; applicability of the data to solving specific problems or analytic needs; and the link between data files and the software created to manage or operate them. These links reveal the presence of accompanying documentation, the bibliographic citation of accompanying documentation, software compatibility, and linkage with other files or programs.” Further, Nasatir stresses “[T]he more that users from all kinds of institutions contribute to and access a MRDF database, the more its economic viability will be assured… [and that] “communicating the accessibility of usable data [is] an integral part of librarianship.”


Commentary: “An examination of AACR2 Chapter 9 and the CCDA Guidelines for using AACR2 for Cataloging Microcomputer Software (Chicago: ALA, 1984) for catalogers not familiar with microcomputers. Includes seven full descriptive cataloging examples of microcomputer software using these recently developed guidelines.” Sue Dodd and Ann M. Sandberg-Fox’s Cataloging Microcomputer Files: A Manual of Interpretation for AACR2 is widely quoted. Paden states the guidelines for using AACR2’s Chapter 9 for MRDF’s were “written in the mid ‘70s before microprocessors and microcomputers”
were fully developed and available on the scale they are today.” The article is a snapshot of the time and how changes in technology impacted libraries. These machines were so new, and they required a variety of software to be installed by the user in order to operate, unlike the early 21st C. when most machines come pre-equipped. Libraries maintained copies of each version of programs that would then be installed as needed. There was confusion on exactly what to catalog, since one program usually contained several files to install and operate. “Only a cataloger with considerable experience in computer software could be expected to determine the nature of each file in a program.” The section on how to describe physical characteristics is quite charming, with its directions on how to measure various storage media as well as their containers.


Commentary: In a volume containing articles on the National Archives in the U.S., National Technical Information Service (NTIS), National Archives of Canada, Roper Center, and an article by Margaret Hedstrom on state archives, Wenzel describes a project, carried out at the Data and Program Library Service at the University of Wisconsin - Madison to enhance access to their collection of machine-readable data files. Project goals were based on the problems in access described by Dodd and others. Three levels of access are discussed. The archive level, study level and variable level each serve to identify the organization in which data are housed, provide broad descriptions of individual studies, and to document each variable within an individual dataset. This article documents early attempts by domain-specific data archives to develop and make visible online catalogs of their holdings.


Commentary: “In the past many libraries have been reluctant to acquire MRDF, as they presented a number of obstacles. To some they seemed prohibitively expensive; to others, they fell outside the library’s purview since they did not appear in bibliographies, abstracting and indexing services, or even databases. In addition, they require a computer and a degree of technical expertise to use. …Machine-readable data files present a myriad of collection development challenges.” The author suggests readings by Sue Dodd and Judith Rowe, and participation in IASSIST as key to understanding the research and kind of data used by quantitative social scientists. In describing how libraries would manage data acquisitions, Wittenborg addresses the use of bibliographic links between data and documentation with attention toward version control, and level of processing (curation). The article concludes with an admonition: “New information and new tools appear with astonishing frequency and the knowledge and tactics one has taken pains to acquire become outdated with alarming speed. …The selector must simply make a great effort to keep in touch…”

References


NOTES
1. Libbie Stephenson is Director of Social Sciences Data Archives at University of California, Los Angeles. libbie@ucla.edu Please contact Ms. Stephenson with any additions or corrections to this list of works

Sue A. Dodd’s Lasting Influence: Libraries, Standards, and Professional Contributions
by Ann S. Gray

Abstract
This article uses the author’s recollections and some of Sue A. Dodd’s own publications to provide a brief overview of various activities leading up to the publication of her *Cataloging Machine-Readable Data Files: An Interpretive Manual* in 1982. Of particular importance was the Airlie House Conference on Cataloging and Information Services for Machine-Readable Data Files of 1978. The article also highlights events that accompanied the development of cataloging rules for data files in the 1970s. It concludes with the author’s memories of Dodd’s commitment to professional involvement, the development of standards, and the role of libraries.

**Keywords:** Libraries, data libraries, library cataloging standards, data classification.

In 1977 Sue A. Dodd obtained a Master of Science Degree in Library Science from the University of North Carolina, Chapel Hill. It is unlikely that Dodd, who had undergraduate and graduate degrees from the University of Kentucky, needed the MLIS for her employment. She had been a Data Librarian at the Institute for Research in Social Science at the University of North Carolina for ten years and was also the Associate Director of the Louis Harris Data Center, which was also housed in the Institute. She was probably supported and encouraged to study Library Science by Richard Rockwell who was the Director of the Institute’s Data Library from 1969 until 1976. She had also been engaged in several efforts to bring bibliographic standards to social science data files.

Dodd had been a panelist in the session on “Problems on Inventorying Data: Classification Schemes” at the 1974 Toronto Conference on Data Archives and Program Library Services. She would become the US Chair for the IASSIST Action Group on Classification of Data. In any case, her study of Library Science demonstrated a desire to obtain professional insight in the classification of data files.

In March of 1978 Dodd was the Co-Chairperson/Technical Issues Session Leader for the Conference on Cataloging and Information Services for Machine-Readable Data Files held at Airlie House in Warrenton, Virginia, supported by a grant from the [U.S.] National Science Foundation. Her contributions to the conference included a report of the working group titled “Characteristics of Machine-Readable Data Files” and a reprint of her *Drexel Library Quarterly* (January 1977 No. 1) article “Cataloging Machine-Readable Data Files – A First Step.” That article was probably used by the Technical Issues working group as a starting point for their discussions. Dodd

Their final report would lay the foundation for a chapter on what were called Machine-Readable Data Files.

In 1977 Sue A. Dodd obtained a Master of Science Degree in Library Science from the University of North Carolina, Chapel Hill. It is unlikely that Dodd, who had undergraduate and graduate degrees from the University of Kentucky, needed the MLIS for her employment. She had been a Data Librarian at the Institute for Research in Social Science at the University of North Carolina for ten years and was also the Associate Director of the Louis Harris Data Center, which was also housed in the Institute. She was probably supported and encouraged to study Library Science by Richard Rockwell who was the Director of the Institute’s Data Library from 1969 until 1976. She had also been engaged in several efforts to bring bibliographic standards to social science data files.

Dodd had been a panelist in the session on “Problems on Inventorying Data: Classification Schemes” at the 1974 Toronto Conference on Data Archives and Program Library Services. She would become the US Chair for the IASSIST Action Group on Classification of Data. In any case, her study of Library Science demonstrated a desire to obtain professional insight in the classification of data files.

In March of 1978 Dodd was the Co-Chairperson/Technical Issues Session Leader for the Conference on Cataloging and Information Services for Machine-Readable Data Files held at Airlie House in Warrenton, Virginia, supported by a grant from the [U.S.] National Science Foundation. Her contributions to the conference included a report of the working group titled “Characteristics of Machine-Readable Data Files” and a reprint of her *Drexel Library Quarterly* (January 1977 No. 1) article “Cataloging Machine-Readable Data Files – A First Step.” That article was probably used by the Technical Issues working group as a starting point for their discussions. Dodd

Their final report would lay the foundation for a chapter on what were called Machine-Readable Data Files.

In 1977 Sue A. Dodd obtained a Master of Science Degree in Library Science from the University of North Carolina, Chapel Hill. It is unlikely that Dodd, who had undergraduate and graduate degrees from the University of Kentucky, needed the MLIS for her employment. She had been a Data Librarian at the Institute for Research in Social Science at the University of North Carolina for ten years and was also the Associate Director of the Louis Harris Data Center, which was also housed in the Institute. She was probably supported and encouraged to study Library Science by Richard Rockwell who was the Director of the Institute’s Data Library from 1969 until 1976. She had also been engaged in several efforts to bring bibliographic standards to social science data files.

Dodd had been a panelist in the session on “Problems on Inventorying Data: Classification Schemes” at the 1974 Toronto Conference on Data Archives and Program Library Services. She would become the US Chair for the IASSIST Action Group on Classification of Data. In any case, her study of Library Science demonstrated a desire to obtain professional insight in the classification of data files.

In March of 1978 Dodd was the Co-Chairperson/Technical Issues Session Leader for the Conference on Cataloging and Information Services for Machine-Readable Data Files held at Airlie House in Warrenton, Virginia, supported by a grant from the [U.S.] National Science Foundation. Her contributions to the conference included a report of the working group titled “Characteristics of Machine-Readable Data Files” and a reprint of her *Drexel Library Quarterly* (January 1977 No. 1) article “Cataloging Machine-Readable Data Files – A First Step.” That article was probably used by the Technical Issues working group as a starting point for their discussions. Dodd

Their final report would lay the foundation for a chapter on what were called Machine-Readable Data Files.
often said that her award winning book *Cataloging Machine-Readable Data Files: An Interpretive Manual*, published in 1982, was a result of the Airlie House conference, but her recognition that library procedures for bibliographic control could be applied to data files preceded that event.

The *Drexel Library Quarterly* (DLQ) article mentioned above provides a fairly comprehensive history of libraries' involvement in cataloging and data files, beginning with an American Library Association/Resources & Technical Services Division/Cataloging and Classification Section (ALA/RTSD/CSS) ad hoc subcommittee established in 1970. This subcommittee labored for six years and worked with various groups engaged in cataloging efforts for data files. Their final report would lay the foundation for a chapter on what were called Machine-Readable Data Files or MRDF in the second edition of *Anglo-American Cataloging Rules (AACR II)*, published in 1978. (Although the Rules were published in 1978, they would not be implemented until 1981 for reasons that had nothing to do with the inclusion of MRDF.)

The same article also details the work of the IASSIST Action Group on Classification and its goals. In 1976 Dodd had prepared a “Working Manual for Cataloging Machine Readable Data Files” based on her interpretation of the ALA subcommittee’s recommendations. The IASSIST Action Group’s first project involved establishing the feasibility of cataloging data files by having several institutions actually catalog some data files, ideally ones unique to their collection. Dodd gave each a copy of her Working Manual. This project was probably undertaken in 1976, as Dodd cites an Action Group memorandum of that date in the DLQ article.

In the 1970s there were a number of important changes in standards for bibliographic description of print materials. In 1967 the *Anglo-American Cataloging Rules (AACR1)* were implemented at the Library of Congress. AACR1 was not that different from the Library of Congress’s Rules for Descriptive Cataloging, published in 1949, and allowed libraries to continue to use forms for older entries even though they conflicted with AACR1. This was called superimposition. Only newly established entries would use the new forms set by AACR1. In 1974 a revised International Standard Bibliographic Description (ISBD) was published for monographs. Standards and revisions for other types of materials would follow. The ISBD provides standards for the form and content of bibliographic descriptions. In 1978 the second edition of the Anglo-American Cataloging Rules (AACR 2) was published and it incorporated the revised ISBD for both monographs and serials and included other types of materials, including machine-readable data files. But the change that caused many libraries to refuse to implement AACR 2 was that superimposition was no longer allowed. For example, “U.S.” became “United States.” “United States. Department of Commerce. Bureau of the Census” became “United States. Bureau of the Census.” No matter how many cards for older Census Bureau publications were in the catalog, all new publications would have to use the new format. Furthermore, a serial—which might often change its title or publisher—would now be cataloged using the current title, omitting designations such as Bulletin or Magazine. In the age of card catalogs these changes would result in very different locations of the same materials published in different years or the library would have to re-catalog all of the older material to the new standard. Librarians referred to this as “Desuperimposition.” Considering the problems this would cause for libraries, cataloging of MRDF was a minor concern and probably involved people who had not cataloged using the older forms, persons new to the field.

Of equal importance was the development of the Machine Readable Cataloging (MARC) programs at the Library of Congress. In 1966 the Library of Congress launched its pilot project for the MARC system. Using feedback from this test, *The MARC II format: a communications format for bibliographic data* was published in 1968. It would be adopted as a standard by several American Library Association divisions as well as other agencies. In 1971 it was given the American National Standard designation ANSI Z39.2-1971. It is not necessary to go into the details of these standards, but it is important that they exist. Standards are altered over time, but they provide a framework and a community that uses and applies them to make changes. ANSI Z39 and other communication standards provide a means to encode, organize, and retrieve information in a common or shared environment.

In 1967 a group of libraries in Ohio formed an organization to share cataloging and catalog information using a computerized network. By the early 1970s this system, using the MARC standards, was operational. It provided a shared database of catalog information and the production of catalog cards. The Ohio College Library Center would expand its services to libraries outside of this network and change its name to OCLC. Having catalog records in a machine-readable format lead to the end of the card catalog.

About the time the Dodd Manual was published, microcomputers, as they were called, were becoming common. The focus of the Manual had always been social science data files. But school librarians were acquiring computer programs and other types of files for use on microcomputers, and thus there was a need to adapt AACR 2 Chapter 9 for cataloging those types of materials. Dodd worked with Ann M. Sandberg-Fox on a follow-up manual for cataloging microcomputer files.

In 1979 I began my own career as a student in the University of North Carolina School of Library and Information Technology. Before that time I had worked for a number of years as a cataloger in a small academic library using the OCLC system. There was a version of the MARC system for use at the Library School and I learned how to enter, change, and retrieve text as well as how to set up database structures and options, and all the system controls that went with a batch process using IBM 360 machines running MVS. The MARC programs were very flexible; it was fairly easy to define fields, subfields, indicators, and output formats. But like many computer systems of the day it was merciless if there was an extra space or extra slash mark or missing comma. It was command driven and one had to know the commands. Because I had somewhat mastered the system, I was offered a temporary position at the Institute for Research in the Social Sciences Data Library where the MARC programs had been used for many projects. The job was to catalog the 1970 U.S. Census Files based on a machine-readable version of the Census Bureau’s *Directory of Data Files* (Abramowitz and Aldrich, 1979). The machine-readable version of the Directory contained print commands which could be used to identify separate sections as containing entry information, such as which part is the title, which part is the edition statement, the collation, notes, etc. I would use the MARC system, design the fields, and add information if needed. There I met Sue Dodd who was finishing up her Manual.
Sue and I, sometimes with others, would get together every day to review her interpretation of AACR2 Chapter 9. She was very keen to know how a working cataloger would use the manual and what questions such a person might have. Although her preliminary Working Manual had been used by the IASSIST Action Group project, she had not had the opportunity to actually talk to the catalogers since the project had been handled through the mail and only a few problems had been found. I had a lot of experience cataloging monographs and some training in cataloging serials. I don't recall finding any problems with her interpretation and when she asked if something should be done differently, I always concurred with her decisions.

Often Sue would explain to me why bibliographic control was needed, why libraries were important to social science data files, and who made what contributions to this effort. Later I would have a staff position at the Data Library and work with Sue for three years where my education would continue. The primary purpose of the IASSIST Action Group on Classification had been to facilitate access and promote the use of social science data. Bibliographic control would provide authentic identification of specific studies.

Within social science publications, data sources were not cited as published works, but were often mentioned in the text using general and ambiguous names, such as 'the Michigan study' or 'Census data.' Writers could not be expected to cite their data if no publication information was available. Cataloging data files was not the only tool Sue promoted. She often spoke of cataloging in production efforts like those of Patrick Bova in the documentation for the General Social Survey. Another project she was enthusiastic about was the generation of catalogs of data holdings. She had copies of all she could locate.

Sue was almost unique in her support of libraries as important resources for data. She believed that libraries wanted to be involved in data services and that they should be involved. In 1981 there was little evidence to support her opinion but she would be proved correct. Sue also instructed me in two other principles she held dear – the importance of standards and the value of professional involvement.

Sue was a generous mentor and an important contributor to the professional development of data services within IASSIST and library organizations such as the American Library Association and the Research Library Group. These professional communities allowed her to be part of the process by which standards were advanced. She understood that standards would only be applied if their complexity was justified by their utility.

In library catalogs the term Machine-Readable Data File for type of material was replaced by computer file which was replaced by electronic resource. But Sue Dodd's legacy is not lesser because of changing terminology. Besides, I don't think the transition from MRDF to electronic resource is an improvement. I think Sue would agree with me.

**References**


---

**NOTES**

1. Ann S. Gray worked in data services at the University of North Carolina, Chapel Hill, Institute for Research in Social Science; Cornell University, Cornell Institute for Social and Economic Research; and Princeton University, Firestone Library. She has been a member of IASSIST since 1984. She is retired.
A Legacy of Inspiration and an Enduring Smile

by Peter Burnhill

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 the action 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


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’. 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 after the section on 'Documentation' and then again 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. 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.

Although not formally published that article is now, belatedly, in (Burnhill, 1985). In what now looks like a 'use case workflow', 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 Robbins, 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.

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: Alison was developing the online information service ‘datalib’ enabling users to navigate a form of hypertext in ‘view’ (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 that received had no header file with a descriptive title. We were asked to suggest the Data Library as the basis of that bid. That event might signal the date when my energies finally shifted away from the sharp focus on the social science data file. The seminar at City University was interesting, attracting a wide variety from the library world as well as the Data Archive at Essex. 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 seminar at City University was interesting, attracting a wide variety from the library world as well as the Data Archive at Essex. 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 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 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.  

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’.  

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’ 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 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).
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 catalogue29 as well as a set of services geared at benefitting researchers, students and their teachers at the University of Edinburgh.30 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 MANTRA31, an online course designed for researchers and 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](image-url)

**Figure 1 - Keepers Registry**

<table>
<thead>
<tr>
<th>SERVICES: user requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-J Preservation Registry Service</td>
</tr>
<tr>
<td>Data dependency</td>
</tr>
<tr>
<td>E-Journal Preservation Registry</td>
</tr>
<tr>
<td>METADATA on e-journals</td>
</tr>
<tr>
<td>METADATA on extant e-journals</td>
</tr>
<tr>
<td>ISSN Register</td>
</tr>
<tr>
<td>Digital Preservation Agencies e.g. CLOCKSS, Portico, BL, KB, UK LOCKSS Alliance etc.</td>
</tr>
</tbody>
</table>

...
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 (SUNCT), 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,1 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. 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 PEPJS 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. 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.2

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 when triggered as orphaned.
3. Publish the terms of access for the archived content (now and (and how)
4. Tell someone what you are doing and what you hold

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, p353). 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) 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, 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. 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 URL 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


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>


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.


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/>


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/insi-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,
Keywords from Vol 37 No. 1-4. Courtesy Tagxedo.com
Abstract
The work and life of Sue A. Dodd had influence in its own right and her work was adapted and incorporated by others just as her work was influenced by others and part of a general evolution of social science metadata. From her focus on the catalogue description of machine-readable data that made users able to reference and identify data files as a research source, the description of social science data files gained further momentum. This paper centers on the fundamentals of social science data and their relation to metadata. There are levels of metadata in typical social science where the study, the variables of the study and the codes of the variables define a hierarchy. For each level there are many potential descriptive items that can be part of the full metadata. The work was initiated in the US but there was also work carried out in Europe through the described period, mostly centered around 1975–1995. All of this can be considered the foundation of social science data metadata description that later evolved to become the work carried out within the Data Documentation Initiative (DDI).

Keywords: DDI, Data Documentation Initiative, metadata, social science, study description, codebook.

Introduction
Sue A. Dodd identified the vacuum of library catalogue description for the new and growing area of machine-readable data (Dodd, 1979), and provided guidance for using a standard bibliographic format to fill this vacuum (Dodd, 1982). In following years she continued to communicate, discuss and elaborate upon the guidance. Some of the other papers in this collection in the IASSIST Quarterly (IQ) will bring more focus to the writings of Sue Dodd and some papers will elaborate on the work carried out within the Data Documentation Initiative (DDI). This paper sees the influence of Sue Dodd as her work was adapted and incorporated, highlighting some of the European work on the study description for social science data during the period 1975 to 1995.

For a tour of the developments in data archiving combined with the technical and political – as well as personal - developments I’ll recommend ‘The Decades of My Life’ by Judith Rowe (1999).

The digital age: Useful data are useless without documentation
The digital age signifies that everything is stored as numbers. Obviously when only a number is communicated – ‘42’ is my favorite example - it cannot alone carry any meaning for the recipient. A number has to be wrapped in explanation to convey any meaning. We know we have been fooled - and we find the nonsense amusing - when the computer Deep Thought after seventy-five-thousand human generations of calculating produces the answer Forty-Two to ‘the great question of Life, the Universe and Everything’ (Adams, 1986, p. 128).
Data, information, knowledge

The machine-readable data file consists of nearly endless series of numbers. Let us visualize the data file as a database table (survey file of a questionnaire) consisting of many attributes (variables as columns) concerning entities (individuals as rows).

Many information scientists have attacked the problem of having more precise concepts for categories or levels of information. In everyday life we might use the term ‘data’ intermingled with the term ‘information’ without much bother. I must admit that I personally don’t have any problem with the less rigid use of the concepts. However, there are insights to be gained when entering a more meticulous definition of the terms. When working within systems analysis in Britain the researcher Peter Checkland proposed some useful distinctions for his methodology of soft systems development. Data are viewed as an unordered, formless, disorganized pool of facts. (Checkland actually used the term ‘cloud’ that now carries the meaning of organized access to safe storage always available.) ‘Information’ is facts situated in context resulting in ‘meaningful facts’. In relation to the data file the documentation is delivering the context. ‘Knowledge’ is larger, longer living structures of meaningful facts. In our analogy this relates to the use of the data file mostly exemplified as the relationship between variables. In getting to knowledge the first issue is to select the individuals and the variables for our data collection. Thus we select the data relevant to us. Checkland proposed the term ‘capta’ for the selected data (Checkland and Holwell, 1998, p. 90) to be distinguished from the disorganized pool of data. The context in the form of the description of the selection process – concerning both the selection of entities and the selection of attributes - is a necessity for creation of ‘information’ from ‘capta’.

These descriptive entries are the metadata or data description. I did warn you that I might not be rigid and consistent in my use of these concepts. The term ‘capta’ is in my view a way to understand what we normally call ‘data’. When we have data we require data documentation in order to produce information and knowledge.

Multiple benefits of data documentation

During the 1960s and 1970s the research data file became a regular resource available to other users for secondary analysis and the benefits of data sharing and data documentation was addressed consistently in the 1980s (Rasmussen and Grant, 2007, p. 60). A conference in 1979 resulted later in a comprehensive report on ‘Sharing Research Data’ supported by the National Research Council (USA) with extensive discussions and papers and a leading chapter of ‘Issues and recommendations’. The number one recommendation is ‘Sharing data should be a regular practice’ (Fienberg et al., 1985, p. 25). In the journal American Sociological Review the benefits of data sharing published in the report was summarized as:

‘reinforcement of open scientific inquiry; the verification, refutation, or refinement of original results; the promotion of new research through existing data; encouragement of more appropriate use of empirical data in policy formulation and evaluation; improvement of measurement and data collection methods; development of theoretical knowledge and knowledge of analytical techniques; encouragement of multiple perspectives; provision of resources for training in research; and protection against faulty data’ (Hauser, 1987).

The benefits of sharing of machine-readable data are parallel to the benefit of having libraries sharing human-readable material. Making the sharing possible implies the benefits of metadata describing the data file. The value of sharing data can be accredited to several dimensions of arguments:

Actors

Sharing data primarily implies the use of the data as secondary data when data are reused for other than the intended purposes by other researchers or by students for educational purposes. Data archives have also often experienced the original investigator(s) among the requestors for their own dataset because the archives had not only preserved but also value-added to the dataset by elaborated and accessible metadata.

Resources

Collecting data is a money intensive action. Research data collection is often financed by public money. Sharing the data is the most cost-efficient way to carry out science. Furthermore, some retrospect research initiatives are only possible to realize through secondary analysis often of an extensive kind where several earlier data sources are being used in combination to produce a more accurate account.

 Naturally there are also costs involved in sharing data and producing useful metadata. It is possible to argue that not all collected data are worth the extra cost. However, when research projects are financed in competition (e.g. from types of science funds) it would be counter-intuitive if the board would not regard the future collected data as sufficiently valuable. The National Science Foundation (USA), the Economic and Social Research Council (UK) and the Danish SSF (Social Science Foundation) – and probably many more like these funding agencies - all had clauses in the contracts for archiving and documenting research data for reuse when I investigated this in the late 1990s (Rasmussen, 2000, p. 169).

Controls

A popular issue of research data being publicly available is an issue of being able to control that science is not infected with fraud. However, it is also mentioned in the citation above that the public through data access can control administration and evaluation of policies. Furthermore the sharing of data also supports ‘multiple perspectives’. This is regarded as fundamentals of having a democratic and free science.

The short Hauser entry above reminds us that there is a distinction between survey data being ‘public’ and ‘usable’. This is a distinction depending upon metadata. Hauser recommended that journals (e.g. the ASA) would be keeping the tabulations related to the published papers in the journals (this was more than 25 years ago and the recommendation was practical and proposed the technical solution of storage on floppy disks). The recommendation also brought an attention for a recommended format. However, the discussion on what to archive and in which formats was in the 1980s already a continuing discussion and system decisions were implemented in data archives around the world and will be addressed later in this paper.

The general improvement and development of research can also be placed under the dimension of control. Metadata description of a research dataset will act as an evaluation of the study, e.g. a survey is described in terms of the population, the selection
process, the nonresponse, the response rate etc. These items will act as a checklist for new and not so experienced researchers as well as demand consistency and thus easier access to the data.

Archives and archivist collaboration

Some twenty years before the comprehensive reports on sharing of research data (Fienberg et al., 1985) and of social science data (Sieber, 1990) the sharing of data had already been implemented in institutionalized form by the establishment of the ICPR (Inter-university Consortium for Political Research later ICPSR, Inter-university Consortium for Political and Social Research at the University of Michigan). Before that the Roper Center had been established as an archive for Gallup and other commercial public opinion polls since shortly after the end of World War II. With the eye on comparison of national statistics there were three (UNESCO-) conferences in the early 1960s also addressing data archives (Rowe, 1999).

The concept of data archiving was also institutionalized in Europe by the creation of several national social science research data archives from the mid 1960s and onward. In Germany the Zentralarchiv für Empirische Sozialforschung (later included in GESIS (Leibniz Institute for the Social Sciences)), in UK the UK Data Archive, in the Netherlands the Steinmetz archive (later included in DANS (Data Archiving and Networked Services)), in Norway the NSD, and in Denmark the Danish Data Archives. Many data archives incorporating data files from research in many other European countries followed quickly after.

When the IASSIST1 was founded there were a sufficient number of individuals within the profession of data librarians, data archivists and other advocates, researchers and technicians who were sharing machine-readable data to form an organization that has lived and has had influence for a span of time that was in older days a lifetime. The acronym IASSIST was created in advance of the name: International Association for Social Science Information Services and Technology (Geda, 2006). Both the acronym and the name continue to capture very well the focus of the association. The sharing of knowledge was institutionalized internationally through the IASSIST conferences - and the included workshops - as well as through workshops and the meetings of official representatives (ORs) under the auspices of the ICPSR.

Soon after the realization of IASSIST, the European CESSDA (Consortium of European Social Science Data Archives) was formed as an umbrella organization for the European research data archives. Issues were trans-border sharing of data and privacy in the different countries. CESSDA was also successful in knowledge transfer between individuals through themed seminars many of which addressed the issue of metadata and discussions on the use of different standards.

Metadata and levels of documentation

The fundamental documentation of a data file as a whole is the identification of the data file as a research source. When we talk about the ‘American National Election Study, 2004: Panel Study’ others will be able to locate the study2 (and they will find a shorter form of identification as ‘ICPSR 4293’). Similar to citations from literary sources there should be a way to accurately identify the research data source. The documentation of the data source makes it possible for secondary users to give positive credit to the people behind the data source. Among the most fundamental aspects of scientific research is the possibility of inter-subjectivity that is the closest thing to the unattainable objectivity. Documentation has the potential to introduce discussions on the methods used and the research decisions. The researcher describes how the data were created and makes critical evaluation possible. Provided the documentation reveals valid methods, the replication of the procedures described in the documentation should ideally lead another investigator to the same result.

Data equals documentation - documentation equals data - metadata

For some of us the rectangular data file was - and is - the ‘normal form of social science data. The term ‘normal form’ brings us to relational databases and that is no coincidence. All kinds of complicated database structures are possible with the methodology of relational databases relating rectangular tables. Others may be interested in more exotic examples of data for social scientific analysis, such as artifacts like images and sound bites. Whatever type of data is to be analyzed using a systematic method, you have to define what your specific interest is and define and select your ‘capta’.

In other words, the systematic and comprehensive documentation of a collection of machine-readable files can become data for a researcher investigating the collection. For instance a research objective could be to compare surveys carried out at different periods of time or at different geographical locations. The description of data is called metadata as it is ‘data about data’. We should note that metadata are not only ‘about’. Metadata are also genuine data that can be analyzed.

The levels of data documentation

The primary user of a rectangular social science data file needs information on the variables (columns) as well as explanations of the codes as exemplified in the information:

<table>
<thead>
<tr>
<th>V6</th>
<th>Sex of respondent</th>
<th>1. ‘male’</th>
<th>2. ‘female’</th>
</tr>
</thead>
</table>

Figure 1. Codebook documentation of a variable with codes (a minimum example).

This codebook with information at the variable level and the code level is elementary yet helpful for the analyst who already is familiar with the overall background for the data file.

However, for the secondary analyst ‘(i)nformation about variables is useless unless the population, sample, and sampling procedures are described’ (Blank and Rasmussen, 2004, p. 307). For this reason, many more precise items were introduced for study level description early in the history of data archiving and a standard for the study level description was agreed upon (Nielsen, 1974 and 1975). It was discussed, refined and presented at several meetings, workshops, and conferences in the 1970s. However, although there was this continued presence around the ‘standard study description’ the standard was primarily the foundation for local implementations and never achieved the status of an international ‘de facto standard’. More important was the international agreement concerning the items described in the standard. Several of these items were also implemented in other archives – in other local formats and systems – and the items were later included in the development of the DDI-standard. They are listed below:
Zentralarchiv in Cologne (now GESIS) was in the forefront of the revolution happens when the machine-readable documentation of machine-actionable metadata.

There was a great effect when documentation of machine-readable data had the positive experience caused by ‘taking its own medicine.’ Naturally the decision on which items to include into the documentation was a very important and first decision. However, when that was settled and the human could read the documentation there was a revolution in having a well-formed documentation not as typed pages but in machine-readable form.

The revolution happens when the machine-readable documentation becomes machine-actionable. When metadata collections were formed several data archives were looking into retrieval systems for supporting secondary researchers in their quest for finding appropriate data. In Europe the German archive Zentralarchiv in Cologne (now GESIS) was in the forefront of building retrieval systems. Another earlier use of rigidly structured metadata was the translation of OSIRIS codebooks into the - more reduced - control language used by other statistical packages like SPSS and SAS. As system files were dependent upon the configuration platform of both machine and operating system the character-based solution of exchanging old-fashioned card images in the form of lines of text was the most general solution. Furthermore, at the codebook level the machine-actionable documentation meant that a machine - e.g. software on a computer - would be able to read the documentation and be able to automatically interpret and perform calculations and controls on the data file. In addition, the machine-actionable documentation could, with great benefit, be created prior to the data file. The documentation could by a machine be the foundation for data collection, e.g. generating the screens and software for data-entry for telephone interviewing (CATI, computer aided telephone interviewing).

Turning the complete documentation - including the study level – into machine-readable form meant that studies could be searched effectively and with the Internet the accessibility to data files increased tremendously both regarding the number of studies as well as regarding the speed the users could access the selected information. Having documentation in machine-readable form implies that there should be a defined format - a standard.

**Standards of data documentation**
The old joke about standards goes: ‘Standards must be good since there are so many to choose from.’ In this section some of the actually used standards for documentation of social science research data before the DDI are briefly described.

Machine-readable Cataloguing - in short MARC - was developed at and institutionalized by the Library of Congress as a computerized library card format to build a library catalogue that was machine-readable in contrast to the paper cards traditionally used in libraries. The latest MARC format (MARC 21) was finalized before the millennium. Some can consider the format as old-fashioned and related to outdated technology. However, the format is still very much in existence in libraries all over the world.

In regard to the legacy of Sue Dodd, the MARC format for MRDF (machine-readable data files) stands central because her 1982 manual provided guidance for applying the MARC/MRDF format to create standard catalog records for MRDF. Her guidance based the catalog record upon file or study-level documentation. Standards for documentation were discussed continuously throughout the period. As mentioned earlier, there were during the 1970s international workshops in Europe on the documentation at the study level. IASSIST accommodated for many years sessions on documentation with presentations, proposals, and discussions at the international yearly conferences. Further information on the study level documentation standards standards including MARC and ‘Dublin Core’ is found in the Green and Humphrey paper in this issue of the IQ.

**Machine-readable documentation in use**
At the IASSIST conference in 1993 an action group for ‘Codebook Documentation of Social Science Data’ was formed. In 1994 I carried out an investigation by mail questionnaire in order to obtain an overview of the amount and kind of documentation that existed at archives. The investigation also obtained preferences of documentation among the professionals. This was reported as a presentation at the IASSIST conference in 1995 as well as in the IQ (Rasmussen, 1995). The paper presented a snapshot of the situation now 20 years ago. The majority of the studies were then without any machine-readable documentation. The reported machine-readable datasets were from 19 answering archives. The distribution of datasets by type of machine-readable documentation is shown in Table 1. The different machine-readable formats or level of standards are explained below demonstrating the development of metadata for social science.

| 001- | General information documentation level, subject cluster, keywords |
| 101- | Identification and acknowledgements bibliographic reference, archive, primary investigator(s) and other references |
| 201- | Analysis conditions abstract, kind of data, data sources, type of unit, number of units, size of dataset, time dimensions, universe, selection, sampling, data collection instruments, weighting |
| 301- | Reuse of data Data representation, data cleaning and controls, access conditions |
| 401- | References to relevant publications primary publications, secondary publications, analysis results, references to other studies (data files) |
| 500- | Background variables personal, (age, gender, ethnic group ...), household information, employment, occupation, income, education, mass media, ... |

*Figure 2. Overview of a selection of items included in the Study Description as used at the Danish Data Archives (summarized from Rasmussen, 2000, p. 287-352, and Rasmussen, 1981).*
The availability of a dictionary implied that the structure of the data file was reflected in the proofed dictionary and data could be loaded into standard packages like SAS, SPSS or OSIRIS. This brought down the time involved in setting-up a system for the secondary analysis as well as increasing the accuracy. The data quality could be greatly deteriorated by the hand-to-hand passing on of instructions. With the availability of a machine-actionable dictionary the secondary analyst would spend less time on controls and more on analysis.

At this level - see below the higher level ‘dictionary-plus’ - only column information existed and often in a very restricted form. For instance variable labels would often only be able to carry 24 characters of documentation which people did not find sufficient. Furthermore, this format did not include any information about the codes and categories. The user would for instance need scanned images in order to find explanation for the codes actually encountered in a variable.

### Dictionary-plus

For the category of dict+ ‘the plus indicates that the documentation comprises category labels in OSIRIS or in the form of ‘value labels’ in SPSS or ‘user formats’ in SAS. Often the documentation was stored in a system proprietary format and often the packages were forcing a ‘lock-in’ on the users so you could not for example directly analyze an SPSS system file with the SAS package. The situation has improved but at that time even the change of version of system files within SPSS presented a problem with backward compatibility. Having documentation embedded in system files also presented the archives with a load of migration tasks as information could be lost if files were left in old system formats just as they would be lost if they were left on old media. As Rothenberg phrased it: ‘digital information lasts forever - or five years, whichever comes first!’ (1995).

### Dictionary plus codebook

Through the period in focus here (1975-1995) the SAS and SPSS packages were the popular tools for analysis although they had no support for documentation at the study level apart from a filename and a short title for the study. The amount of study level documentation was very similar to the restricted documentation of variables and categories: a name or value and a label of limited length (Grant, 1993; Rasmussen, 1989).

The OSIRIS documentation format was early regarded outdated by being tied to a card-image format of 80-characters - an inheritance from physical punched Hollerith cards. When using physical cards placed in sequence it was very important to be able to reconstruct the sequence (with a counter-sorter) in case a stack of cards got dropped on the floor or was mixed into another stack. The OSIRIS codebook layout was originally mostly a format for storing the electronic information for later printing a more nicely formed codebook. However, the OSIRIS format was remarkable by being able to store unlimited amounts of text because a description of a variable could occupy multiple lines; this was also similar for comments, explanations, and code description. In OSIRIS different types of documentation were identified through an alphabetical “tag” in the first field or column of each card. (See figure 3).

Remarkably, the OSIRIS format included features for description at the study level (S-cards). OSIRIS was developed over a period of time and I’m here referring to the OSIRIS III (ICPSR, 1973). Structure of the description at the study level was available as the format even included further distinctions on the study level for entries as title, original archive, etc. Furthermore some ‘meta-metadata’ were possible as the general structural principles of the codebook-layout could be described within the standard itself using a meta explanation type card (E-cards). These features and other parts of the OSIRIS format were extended at the Danish Data Archives and the German Zentralarchiv (Rasmussen, 2000, p. 351). These were extensions for more elaborate description, data checking, and retrieval. The harsh backside was that the OSIRIS system in itself in the analysis tasks – including tabulations – totally ignored all available information apart from the limited dictionary information supplying variables with a number and a 24-character label plus some format information as well as information on missing data.

### Table 1. Datasets with machine-readable documentation (Rasmussen, 1995).'

<table>
<thead>
<tr>
<th>Dataset Type</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scanning</td>
<td>505</td>
</tr>
<tr>
<td>Text</td>
<td>1,943</td>
</tr>
<tr>
<td>Dict</td>
<td>4,520</td>
</tr>
<tr>
<td>Dict+</td>
<td>3,656</td>
</tr>
<tr>
<td>Dict + Codebook</td>
<td>2,003</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>12,627</strong></td>
</tr>
</tbody>
</table>
Although the OSIRIS format was old-fashioned judged by the card-image layout, it was thus also very farsighted. The original OSIRIS format and the extensions - including some developed outside of ICPSR - had the capability to include a very high degree of the relevant documentation compared to the offerings by other statistical packages like SPSS and SAS. The documentation types brought attention to the levels and structure of documentation and to the comprehensive items that were later developed into the DDI. The letter in the first column of the lines of OSIRIS documentation was a very early markup of documentation through the 'tagging' by first-column letter. Furthermore, the rigid card-image and fixed-columns OSIRIS format was kept as the archival format, but it was generated from freely typed input with software generating the variable numbering and the required different indentions based on the tagging for the relevant card-type (‘Q’, ‘X’, ‘C’ etc.)

Towards a standard

Some months after the formation of the IASSIST action group on codebooks a CESSDA seminar was held in Gothenburg in August 1993 on ‘Variable Level Documentation’. The following year another CESSDA seminar on ‘Networking and Internet’ was held in Grenoble. Both of these workshops also had non-European participation, most significantly the participation by staff of the ICPSR.

The discussions not only collected the sum of identified elements that from the viewpoint of archives were considered important to include in a standard documentation package but also brought to attention the many functions that the documentation should support. They also introduced carefulness towards what could be termed independence. This independence was the guarantee that a standard could evolve and not be locked, as well as being available to all.

Functions of the documentation

This paper has mentioned how documentation first of all delivered the printed study description and the codebook in a well-formatted human-readable form. This is believed to continue to be a relevant use though the human might read the information from another device than paper. Another important function is that collections of documentation – especially in the form of well-structured computer files – are searchable. It has also been mentioned that documentation can deliver input for the validation of data previously collected and also control data being collected. Lastly, the ultimate use of documentation is in the analysis and the presentations from statistical software of the documented data.

But do users really need all the bells and whistles delivered by the DDI? When huge commercial companies like SPSS (now IBM SPSS) and SAS deliver only a minimum of documentation facilities for a dataset should that not be taken as a sign of what the user community is interested in – and as a sign of what quality level of documentation the community is willing to pay for? OSIRIS is still in existence and can be found as MircOsiris for MS Windows. However, this micro-version does not support the unlimited and elaborate OSIRIS codebook format. MircOsiris has only the minimum documentation facilities as described in Figure 1 above.

I believe that minimum documentation presents a one-eyed view with a narrow focus on machinery for analysis of your own data. The limited documentation will prove to be a problem for even the primary researcher if and when an older dataset is to be re-analyzed. Naturally the problem will be even greater for the secondary analyst. In commercial settings they know within data management that elaborate - and often very expensive - documentation and management systems are necessary for gaining profit of the data warehouse.

Independent documentation

When developing a new documentation format as set forth by the Data Documentation Initiative it was considered very important that the standard be independent of commercial interests. Public archives and university libraries would not be able to afford to tie themselves to a storage format that could imply a yearly user fee. Independence was also the term used in connection to which systems should be allowed to analyze data described in the DDI-format so there should be no licensing. The DDI-format should also be independent of operating system platforms. It might be possible to obtain financial support for the development if a company could obtain rights for a proprietary format and system. However, data archives regularly service many users who have distinct preferences for this or that system. Therefore the solution should be that the documentation format is open for use by all software developers or vendors. The archives already have great expertise in converting existing codebook/dictionary documentation to the reduced description used by SAS and SPSS. When developers at archives would use the DDI-format the software for conversion to users' preferred software would naturally follow.

<table>
<thead>
<tr>
<th>S</th>
<th>study level</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>meta explanation</td>
</tr>
<tr>
<td>T</td>
<td>dictionary, sub-structured including missing data and format</td>
</tr>
<tr>
<td>Q</td>
<td>variable description, question in questionnaires</td>
</tr>
<tr>
<td>K</td>
<td>continuation</td>
</tr>
<tr>
<td>X</td>
<td>explanation</td>
</tr>
<tr>
<td>C</td>
<td>code value and label, sub-structured</td>
</tr>
<tr>
<td>B</td>
<td>for grouping of more Cs (higher level categories)</td>
</tr>
<tr>
<td>F</td>
<td>frequencies, attached to the C</td>
</tr>
<tr>
<td>J</td>
<td>temporary comments</td>
</tr>
<tr>
<td>G</td>
<td>note number</td>
</tr>
<tr>
<td>M</td>
<td>note text</td>
</tr>
</tbody>
</table>

Figure 3. The original card-types of the OSIRIS-codebook (summarized from Rasmussen, 2000, p. 340-352; originally in ICPSR, 1973).
Along came the Internet with a markup language

The Internet was early on seen as being of main interest to archives as a media for users searching for data and thus identifying relevant potential datasets. Later the Internet also became a media for direct delivery of data. And later again the client could get thinner and direct analysis would be performed on the network servers. The Internet was as such very promising for much faster and easier identification and access to data sources as well as much cheaper distribution and easier analysis.

The use of the Internet was accelerated with the popularity of the World Wide Web. For a coming standard of data documentation the display of documents through the use of HTML (Hyper Text Markup Language) was very stimulating. Further stimulation came from the Text Encoding Initiative (TEI) that used SGML (Standard General Markup Language) for marking up documents. Consequently the proper move would be to make a document type definition (DTD) of data documentation using SGML. Early on some were experimenting with HTML for their documentation but realized that if HTML was to become the standard they would tie the documentation to the presentation and thus commit an offence against the general principle of independence. Quite similar to the reduction by software of documentation to SAS or SPSS format - or any other format - it would be easy to reduce a complete standard data documentation like the DDI to an HTML page or to several selected forms of preferred HTML presentations.

Another technological development from the Internet further paved the way for an effective solution for data documentation. The introduction of XML (Extensive Markup Language) implied easier and more flexible work than SGML and with a strong connection to the Internet. The slogan from Jon Bosak from Sun Microsystems - one of the founders of XML – was that ‘X’ML gives Java something to do (1997). Internet, Java, XML – all worked together.

Conclusion

Hopefully this paper has demonstrated that the origins of the DDI evolved from work related to social science data documentation issues by several institutions and people in the decades before the emergence of the Data Documentation Initiative. During the period described in this paper several principles and levels of documentation were identified and further refined. Re-using data is a community effort and the community encases the world. This was getting more and more attention with the fast expansion of the Internet, with the spread of the World Wide Web. Local inventions were put together and further developed in a distributed effort. That further story is another paper!

Actually the development of the DDI is discussed in several papers in this special issue of the IQ. Ann Green and Chuck Humphrey describe the early years and Mary Vardigan offers a paper on the later years of development as well as a very useful DDI-timeline.

References

Blank, Grant (1993) Codebooks in the 1990s; or, Aren’t you embarrassed to be running a multimedia-capable, graphical environment like Windows, and still be limited to 40-byte variable labels?. Social Science Computer Review 11(1): 63-83.
Green, Ann and Humphrey, Chuck (2013) Building the DDI. IASSIST Quarterly 37.
Nielsen, Per (1974) Report on Standardization of Study Description Schemes and Classification of Indicators. Copenhagen, Danish Data Archives.
Nielsen, Per (1975) Study Description Guide and Scheme. Copenhagen, Danish Data Archives.
NOTES
1. Karsten Boye Rasmussen is an associate professor of IT and organization and a data scientist at the University of Southern Denmark. He worked in data archiving from 1974-1998. Email: kbr@sam.sdu.dk.
2. <www.iassistdata.org>
3. <https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/4293>
Building the DDI
by Ann Green and Chuck Humphrey

Abstract
This paper describes the context, motivation, and requirements behind the design and development of the first version of the Data Documentation Initiative (DDI) metadata community specification, with an emphasis upon the process of creating the initial element set for the "study level" of DDI version 1. We also offer a framework for understanding the infrastructural changes that contributed to the establishment of the DDI. By taking a close look at the confluence of influences on the earliest efforts to design and build the DDI, we can better understand what essential elements of metadata are necessary to support independent use of social science data over time.

Keywords: Metadata, data description, documentation, social science data, DDI

Introduction
Initially, the DDI was built to provide a structured framework for metadata describing social science surveys, but has since grown to include a broad range of data types. From the beginning, it was clear that over time the DDI would need to provide a metadata structure for longitudinal data, comparative data across geographies, panel studies, aggregate data, and administrative records. By taking a close look at the confluence of influences on the earliest efforts to design and build the DDI, we can better understand what essential elements of metadata are necessary to support independent use of social science data over time.

By the 1980's, the types of descriptive information necessary for researchers to use data created by someone else were well established. Such documentation needs to describe the content, quality, and features of a dataset, which in turn provides an indication of its fitness for use. The social sciences benefit from a long history of large-scale databases that were accompanied by extensive documentation (some examples in the United States are the General Social Survey, the American National Election Study, The California Poll, the Health Interview Survey and in Canada there are the public use microdata files for the Census of Population beginning in 1971). These databases were designed for sharing and wide use, so the documentation needed to support the independent use of the data without making it necessary to return to the producer of the data to make sense of things such as the methodology, coding of variable, question text, interviewer characteristics, sampling, etc. (CCSDS, 2012). Data producers, primarily in government agencies and large research centers, were expected to publish printed documentation about the data they collected and disseminated to the world. At the time, almost all of this documentation was in printed form.

Data archives across the US, UK, Europe, and Canada, which had already begun collecting and taking stewardship of large collections of social science data files and accompanying documentation, were also building catalogues and access systems of technical capacity far ahead of their time. Data professionals began to build a community of expertise to support the many services related to curating and preserving voluminous collections of social science data. Libraries began collecting and cataloguing these materials to support their designated research communities as the reuse of large data collections became a standard practice in the social sciences. It was at this time that a shift from paper documentation to machine readable alternatives presented new capabilities for expanding the role of descriptive information and hyperlinking the intellectual components of that information. What was needed was a way of structuring this information so that it was both human readable and machine actionable.

Challenges to make data independently usable continue today and as such, the research community needs to be encouraged to continue its commitment to produce and distribute information about the data they collect, share, reuse, analyze, replicate, and publish. As important now as in the 1960's, metadata are used to locate and review data for fitness of use, to have transparent access to methods and sampling, to understand the capacity for linking data, and to create maps, visualizations, or mine large data collections. Metadata are integral to all data manipulation functions. Without structured, complete, and accessible metadata these challenges cannot be met.

In this review, we make connections between metadata activities today and developments in the past where the rich history of documenting social science data has been overlooked. We also offer a framework or model for understanding the
infrastructural changes that contributed to the establishment of the DDI. Finally, we explore the importance of cataloguing and citation standards, study description guidelines, and how information in ‘codebooks’ could be best represented in the DDI. Throughout this exploration, we point to the contributions of the many participants in the DDI’s early development, especially acknowledging Sue Dodd’s influence and engagement.

The recent flurry of interest around widely promoting data citation and attribution to entice researchers into sharing their data has been largely unconnected to the history of cataloguing machine-readable data files. In the introductory chapter to the National Research Council’s report (2012), For Attribution – Developing Data Attribution and Citation Practices and Standards, Christine Borgman acknowledged that the current debate around drivers behind data citation and attribution has failed to recognize long-standing cataloguing practices for research data. Simply put, if research data files can be catalogued, they can be cited.

We have had standards for cataloging data files since the 1970s. Objects that can be cataloged also can be cited. Similarly, data archives have been promoting data citation practices for several decades. However, over this same period, very few journal editors required data citations, disciplines did not instill data citation as a fundamental practice of good research, granting agencies did not reward the data citations of applicants, tenure and reward committees did not recognize data citations in annual performance reviews, and researchers did not take responsibility for citing data sources. (National Research Council, 2012, p. 1)

While the potential for widespread adoption of data citation practices has been present for several decades, the uptake has been slow largely because the production of catalogue records has been primarily associated with print objects. It is significant that the rules for cataloguing machine-readable data files were openly embraced by the Social Science data archive community in the late 1970’s and early 1980’s. However, the wider library community was slower to adopt these practices. For example, an OCLC project in 1990 generated catalogue records for the complete set of ICPSR Class I codebooks in print, which reflected the bias at that time for print over digital objects. Nevertheless, an increasing volume of descriptive information about machine-readable data files and accompanying documentation catalyzed the change from cataloguing to metadata.

Over the past three decades, the production and use of descriptive information supporting the discovery, access, usability and preservation of research data has fundamentally changed. We find it helpful to think of this transformation in terms of a shift in research infrastructure. Paul Edwards, Steven Jackson, Geoffrey Bowker, and Cory Knobel (Edwards, et. al., 2007) provide a model that characterizes such changes in research infrastructure. Adoption of this model requires seeing metadata as a component of research infrastructure, which is itself a mental shift.

Figure 1 represents the Edwards, et. al. distribution of cyberinfrastructure solutions that take shape across the dimensions of global-local and social-technological contexts. The authors insist that building cyberinfrastructure is not a case of selecting an end point along these dimensions but one of choosing from the distribution of solutions that are availed across these factors.

The following example about the choice of infrastructure to support guest Internet access on a local campus will illustrate the application of this model and help pave the way to discussing changes in infrastructure options for research metadata. A typical university service providing visitors with wireless Internet access requires visitors to be issued a guest account and password. A visitor will need to go to the campus office where the person responsible for issuing guest accounts is located, to show some identification, and to sign an Internet use agreement before receiving an account and password. This specific solution to provide guest Internet access is defined completely by local organizational practices, social norms, and technology and is characterized as a Local-Social solution in the above model.

An alternative solution can be found through institutional membership in eduroam®. Wireless network access is available to anyone at an eduroam site as long as the person is from an institution that is part of the eduroam network. Using the credentials from one’s home institution, the eduroam service
authenticates a guest by automatically verifying their account with their home institution. Local policies can also be configured on eduroam servers to make additional resources available to guests, such as, printers or access to licensed databases. In terms of the model, eduroam is a Global-Technological infrastructure solution. This service is available in over 65 countries worldwide and is governed through of confederation of national organizations.

The array of infrastructure solutions at any one time is in flux. For example, as norms around privacy and confidentiality in today’s digital world swing, the range of social solutions will expand or contract. As interconnectivity using trusted, standards-based protocols shrinks the world, new global solutions become available. New local possibilities emerge as individual institutions develop policies, procedures, and guidelines around digital asset management. Finally, rapid changes in information technology are constantly resulting in new ways of doing things. The story of the movement from cataloguing to metadata is expressed both in the changing array of infrastructure solutions over time and the dominant solutions that have emerged.

The solutions for cataloguing machine-readable data files (which AACR2 now designates as an electronic resource) began largely within the Local-Social context in the late 1970’s when a group of university libraries began producing their own MARC records for research data. In the 1980’s, the ICPSR began circulating study-level MARC records of their data holdings, pushing catalogue infrastructure toward a Global-Social solution. Increased automation by ICPSR in the production of MARC records and the OCLC becoming a distributor of MARC records for ICPSR holdings moved cataloguing support for this collection into the space of Global-Technological infrastructure. However, with increased use of DDI metadata since 2000, MARC records for social science research data have tended to be produced through a crosswalk between the DDI and other record formats, such as Dublin Core or MARC21 XML. Catalogue records can now be derived through other forms of metadata, largely making the practice of cataloguing research data unnecessary.

Catalogue records for research data have been useful for study level discovery purposes but the quest has long been for variable level discovery. In the 1980’s the ICSPR introduced a variables database searchable through SPIRES for a subset of studies. This database demonstrated the value of variable level metadata but the workflow to produce such metadata was dependent on special, extended processing, namely, the generation of ICPSR Class I studies and OSIRIS dictionaries.

The array of infrastructure solutions to facilitate variable level discovery changed throughout the 1980’s and into the 1990’s. Coming out of the 1970’s, documenting research data was treated primarily as a publication process. This information was assembled into a printed report, which was commonly referred to as a codebook. These volumes often contained sections dedicated to a technical description of the study and method of data collection, a detailed listing of the variables, their codes, and their layout in a data file, a copy of the data collection instrument, and any other contextual information providing background to the data. The production infrastructure for such documentation tended to be in the form of solutions that were local, social, and technological. They entailed some automation with a substantial amount of human resources within a local operation.

The mindset of this era was one of assembling as much information as possible and organizing it in a printed booklet. Subsequent use of this metadata, however, often required reentering information for other automated functions. For example, record layouts of variables would be rekeyed for multiple statistical packages, which would often be repeated locally across the many universities receiving copies of the same study and its data.

An important change in the practices around metadata production occurred with a growing acceptance of reusing digital information for many purposes. This practice of enter-once-and-reuse-for-many-purposes pushed the design of data documentation into incorporating digital content that is structured consistently, well defined, and universally sharable.

Concurrent with this view toward reuse of digital information was the introduction of dynamic digital texts. Project Xanadu and subsequent hypertext initiatives demonstrated the power of automating connections between bodies of text. HyperCard (produced by Apple) in the late 1980’s popularized the mapping of relationships in digital text and in the 1990’s the World Wide Web became the most successful implementation of hypertext. The Web also proved the utility of linking key descriptive elements within a document without necessarily linking to the whole document: targeted reuse of specific information elements.

The technology around hypertext coincided with the development of structured conceptual data models that supported the identification of key information elements. The introduction of SGML digital texts employing mark-up tags allowed designating text to a specific structural element. Furthermore, SGML supported the description of conceptual layers of digital information. When hypertext and mark-up languages converged with the Web, the power of describing layers of information in ways that could be linked and reused substantially altered our understanding of metadata for research data. The application of entering information digitally once and then reusing it for multiple purposes through conceptual linkages is now a dominant technological solution in metadata infrastructure.

All of the technological changes leading up to the uses of digital text on the Web shifted the array of solutions for metadata infrastructure from Social to Technological. In addition, the array of solutions has been expanded through two factors driving solutions from Local to Global infrastructure. First, metadata standards for research data were needed to define the key information elements in data documentation. The widespread acceptance and use of a standard such as DDI pushed metadata toward global solutions. Standards enable information to be appropriately compared with predictable outcomes. Second, the ultimate reuse of metadata occurs when this information is turned into machine actions. The movement from the systematic use of metadata to describe elements within a conceptual model to invoking machine actionable workflows is one headed toward creating global data interoperability.

Our thinking about description, discovery, access, usability and preservation has been altered through changes in metadata infrastructure. We are now being challenged about what should be described, how to structure the description, the purpose for which the content can be used, and the workflow processes that this information can drive.
DDI Committee 1995 – 1997

The first meeting of what was to become the Data Documentation Initiative Committee was held in conjunction with the annual meeting of the International Association of Social Science Information Service and Technology (IASSIST) in Quebec City, Quebec, in May of 1995. (See the accompanying *IASSIST Quarterly* article by Karsten Boye Rasmussen for some data description activities prior to this meeting.) At that time, Merrill Shanks, Professor of Political Science at the University of California at Berkeley, was named Chair of what was the first iteration of the DDI Committee: the ICPSR Committee on Survey Documentation (later to be called the Committee on Data Documentation). Invitations to join the committee were sent in February 1995.4

The charge to the committee⁵ made it clear that this was to be an international effort with inclusive involvement across research, survey, and data professional organizations, representing the interests of data producers, archives, distributors, and end-users of social science data.

“As you probably know, this will be hard work; it will require production of a DTD [Document Type Definition, ed.] that meets the needs of ICPSR and that can be implemented immediately in production. I am even more ambitious than that. I would like for the product of this committee to provide a standard that is acceptable to IASSIST, APDU, and our European partner archives as well. It should fully conform to SGML, and this should be tested with standard SGML software. A document should be produced (and made available on the World Wide Web) defining this DTD and making it available to the entire community.”*⁶

It is important to note that the committee was encouraged “to continue to consult with other interested parties concerning both the short-term and long-term goals (or content) for our evolving DTD, and we should keep each other informed of any new development or second thoughts concerning our initial agreement in Quebec.” This was, from the start, seen as a community effort and the committee members were charged with the responsibility of consulting with interested parties.

Some of the larger goals surrounding the development of the DDI were to: come up with a non-proprietary format that was ‘preservation friendly,’ streamline the process from data collection to metadata production; develop metadata authoring tools that are usable for specific purposes; produce and distribute software converters to automate the transport of metadata to varying formats; develop cross-walks and supporting linkages; make it easier to integrate DDI metadata into various systems for resource discovery and statistical analysis; improve linkages between data and metadata; analyze more than one study at a time; offer cross-domain searching and integration; and better integrate geospatial analysis and statistical analysis. (Green, 1999)

Committee members and others from the social science community were given the task of developing a draft list of elements for the first version of the DDI (which was initially rendered in SGML in April 1996). Two subcommittees were established to make recommendations and clarify issues; one of these subgroups concentrated on the different kinds of study level information, while the other focused on the detailed specifications at the variable level. David Barber, at the University of Michigan Library, was charged with combining the suggestions from both subcommittees and with developing the first DDI DTD. It is important to note that the DDI was built to contain descriptive information not only about the variables and coding in a data file, but also to include descriptive information about the study itself and to provide a tagged structure for potentially all of the elements that were deemed important to fully documenting data.

Defining the DDI Study Level Information

The study level subcommittee, led by Ann Green and Mary Vardigan, coordinated the development of the elements for the study level description with the help of Sue Dodd, Karsten Boye Rasmussen, Laine Ruus, Bill Bradley, Carolyn Geda, Pat Vanderberg, Bridget Winstanley, Atle Alvheim, Rolf Uhrer, Richard Rockwell, Merrill Shanks, David Barber, and John Brandt. Their goals were to develop a common core of elements that could be understood and applied across communities of data producers, survey collecting agencies, libraries and archives, researchers, and software developers. The DDI was constructed from existing standards and guidelines that had been in use, in some cases, for decades at data archives in the US, Canada, the UK, and Europe.

The DDI also was intended to support new applications for Study Description Information: to enhance the ability to search, display, and manipulate metadata; to provide a means of discovering that a data set exists and how it might be obtained or accessed; to document the content, quality, and features of a data set and so give an indication of its fitness for use; to supply information for statistical analysis software; and to provide information for citations, cataloguing records, and electronic headers.

The major challenge in developing the structure and individual elements for the study level portion of the DTD was to incorporate the concepts and parts of traditional printed codebooks and also to build compatibility with computer-generated data collection and documentation processes. At the same time the subcommittee gathered elements for the intellectual description, they needed to examine the processes and output of computer-generated surveys to understand the relationship between the survey instrument and the production of study-level descriptive information.

Even though codebooks describing datasets did not at the time have strict standard structures or formats, there was a standard set of intellectual content outlined in guidelines and reviewed in the major social science documentation literature. It was critical that these intellectual components be the defining force behind the “distillation” process of producing the individual elements in the codebook DTD. The goal was to distill a set of elements that were comprehensive and flexible, and capable of producing pieces that are compatible with automated methods of producing codebooks, as well as feeding into systems that describe, cite, catalogue and locate datasets.

The procedure for identifying and defining the study level elements for the DTD included reviewing the following five kinds of resources, each of which is described in detail below:

1. Review cataloguing and citation guidelines
2. Study the ways in which social science data have been described by data archives.
3. Examine Data Archive and Data Producer Guidelines
4. Examine the pieces of existing print and machine-readable codebooks
Bibliographic Citation so that a complete citation could be carried with the documentation. That meant that there was to be a citation to the document itself (the DDI instance), a citation to the study being described (study description), and detailed identification of the particular file/s that made up the physical object being described (file description).

This was part of the motivation to produce the DDI as a modular entity, with components that clearly articulated and integrated these separate conceptual pieces.

Dodd also addressed the need to verify authenticity. She wrote: “the study number supplied by the producer and the archival number supplied by the distributor and archive may be different. This difference should be noted. There can be an original study number (e.g. Harris A019) and an archival study number (e.g. ICPSR 7657). They represent the same data, but different distributors and archives.” It may seem obvious now, but at the time it helped articulate the importance of retaining all distributor and archive assigned study numbers, a key component of trust that content is what it purports to be.

Defining citation principles for data has become a popular topic (CODATA, 2013), but data archives have been promoting data citation practices for approximately forty years, and have for decades included citations to data within data documentation. Since the 1980's, libraries have been producing bibliographic records containing the basic information for how data should be cited in a publication (Mooney & Newton, 2012). The DDI was built upon this history of promoting data citation, of cataloging data, and of including data citations in documentation. The history of data attribution and citation has always been at the core of the DDI.

Cataloging and Citation Guidelines

- MARC-MRDF: the work of the American Library Association Sub-Committee on Rules for Machine-Readable Data Files. Local variations of MARC format have been developed in Canada, the United Kingdom, Sweden, etc.
- ISBD-Computer Files: The International Standard Bibliographic Description for computer files. Recommended by the Working Group on the ISBD set up by the International Federation of Library Associations (IFLA) Committee on Cataloging. (IFLA, 1990)
- GILS: Government Information Locator System: These locators provide users with descriptive, location, and access information for a wide range of (U.S.) Federal government information resources. Compliant with Z39.50, a standard way for two computers to communicate for the purpose of information retrieval and facilitates the use of large information databases by standardizing the procedures and features for searching and retrieving.
- ISO 690-2: Draft Standard for Bibliographic References to Electronic Documents ISO 690-2 is a standard in review for the content, form and structure of bibliographic references to electronic documents, being developed by ISO Technical Committee 46, Subcommittee 9.
- Dublin Core: OCLC/NCSA Metadata Workshop, Online Computer Library Center 1996: University of Warwick/ UK Office for Library and Information Networking OCLC/NCSA Metadata Workshop recommendation for core data elements for discovery and retrieval of Internet resources by a diverse group on Internet users. Listed data elements with possible equivalents in USMARC.

2. Examine the ways in which social science data have been described by data archives

The DDI subcommittee also examined descriptive metadata contained in study descriptions and data catalogs to identify key descriptive material that should be included in the ideal...
The best practices for how to prepare data for archiving have been around for over three decades. It was essential that the DDI support the best practices of preparing and documenting data as described in these guides.

### Data Description

- **Standard Study Description**: developed by and for data archives, adopted by several members of the Council of European Social Science Data Archives in 1974 and endorsed by the Council of European Social Science Data Archives. Further details regarding the origins of the study description can be found in: Nielsen, Per. *Report on Standardization of Study Description Schemes and Classification of Indicators*, Copenhagen: DDA, September 1974, 62 pp. Nielsen, Per. *Study Description Guide and Scheme*, Copenhagen: DDA, April, 1975, 55 pp.

- **ICPSR Study Description "Template" Manual**: "Every new or revised ICPSR study requires a study description which is written by the staff member who processes or evaluates the study. These descriptions follow a strict format, called a template, which insures that standard information is recorded for each study. The template consists of named fields into which the staff person enters appropriate information about the study. Completed templates ultimately reside in an online SPIRES database."

- **Essex ESRC Data Archive Study Description outline**, supplied by Bridget Winstanley

- **Federal Geographic Data Committee (FGDC) Subcommittee on Cultural and Demographic Data (SCDD) Content Standards for Cultural and Demographic Data Metadata (C&DD Metadata Standard)** Specifically the Crosswalk.

- **Geda, Carolyn**: 1980, ICPSR, Data Preparation Manual

- **Essex Survey Research Center Data Archive: Documentation Guidelines Committee (presentation to IASSIST in 1994)**

### Guidelines for Preparing Data

- **Roistacher, Richard**: *A Style Manual for Machine-Readable Data Files and their Documentation* with Sue Dodd, Barbara Noble and Alice Robbin. (Roistacher, 1980). Note that numerous data archives were established in the 1980s and 1990s used this manual. The style manual was influential in the development of standardized documentation of data files.


- **Essex Survey Research Center Data Archive: Documentation Guidelines Committee (presentation to IASSIST in 1994)**

### 4. Examine the pieces of existing codebooks in print and machine-readable formats.

An important step in the process of building the DDI was to review the components of standard printed codebooks, which included full bibliographic identification with a standard citation; an abstract; descriptive and contextual materials, such questionnaires, statements of methodology, appendices and glossaries, and coding schemes for things like geographic entities, topical recoding, or occupation and industrial classifications.

It was also important to examine how statistical packages and data archives at the time included metadata in the system files of these packages and programs. The most comprehensive statistical package, in terms of metadata, was OSIRIS (Rattenbury & Pelletier 1974), a set of computer programs that also included descriptive information. The OSIRIS codebook was part of that system and carried structured information about the survey instrument, file descriptions, and the elements making up a bibliographic reference. One of the goals of the DDI committee was to come up with a replacement of the OSIRIS Codebook / Data Dictionary format. Of course the DDI became more than simply a replacement for OSIRIS as it grew to support the entire lifecycle of data.

Two other information systems influenced the construction of the DDI: One is a suite of software developed at the University of California at Berkeley, CASES (Computer Assisted Survey Execution System) and CSA (Conversational Survey Analysis). Codebooks were produced as a by-product of computer assisted interviewing software (CATI/CAPI) and were integrated into accompanying analysis software. Not only was it useful to examine this system to understand how the survey process intersected with the documentation process, but these tools were in use by researchers and government agencies who could benefit from incorporating the DDI into their survey process and the subsequent dissemination of data and documentation to their user communities. Another system that informed the development of the DDI was Health Canada's DDMS System (Data Dictionary and Documentation Management System). DDMS was a PC-based package for producing social science data dictionaries and documentation and for managing research outputs from Health Canada. This tool furthermore interoperated with metadata registries.

---

**3. Examine Data Archive and Data Producer Guidelines**

The best practices for how to prepare data for archiving have been around for over three decades. It was essential that the DDI support the best practices of preparing and documenting data as described in these guides.
A close review of each of these systems provided insights into what the DDI needed to contain to be integrated into and support the processing of surveys and the management of documentation within analysis systems and metadata registries.

5. Review other encoding guidelines in development at the same time.

Other developing standards at the time, in particular the TEI for encoding digital texts and the EAD for encoding archival descriptions, informed the development of the DDI. They were especially influential because they addressed similar objectives, were based upon community standards, and initially used the SGML framework. All of these initiatives were working on file header standards and the DDI incorporated those guidelines.

Encoding Guidelines
- TEI: Text Encoding Initiative DTD for SGML: markup for primary materials (note that the TEI was used to some extent in producing the DDI XML version 1.6X dated December 27, 1997)
- EAD: Encoded Archive Description DTD for SGML: markup for metadata describing primary archival materials
- File header information, drawing upon other encoding standards. This set of elements contain information about the marked up DDI instance itself. File headers support resource discovery and establish bibliographic identity of the DDI file itself. The DDI Document Description component of the DDI is essentially the “header.”

Refining the elements and moving from SGML to XML: 1995 – 1997

The DDI committee met in October 1995 and again in April 1996 to examine a sample SGML DTD prepared by John Brandt and his colleagues at the University of Michigan Library. At a meeting in October 1997, subcommittees were formed to conduct a review of the elements of the DTD and to address the issue of handling aggregate data in the DTD. In December 1997, the DTD was made compliant with XML (Extensible Markup Language) by Jan Nielsen of the Danish Data Archives. (Nielsen, 1998)

Impact
As we have shown by describing the beginnings of the DDI, the specification was developed at a time of extraordinary shifts in research infrastructure and information science. The creators of the DDI were aware of these shifts and committed to producing a solution to the metadata challenge that could build upon the strengths of data description from, as in the Edwards et. al model, a highly Social and Local context as well as meeting the Technical and Global demands and capabilities of the time. However, with the production of the DDI came new dependencies to find solutions to capture and produce DDI compliant metadata and to take advantage of the constantly evolving technical capabilities and a rapidly changing research environment.

The vision of the DDI and the metadata produced through its application went beyond merely structuring information necessary for using data. Metadata was also seen as the connection between data producers and data users and the technical solutions required to meet the challenges of transferring knowledge in structured formats. As Jostein Ryssevik wrote (Ryssevik, nd):

“Whereas the creators and primary users of statistics might possess ’undocumented’ and informal knowledge, which will guide them in the analysis process, secondary users must rely on the amount of formal meta-data that travels along with the data in order to exploit their full potential. For this reason it might be said that social science data are only made accessible through their metadata. The metadata provides the bridges between the producers of data and their users and convey information that is essential for secondary analysts.”

But building that metadata bridge is a difficult task. At the time the DDI began, there were, and continue to be, major challenges in collecting and distributing metadata. The most obvious and disconcerting fact is that information about data, its context, and its content, is not recorded or is inadequately stored – for example in unstructured and incomplete ‘read me’ files. The commitment, workflow tools, and production of what needs to accompany data for informed use have not been widely or enthusiastically embraced across research teams. The reasons are primarily due to time and resource constraints (Tenopir et al., 2011), but also a lack of integrated tools that capture metadata throughout the research lifecycle and that package the information in ways to support the sharing and archiving of data.

Recent requirements to share and preserve data have created a new conversation about research data management, yet at the same time data sharing platforms accept data without verifying the quality of the documentation. The norm with incoming data is not to review or to check for adequate metadata that would support reuse and replication. In spite of the presence of metadata specifications across many disciplines and detailed guidelines in preparing data, the challenges of producing and distributing good quality structured documentation continue to impede the reuse, replication, and sharing of data. The promise of the DDI cannot be met as long as metadata are not being captured.

Another challenge is to explore how the DDI could be incorporated into tools that capture metadata throughout the full lifecycle of the research process. As Alice Robbin wrote “(d)ata documentation, the descriptive text accompanying a file, is the key to understanding its quality” and “should be prepared at the time of a file’s creation and may contribute significantly to future use of the data …”(Robbin, 1981). Lifecycle models developed soon after the DDI emerged made it clear that metadata production was not simply a process that happened at the end of a research project (Green and Kent, 2002; DDI, 2004; Humphrey, 2006). The idea of the metadata lifecycle, and its intersections with the research lifecycle, has become a common element in publications, conference presentations, and metadata modeling efforts. Mary Vardigan’s article in this issue carries our story forward into the development of the lifecycle model of the next version of the DDI.

The DDI specification is dependent on new technological developments to reach its potential capacity. We draw particular attention to:
- interoperate with other metadata systems for resource discovery and cataloging systems;
- establish software for parsing, validating, viewing, searching, manipulating, authoring and converting;
- exploit the ability to link to with other digital objects;
- take advantage of non-proprietary and platform independent metadata in preservation systems;
• integrate descriptive metadata into analysis, visualization, mapping, data mining,
• realize interoperability with other data and the promise of open data, especially the demands related to comparability, privacy, authenticity, and attribution; and
• inculcate into the habits and workflow of research and data production systems tools and incentives for creating metadata.

Responses to such challenges are best met through a concerted effort to enrich the evolving array of solutions identified within the metadata infrastructure model described above. As a community, we especially want to exploit technologies that are flexible and responsive to local requirements, to incorporate social drivers and habits, and to have a clear goal of meeting global requirements for shared and open data. This can be done, just as the DDI was created, by incorporating potential solutions, carefully articulated requirements and expertise from across the communities of data producers, researchers, data archives, and institutional repositories.

References

CODATA/ITSCI Task Force on Data Citation. (2013). “Out of cite, out of mind: The Current State of Practice, Policy and Technology for Data Citation.” Data Science Journal 12, 1-75. <http://dx.doi.org/10.2481/dsj.OSOM13-043>


Notes

1. Ann Green, Digital Lifecycle Research & Consulting (dlifecycle@gmail.com), is an independent consultant working in the areas of research data management and digital preservation. Chuck Humphrey (chuck.humphrey@ualberta.ca) is the Research Data Management Services Coordinator in the University of Alberta Libraries and the Academic Director of the Alberta Research Data Centre.


3. See <https://www.eduroam.org>

4. Memorandum from Merrill Shanks, Chair, to Fellow Members of the ICPSR Committee on Survey Documentation, hand dated June 7, 1995

5. Letter of invitation from Richard Rockwell


7. A draft revision of AACR2 chapter 9 (renamed: Computer Files) was published in 1987.

8. Email from Ann Green to Mary Vardigan, March 12, 1996. Subject: dtd changes. Enclosed are communications between Sue Dodd and Ann Green in regard to the review of study level elements.


10. For information about the history of the EAD see: <http://www.dlib.org/dlib/november99/11pitti.html>

11. See the DDI lifecycle illustration here: <http://www.ddialliance.org/what/>
The DDI Matures: 1997 to the Present

by Mary Vardigan

Abstract
The Data Documentation Initiative (DDI) began in 1995 with a small international group coming together with a focus on social science metadata. The group moved quickly to develop a first specification, around which a community of practice emerged. That community, along with the DDI specification itself, has evolved over the last two decades to reflect developments in the social sciences, technological advances, and innovation in research practice. This paper chronicles the history of the DDI from its instantiation in XML in 1997 to its current status as the de facto standard for documenting data in the social and behavioral sciences.

Keywords Documentation, metadata, standards

Note of Acknowledgment
The centrality of good documentation to effective social science has been a key tenet of the IASSIST vision throughout its history. Since hosting the inaugural meeting of what was to become the Data Documentation Initiative in Quebec City in 1995, IASSIST has nurtured and promoted the DDI effort every step of the way. Indeed, the DDI story cannot be told properly without acknowledging the support of the IASSIST community through the years. The pioneering efforts and leadership of Sue Dodd deserve special recognition. Her work in data citation and structured metadata inspired a community and set the stage for developments like DDI. Fittingly, in 1993 Sue received the first ICPSR-sponsored Warren Miller Award for Meritorious Service to the Social Sciences, which recognizes contributions to essential infrastructure. We continue to benefit from her foundational work and wisdom.

Introduction
When our story left off (see Ann Green and Chuck Humphrey’s article in this issue), the Data Documentation Initiative (DDI) specification had just been translated from SGML to XML in 1997, and it was moving toward official publication. There was a strong sense in the community that DDI XML, given its rich and structured nature, could be used to drive process, and that its coverage could and should be extended to document more complex datasets.

What happened next? How did the DDI Committee go on to enhance and augment the specification to meet rising expectations? How did the community of practice grow to encompass users in over 70 countries around the world? And how did infrastructure surrounding the DDI, including sustainable support for the organization itself and tools to make use of DDI, come into being? This paper describes the ways in which the DDI initiative addressed these challenges from 1997 to the present, ending with a view into what the future holds for DDI as we move forward.

Publishing DDI Version 1
As we have seen, DDI began as a volunteer effort, drawing on metadata expertise and interest from across the social science research community. In-kind contributions made it possible for DDI to be instantiated as a specification with a user community actively working around it.

However, in-kind contributions cannot provide the type of sustainable structure needed to fund face-to-face meetings and development work, and thus the DDI Committee and its founders decided to pursue external funding streams. In 1997 the Inter-university Consortium for Political and Social Research (ICPSR) applied to the National Science Foundation (NSF) under
its “Infrastructure in the Social Sciences” program and received an award that included funds for enhancing and testing the DDI specification. Having convened the first DDI Committee two years earlier, ICPSR was the “home” for DDI, providing administrative and substantive support.

The beta-test of the DDI DTD began in March 1999 and continued until August. Betatsters’ received financial support to test the specification and to report on their findings.

At the conclusion of the beta-test, a list of changes suggested by the testers was compiled and subsequently reviewed at a meeting of the DDI Committee held in October 1999. Version 1 of the DTD, incorporating these changes, was published March 24, 2000.

Extending Version 1

With Version 1 published and external support for DDI activities ending, the DDI Committee needed to find new funding sources. While an independent review funded by NSF found that DDI was a “worthwhile scientific effort that filled an urgent need for standardization of social science technical documentation and interoperability” (indeed, one evaluator termed it “a strategic component of the infrastructure necessary to support the exchange of structured social research survey data” [ICPSR 2001]), it was difficult to obtain funding for this type of endeavor. Health Canada came to the rescue, providing a substantial amount of financial support during 2001-2002 to enable the DDI Committee to meet and to make improvements to the specification, most notably additions related to aggregate data and geography.

Aggregate data was the focus of a small Working Group meeting in April 2001 in Voorburg, the Netherlands. Agreement was reached during that meeting on a draft aggregate data model (also known as “nCubes”), which was reviewed by the Committee at a meeting in Washington, DC, held in June 2001. Committee members began testing the new Version 1.02 of the DTD, with the extension describing aggregate/tabular data. Several other changes were made to study- and variable-level elements. By March 2003, Version 2 of the specification was published, with these enhancements:

- Aggregate extension elements (nCubes and location map)
- Internal formatting elements from the TEI specification to permit formatting within elements
- New geographic elements: Geographic Bounding Polygon, Polygon, Point, G-Ring Latitude, G-Ring Longitude, and Geographic Map

The DDI Alliance Emerges

It was becoming clear that generating external funding to support continued development of the DDI specification would be challenging so another approach was considered: becoming a self-sustaining membership Alliance, modeled along the lines of the successful World Wide Web Consortium.

In June 2002, the DDI Committee met in Storrs, CT, in conjunction with the IASSIST meeting. The main focus of this meeting was a draft charter, written by Richard Rockwell, to create a DDI Alliance – a new membership structure and funding base that would provide support so that the initiative could continue. The charter document provided for an Expert Committee with representation from the DDI Alliance membership, with each member of the Committee having a vote and thus a say in the future of the DDI. A Steering Committee to provide oversight was also established via the charter.

The final meeting of the original DDI Committee was held in February 2003, in Washington, where the group approved numerous changes to the DTD leading to the publication of DDI 2.0 (see above). An Open Meeting of the DDI Alliance was held in conjunction with IASSIST in Ottawa in May 2003. Meeting participants discussed the new Alliance structure and elements of a Strategic Plan for the next three years of the Alliance.

Toward a Lifecycle Specification

Meanwhile, expectations around what the DDI could document were growing. A page on the DDI Web site (“About the Specification”) in 2003 provides this vision for the DDI:

“The DDI aims to be the foundation for collection, distribution, use, and archiving of many future data collection projects in the social and behavioral sciences, across institutions, countries, and disciplines. It also aims to be the basis for retrofitting documentation of older studies for improved ease of use and stronger guarantee of archival preservation.”

At its first meeting in 2003, the new DDI Expert Committee picked up this vision and set an agenda for the future that was ambitious and comprehensive.

First, the committee discussed the need for a data model. It was generally agreed that the XML Document Type Definition (DTD) for the DDI had limitations: it was not as modular and easily extensible as it should be and it had not been thoroughly reviewed for internal logic. Having a model, most likely in Universal Markup Language (UML), to reflect the underlying design and structure of the specification would represent a big step forward. With such a model, the DDI could be expressed as XML Schema, RDF, a DTD, or possibly other formats. The Health Canada/Nesstar partnership had already done some work on a data model for its DAIS/neststar software that harmonized the DDI with ISO 11179.

Employing XML Schemas to express DDI was also discussed. Moving the DTD to a Schema to take advantage of the modularity in Schemas, the capability for local extension, and the flexibility of namespaces was considered essential to the DDI’s continuing evolution. ICPSR and Harvard-MIT Data Center had been working on a Schema version of the DDI that incorporated all of the documentation found in the Tag Library as well as the DTD comments.

It was pointed out that the Alliance could not abandon the DTD since a lot of markup had been done that was compliant with Versions 1 and 2. The Alliance discussed the need to proceed on parallel tracks, moving the DTD along from Version 2.0 to subsequent iterations in that development line at the same time that a modular Version 3 was developed.

The group also reviewed the Statistical Data and Metadata eXchange (SDMX), a project to develop an interchange format for time series data and metadata. The SDMX initiative was viewed as a natural partnership, and it would come to be seen as complementary to DDI (Gregory and Heus, 2007). Aligning with the MetaDater and MADERA projects in Europe was another topic raised.
To carry forward this ambitious program of work, new working groups were formed:

- Structural Reform Working Group (later called the Technical Implementation Committee [TIC] and now the Technical Committee [TC]), which would take on the task of “schematizing” DDI
- Substantive Content Working Group, broken out into:
  - Group 1: Aggregate Data, Geography & Time
  - Group 2: Comparative Data/Families of Datasets
  - Group 3: Complex Files
  - Group 4: Instrument Documentation
- Usability and Outreach Working Group

This ambitious agenda set the stage for DDI 3, which was five years in the making. At its meeting in 2004 in Madison, Wisconsin, the Expert Committee discussed the coverage and scope for DDI 3 and introduced the concept of a lifecycle model. This model (see Figure 1) was innovative for its time; subsequently the notion of the data lifecycle became an integral part of the discourse around research data management. The DDI Lifecycle approach would come to influence the Generic Statistical Business Process Model (GSBPM) used by national statistical institutes as a framework for data production (Vale, 2010).

Not everyone was on board with a move to this lifecycle model. A considerable investment had been made in DDI 2.X and people were understandably reluctant to support a new specification that would in effect shift attention and resources away from this version. However, in 2005, at the meeting in Edinburgh, the Alliance ratified the lifecycle model and DDI 3 began to take shape.

A public review of DDI 3 took place in 2007 and the specification was published in 2008 as XML Schemas. It was a radical departure from DDI 2.X in many ways. First, it was designed to be used by developers with machine-actionability in mind. While DDI 2.X could be understood and implemented by data librarians, DDI 3 was more complex, often requiring a higher level of technical expertise.

The specification itself was designed to be modular and to document and manage different stages of the data lifecycle. It was predicated on the principle of reusing metadata to eliminate costly redundancies and support explicit comparison (Vardigan, Heus, and Thomas, 2008). As Green and Humphrey note, “enter once and use many times” is a powerful paradigm, which DDI 3 exploited through referencing. As an example, response categories can be defined once and then used multiple times by both questions and variables.

DDI 3 also aligned with several other metadata standards including ISO 11179, SDMX, geographic and spatial standards, Dublin Core, and others

The Community Expands

As the DDI community of practice began to integrate DDI 3 into its work after 2008, interest from new audiences, including national statistical institutes (NSIs), other data producers, and developers and implementers, became evident and new DDI projects sprang up, many of which were discussed in IASSIST presentations, workshops, and posters. Uptake of DDI 2.X continued as well, resulting in DDI spreading across the globe. As a result of the World Bank-supported International Household Survey Network (IHSN) program and its incorporation of DDI into documentation tools, DDI came to be used in over 70 countries, many in the developing world (see Figure 2).

Bringing users together

DDI users were eager to meet in a forum where ideas, innovations, and knowledge of DDI could be shared. Led by Joachim Wackerow of GESIS-Leibniz Institute for the Social Sciences and Nikos Aiskitas of the Institute for the Study of Labor (IZA) in Germany, the first European DDI User Meeting (EDDI), cosponsored by GESIS and the IZA, took place in Bonn, Germany, in December 2009. Subsequent meetings took place in Utrecht, Netherlands; Gothenburg, Sweden; Bergen, Norway; and Paris, France, with the 2014 EDDI slated to take place in London.

In 2013 the concept of a DDI user meeting spread to North America, with the first NADDI conference taking place at the University of Kansas, organized by Larry Hoyle with funding from the Alfred P. Sloan Foundation. NADDI 2014 will be held in Vancouver.

Teaching about DDI

DDI training, which had been taking place around IASSIST and in other venues since 2001, was expanded in response to DDI 3, with Joachim Wackerow of GESIS organizing annual training and workshops at Schloss Dagstuhl, Leibniz Center for Informatics, an IT retreat center in Wadern, Germany, during the last quarter of each year. The first such training took place in 2007. Typically, training in DDI 3 is held for a week, followed by a workshop on a dedicated topic.

Developing DDI tools

Tools are key to a metadata standard’s success: if markup cannot be produced efficiently, a standard may not find an audience. DDI owes much of its success to the parallel development of the DDI tools.

![Figure 1: DDI Lifecycle Model](image1)

![Figure 2: Organizations using DDI](image2)
markup and data analysis tool Nesstar, created by the Norwegian Social Science Data Service (NSD) for use with DDI Codebook (NSD, 1999). Nesstar Publisher formed the basis for a toolkit designed to assist data producers in developing countries in documenting and disseminating data (see the IHSN discussion above). Dataverse Network, developed by Harvard-MIT, also adopted DDI, both as a standard for the study-level metadata entered at deposit and as a foundation for variable-level analysis. Support for DDI 2 was also incorporated into the Survey Documentation and Analysis (SDA) online analysis system created at the University of California, Berkeley.

Other innovative tools were developed with a focus on DDI Lifecycle. For example, the Michigan Questionnaire Documentation System (MQDS) exports DDI Lifecycle from Blaise Computer-Assisted Interviewing software. StatTransfer, a commercial product that transfers data among software packages, now also exports DDI 3.

The Danish Data Archive developed a DDI editor and Colectica developed a suite of tools that includes software to produce, view, and edit DDI Lifecycle metadata with an interface to CAI tools. Colectica for Excel was also added, permitting researchers working in Excel to create study descriptions and document their data at the variable level.

Database tools for DDI like Questasy, developed at CentERdata in the Netherlands, added new options for DDI users. The recently released Sledgehammer tools suite developed by Metadata Technology facilitates the transformation of data across formats and enables the extraction and generation of DDI metadata. Developers of many of these DDI-enabled tools have begun to meet periodically during the year at conferences to share ideas and to keep each other informed as new tools are created.

**Stepping Back: An Evaluation Takes Place**

The flurry of activity after DDI Lifecycle was released and the increasing diversity and expectations of new audiences led the DDI Alliance to initiate an open and independent review of the DDI initiative to inform its evolution going forward. There was a sense that the Alliance had matured since its inception in 2003 and that the organization needed to restructure to align with its accomplishments in order to be equipped to address new challenges.

Thus, in 2010, at the request of the DDI Alliance members, the DDI Steering Committee initiated a thorough and independent review of DDI governance and IP issues. The Steering Committee contracted with Breckenhill Inc. to provide a review under the following terms of reference:

- Clarify the intellectual property rights to the DDI specification and how the Alliance can best protect its IP
- Consider alternatives to the current Alliance governance structure
- Review the structure of host institutions and associations described in the Bylaws with a view toward opening up the Alliance to others to participate in governance
- Review the Bylaws and rewrite to be more specific on the above points
- Provide guidance on having a Constitution that does not change and Bylaws that are easier to change, separating the mechanism for revising the specification from the Bylaws
- Review the membership agreement and suggest content
- Suggest content of a contributor agreement for those contributing products to the Alliance
- Review the current Conflict of Interest form used by the Alliance and provide guidance on how the Alliance should approach this broad area

After interviewing a large group of DDI stakeholders and consulting widely on legal issues, Breckenhill provided a report to the Steering Committee in May 2011 detailing the findings related to the above questions (Breckenhill Inc., 2011). In response to the review, the Alliance drafted a new Charter and Bylaws, which went into effect in July 2013. These new Bylaws outlined an organization that is broadly representative of the membership and structured to support the effective development of the DDI specifications. There is an Executive Board elected by the member representatives, a Scientific Board that oversees the substantive development of the DDI specifications, and a Technical Committee that creates and stewards the specifications and ensures their usability.

The revised Bylaws also allow for the DDI Alliance to be instantiated within the University of Michigan as an organizational host. This arrangement permits the U-M to protect the intellectual property of the Alliance and provides a home for the DDI Alliance Secretariat through the Inter-university Consortium for Political and Social Research (ICPSR).

DDI also rebranded its specifications after the review to give an indication of their scope: DDI 1.X and 2.X became DDI Codebook, while DDI 3.X became DDI Lifecycle.

**DDI Moving Forward: What Lies Ahead**

**Attracting new audiences**

With the publication of DDI Lifecycle, there was a surge of interest in DDI by many of the national statistical institutes and organizations around the world, and the DDI Alliance now counts the U.S. Bureau of Labor Statistics, Statistics New Zealand, the Australian Bureau of Statistics, the French National Institute of Statistics and Economic Studies, the Food and Agriculture Organization of the United Nations, and Eurostat as DDI members.

The DDI Alliance began working with NSIs in some notable ways with important synergies emerging. The SDMX-DDI Dialogue project helped to surface the similarities and differences between the two standards (many NSIs mandate the use of SDMX), and the DDI Alliance formally endorsed a collaboration with the SDMX community to enable the two standards to work together.

The Alliance has also supported development of the Generic Statistical Information Model (GSIM), the first internationally endorsed reference framework for statistical information that NSIs are using to inform the modernization of the production of official statistics. The Alliance has made an offer of support to work together on an implementation model for GSIM.

**Integrating with the semantic web**

Work is under way on two RDF (Resource Description Framework) vocabularies, the DDI-RDF Discovery vocabulary for publishing metadata about datasets into the Web of Linked Data, and XKOS, an RDF vocabulary for describing statistical classifications, which is an extension of the popular SKOS (Simple Knowledge Organization...
System) vocabulary. The public review of both vocabularies is planned for 2014.

**Developing the next-generation DDI**

While the existing DDI Codebook and Lifecycle specifications continue to be fine-tuned (DDI Codebook Version 2.5.1 [in Schema form] and DDI Lifecycle 3.2 were published in early 2014), the Alliance has begun another ambitious project – to create a DDI specification based on an information model. The Alliance supports this move to a model-based specification as it will provide greater flexibility: the model can be expressed in a variety of technical formats including XML Schema, RDF/OWL Ontology, relational database schema, and other languages. Also, having a model will make it easier to understand the specification, to interact with other disciplines and other standards, to develop and maintain it in a consistent and structured way, and to enable software development that is less dependent on specific DDI versions. Interestingly, creating a data model was a component of the original agenda for development of DDI 3, so the initiative has come full circle.

The Alliance has other goals for this new model-based DDI: this is an opportunity to respond to community expectations by creating a new version of the specification that can transcend traditional disciplinary barriers to document data about humans and their impact more broadly. As an example, while data collection instruments in the social sciences have traditionally been surveys, we can also view blood pressure gauges and magnetic resonance imaging (MRI) scans as new types of instruments that capture and export data. There is also a growing emphasis on documenting data from administrative registers and various Internet sources. In addition, the next-generation DDI will ultimately add coverage in several new areas:

- Abstraction of data capture/collection/source with “plug-ins” to handle different types of data
- New content on sampling, survey implementation, weighting, and paradata
- New content pertaining to qualitative data
- Framework for data and metadata quality
- Framework for access to data and metadata
- Process (work flow) description across the data life cycle, including support for automation and replication
- Integration with existing standards like GS BPM/GSIM, SDMX, CDISC, Triple-S
- Disclosure review and remediation
- Data management planning

Work on the model began in October 2013 when a group convened at Schloss Dagstuhl to focus on gathering requirements for and modeling this next-generation DDI. As part of a paper summarizing the requirements (DDI Working Paper Series No. 4), the group articulated a set of design principles for the information model that reflect what the Alliance has learned over the years about effective standards and their development:

1. Simplicity – The model is as simple as possible and easily understandable by different stakeholders.
2. User Driven – User perspectives inform the model to ensure that it meets the needs of the international DDI user community.
3. Terminology – The model uses clear terminology and when possible, uses existing terms and definitions.
4. Iterative Development – The model is developed iteratively, bringing in a range of views from the user community.
5. Documentation – The model includes and is supplemented by robust and accessible documentation.
6. Lifecycle Orientation – The model supports the full research data lifecycle and the statistical production process, facilitating replication and the scientific method.
7. Reuse and Exchange – The model supports the reuse, exchange, and sharing of data and metadata within and among institutions.
8. Modularity – The model is modular and these modules can be used independently.
9. Stability – The model is stable and new versions are developed in a controlled manner.
10. Extensibility – The model has a common core and is extensible.
11. Tool Independence – The model is not dependent on any specific IT setting or tool.
12. Innovation – The model supports both current and new ways of documenting, producing, and using data and leverages modern technologies.
13. Actionable Metadata – The model provides actionable metadata that can be used to drive production and data collection processes.

**Conclusion**

From its modest start in Quebec City in 1995 with 23 individuals around the table, the DDI initiative has accomplished some important objectives, producing two development lines to document social science research data. The work continues, with a new, more ambitious goal: to spread the next-generation DDI across the social and behavioral sciences and into new communities to ensure the effective documentation of research data and its future use.

DDI will mark its 20-year anniversary in 2015. With almost two decades of experience, the DDI community has learned a lot about metadata standards development, and the lessons learned can inform what lies ahead. The journey is sure to be interesting and we welcome fellow metadata travelers, both within IASSIST and beyond.

**References**


About the MADIERA Project: <http://www.ist-world.org/ProjectDetails.aspx?ProjectId=2bbe31ebd47c4eda7c4465b7e9ee8ca5SourceDatabaseId=9cd97ac2e51045e39c2ad6b86dce1ac2> (Accessed February 12, 2014)

About the Metadater Project: <http://www.ist-world.org/ProjectDetails.aspx?ProjectId=2bbe31ebd47c4eda7c4465b7e9ee8ca5SourceDatabaseId=9cd97ac2e51045e39c2ad6b86dce1ac2> (Accessed February 12, 2014)


IASSIST QUARTERLY 2013

DDI Tools Catalog: <http://www.ddialliance.org/resources/tools>
(Accessed February 12, 2014)


<http://dx.doi.org/10.3886/DDIWorkingPaper04>


<http://dx.doi.org/10.2218/ijdc.v3i1.45>

Notes

1. Mary Vardigan is an Assistant Director at the Inter-university Consortium for Political and Social Research (ICPSR) and Director of the DDI Alliance. She can be reached at vardigan@umich.edu.

2. Beta-esters of the first DDI specification included the following institutions:
   • Centre for Comparative European Survey Data (CCESD) – Contact: Richard Topf
   • Danish Data Archive – Contact: Nanna Floor Clausen
   • The (UK) Data Archive – Contact: Ken Miller
   • Harvard-MIT Data Center – Contacts: Michael McDonald, Micah Altman
   • NMI-Steinmetz Archive – Contact: Repke de Vries
   • Norwegian Social Science Data Services (NSD) – Contact: Jostein Ryssveik
   • University of California, Berkeley, Survey Research Center – Contact: Juthe Theresa Cheng, Jeff Royal
   • University of Giessen – Contact: Karsten D. Wolf
   • University of Ljubljana, Social Science Data Archive – Contact: Janez Stebe
   • University of Michigan, Harlan Hatcher Library – Contacts: Bonnie Dede, JoAnn Dionne, Lynn Marko, Patricia Dragon
   • University of Minnesota, Machine Readable Data Center – Contact: Wendy Treadwell
   • University of Warsaw, Institute for Social Studies – Contacts: Pawel Morawski and Jacek Szamrej
   • University of Wisconsin-Madison, Data and Program Library Service – Contact: Cindy Severt

3. Members of the DDI Committee when it met for the last time in February 2003 included: Bjorn Henrichsen, Chair, Norwegian Social Science Data Services; Micah Altman, Harvard University; Atle Alvheim, Norwegian Social Science Data Services; Grant Blank, American University; Ernie Boyko, Statistics Canada; Bill Bradley, Health Canada; Cavan Capps, Bureau of the Census; Bill Connett, University of Michigan; Cathryn Dippo, Bureau of Labor Statistics; Pat Doyle, Bureau of the Census; Dan Gillman, Bureau of Labor Statistics; Peter Granda, ICPSR, Ann Green, Yale University; Peter Jofits, ICPSR; Ken Miller, ESRC Data Archive; Tom Piazza, University of California, Berkeley; Karsten Boye Rasmussen, University of Southern Denmark; Richard Rockwell, The Roper Center; Jostein Ryssveik, Norwegian Social Science Data Services; Merrill Shanks, University of California, Berkeley; Peter Solenberger, University of Michigan; Wendy Thomas, University of Minnesota; Rolf Uher, Zentralarchiv für Empirische Sozialforschung; Mary Vardigan, ICPSR

Appendix: Past Chairs, Vice Chairs, and DTD and Schema Authors

DDI Committee Chairs
- Merrill Shanks, University of California, Berkeley: 1995-2002
- Bjorn Henrichsen, Norwegian Social Science Data Service (NSD): 2002-2003

DDI Alliance Expert Committee Chairs and Vice Chairs
- Tom Piazza, University of California, Berkeley: 2003-2005
- Hans Jorgen Marker, Danish Data Archive (DDA), Chair, and Ron Nakao, Stanford University, Vice Chair: 2005-2010
- Chuck Humphrey, University of Alberta, Chair, and Mari Kleemola, Finnish Social Science Data Service (FSD), Vice Chair: 2010-2013

DDI Executive Board and Membership Chair and Vice Chair
- Gillian Nicoll, Australian Bureau of Statistics, Chair, and Ron Nakao, Stanford University, Vice Chair: 2013-

DDT and Schema Authors (in order of contributions)
- David Barber, University of Michigan
- John Brandt, University of Michigan
- Ann Green, Yale University
- Paul Schaffner, University of Michigan
- Nancy Vlahakis, University of Michigan
- Daniel Pitti, University of California, Berkeley
- Jan Nielsen, Danish Data Archive
- Jerome McDonough, University of California, Berkeley
- Perry Roland, University of Virginia
- Sandra Ionescu, ICPSR
- Mark Diggory, Harvard-MIT Data Center
- Wendy Thomas, University of Minnesota
- Arofan Gregory, Metadata Technology
DDI Timeline

by Mary Vardigan

Abstract
This timeline lists key developments and events in the history of the DDI, beginning with selected foundational developments that set the stage for DDI and facilitated its creation. Note that while many archives have played a significant role in DDI history, only the ones established early on are listed here.

1946
Roper Center established.

1960
Zentralarchiv established.

1962
Inter-university Consortium for Political [and Social] Research (ICPSR) established.

1964
Steinmetz Archive established.

1965-1968

1967
United Kingdom Data Archive (UKDA) established.

1967
OSIRIS (Organized Set of Integrated Routines for Investigations with Statistics) statistical software package developed at University of Michigan, Institute for Social Research.

1968
First version of SPSS (Statistical Package for the Social Sciences) released.

1970
American Library Association appoints Subcommittee to Recommend Rules for Cataloging Machine-Readable Data Files.

1971
Norwegian Social Science Data Service (NSD) established.

1974
International Association for Social Science Information Services and Technology (IASSIST) formed following the Conference on Data Archives and Program Library Services, held in conjunction with the 8th World Conference of Sociology in Toronto, Canada.

1974
Report issued on "Standardization of Study Description Schemes and Classification of Indicators," a product of a meeting held in Copenhagen, Denmark, at the Danish Data Archives with nine archives from six countries in attendance.

1975
IASSIST establishes Classification Action Group.

1976
Council of European Social Science Data Archives (CESSDA) established.

1978
Anglo-American Cataloguing Rules, Second Edition (AACR2) published by the American Library Association (includes a new Chapter 9, for machine-readable data files).

Selected foundational developments that set the stage for DDI

1979

1980
A Style Manual for Machine-readable Data Files and Their Documentation published by R.C. Roistacher with contributions from Sue Dodd, B.B. Noble, and Alice Robbin.

1982
1985
Sue Dodd and Ann M. Sandberg-Fox publish *Cataloging Microcomputer Files: A Manual of Interpretation for AACR 2.*

1985

1985
NSFNet, the precursor to the Internet, created.

1986
SGML (Standard Generalized Markup Language), descended from IBM's Generalized Markup Language (GML) developed in the 1960s, becomes an ISO standard.

1989
World Wide Web invented.

1989
International Federation of Library Associations (IFLA) recommends an International Standard Bibliographic Description for Computer Files (ISBD/CF).

1993
IASSIST forms Action Group for “Codebook Documentation of Social Science Data.”

1993
CESSDA holds seminar on “Variable Level Documentation” in Gothenburg, Sweden.

1993
Dublin Core Metadata Initiative established.

1995
First SGML Codebook Committee, constituted by ICPSR Director Richard Rockwell, meets in Quebec City, Canada. Members develop a draft list of codebook elements.

1996
First DDI specification prepared at University of Michigan Library. *An SGML Document Type Definition (DTD) was produced by David Barber and John Brandt (University of Michigan), Ann Green (Yale University), and the DDI Committee.*

1996
First Working Draft of XML (eXtensible Markup Language) published.

1997
U.S. National Science Foundation (NSF) funding received to enhance DDI and betatest it. This award (SBR-9617813) funded DDI development and ICPSR codebook digitization. *Final Report (pdf).*

1997
SGML DDI specification translated to XML. This work was done by Jan Nielsen, Danish Data Archive.

1998
Committee meets in New Haven, CT. Prepares for betatesting.

1999
Betatest of DDI DTD takes place (see related article for testers’ names).

2000
DDI Version 1 (DTD-based) published. *View Version 1 and successive iterations*.

2001

2001
Funding for DDI development received from Health Canada during 2001-2002. Covers costs of meetings until Alliance established.

2001
Working group on aggregate data meets in Voorburg, Netherlands. Group develops a proposal for DDI coverage of aggregate/tabular data.

2002
Bjorn Henrichsen, NSD, becomes Chair of the DDI Committee.

2002
Committee meets in Storrs, CT, to draft DDI Alliance charter. *View the Charter*.

2003
Final meeting of original DDI Committee held in Washington, DC (see related article for members’ names).

2003
DDI Alliance established with Tom Piazza, UC-Berkeley, Chair. View *About the Specification* from original Alliance Web site, March 2007.

2003
DDI Alliance Steering Committee meets for the first time. *Minutes*.

2003
DDI 2 published. DDI now covers aggregate data and geography. *View Version 2 and successive iterations*.

2003
DDI Version History.
2003
DDI Expert Committee meets for first time in Ann Arbor, MI. Committee discusses transition from DTD to Schemas; New working groups on Structural Reform and Substantive Issues formed

2004
DDI Expert Committee meets in Madison, WI. Committee discusses requirements for Version 3

2005
Hans Jørgen Marker, DDA, becomes Chair of the DDI Expert Committee; Ron Nakao, Stanford, Vice Chair.

2005
DDI Expert Committee meets in Edinburgh, Scotland. Committee ratifies life cycle model and DDI 3 begins to take shape

2006
DDI Expert Committee meets in Ann Arbor, MI. Committee approves the scope and timeline for Version 3

2007
Public Review of DDI 3 takes place.

2007
DDI Expert Committee meets in Montreal, Canada. Committee approves Candidate Draft of DDI 3

2007
First DDI Training Workshop takes place at Schloss Dagstuhl in Wadern, Germany.

2008
DDI 3 published as XML Schemas.

2008
Expert Committee meets at IASSIST in Palo Alto, CA. Plans for tools development discussed

2009
DDI Lifecycle 3.1 published.

2009
First European DDI Users Conference (EDDI) held in Bonn, Germany.

2009
DDI Expert Committee meets in Tampere, Finland. Committee discusses tools and outreach to NSIs

2010
Chuck Humphrey, University of Alberta, becomes Chair of the DDI Expert Committee; Mari Klemola, FSD, Vice Chair.

2010
Second EDDI Conference held in Utrecht, Netherlands.

2010
DDI Developers Group meets for first time in Utrecht, Netherlands. Rebranding DDI 2 and 3 as DDI Codebook and Lifecycle approved

2011
External review of DDI takes place.

2011
DDI Alliance publishes first set of controlled vocabularies.

2011
Expert Committee meets in Vancouver, Canada. Results of external review discussed

2011
Third EDDI Conference held in Gothenburg, Sweden.

2011
DDI Alliance releases Tools Catalog.

2012
DDI Alliance establishes agency registry.

2012
DDI Expert Committee Meets in Washington, DC. Plans for a model-based specification discussed

2012
DDI Codebook 2.5 published as XML schemas.
2012
Fourth EDDI Conference held in Bergen, Norway.

Program

2013
DDI RDF Discovery and XKOS vocabularies published.

Description

2013
First North American DDI Users Conference (NADDI) Held in Lawrence, KS.

Program

2013
DDI Membership/Scientific Board Meets in Cologne, Germany. Transition to new governance structure discussed

Minutes

2013
First DDI Executive Board (successor to Steering Committee) meets with Gillian Nicoll, Australian Bureau of Statistics, as Chair and Ron Nakao, Stanford University, as Vice Chair.

View members and minutes

2013
DDI "Sprints" launched to work on model-based DDI.

More information

2013
Fifth EDDI Conference held in Paris, France.

Program

2014
DDI Lifecycle Version 3.2 published.

References


Notes
1. Mary Vardigan is an Assistant Director at the Inter-university Consortium for Political and Social Research (ICPSR) and Director of the DDI Alliance. She can be reached at vardigan@umich.edu.
2. Members of original SGML Codebook Committee:
   • Merrill Shanks, UC-Berkeley, Chair
   • Richard Rockwell, ICPSR
   • Atle Alvheim, NSD
   • Martin Appel, Census
   • David Barber, Michigan
   • Grant Blank, University of Chicago
   • Bill Bradley, Health Canada
   • Pat Doyle, AHCPR
   • Terry Finnegan, National Center for Supercomputing Applications, University of Illinois
   • Peter Granda, ICPSR
   • Ann Green, Yale
   • Stephan Greene, Maryland
   • Lynn Jacobsen, Columbia
   • John Price-Wilkin, Michigan
   • Karsten Rasmussen, DDA
   • Rolf Uher, Zentralarchiv, Koeln
   • Mary Vardigan, ICPSR
3. Full URLs for all material linked in the Time Line can be found in References.
Keywords from Vol 37 No. 1-4. Courtesy Tagxedo.com
Abstract
The social science data community is fortunate to have a tremendous group of talented professionals. We work each day to build upon the ideals founded by those that came before us. This article recognizes the efforts of early IASSIST members whose pioneering efforts enabled our work today. In particular the work of Sue Dodd will be acknowledged. This article reflects on the many partnerships and data oriented projects the author has had the good fortune to be a part of over the last ten years in his work at the Odum Institute (Odum, 2014). Many of these projects are work performed under the aegis of the Data Preservation Alliance for the Social Sciences “Data-PASS” (Data-PASS, 2014). These projects are just a small subset of many achievements accomplished by the greater social science data community.

Keywords: Collaborations, Digital Preservation, Data Management, Metadata, Catalog

A Personal Prologue
It is an honor to have the opportunity to reflect on the advancements our social science community has made in recent years toward metadata harmonization as well as the preservation of the materials we all hold so dear. I have had the pleasure of working for the Odum Institute, University of North Carolina UNC, for twenty-one years so it seems fitting for me to reflect on what has been accomplished as the Institute celebrates its 90th anniversary this year. I was fortunate that my service here at the Institute overlapped with the tenure of Sue Dodd, if only for a few years. The Institute has always provided a home for researchers and staff who share a passion for service to the social science community. Sue Dodd exemplified this ideal, and today as we build on her work, the Odum Institute Data Archive is dedicated to serving the social science community and its customers around the world who are seeking critically important data and information to support their research and data management services.

Foundations
Libraries and archives have been organizing information long before the advent of digital records. In early 1970 the investigation into a set of rules to catalog Machine-Readable Data Files, or “MRDF” began. (Dodd, 1982). Recognizing the unique properties of digital materials and having a keen eye towards both the potential challenges and affordances of cataloging these materials, Sue Dodd was instrumental in the evolution of cataloging standards for MRDF that first made their appearance in the second edition of AACR2 published in 1978 (Dodd, 1982). Her work paved the way for the development of many tools that simplify the discoverability, accessibility, and usability of vast amounts of social science data. Today’s advancements would have been tremendously more difficult without the development of standard cataloging requirements and descriptive methodologies used to define these MRDFs.

My early work at Odum was in the information technology arena. I knew nothing of these early foundations and the valuable work of my new colleague Sue Dodd. I did not know that one day I would be tasked with the migration of thousands of...
The Odum Institute was involved with one of the first projects – with the National Archives and Records Administration (NARA), and the Murray & Rasmussen, 2004) led to a shared cataloging, authority control, acquisition systems, private file creations, products and a union list. As we know, these services and products we now take for granted are offered around the world today for a vast amount of social science data.

Behind this mountain of data is a network of researchers, archivists, librarians, information scientists, and administrators like Sue Dodd who work tirelessly to safeguard and provide access to valuable social science data that has helped to guide everything from public policy to education. We owe credit not only to Sue Dodd, but also to the whole of our international social science community for building these remarkable tools and services that continually add to the legacy of pioneers in our field. I value this opportunity to reflect on the enriching collaborations I have been involved with over the past ten years working to fulfill these earlier visions, and I encourage readers to do the same.

**Building a Union Catalog**

The Odum Institute was involved with one of the first projects following the formation of the National Digital Information Infrastructure and Preservation Program “NDIIPP” (Library of Congress, 2014). As part of the newly formed Data Preservation Alliance for the Social Sciences “Data-PASS” (Data-PASS, 2014) led by the Inter-university Consortium for Political and Social Research (ICPSR), we became a member of a voluntary partnership to archive, catalog and preserve valuable social science data that were at risk, in support of the NDIIPP agenda. The early Data-PASS partners – ICPSR, the Institute for Quantitative Social Science at Harvard (IQSS), the Odum Institute, the Roper Center, the National Archives and Records Administration (NARA), and the Murray Center – were not strangers to one another. For many years, we had worked together on projects to provide access to quality social science data for our constituents. This familiarity, combined with a shared common goal, allowed the partnerships to grow and take root. Once Data-PASS was established, the group immediately began to survey the landscape and take action. By building on existing relationships, the Data-PASS partners were able to expedite the process. (Crabtree & Donakowski, 2006). The Data-PASS partners had four primary goals during the NDIIPP project: (1) archive at-risk social science content, (2) build a shared union catalog, (3) provide replicated preservation, and (4) advocate for best practices in digital preservation. We began to identify at-risk content almost immediately and developed strategies and best practices to manage this task. Jointly, we also began to develop a plan to take steps toward building the union catalog envisioned in the early days of MRDF catalog records.

Our strategy was to utilize standard harvesting methodologies like the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH, 2014) to collect metadata records from partner repositories. This would provide a standardized interface and allow the integration of existing and diverse technologies across the partnership into the new common catalog. We were very fortunate that our partners at Harvard IQSS (IQSS, 2014) were already developing open source archive technology that used these standards and thus had experience in this area. The Odum Institute took advantage of Harvard’s success in building and implementing their virtual archiving platform and became one of the first outside implementers of what is today the sophisticated Dataverse Network, DVN (Crosas, 2011). Because the Data-PASS common catalog design was platform agnostic, partners not having adopted the DVN could still contribute to the catalog via a simple OAI-PMH interface server. This low barrier of technical entry was essential to the success of the partnership.

Of utmost importance to the success of the Data-PASS common catalog was the standardization of each partner’s metadata—made feasible by the groundbreaking work of Sue Dodd many years earlier. The Odum Institute during this period was also migrating MARC records to the new DDI standard and looking for a replacement for the soon-to-be outdated version of the Stanford Public Information Retrieval System (SPIRES) database (SPIRES, 2014). This is not to say it was without challenges, but working with our partners at Harvard we were able to complete the migration of the metadata and ingest the contents of the Odum Archive into our newly built Dataverse Network.

Aside from the Odum Institute’s collection, the Data-PASS partnership brought together a wide collection of social science catalog records from six major US-based social science repositories and for the first time allowed discovery of these social science data from a single common catalog. The Dataverse Network has since expanded to include collections from all over the world and continues to grow daily. The power of a standardized metadata catalog record has been exploited to provide discoverability for a vast amount of social science data worldwide. Replicating and preserving the catalog records of our joint institutions was an important first step, but our partnership also sought to create a distributed preservation system that our partners could leverage to provide geographically distributed preservation for the group.

**Collaboration for Preservation**

The formation of the Data-PASS partnership established the groundwork for a distributed preservation project. Building on the success of the union catalog, the group identified the need to distribute our joint content in addition to our metadata as a means to better protect our data—despite disparities in repository size and resources among the partners. This is a challenge many organizations face. The expense of maintaining multiple machine rooms and backup systems in multiple geographic regions is prohibitive for small to mid-sized repositories. I would argue that it is equally a burden on larger repositories that would rather spend their ever shrinking resources in more fulfilling areas. Both of these circumstances were present among Data-PASS partners, which created the need for our preservation system to deal with the asymmetrical size of the collections (Altman et al. 2009).

Rather than reinvent the wheel, we decided to borrow from the work of other NDIIPP partners working in this space. The MetaArchive (MetaArchive, 2014) project had been working on defining Private LOCKSS Networks (PLN) to adopt solutions already implemented at Stanford University (LOCKSS, 2014). We were able...
to build our preservation network using tested strategies. We had additional challenges along the way due to our content types and sizes but by leveraging the work of fellow NDIPP partners we were better prepared to tackle these challenges. The asymmetrical nature of our PLN layered additional challenges on top of our more distributed administration approach. Each partner had primary responsibility for running their independent LOCKSS node, and because we had no one central administrator for the network, it was essential that we developed a reporting structure that would generate audit reports of the network. These tools did not exist, so we sought additional funding to build auditing tools for our PLN.

Trust but Verify

Data-PASS members needed the ability to audit the new preservation network if it were to demonstrate compliance with standards for trustworthy repositories. The members all had diverse plans for preservation of content in place already, but the addition of a remote copy of each repository under the administration of other members is something that not only needed legal policies in place, but also the ability to audit the performance of the network. This prompted the design of the asymmetrical audit system prototype developed by Data-PASS during the NDIPP project extensions (Altman et al., 2009). Follow up funding from the Institute for Museums and Library Services (IMLS) had allowed the prototype to mature into the current open-source offering, the SafeArchive Audit System (SafeArchive, 2012). Utilizing the TRAC audit framework (CRL, 2007) allows the SafeArchive to enable a PLN to define preservation policies in both qualitative and quantitative means. These user-defined policies are stored in a schematized XML format and used to compare the actual performance of the LOCKSS PLN to their policies. The result is an audit report that can be provided for each of the members on the status of their content as it compares to the preservation policies they have specified.

Data Management Services

It seems that today we are living in the “Age of Data Management Enlightenment.” Everywhere you turn governments, funders, publishers and research institutions are seeking assistance for data sharing, data management, and data science (OSTP, 2014). As I reflect on my time here at the Odum Institute, I want to scream “Social Science Archives Already Do This!” When I calm down, I am thankful that the early work on MRDF has positioned the social science data community at the forefront of modern data management. Our community is comprised of many individuals like Sue Dodd who have the insight, ingenuity, and enthusiasm to contribute to new initiatives. Joint efforts to adopt a common metadata standard like the MRDF metadata grandchild, DDI, along with sophisticated approaches for handling confidential data and the experience of building partnership for preservation and access of complex data files, all provide a wealth of expertise as our society embraces open data policies (Data Transparency, 2013) and builds massive indexes of health-related data (NIH, 2014).

We should embrace new partnerships with libraries and library educators as they tackle the monumental task of managing a research output that is growing exponentially. The Odum Institute is currently working with the UNC School of Information and Library Science and the UNC Libraries in a joint effort to design data management curricula that are flexible enough to be delivered as online content via a Massive Open Online Course (“MOOC”) yet grounded enough to allow students to develop local support networks within their own institutions. As a result of this new Curating Research Assets and Data using Lifecycle Education (CRADLE, 2014) grant, we hope students will share their new knowledge and experiences as they enhance their local data management networks.

The international social science data community is graced with many great organizations that are working to educate researchers on proper data management practices. The current data-sharing climate has prompted the research community to seek these services around the world. Journal publishers are encouraging and in some cases are requiring authors to submit data supporting their findings alongside their manuscripts. This push toward such a replication data requirement will provide a solid foundation for future scientific discovery as new research is designed around previous discoveries. The Dataverse Network is working with journal publishers to help satisfy this new requirement. Efforts like the Open Source Journal (OSJ) deposit API for the Dataverse seek to streamline and simplify this process for the authors and publishers (OSJ, 2014).

Where Do We Go From Here: Hello, Big Data

In the spirit of the 1982 Dodd manual (Dodd, 1982), “Which direction do we go from here?” I hereby declare that the social science data community has come a long way in standardizing data descriptions to make data accessible and understandable for secondary use. Pausing to reflect on our past is indeed a worthy exercise, but we should not rest in our efforts to seek improvements for managing the growing collections of data under our stewardship. Sue Dodd would not be surprised that today the data we are entrusted to are increasingly larger and more diverse than those that came before them. The need for tools and services to visualize and analyze new data types has never been greater. Social science is becoming more and more interdisciplinary, and the community will be facing more complex and larger data types like those used in social network analysis and mixed methods studies. Relationships between social science datasets will become increasingly complex and require a complex object model to describe. This is not a revolutionary notion, and new standards like DDI version 3 are already designed to handle these relationships (DDI, 2014). The challenge will be to integrate these new models into large preexisting relationship among data sets within archives.

As the sheer volume of research data becomes much more massive, we will be forced to seek the council of those in other disciplines that have become accustomed to handling dataset in the petabyte range. The Odum Institute has begun working with the Data Intensive Cyber Environment “DICE” group to begin leveraging the iRODS (iRODS, 2014) rules-based grid system. Tools like iRODS that have the ability to manage multi-petabyte collections and apply active policies will be needed as we begin embracing the new and larger data formats in the future. We have initiated the integration of the Dataverse Network and iRODS that seeks to provide data archiving at scale and allow the federated Dataverse Network access to discover the massive amounts of data existing in data grids around the world. Through our work on the National Science Foundation DataNet Data Federation Consortium project we hope to link diverse communities of data users ranging from oceanographic and hydrologic disciplines to temporal dynamics and plant genomics communities (DFC, 2012).

As social science researchers are encouraged or required to share their data, we must always remember our dedication to protecting human subjects. This will require archives to provide tools to...
assist in this process. The Odum Institute is closely monitoring the progress of and learning from projects like the Data Privacy Center at Harvard’s Data Tags initiative, which will be critical in providing new tools to share these data while protecting our human subjects (Privacy Tools, 2014). We should also seek to partner with computer science and data science initiatives like the National Consortium for Data Science (NCDS, 2014) to better understand our security risks and provide input into the next generation of secure data transmission systems.

Data volumes are almost guaranteed to increase exponentially into the future. The social science data community will need to leverage as much as possible automated metadata generation technologies to help reduce the burden on depositors and archive staff. Automated ingest tools that create variable level metadata are already being deployed in tools such as the Dataverse Network. Projects such as the NSF-funded DataBridge project (Rajasekar et al., 2013) seeks to use sociometric analysis techniques used in social networking to help determine relationships between users, data, and methods. These relationships could be used to produce multilevel object relationship models to aid in data discovery and population of DDI 3 object relationship models. Tools like these, combined with advanced commercial indexing of datasets, will be important to the sustainability of data sharing.

If the Odum Institute and other organizations dealing with data are to contribute to Sue Dodd’s legacy, we must recognize that the complex problems we contend with today often warrant complex solutions. These are solutions that likely cannot be generated by any one individual or organization alone. Members of the social science data community must be willing to reach out beyond their own walls to forge partnerships that take full advantage of the vast amounts of talent that are dispersed throughout our community. To answer Sue Dodd’s question today, “Where do we go from here?” I would suggest that “wherever we go, we go together.”

Conclusion

The social science data community has made great advancements over the years since Sue Dodd and others began defining bibliographic control over computer information in the late 1970s. I have been fortunate during the past ten years to work with wonderful collaborators and colleagues to build on the work of early IASSIST members.

Our community has been fortunate to have strong foundations that have placed us ahead of the game. The social sciences are becoming increasingly interdisciplinary, and we should make every effort to help other disciplines that could learn from our experiences. Sharing knowledge will enhance our ability to deliver quality data management and archiving to the diverse social science researchers we will encounter in the near future. Building new relationships takes valuable time and effort, but the rewards are great. We have a wonderful data community and we should promote open exchange of knowledge to other disciplines.

Social science data specialists are seeing the demand for our assistance increase exponentially. Our workflows are becoming increasingly complex with the introduction of innovative data formats and expanding data sizes. Today’s modern services and tools for managing the outputs from social science research are grounded in the early works of IASSIST members like Sue Dodd. Without these tools, we would not be equipped to handle our growing set of responsibilities as data stewards. New challenges for the social science data community evolve everyday. As we design services to address these needs, we should encourage new collaborations, encourage open exchange of knowledge, and build on past experiences. The social science data community has tremendous knowledge and experience in its ranks. We should share these with the world.

Acknowledgements

I first would like to thank Peggy Adams and Libbie Stephenson for asking me to reflect on some of the current projects that are building on the early work of Sue Dodd. It would be overwhelming to mention all of the exceptional colleagues I have worked with on these projects but I feel compelled to mention a few. An overwhelming thank you goes to the Odum Institute that has encouraged me to build these relationships. Working with the world class Odum Institute Archives and Information Science staff has been a great pleasure. Dr. Kenneth Bollen and Dr. Myron Gutmann allowed me to work with the fabulous partners in the early days of Data-PASS. I owe a tremendous debt to these two great researchers and dear friends. I also would like to thank the current Odum Institute director Dr. Thomas Carsey for his support and assistance as we continue to support archive development and data management services. A personal thank you also goes to our Deputy Director Peter Leousis and our current Data Archivist Thu-Mai Christian for all the support.

The projects I have described are the products of many great relationships and wonderful partner organizations. Without the collaborations within the Data-PASS partners few of these projects would have been possible. Many other great partners across the social science data community also deserve mention: Educopia, California Digital Library, LOCKSS, DICE, RENCI, UNC School of Information and Library Science, UNC Libraries, and the Australian Data Archive have been valuable partners in our efforts.

None of this would have been possible without great funders. Many thanks go to the Library of Congress, the Institute for Museums and Library Services, the National Science Foundation, the National Institutes for Health, and the Sloan Foundation. Finally, I wish to thank the wonderful IASSIST community. We have a great history and community on which to build.

References


Notes

1. Jonathan D. Crabtree is the Assistant Director for Information Technology and Archival Research for the Odum Institute for Research in Social Science at UNC Chapel Hill and can be reached at Jonathan_Crabtree@unc.edu.
The Evolution of Data Citation: From Principles to Implementation

by Micah Altman and Mercè Crosas

Abstract
Data citation is rapidly emerging as a key practice in support of data access, sharing, reuse, and of sound and reproducible scholarship. In this article we review the evolution of data citation standards and practices – to which Sue Dodd was an early contributor – and the core principles of data citation that have emerged through a collaborative synthesis. We then discuss an example of the current state of the practice, and identify the remaining implementation challenges.

Keywords: Data Citation; Bibliographic Practices.

Background
Data is, as they say, the new black. Scientific data are increasingly being made available online, and access to large collections of data is increasingly sought for education, science, policy, and commerce. Lowering barriers to discovery and use of these data and increasing our ability to link data with publications have the potential to enable new forms of scholarly publishing, promote interdisciplinary research, strengthen the linkage between policy and science, and lower the costs of replicating and extending previous research.

Many problems arise when research findings become disconnected from the underlying data that forms the evidence for these findings. The most well-publicized of these problems is scientific fraud. Access to data and the documentation of clear connections between the research results and the data facilitate detection of structural fraud both before and after publication. Other problems arising from this disconnect include irreproducibility, lack of reuse and wasted effort collecting new data, a proliferation of unmanaged versions and subsets of the ‘same’ data, and weak incentives for data sharing.

This is why the submission requirements for Science, one of the most cited, read, and respected journals in the sciences, requires that “all data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of Science” and that “citations to unpublished data and personal communications cannot be used to support claims in a published paper” (emphasis added). (Science 2014)

Too often, this proscription, and others like it, have been honored only in the breach. The history of data sharing makes this clear – despite clear recognition of the benefits of data sharing (Fienberg, et al. 1985) many research findings are based on data that is not made available – making this research surprisingly difficult to replicate and even more difficult to extend.

adoption of data citation and in the promotion of data sharing and its benefits.

Furthermore, most research articles fail to provide clear citations to data, or the code necessary to reproduce, reuse, or extend results (CODATA 2013).

Within the social sciences, the vast majority of datasets produced by sponsored research is never deposited or shared (Pienta 2006), and, as a result, reproducing published tables and figures, and directly extending prior results is often difficult or impossible (Dewald, et al., 1986; Altman, et al., 2003; Hamermesh 2007). Similar problems exist in other fields: A recent study by Vines et al. (2014) of a sample of zoology articles found that less than 30% of even the most recent publications made data available, and that research data availability declined rapidly with article age, while loss of data increased. Moreover, a study of articles published in high-impact journals during 2009 showed that only
41% minimally complied with the journal’s own data-sharing policies, and of these only 9% deposited the full primary raw data corresponding to the paper online (Alsheikh-Ali, et al 2011).

The research community has begun to take wider notice of this. And in the past two years a number of efforts have been launched by publishers, funders, professional associations, and organized projects to improve reliability, reproducibility, and data availability across a variety of scientific fields. We are optimistic that these projects will succeed, and if they do a key part of their success is likely to be through better scholarly recognition of data authorship.

There is increasing recognition that researchers are more inclined to share their data when they get credit (Borgman, 2012, p. 1072). Conversely, recent studies also suggest that researchers receive more credit when they share their data (Piwowar & Vision 2013). Publications that shared data from earlier years yielded an increase in citations of up to 30%.

Data citation, which has existed for 40 years in principle, is finally emerging as a pivotal norm for promoting data accessibility and accountability. Robust data citation practices and infrastructure will play a critical role in the widespread adoption of data citation and in the promotion of data sharing and its benefits.

**The Emergence of Data Citation Principles and Practices**

Within traditional print publishing, scholarly citation was widely formalized over a century ago. The first edition of the Chicago Manual of Style, published in 1906 under the title Manual of Style: Being a compilation of the typographical rules in force at the University of Chicago Press exemplified (and helped catalyze) the extent of standardization in scholarly citation. (Pollack, 2006) Within this tradition, a “bibliographic citation” referred to a formal, structured reference to another scholarly work that appeared in the text of a work. Typically, citations were either marked off with parentheses or brackets, such as: “(Altman 1992),” although in some fields footnotes were used. A standard reference entry included author(s), a title, a date, and a publisher (publishing house for books, journal name for articles) (Van Leunen 1992, pg. 186-208). In addition, citations could include “pinpointing” information that identified which part of the cited work was being referenced, typically in the form of a page range. Citations to a single work could be repeated throughout the text. The reference list, typically appearing at the end of the main text, provided more detailed bibliographic information for each work cited in the text. Many variations were used for references to archival sources, correspondence, government documents, and artworks. However, each of these reference formats provided as well as possible at least three elements: author/creator, dates of the work, and the publisher or distributor of the work.

When the first scientific digital data archives were established in the late 1960s, their design focused on issues of access, storage, formatting, costs, and information retrieval (Bisco 1965). Bibliographic standards for cataloging data were developed over the next decade. In 1970 the American Library Association (ALA) formed a subcommittee on Rules for Cataloging Machine-Readable Data Files (MRDF), and tasked it with, among other things,

---

**Figure 1:** A chronology of data citation principles and related systems

---

**Exemplar Systems**

- **ICPR Archive**
- **MARC catalog systems.**

**Core Principles**

- Facilitate description & information retrieval
- Describe data in archives
- Describe as works not media
- Provide author, title, version.

**Key Work**

- [Avram 1975]
- [Dodd 1979]
- [ISBD 1990]
- [ISO 1997]

---

**1999–2003**

- **NESSTAR Virtual Data Center**

**Core Principles**

- Facilitate access & persistence
- Cite research data in all publications that use it.
- Provide actionable URIs
- Use persistent institutions

**Key Work**

- [Altman, et al. 2001]
- [Nyseevik & Musgrave 2001]

---

**2004–2009**

- **TIB DOI Service Dataverse Network**

**Core Principles**

- Facilitate verification & reproducibility
- Provide bit- or semantic- fixity
- Provide granularity

**Key Work**

- [Brase 2004]
- [Buneman 2006]
- [Altman & King 2007]

---

**2009–**

- **DataCite**
- **Data Dryad**
- **FigShare**
- **Data Citation Index**

**Core Principles**

- Facilitate integration
- Include data citations in standard locations in text
- Index data citations in existing catalogs
- Integrate data citation with

**Key Work**

- [Uhlir (ed.) 2012]
- [CODATA 2013]
- [Data Synthesis Group 2014]
brining bibliographic control to MRDF. It was a long time before academic citation practices started to catch up with archiving practices, as summarized in Figure 1.

(In the figure, “Exemplar Systems” indicate key software and or technical infrastructure supporting practices. “Core Principles” summarizes the principles identified for data citation – as described in the text below. “Key Work” indicates the work related to principles of data citation and bibliographic practice – not responsibility for exemplar systems.)

The American Standard for Bibliographic Reference (ANSI Z29.29-1977, aka ASBR) provided a minimal “Data File” type to be used as part of the general material designator element in bibliographic metadata. Dodd (1979) quickly noted the shortcomings of ASBR in practice – notably the inconsistencies in describing the same dataset when presented in different physical formats, and the fact that the general approach conflated specific media with the “intellectual works” temporarily stored in those media. Dodd proposed using existing ASBR elements in a consistent and systematic way to bibliographically describe datasets as intellectual works. The key elements of Dodd’s approach emphasized the use of consistent Title, Author, and Edition (which included date). They were used along with a general media designator of “Machine Readable Data File(s)” (MRDF), which was format and media agnostic.

The recognition of data as a public good was, however, insufficient by itself to support or incentivize data sharing. In general, public goods in the absence of effective norms, regulation, or subsidies will be under-supplied. The state of the art in data citation, as well as in data sharing, did not progress quickly until catalyzed through advances in information technology, open source software development practices, and legal infrastructure. The growing recognition among scholars that data is a fundamental product of research, a trend identified in the National Research Council’s foundational report on data sharing (Fienberg 1985), began to build slowly in momentum through the leadership of individual scholars such as Sieber (1991) and King (1995). Then, rapid advances in Internet and web infrastructure greatly decreased the technical barriers to data sharing. More recently the rapid growth of the Open Software movement generally, together with development of the legal “technology” of robust standardized open licenses, have sparked initiatives in academia to build open tools in support of scholarly access, discovery, collaboration, and research sharing.

Building on these trends, and supported by the NSF Digital Library Initiative (Griffin 1998), Altman, King & Verba developed one of the first open source (and open access) data publishing systems, the Virtual Data Center (Altman et al. 2001). This system successfully fielded the largest federated catalog of social science datasets in the world (Altman et al. 2009). The virtual data center was designed to support persistent access to research data through federated institutional curation. Data citation was deeply integrated into the Virtual Data Center – each dataset managed was assigned a persistent identifier, and a citation. Moreover, the Virtual Data Center was based on the principle that all data supporting published research should be cited, and that these citations and identifiers should be machine-actionable through the web (e.g. through machine-actionable URIs). Nesstar, a system developed in parallel by Rysevsk & Musgrave (2001), and later used by many European archives, also incorporated the concepts of actionable web links, and persistent federated curation – although it did not initially support or emphasize citation.

Incorporating work by Altman, et al. (2003) and Altman and King (2007), the Virtual Data Center incorporated both support for “deep citations” (Buneman 2006) that identify precise subsets of a larger dataset; and for semantic fixity information that enables verification of a dataset using the citation itself. These capabilities were further extended in the Dataverse Network (King 2007), which succeeded the Virtual Data Center. The Dataverse Network has since been adopted by the Harvard University as its data publication infrastructure and is used by hundreds of researchers in dozens of institutions to curate and publish data. (Crosas 2011, 2013)

In parallel work, Brase (2004) lead an initiative to systematically archive datasets associated with research outputs, and to systematically associate these datasets with Digital Object Identifiers (DOI, 1997) – a robust form of persistent identifier used in the publication community. This was the first step toward integration of data citation and data publication into the larger publishing ecosystem.

To summarize, from 1977 through 2009 there were three phases of development in the area of data citation.

• The first phases of development focused on the role of citation to facilitate description and information retrieval. This phase introduced the principles that data in archives should be described as works rather than media, using author, title, and version.

• The second phase extended citations to support data access and persistence. Building upon the principle that research data used in publication should be cited, this phase introduced the principles that those citations should include persistent identifiers, and that the citations should be directly actionable on the web.

• The third phase of development focused on using citations for verification and reproducibility. Although verification and reproducibility had always been one of the motivations for data archiving – it had not been a focus of citation practice. This phase introduced the principles that citations should support verifiable linkage of data and published claims, and it started the trend towards wider integration with the publishing ecosystem.

The importance and urgency of scientific data management and access is now starting to be recognized broadly. Many publishers recognized this, in theory, in 2006, when the “Brussel’s Declaration” put forth the principle that data associated with publications should be openly available. This same year, the U.S. National Science Foundation introduced a policy requiring every grant proposal to be accompanied by a data management plan. Also that same year, data management was the theme of the annual meeting of the Society of Scholarly Publishers, the premier conference in that field. This continues a trend of funders and publishers adopting data publication and management policies. Universities have likewise become involved and have started to develop their own policies requiring data management, while journals, archives, and research libraries are increasingly grappling, largely independently, with the issues of data management.

Even the media has taken note. This is reflected by numerous articles drawing attention to particular high-profile cases of scientific fraud, such as the Stapel affair (e.g., Carey 2011), to increased rates of retractions (e.g. Ionaddis 2005, Steen 2010, Fang
et al 2012), and to the practice of Open Science more generally (e.g., Lin 2011).

The culmination of this trend, thus far, is an increasingly widespread consensus by researchers and funders of research that data is a fundamental product of research and therefore a citable product. The fourth and current phase of data development work focuses on integration with the scholarly research and publishing ecosystem. This includes integration of data citation in standardized ways within publication, catalogs, tool chains, and larger systems of attribution. It is exemplified by systems such as Data Dryad (Vision 2010) and Figshare (Hahnel 2013) which integrate data deposition into publisher workflows, and DataCite and the Thomson Reuters Data Citation Index, which integrate data citations into index and discovery of other published work, and by community standardizations efforts, such as those coordinated by the National Academies (Uhlir 2012), CODATA (2013), and the Data Citation Synthesis Group (2014).

Across these various groups there has been a developing agreement over the years that an essential part of connecting research publications or claims to data is formal data citation that includes a persistent link to guarantee long-term data accessibility. Global persistent identifiers, such as DOIs and Handles, offer a mechanism to provide a permanent link that can be configured to always resolve to a web page from which the data can be accessed, independent of whether the location of that page changes over time. An increasing number of data repositories generate DOIs which can be directly used in a publication to reference the data. However, until now, there has been not only a single set of principles or guidelines for data citations which represents and is in agreement with all these initiatives.

What has emerged in the bibliographic and research community is a substantial core of agreement over the need for citation to support attribution and verification; the recognition that citations must support both human and machine clients; the existence of robust persistent identifiers and the understanding of the core role; and the publication of key reference documents such as the National Academies and CODATA reports.

Converging Data Citation Principles

Given the rise of these parallel, variously implemented initiatives on data citation, as well as the lack of unified guidance for publishers, journal editors, and funding agencies, there was a need for a synthesis set of general recommendations and good practices for data citation. In the summer of 2013, a synthesis group was formed to unify the various recommendations. It came to be known as the Data Citation Synthesis Group. It met weekly from July to November of 2013 to thoroughly deconstruct previous data citation principles defined by CODATA, the Amsterdam Manifesto, and DataCite, and to produce a synthesis set that included the input of more than 25 organizations. During that time, the group met as part of the RDA (Research Data Alliance) conference in Washington, DC in September, in two half days of public workshop. As a result, in November 2013, the proposed Joint Declaration of Data Citation Principles was released to the public for open comment, and finalized at the end of February 2014 (Data Citation Synthesis Group, 2014).

The scope of the synthesis principles is solely to provide data citation recommendations, and does not intend to include detailed specifications for implementation or to focus on technologies or tools or research data repositories. The principles should extend to all disciplines and all types of data. Some of the challenges for specific types of data will be discussed in the next sections. As will be seen below, the Joint Declaration of Data Citation Principles reflect the various efforts described in the last section and a broad convergence on core principles:

1. **Importance.** Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.

2. **Credit and Attribution.** Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.

3. **Evidence.** In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.

4. **Unique Identification.** A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.

5. **Access.** Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.

6. **Persistence.** Unique identifiers, and metadata describing the data, and its disposition, should persist — even beyond the lifespan of the data they describe.

7. **Specificity and Verifiability.** Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.

8. **Interoperability and flexibility.** Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities.

At the time this article was completed, less than a month after the principles had been finalized, they had been officially endorsed by thirty organizations, including many major publishers and data archives. The synthesis group has also committed to a dissemination plan that includes reaching out to a large number of stakeholders from multiple organizations and disciplines for an endorsement of the principles.

We anticipate that the impact of the unified, widely broadcasted Joint Declaration of Data Citation Principles will be substantial and will: change current publication workflows, create new data citation technologies, define new metrics for scholarly impact and recognition, and, more importantly, provide persistent access to the data supporting scientific results to validate and extend previous scientific work. The Principles will facilitate interoperability across existing and new implementations, and will help guide enhancements and new versions of the current implementations. Several data repositories are already compliant, or close to compliant, with these principles (e.g., Dataverse, DataDryad). In section five, we describe, as an example, the Dataverse Network data citation implementation.

A Generic Example

A generic example for a data citation can be represented as:

1. Importance. Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.
2. Credit and Attribution. Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.
3. Evidence. In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.
4. Unique Identification. A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.
5. Access. Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.
6. Persistence. Unique identifiers, and metadata describing the data, and its disposition, should persist — even beyond the lifespan of the data they describe.
7. Specificity and Verifiability. Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.
8. Interoperability and flexibility. Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities.
7,641 astronomy publications from four main astronomy journals, The 2012). This wide array of authors is more common for data as in citations of literature, in some cases the creators are not the same importance as citations of other scholarly records, as stated the support principle two, providing credit and attributions to the scholarly product. We already find these authorship challenges in publications in high-energy physics, such as articles related to the observation of the Higgs Boson having nearly 3,000 authors (e.g., CMS Collaboration, 2012). This wide array of authors is more common for data products than for articles. The Principles and this citation example do not address the authorship problem, but, as described below, the metadata associated with the dataset can allow annotation of various levels of contribution during the creation and processing of the data, and also allows reference to related datasets or other scholarly products.

The year in which the dataset is first published and the title are not directly related to a principle. However, these elements are common in traditional literature citations, and such consistent and informative formats contribute toward giving data citation the same importance as citations of other scholarly records, as stated in principle one.

The Global Persistent Identifier is an essential piece of the citation of a digital object and directly supports principle four. The persistent identifier or URL allows separation of the link given in the citation with the URL to which it resolves, thus guaranteeing that even if the hosting or location of the dataset’s web page changes, the link in the citation will always go to the same dataset page. In a forthcoming article by Pepe, et al (2014), based on a study of 7,641 astronomy publications from four main astronomy journals, we show that 44% of the links in publications from ten years ago are broken. These are regular links to web sites, and not global persistent identifiers. The persistent identifier or URL solves a technical problem, but it is not sufficient without a publisher that supports and guarantees the validity of its persistent identifiers. In the case of data, the publisher is usually the data repository or archive. The more commonly used global persistent identifiers are handles (Sun, et al. 2003) and DOIs (Paskin, 2002). The persistent identifier in the data citation example also supports principles five and six. In support of principle five, the handle or DOI should resolve to a dataset page, which contains sufficient information describing the data and facilitating their reuse. In the rare cases in which the data cannot be made accessible any longer or must be destroyed, the data citation should still be valid. That is, the persistent identifier should resolve to a page with information about the discontinuation of that dataset (principle six). The last element in the generic citation example is the version, subset, or timestamp, which supports principle seven. This element is particularly relevant when citing data. Contrary to most literature publications, a dataset is often altered or expanded with time. The frequency with which a dataset might be changed can vary, from a static dataset that never changes once published, to a dataset that is updated once in a while with a new version, to datasets that are constantly changing, as is the case of dynamic data from meteorological sensors or streaming data Twitter feeds that grow constantly over time. Dynamic and streaming data offer a number of challenges for both citation and replication of published results contingent upon the reuse of a specific version of a dataset. Those challenges are described in section six.

The generic citation example might vary in style from community to community (principle eight), but across all cases it should be considered as important as other citations and should be part of either the standard reference section of a publication or a similar section for data citations, in accordance with principle one. The Data Citation Synthesis Group also recommends that when a published claim is made based on the data, enough information should be provided in the text to identify the data citation listed in the reference section, in the same fashion as other citations. When the published work makes a claim based on a subset of the data, specific information about the subset should be referenced by that claim.

Due to the possible complexity of such a citation, it is not always feasible to include in the reference section all the information needed to fulfill the core data citation principles. For this purpose, as stated in principle four, an important component of any data citation is machine-readable metadata that is bound to the data citation and persists with it. For example, the DataCite metadata schema and ontology (DataCite 2013) describe a detailed set of fields that may be used to complete a data citation. Typically, additional fixity and provenance information is required to support the verification requirements – such that future users of the citations can ensure that the data they use is identical to that cited. Such information might include bit-level fixity information (such as a MD5, SHA-256 or other cryptographic hash), or preferably, where available, semantic fixity information (such as a UNF or perceptual fingerprint).

Additional information on contributors will be required to fulfill the attribution requirements wherever the authors explicitly listed in the reference are ambiguous or incomplete. Unstructured metadata such as a contributors list may fulfill the bare legal requirements for attribution; however, structured name authority or identifiers such as ORCIDs (Open Research Identifier) or ISNIs (International Standard Name Identifiers) are much preferred, because they facilitate scholarly attribution (credit). This information can be embedded in published documents in machine-accessible form, included in the metadata stored with the DOI or other persistent identifier by its resolver service, or stored in an associated community index, such as CrossRef or DataCite. Such metadata should also be presented through the landing page provided to humans when the persistent identifier for the data is resolved.

Implementing the State of the Practice

Data repositories, or data publishers, are often responsible for implementing and generating data citations for the datasets hosted within them. As noted above, there are a number of repositories that are already generating data citations upon deposit of a dataset, and those citations are often compliant with the principles above (e.g., Dryad, Dataverse, Figshare). The generic data citation example in section four is based on the citation format generated by the Dataverse Network software application. This application is a data repository platform that allows organizations to host dataverses, where each dataverse
contains datasets, and where each dataset contains data files and metadata. A dataverse is, in essence, a virtual archive, which can be branded and administered individually, giving control to the data owner or distributor, while its data and metadata are stored by the repository in accordance to professional archival practices, metadata standards, and preservation formats (King 2007, Crosas 2011). The software is open-source and developed at the Institute for Quantitative Social Science at Harvard University (King 2014). The Harvard Dataverse is one of the Dataverse Network instances open to all researchers and to all data types. It supports a variety of types of dataverses, from journal dataverses, to dataverses for individual researchers, to dataverses for data associated with an institutional department (Crosas 2013). In this section, we describe the implementation of data citation as it is built in Dataverse version 4.0.

When a new dataset is added to a Dataverse, the required metadata fields that must be entered by the depositor include the author(s) or producer organization and the dataset title. In addition, an extensive set of metadata fields are provided, some required and others optional. The citation metadata supported by Dataverse maps closely to the DataCite metadata, and can also be mapped to the format developed by the Data Documentation Initiative (DDI, <http://www.ddialliance.org/Specification/> ) and Dublin Core Metadata Initiative Terms (DCTERMS, <http://dublincore.org/documents/dces/>). The dataset, when created, is in a draft form that is unpublished, and data files and additional metadata can be added at a later time. Upon dataset creation, however, even if the dataset is not yet published, a draft data citation is instantly generated following these steps:

1. Authors and title are obtained from the metadata fields entered by the data depositor. If instead of individual authors, a producer (organization or institution) is entered, the producer is used in place of the authors.
2. In the draft citation, the year is automatically populated by the year of deposit. At the time when the dataset is released, the final citation is updated with the year of the released or published date, which is often, but not always, the same year the dataset was deposited.
3. The Dataverse Network software supports both handles and DOIs as persistent identifiers. If a Dataverse Network is configured to use handles, each handle is registered to the Handle System. The Harvard Dataverse is configured to use DOIs, which are registered to DataCite through the EZID API (<http://ezid.dlib.org/home/documentation>). Upon deposit, the dataset is registered with status “reserved”, an option provided by the EZID API. When the dataset is released, the status becomes “public.” This means that the DOI at that point resolves to a public dataset page, which includes description information about the dataset, as well as information on how to access the data. Even when data cannot be completely open, and one or more data files in the dataset are restricted due to data user agreements or confidential information, the DOI resolves to a dataset page where access can be requested.
4. The publisher or data repository element in the citation is automatically populated as the repository name, in this case, the Harvard Dataverse. If additional distributors or archives are responsible for those data, they can be listed in the dataset page, as part of the additional metadata.
5. The Dataverse Network software supports versioning of datasets because, unlike traditional literature publications, data are often updated even after being published. The data citation generated by Dataverse includes the version of the dataset. When the dataset is released, the version in the citation is set to 1. If the dataset metadata or files are updated in the future, a new version is created, and a new citation, with the same DOI, but a new version number, is created. This allows reference to a specific previous version, and access to that version from the dataset page within a dataverse. It is important to note that a DOI or other persistent identifier is not equal to a data citation.

The data citation generated by the Dataverse Network software also supports Universal Numerical Fingerprints (UNF) for tabular datasets (Altman and King, 2007). The UNF guarantees fixity; it is a unique fingerprint on the semantics of a dataset. That is, even if a dataset changes format, if the data values remain the same, the UNF remains the same. When a UNF cannot be calculated, the Dataverse calculates bit-level fixity information (the MDS) of the data file(s) contained in the dataset.

The Dataverse Network implementation is fully compliant with the data citation principles discussed throughout this article. However, it does not support, in its current form, dynamic or streaming data. This is discussed in more detail in the next section.

Remaining Challenges
At the broadest conceptual level, the substantial remaining challenges for implementing robust data citation systems fall into three categories:

- **Challenges of provenance.** Provenance includes the chain of ownership of an object, and the history of transformations applied to it. Models of provenance have strong implications for how data citation is integrated into the data curation workflow.
- **Challenges of identity.** These theories involve defining data themselves, the identity of data and how to define equivalence and derivation relationships, and the granularity and structure of data. Theories of data have strong implications for determining what should be cited.
- **Challenges of attribution.** Attribution plays a key role in the incentives for citation. Models of attribution have strong implications for determining the presentation of data citations.

Provenance is a particularly important concern because many data citations are used to document a direct evidentiary relationship between a published assertion and the underlying evidence that supports it. However, supporting this evidentiary relationship does not require recreating or establishing the entire provenance chain – and much of provenance can be considered as orthogonal to citation, as Groth (2012) argues. Notwithstanding, as Smith (2012) points out, enabling readers to establish authenticity of the cited object is an important use for citation and requires that citation be connected to provenance information. The maintenance of this connection and of the associated provenance information is a major challenge for developing reliable citeable scientific workflows.

Identity is close to the heart of creating a citation. To cite something requires it to be identified – the citation should enable
Although there are no complete solutions to these problems, a number of practical questions emerge:

- **The equivalence question.** How does one determine whether two data objects, not bitwise identical, are semantically equivalent (interchangeable for scientific computation and analysis)?
- **The versioning question.** How does one unambiguously assign, at the time of citation, a ‘version’ to a data object, such that someone referencing the citation later can retrieve or recreate the data object in the same state that it was at the time of citation?
- **The granularity question.** How does one unambiguously describe components and/or subsets of a data object for purposes of computations, provenance, and attribution? How does one incorporate this granularity with a bibliographic data citation to create a “deep” citation?

Natural corollaries to these questions involve considerations of scalability. For example, how does one track and recreate versions of databases. Moreover, open annotation frameworks and ontologies are being developed to allow interoperable annotation of digital objects that define spatial (logical) and temporal granularity which might be used generally to complement bibliographic data citations and support deep citation (Van de Sompel, 2012).

A third challenge is that of attribution. Citation should support unambiguous attribution of credit for all contributors. As the scale of the data increases, and more people contribute to its creation and maintenance, practical challenges with attribution arise. These include supporting attribution for contributors that may number in the hundreds of thousands in crowd-based citizen science (e.g. Wiggins and Crowston 2011), distinguishing among different contributor roles (IWCSA 2012), and capturing the nature of the relationship between the cited and citing objects (e.g. Cronin 1984).

**Summary**

Scientific data are increasingly being made available online. Lowering barriers to discovery and use of these data, and increasing our ability to link data with publications have the potential to enable new forms of scholarly publishing, promote interdisciplinary research, strengthen the linkage between policy and science, and lower the costs of replicating and extending previous research. Robust data citation practices and infrastructure will play a critical role in achieving these outcomes.

Bibliographic standards for cataloging data developed gradually from the early days of data archives but it was a long time before academic citation practices started to catch up with archiving practices. Over four decades ago, however, several core principles for data citation and bibliographic description were recognized – in part based on the pioneering work of Sue Dodd. For the next 25 years, data citations had little attention from or impact on either the scientific or library community – despite the fundamental soundness of many of the early principles and the implementation of citation practices by selected major data repositories. More recently data citation principles and practices have made a resurgence – fuelled both by advances in web and network technologies and by a growing public and scientific recognition of the importance of scientific reproducibility, data sharing, and reuse. Recently, a wide convergence on principles has emerged, and the deployment of production infrastructure to support data citation across the research lifecycle is rapidly advancing.

Key enablers of a successful synthesis process have included a substantial core of agreement concerning the need for citation to support attribution and verification; the recognition of the need for citation to support both human and machine clients; the existence of robust persistent identifiers and the understanding of their core role; and the publication of key reference documents such as the National Academies and CODATA reports.

A number of central challenges remain, particularly related to the frontiers of data – big data, complexly structured data, dynamic data, and data in changing formats. These are being addressed gradually through groups such as RDA and through state-of-the-practice development of systems such as the Dataverse Network.

**Acknowledgments**

We would like to thank the members of the Data Citation Synthesis Task Group and of the Co-Data Data Citation Working Group for commentary on this paper and on the ideas leading into it: Amy Brand, Amylee Kenall, Andras Rauber, Anita deVlaard, Bonnie Carolí, Christinge Borgman, Dan Cohen, David Shotton, Eefke Smit, Elizabeth Arnaud, Elizabeth Iorns, Fiona Murphy, Franciél Linares, Giri Palanisami, Hannelore Vanhaverbeke Heige Sagen, Hylke Koens, Ivan Herman, Jan Brase, Jianhui Li, Jo McCrynty, Joan Starr, Joe Hourke, John Helly Maren Morgenroth, Kathleen Cass, Kerstin Lehner, Koji Zettus, Mark Hahnel, Mark Parsons, Martie van Deventer, Maryann Martone, Michael Diepenbroek, Michael Wit, Mustapha Mokrane, Natalia Moanola, Paul Groth, Paul Uhlir, Phil Archer, Purneet Kishor, Ruth Duerr, Sarah Callaghan, Simon Hodson, Stefan Proell, Stephanie Hagstom, Tim Clark, Tim Smith, Todd Carpenter, Vishwas Chavan, Yannis Ionnasis, Yasuhito Murayama

**References**


CODATA/ITSCl Task Force on Data Citation, (2013). “Out of cite, out of mind: The Current State of Practice, Policy and Technology for Data Citation.” Data Science Journal 12. 1-75. <http://dx.doi.org/10.2481/dsj.OsOM13-043>


Piwowar, H. and T. Vision (2013). “Data Reuse and the Open Data Citation Advantage” PeerJ.


Notes

1. Authors are listed alphabetically; the authors have made equal contributions to this work.

2. Or more precisely, in some cases it is a “club good” – nonconsumptive and only partially excludable.

3. Efforts in this area have been made by CODATA, as part of an extensive report on data citation (CODATA 2013), DataCite principles, DCC as part of the core guidelines on data curation, Harvard’s Institute for Quantitative Social Science through a data citation workshop hosted in 2012, and Force11 in the form of the Amsterdam Manifesto for Data Citations Principles, born at the Beyond the PDF 2 conference in 2013 (Crosas et al, 2013), among others, and by multiple research data repositories that offer to generate data citation upon deposit of a dataset (such as Dataverse, DataDryad, Figsshare, and the Inter-University Consortium for Political and Social Research (ICPSR)).

4. This section in part summarizes and updates section 7.2 in the CODATA report (2013), which was originally written by one of the authors of this article.

5. See Buneman (2006) for fundamental work in this area; also Proll and Rauber (2013) for a more recent approach.

6. Cronin (1984) reviews over 10 different proposed taxonomies of citation types and roles, some of which identify dozens of individual relationships.
A Practical Approach to Data Citation: The Special Interest Group on Data Citation and Development of the Quick Guide to Data Citation

by Hailey Mooney

Abstract

The Special Interest Group on Data Citation (SIGDC) carries on the work of IASSIST established in the early years of the organization by Sue Dodd by advocating for the standardized citation of datasets. A major accomplishment of the SIGDC is the development of the Quick Guide to Data Citation. This educational document simplifies the proliferation of data citation guidelines by presenting the five common elements of citation and offering examples in APA, MLA, and Chicago style formats. It can be used as a tool in education and advocacy to foster growth in data sharing behavior and research data management practice.

Keywords

Data citation, bibliographic references for data, data sharing, history of IASSIST

Introduction

Although many students find mastering citation formatting to be a tedious venture, as information professionals we recognize the significance of the citation. Librarians appreciate the predictable structure of bibliographic references for the ability to easily look up known items. Research faculty rely on citation counts for merit review. Scholars seek out the complex connections that make up the scholarly conversation. Data sharing advocates bemoan the failure to directly integrate datasets into the conversation and look to the citation as savior. If only we could get researchers to treat the contribution of re-usable data itself as equally valuable to the description and analysis of that data, then the cause of research data management and the curation of data would be furthered. IASSIST is an organization founded on the value of curation, preservation, and access to data. So it comes as no surprise that advancing the cause of data citation has been embraced by the membership since the beginning.

SIGDC authored the Quick Guide to Data Citation

As Margaret O’Neill Adams reports, the Classification Action Group was a strong component of the early IASSIST organization and Sue Dodd’s Working Manual for Cataloging Machine-Readable Data Files and later book, Cataloging Machine-Readable Data Files: An Interpretive Manual, were a crowning achievement for the organization (Adams, 2006). A clear intended
The impetus for the Quick Guide came from the reality that many data files, providing for identification and access via bibliographic references was an essential component of the Classification Action Group’s activities. Dodd published a concise set of guidelines for the citation of data in the *Journal of the American Society for Information Science* in her capacity as chairperson of the IASSIST U.S. Classification Action Group (Dodd, 1979).

In specifying the need for data citation, Dodd explains that data files are cited irregularly and that in many cases, “the information provided is not sufficient to indicate the proper source of the data and consequently an interested party has to spend a considerable amount of time determining additional information on the availability of a particular data file that could easily be provided by a full and proper bibliographic reference” (Dodd, 1979, p. 78). Over the years, the topic of data citation has continued to receive intermittent attention as part of the conversation within IASSIST, with a focus on the need to further the practice through the development of norms and guidelines (e.g., Altman & King, 2006; Drolet, 2005; Hankinson, 1988).

Unfortunately, despite the championing of Dodd and others the status quo of lackluster data citation behavior has yet to significantly change (Mooney & Newton, 2012). The recent attention within the profession on data management planning, including priming researchers to participate more broadly in data sharing, has placed renewed emphasis on the importance of citing data as first-class scholarly objects within the literature. Especially with the emergence of DataCite (<http://www.datacite.org>) and the participation of IASSISTers in the organization, IASSIST experienced a rekindling of attention to data citation. This coalesced into the formation of the Special Interest Group on Data Citation (SIGDC).

**SIGDC & the Quick Guide to Data Citation**

The SIGDC was started in late 2010 with a goal to promote awareness of data-related research and scholarship through data citation. To that end, SIGDC has sponsored several IASSIST conference sessions and posters. The group has sent advocacy letters to the major style guides imploring them to include instructions on best practices for the citation of data. Most notably, SIGDC authored the Quick Guide to Data Citation (International Association for Social Science Information Services and Technology, Special Interest Group on Data Citation, 2012), an instructional pamphlet that distills the main elements of a data citation from the key existing and proposed standards and provides an example citation in the three major style formats. The core team behind the development of the Quick Guide to Data Citation included Robert Downs, Michele Hayslett, Hailey Mooney, and Michael Witt. Elizabeth Moss, Mary Vardigan, and other SIGDC members also contributed to its final form.

In crafting the example citations within the Quick Guide, the authors drew from examples and standards from the scholarly literature (Altman & King, 2007; Dodd, 1979; Mooney & Newton, 2012), data producing organizations (e.g., GESIS, 2013; Green, 2009, International Polar Year Data and Information Service, 2008; Socioeconomic Data and Applications Center, n.d.; Statistics Canada, 2009; UK Data Archive, n.d.), and other stakeholder groups (Ball & Duke, 2011; DataCite Metadata Working Group, 2013).

So, although style guides and normative practices do not universally embrace data citation, a relative proliferation of various data citation suggestions exists across disciplinary areas. The goal of the Quick Guide then was to distill these guidelines to the essential elements and provide practical examples in common referencing styles. Distribution of the Quick Guide in printed brochure format at the Annual Conference provided IASSISTers with a means to easily provide data citation advocacy materials in orientation packets or other venues; the online version can also easily be printed out. Librarians are encouraged to use the Quick Guide as a basis for LibGuides that promote data citation best practices to their communities. It is a practice-orientated tool that simplifies what can be a complex endeavor for ease of use by researchers and scholars.

The final set of the minimal baseline data citation elements recommended by the Quick Guide includes:

- **Title**: Complete title of the dataset, including the edition or version number. Ultimately, the desire for simplicity won out in keeping to a set of just five citation elements. Edition and version are discussed within the Title element and data retrieved is part of the definition for Electronic Location or Identifier. These were seen to be closely tied with the other elements in aiding for precision of recall.

- **Author**: Name(s) of each individual or organizational entity responsible for the creation of the dataset.
- **Date of Publication**: Year the dataset was published or disseminated.
- **Title**: Complete title of the dataset, including the edition or version number, if applicable.
- **Publisher and/or Distributor**: Organizational entity that makes the dataset available by archiving, producing, publishing, and/or distributing the dataset.
- **Electronic Location or Identifier**: Web address or unique, persistent, global identifier used to locate the dataset (such as a DOI). Append the date retrieved if the title and locator are not specific to the exact instance of the data you used.

Areas of debate during the development of these citation elements included the issue of whether edition/version and date retrieved should be elevated to a separate element. This information, especially the date retrieved, could be critical given the need to be precise about the exact dataset used in an analysis and the potential for some datasets to have multiple updates over time that may or may not be reflected in the assignment of a new edition or version number. Ultimately, the desire for simplicity won out in keeping to a set of just five citation elements. Edition and version are discussed within the Title element and data retrieved is part of the definition for Electronic Location or Identifier. These were seen to be closely tied with the other elements in aiding for precision of recall.

It should be noted that Dodd’s conception of the “imprint” (Dodd, 1979), or producer and distributor statement, was a major influence in the conception of this citation component. This accounts for the somewhat unique characteristics of research datasets in that they may originally be produced and published at one institution, but trusted to another institution (the data archive) for curation and dissemination. While it is important to acknowledge the original producer, for purposes of retrieval the dataset distributor is a crucial piece of information.

In crafting the example citations within the Quick Guide, the General Social Survey was chosen as an emblematic social science dataset. Examples were provided using the general guidelines for
formatting from the “big three” style guides in regular academic use: APA, MLA, and Chicago.

- **APA (6th edition)**  

- **MLA (7th edition)**  

- **Chicago (16th edition) (author-date)**  

Since the time the Quick Guide was published in 2012, the data citation landscape has continued to evolve as the conversation endures around required citation elements, best practices, and author behavior (e.g., CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013; FORCE11, Data citation Synthesis Group, 2013). Educating scholars on the benefits of data citation is crucial in the move to support funder-mandated data sharing practices within the profession is a core part of the IASSIST mission. As a practical synthesis of many voices, the Quick Guide continues to stand as a valuable resource to information professionals, scholars, and researchers looking for a simple and direct way to cite data.

**References**


**Appendix**

Quick Guide to Data Citation (Pages 74-77)

**Notes**

1 Hailey Mooney is the Data Services Coordinator and Social Sciences Librarian at Michigan State University Libraries. She can be reached by email: mooneyh@msu.edu.
Appendix
Quick Guide to Data Citation

ABOUT IASSIST

IASSIST is an international organization of professionals working with information technology and data services to support research and teaching in the social sciences. Its 300 members work in a variety of settings, including data archives, statistical agencies, research centers, libraries, academic institutions, government departments, and non-profit organizations.

If you have questions about citing datasets, consider consulting with your local librarian or writing lab.

CITING DATA

Citing datasets used in published research is just as important as citing journal articles, books, and other sources that contributed to the research.

By citing your use of a dataset, you are supporting the reproducibility of your research and attributing credit to those who provided the data—including datasets that you have created yourself. Citations also allow for tracking reuse and measuring impact.
Quick Guide to Data Citation

identify • retrieve • attribute

Instructions for citation styles do not consistently provide examples for dataset citations. This guide will help you determine the citation elements to include. Refer to your author guidelines or style guide to properly arrange and format these citation elements. Many data providers also recommend their preferred citation or supply an example.

Be sure to provide enough information in your citation so that the reader can identify, retrieve, and access the same unique dataset you have used.
ELEMENTS OF DATA C ONTENT

- **Author:** Name(s) of each individual or organizational entity responsible for the creation of the dataset.

- **Date of Publication:** Year the dataset was published or disseminated.

- **Title:** Complete title of the dataset, including the edition or version number, if applicable.

- **Publisher and/or Distributor:** Organizational entity that makes the dataset available by archiving, producing, publishing, and/or distributing the dataset.

---

FOR EXAMPLE

Arrange these elements following the order and punctuation specified by your style guide. If examples for datasets are not provided, the format for books is generally considered a generic format that can be modified for other source types.

**APA (6th edition)**

Citation

- Electronic Location or Identifier: Web address or unique, persistent, global identifier used to locate the dataset (such as a DOI). Append the date retrieved if the title and locator are not specific to the exact instance of the data you used.

These are the minimum elements required for dataset identification and retrieval. Fewer or additional elements may be requested by author guidelines or style manuals. Be sure to include as many elements as needed to precisely identify the dataset you have used.

MLA (7th edition)

Chicago (16th edition) (author-date)

Modified for the IASSIST Quarterly. Original is found at:
http://www.iassistdata.org/sites/default/files/quick_guide_to_data_citation_high-res_printer-ready.pdf
Toronto Welcomes You To The 40th Anniversary Conference of IASSIST
The International Association for Social Science Information Services & Technology

June 3 - 6, 2014

The “official” hashtags are #iassist2014 #iassist40 — spread the word!
The International Association for Social Science Information Service and Technology (IASSIST) is an international association of individuals who are engaged in the acquisition, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, database librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompasses hard copy as well as machine readable data.

Paid-up members enjoy voting rights benefit from reduced fees for attendance at regional and international conferences sponsored by IASSIST. Join today by filling in our online application:

http://www.iaassistdata.info/

Online Application

IASSIST Member ($50.00 (USD))
Subscription period: 1 year, on: July 1st
Automatic renewal: no

Please fill in the information our Online Form

The application is in USD, however, we do accept Canadian Dollars, Euro, and British Pounds as well.

The membership rates in all currencies as well as the Regional Treasurers who manage them are listed on the Treasurers page.