Building the DDI

by Ann Green and Chuck Humphrey[1]

Abstract
This paper describes the context, motivation, and requirements behind the design and development of the first version of the Data Documentation Initiative[2] (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.

From cataloging to metadata and shifts in research infrastructure

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[3]. 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

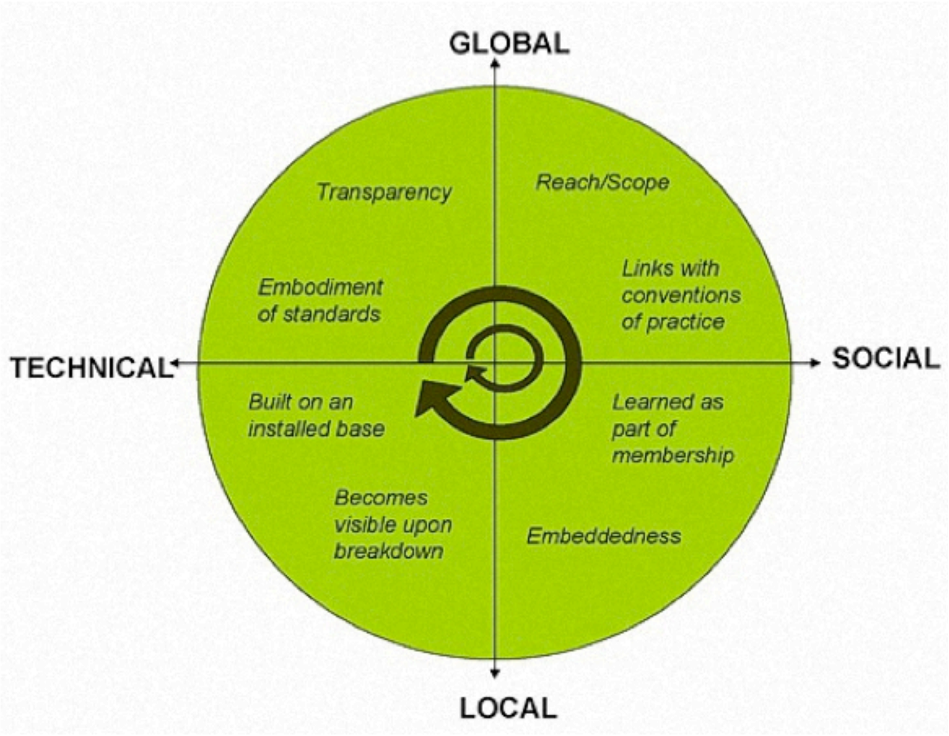
GLOBAL
Transparency
Reach/Scope
Embodiment of standards
Links with conventions of practice
TECHNICAL
SOCIAL
Built on an installed base
Learned as part of membership
Becomes visible upon breakdown
Embeddedness
LOCAL

Figure 1: A Distribution of Cyberinfrastructure Solutions

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 committee[5] 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 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

5. Review other encoding guidelines in development at the same time, and Electronic header standards

1 Review cataloguing and citation guidelines

The first step was to review the standards that establish rules for producing cataloguing records for the study in an online public catalogue or online retrieval system, with an emphasis upon intellectual ownership and identification. This information serves as the basis of a standard bibliographic citation.

The cataloguing of machine-readable computer files was well established by the early 1980's. The *Anglo-American Cataloguing Rules*, Second edition (*AACR2*) was published in 1978 and included a new chapter for machine-readable data files (chapter 9).[7] The Library of Congress, the National Library of Canada, the British Library, and the Australian National Library adopted AACR2 in January 1981. ALA published an interpretive manual by Sue Dodd focusing on machine-readable data files in 1982. This was followed by a manual for cataloguing microcomputer files in 1985 by Dodd and Ann Sandburg-Fox. Prior to these works, Dodd had already published on cataloguing standards in the *ASIS Journal* (Dodd, 1979a) and the *Journal of Library Automation* (Dodd, 1979b). Her manual on *Cataloging Machine Readable Data Files* was one of the founding documents for the bibliographic components of the DDI standard and other citation standards to come.

The integration of bibliographic citations within data documentation is not new. In March, 1996, as part of the review of the proposed DDI elements, Sue Dodd encouraged via email[9] that there be "a better distinction between required bibliographic information denoting intellectual identification (aka citation) and data abstract information (aka study description)." Her recommendation was to "disconnect them (conceptually)." This was an important distinction, which clarified the necessity for the DDI instance to carry within it a distinct reference to the intellectual identity of its creators, producers, and distributors.

From the beginning it was clear that *citations to data are an essential aggregation of descriptive elements best compiled into a standard format*. An element was added to the DTD called "Bibliographic Citation" so that a complete citation could be carried within the documentation instance.

It was also important that the DDI include elements that could be compiled for multiple purposes (for example, compiling a citation in an alternate format, forming a title page if the codebook was printed or rendered as a separate document, or mapping intellectual identification to other metadata standards, e.g. MARC or Dublin Core). This flexibility was illuminated by Dodd's insight into the relationships among instances of documentation, and the necessity for the ability to carry the requisite information to create various output formats from a single DDI instance.

Dodd also noted "You might want to include information on the difference between citations for the documentation and citations for the computer file – provided they are separate (and not "packaged" together)." This was one of the key challenges of creating the DTD – to clearly define all of the objects and intellectual components being described by the DDI instance. 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 distributer 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 comprehensive codebook. The focus was on capturing the **intellectual content** of the pieces rather than their variant names/labels.

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.
- Ruus, Laine. University of Toronto. "**A comparison of major descriptive systems in use to describe computer-readable data files**," 2nd edition, September 1992. "The objective of this document is to track major systems currently in use for the description of computer-readable data files. Identifying the comparable fields in the descriptive systems, as well as the difference, will serve to support recommendations…to satisfy national and international requirements for the formal description of computer-readable data files" (p. 2) Each field (element) was described in numeric MARC tag number order, indicating if the particular field was mandatory, optional, and repeatable. That list is followed by definitions of each field, from each source. The compilation represents a very time consuming and precise effort with international scope. The comparison included the following: the Canadian union list of machine readable data files at the University of Alberta; Statistics Canada catalog; Cataloging computer files in the UK: a practical guide to standards, edited by Peter Burnhill and Ray Templeton; The Treasury Board of Ottawa's guide to structure data model data dictionary; NASA's directory interchange format manual; ISBD international standard bibliographic description for computer files; RLG's Machine-readable data files memory aid; Health and Welfare Canada's Microdata set documentation: reference guide; Danish Data Archives' Study Description completion guide; ICPSR's Template Manual; and the US Library of Congress USMARC concise formats for bibliographic, authority, and holdings data.

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.

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 being established in the 1980s and 1990s used this manual. The style manual was influential in the development of standardized documentation of data files.
- Geda, Carolyn: 1980, *ICPSR, Data Preparation Manual.*
- Collins, Patrick, 1996, *Depositing Data With the National Data Archive on Child Abuse and Neglect: A Handbook for Investigators.*
- US Bureau of the Census, *Statistical Research Division, Statistical Design and Methods Extension to Cultural and Demographic Data Metadata: CDDM draft standard* 1995. An extremely detailed table of contents and definitions of proposed documentation of census or surveys.
- 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.

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[9] for encoding digital texts and the EAD[10] 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[11] 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 describe 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

American Library Association. (1978). *Anglo-American cataloguing rules second edition: chapter 9: computer files: draft revision, Volume 9, Michael Gorman,* Joint Steering Committee for Revision of AACR, American Library Association.

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>

Data Documentation Initiative (DDI). (2004). DDI Version 3.0 Conceptual Model Structural Reform Group. Background of the Conceptual Model. Working Draft. Retrieved from <www.pop.umn.edu/~wlt/arofan/ConceptModel-Review-0610.doc>

Dodd, S. A. (1979a) "Bibliographic Reference for Numeric Social Science Data Files: Suggested Guidelines," American Society for Information Science Journal 30:2, p. 77-82.

Dodd, S. A. (1979b) "Building an On-Line Bibliographic/MARC Resource Data Base for Machine-Readable Data Files," Journal of Library Automation 12:1, 6-21.

Dodd, S. A. (1982). Cataloging Machine-Readable Data Files: An Interpretive Manual. Chicago: American Library Association.

Dodd, S. A. & Sandberg-Fox, A. M. (1985). Cataloging Microcomputer files: A Manual of Interpretation for AACR2. Chicago: American Library Association.

Edwards, P. Jackson, S., Bowker, G., & Knoble, C. (2007). "Understanding Infrastructure: Dynamics, Tension, and Design." Report of a Workshop on History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures. Retrieved from <http://deepblue.lib.umich.edu/bitstream/2027.42/49353/3/UnderstandingInfrastructure2007.pdf>

Green, A. (1999). "Why the Data Documentation Initiative is so Important." Presentation to the Association of Public Data Users. October, 1999.

Green, A. & Kent, J-P. (2002) "The Metadata Life Cycle." In: Kent, J-P., ed. MetaNet: A Network of Excellence for Harmonising and synthesizing the development of statistical metadata. MetaNet Work Package 1: Methodology and Tools. The MetaNet Project. P. 29-34. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.203.5184&rep=rep1&type=pdf>

Humphrey, C. (2006). "e-Science and the Life Cycle of Research." Retrieved from <http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc>

Mooney, H, Newton, M.P. (2012). "The Anatomy of a Data Citation: Discovery, Reuse, and Credit." Journal of Librarianship and Scholarly Communication 1(1):eP1035. <http://dx.doi.org/10.7710/2162-3309.1035>

National Research Council. (2012). For Attribution -- Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop. Washington, DC: The National Academies Press.

Nielsen, J. (1998). "From OSIRIS to XML: Markup and Internet Presentation of Structured Data Documentation." Unpublished thesis.

Rasmussen, K. B. (1995). "Documentation - what we have and what we want: Report of an "enquete" of data archives and their staff." IASSIST Quarterly 19(1). Retrieved from <http://www.iassistdata.org/downloads/iqvol191rasmussen.pdf>

Rasmussen, K. B. (2013) "Social Science Metadata and the Foundations of the DDI." IASSIST Quarterly 37.

Rattenbury, J. & Pelletier, P. (1974). Data processing in the social sciences with OSIRIS. Ann Arbor : Survey Research Center, Institute for Social Research, University of Michigan. Retrieved from <http://babel.hathitrust.org/cgi/pt?id=mdp.39015008207063;view=1up;seq=7>

Robbin, A. (1981) "Strategies for improving utilization of computerized statistical data by the social scientific community." Social Science Information Studies 1, p 89-109.

Roistacher, R. C. with contributions from Dodd, S.A., Noble, B. B., & Robbin, A. (1980). A style manual for machine-readable data files and their documentation. Washington, D.C. U.S. Dept. of Justice, Bureau of Justice Statistics. Grant no. 78-SS-AX-0028 awarded to the Bureau of Social Science Research. Report no. SD-T-3. NCJ-62766.

Ryssevik, J. (nd) "The Data Documentation Initiative (DDI) metadata specification." Retrieved from <http://www.ddialliance.org/sites/default/files/ryssevik_0.pdf>

Tenopir, C. A., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). "Data sharing by scientists: Practices and perceptions." PLoS ONE 6(6). doi:10.1371/journal.pone.0021101 Retrieved from <http://www.plosone.org/article/info:doi/10.1371/journal.pone.0021101>

Vardigan, M. (2013) Vardigan, Mary (2013) "The DDI Matures: 1997 to the Present. " IASSIST Quarterly 37.

Vardigan, M. (2013) "Timeline." IASSIST Quarterly 37.

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.
2. Data Documentation Initiative. <http://www.ddialliance.org>
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
6. Memorandum to Ann Green, dated February 14, 1995 from Richard Rockwell, cosigned by J. Merrill Shanks, Peter Granda, and Mary Vardigan. Subject: Invitation to serve on ICPSR committee.
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.
9. For information about the history of the TEI see: <http://www.cs.vassar.edu/~ide/papers/teiHistory.pdf>
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/>

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