Maximizing the Search Potential of Social Science Codebooks Through the Application of the Codebook DTD

by Wendy Treadwell *

Data libraries and archives have been working with digitized materials longer than most general libraries and archives. However, we have been slow to develop a means of making our collections searchable in an electronic manner. This has been due primarily to the fact that our metadata (codebooks and data dictionaries) have not been available in a machine processable format. The complexity of the material and the need to be able to identify specific structural elements and their contents made even well enhanced bibliographic records inadequate to the task.

Researchers seeking data for secondary analysis have a distinct set of needs. They need the ability to:

- **Search across multiple collections in multiple locations.** With bibliographic records they are able to achieve this at the level of basic information, but cannot do so consistently or at the level of information needed.

- **Search heterogeneous collections**. In other words, they do not necessarily wish to search one system for data and another for related materials.

- **Drill down into individual collections and documents for more detailed information (in particular, detailed information regarding the variables in the data set).** The importance of being able to search at the variable and variable response category level is made clear in the following example. A great number of data sets, particularly those aggregated to small geographic levels, use age cohorts. Data published prior to 1980 frequently used upper age cohorts of '65 years and over'. This practice made them unsuitable for researchers examining the relationship between age and socioeconomic factors within the over-65 population. This piece of information was available only by looking at the response categories for the variable age. When most codebooks were only available in hard-copy, researchers would lose valuable time obtaining codebooks and data only to find that the data set was unusable for their purposes.

- **Search both the metadata and the object**. In the case of text documents this means the full-text of the document as well as its bibliographic or other metadata material. In terms of data this means examining both the data file documentation and the data itself.

- **Obtain or manipulate the file contents**.

The goal of the Data Documentation Initiative (DDI) group was to address the needs listed above. They needed to develop a machine readable and machine processable codebook which would fulfill both archival requirements and serve as a source for inquiry. The XML tagged codebook developed by the DDI addresses each of the issues noted above.

Searching across multiple collections becomes possible using a uniform configuration for the codebook. Creating centralized depositories for codebooks or search engines that can search multiple locations now become options.

By using XML tags, the DDI has adopted a tagging scheme commonly used in text documents of various types. Systems which can parse an XML DTD can search through often familiar layers of information. Many attributes of higher level metadata were retained, such as descriptive bibliographic fields. These were then mapped to commonly used schema like the Dublin Core. This makes searching across types of material more efficient. By providing links between the DDI tagged document and related materials, the codebook can also become a central hub through which other materials are identified and obtained.

Of course, the most important feature of the DDI DTD is that it identifies specific structural elements and their attributes. This allows the searcher to drill down into individual collections and documents for more detailed information. The extent of the tagging provided makes it possible to create specialized search engines which can address the eccentricities of both the researcher and the materials being searched.

The merger of the data and the metadata of the document (codebook) into a single unit results in the entire document becoming a resource for discovery. Researchers are no longer as dependent upon the descriptive skills of the cataloger or archivist to capture the concepts important to the individual researcher in a controlled language. External tools can become the driving force for relating past terminology with future terminology and past conceptual structures with future use and perspective.

Finally, the DDI tagged codebook provides all the information needed to create systems to obtain and/or manipulate data file contents. The identification of elements and attributes in a structured tag provides for both machine understanding and processing. This is a feature that has been absent from many earlier attempts to make codebooks machine readable.

The availability of tools such as the Generalized Record Structure 2 (GRS2) within Z39.50 protocol make DDI tagged codebooks potentially accessible through the same tools used for searching other tagged documents with DTD's . The GRS2 is designed to pass information regarding structure of materials using DTD's and structured tags within Z39.50 compliant systems. It provides the ability to map information such as a query from one set of tags to another. For example, the DDI DTD includes mapping information to Dublin Core elements (fig 1).

The GRS2 would be used by one system to inform another system that it was using the DDI DTD instead of the Dublin Core DTD and that information contained in a specific Dublin Core element should be dumped into the search parameter for the following DDI DTD element.

In addition, the parent, sibling and child nodes of the identified element could be obtained and transferred along with the contents of the element based on the hierarchical information available through the DTD. This would allow for the transfer of variable information with

Fiqure 1

DC ELEMENT	DDI Codebook Element
Title	1.1.1.1 titl (Title of Documentation)
Creator	1.1.2.1 AuthEnty (Authoring Entity)
Subject	2.2.1.1 keyword (Keywords) 2.2.1.2 topcClas (Topic Classification)
Description	2.2.2 abstract (Abstract)
Publisher	1.1.3.1 producer (Producer) [NOTE: The Dublin Core specifies that the publisher should be "the entity responsible for making the resource available *in its present form*" (emphasis added). For a DDI codebook the publisher should be the entity responsible for making the *electronic* version available.
Contributor	1.1.3.2 othId (Other Ident. & Acknowl.)
Date	1.1.3.3 prodDate (Date of Production) [NOTE: Theoretically, the DC Date element should refer to the date the electronic resource (e.g., the DDI version of the codebook) was created, not any preceding paper version.]
Type	DOES NOT MAP TO ANY DDI CODEBOOK ELEMENT Suggested DC Type: "Text.x-Codebook"
Format	DOES NOT MAP TO ANY DDI CODEBOOK ELEMENT Suggested DC Format: "text/xml" [NOTE: use of MIME type text/xml based on Internet Draft by E.J. Whitehead, Jr. of U.C. Irvine, and M. Murata, of Fuji Xerox Info. Systems.]
Identifier	Suggested DC Identifier: URN for DDI Codebook, if applicable. Alternatively, use the IDNo element within the Document Description citation element.
Source	[NOTE: If a DDI electronic codebook has been produced as the *original* documentation for the data from a study, the DC source element does not apply. If the DDI electronic codebook has been derived from a pre-existing version, then the DC Source refers to bibliographic information regarding this previous paper version. In this case, Source would map to the MARCURI on the docSrc element, or alternatively, to the IDNo element within the docSrc element. [NOTE: Use of the DC Source element is deprecated. The DC Relation element is now preferred.]
Language	xml:lang attribute for codeBook element
Relation	partially maps to 1.4 docSrc (Documentation Source). No mapping currently exists for the relation type component.
Coverage	2.2.3.1 timePrd (Time Period Covered) 2.2.3.2 collDate (Date of Collection) 2.2.3.3 nation (Country) 2.2.3.4 geogCover (Geographic Coverage) 2.2.3.7 universe (Universe)
Rights	1.1.3.2 copyright (Copyright)

Dublin Core to DDI DTD mapping suggestions created 7/1/98 by Jerome McDonough, U.C. Berkeley Library Systems Office.

datafile, question and location information attached as a structured unit of information.

```
<dataDscr ID=da8425>
<var ID='V25' name='empl'>
        <location StartPos='45' EndPos='45' width='1'/>
        <labl>Employment Status</labl>
        <qstn ID=Q20>What is the current employment
        status of this person?</qstn>

DATA FILE: da8425

Variable:      Start   End    Width

V25    empl    45      45     1

Employment Status

What is the current employment status of this person?
```

All of these features provide opportunities to develop a range of tools without creating a specialized or unique infrastructure of information. The potential benefits to the researcher are enormous. Possible system features could include:

- Multiple search systems addressing different levels of searches
- The ability to pass information from one level of search to another
- Multiple templates for displaying search results
- The ability to move from the metadata to data manipulation and/or display
- The ability to follow independent tangents of searching through internal links to related materials

The similarity between the DDI DTD and the DTDs of other format types allows for a certain level of cross searching between heterogenous document types. Specific search engines could be developed which exploited this upper level metadata, making it possible for the researcher to cast a wide net for related information. This could be an upper level of a tiered search approach.

The development of special search engines within specific collections or object types would allow libraries and archives to exploit the unique features of their holdings. Special relationships between collection pieces and unique terminology could be featured. Tools such as a dynamic thesaurus or customized dictionaries could be incorporated. Because both the generalized tool and the specialized tool are addressing the same underlying collection multiple search engines which exploit particular research approaches could become common. We would no longer be limited to trying to create one tool that works for everyone.

An excellent example of this potential is found in the following three search systems: NESSTAR, ILSES, and GESINE. All address, or eventually intend to address, data collections held at the Zentralarchiv für Empirische Sozialforschung an der Universität zu Köln (ZA) in Köln, Germany. NESSTAR[1] is a search engine, combined with a data manipulation (basic statistics) and extraction tool. It accesses data held at various archives whose metadata has been tagged to the DDI standard. Using structured metadata, NESSTAR allows the user to identify appropriate data sets by querying the variable descriptions, questions, and study description material. It provides options for running real-time calculations on selected variables to further determine applicability and then allows for data extraction according to the access rules of the governing archive. NESSTAR makes searching across archives, exploring data sets and obtaining data on-line a one-stop operation.

ILSES[2] addresses the collection of data and related materials at ZA. ILSES currently does not address the collection through the DDI compliant metadata. There are plans to use this approach in the future. DDI compliant metadata will be accessed directly by the search system or it will serve as a transport format for entering new materials into the system and exporting information from the system to the end-user. ILSES provides access to related literature as well as the data sets and metadata files. The user can approach the collection from either direction. The user has the option of downloading complete data files or customized extracts within the limits of the archives access restrictions. The focus of this system is narrower than NESSTAR in that it addresses only a single collection of data. However, providing the context of related materials and publications provides a better conceptual appreciation of the unique features of ZA's complete collection of materials.

GESINE[3] is not a data extraction engine nor does it currently address data collections. GESINE provides access to the collection of social science information found at IZ which, like ZA, is part of GESIS (Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen). It currently addresses descriptive information housed in an ORACLE database and performs full-text searches on the documents in their database. There are plans to include ZA study descriptions in this database. This would position them to include options for fully searching DDI compliant metadata files. The value of including full-search capabilities for ZA metadata files would be great. Currently the two other systems, NESSTAR and ILSES, take a data user's approach to the data discovery process. Linkages to related literature within ILSES move from the data collection out to works that are based on the analysis of the

data or related to its collection. Expanding the search capabilities of GESINE to include DDI compliant metadata files would allow the user to search for data within the broader context of social science research. Two of the specialized tools within GESINE, the person/institute search and the graphical search and display system, would provide major enhancements unavailable elsewhere. Access to ZA data collection through GESINE would bring these data sets to the attention of a wider audience, those not aware of the separate systems available data searching. If all of these systems were capable of accessing the DDI format, the user may be able to switch systems, without reentering the search parameters in the new system. For example, the ability to switch systems would allow GESINE users to extract data found initially through GESINE through the ILSES or NESSTAR systems. ILSES or NESSTAR users would also be able to expand their search to a broader range of related materials through GESINE.

The use of a standard underlying structure provides the option for integrating multiple search approaches. A researcher would begin his or her search with a general search engine. Later, as a subset of material or a specific collection was identified, the researcher could switch to a more specific search engine that exploited the features of a certain type of material, area of study, or research approach. All researchers should have the option of choosing the search engine that he or she prefers and that most closely matches their own approach to inquiry.

The ability of GRS2 to pass search parameters between systems means the researcher would not have to be limited to using a single search engine during their inquiry. They should be able to move search parameters between systems which can map from one structure to the other.

Reaping the full benefits of the DDI DTD requires adherence to a set of both design and application principles. First, a level consistency in the development of DTDs across heterogeneous document types must be maintained. This is particularly important for the upper level metadata that would be searched in broad cross collection systems. Second, there needs to be some level of structured language developed and maintained within similar document types or disciplines to identify implied information. Third, there needs to be consistent application of the DTD and tagging nodes within the data community. Finally, we must create the tagged codebooks in the DDI DTD format. Without them, there is nothing to warrant the development of specialized search engines and the ability to address these documents in generalized search systems. This means that producers in the data community need to commit to the DTD and produce documentation in this format. This does not preclude production in other formats, but commits the producer to providing a DDI DTD tagged codebook as one of its format options. Data librarians and archivists must also find a means of translating their existing collections of legacy documents into the new format. Given the variety of documentation in terms of format, layout and quality, this is a massive undertaking. It should be viewed as a means of preserving not only the codebook information, but of preserving and in many cases creating access to the data.

[1] NESSTAR (Networked Social Science Tools and Resources) Developed by the Norwegian Social Science Data Service, the Data Archive at the University of Essex, and Danish Data Archives http://www.nesstar.org

[2] ILSES (Integrated Library and Survey-data Extraction Service) A product of the ZentralArchive and NIWI.

[3] GESINE (Integriertes socialwissenschaftliches Informationssystem) A product of the Informationszentrum Sozialwissenschaften (IZ), Bonn, Germany, http://www.bonn.iz-soz.de

* Wendy Treadwell, Coordinator, Machine Readable Data Center, University of Minnesota, 2 Wilson Library 309 19th Avenue South, Minneapolis, MN 55455. 612-624-4389, wendy@mrdc.lib.umn.edu