# Expert Committee Meeting

Saturday, May 29, 2004
Madison, Wisconsin

**Present:**
Hans Jorgen Marker (Danish Data Services), Vice Chair; Atle Alvheim (Norwegian Social Science Data Services [NSD]); Pat Doyle (U.S. Census Bureau, Demographic Surveys Division) - by telephone; Ilona Einowski (University of California, Berkeley, UCDATA Archive); Arofan Gregory (AEON Consulting); Reto Hadorn (Swiss Data Archive, SIDOS); Carol Haney (SPSS Inc.); Pascal Heus (World Bank); James Jacobs (University of California, San Diego); Ryan Johnson (Washington State University); Mari Kleemola (Finnish Data Services); Julie Linden (Yale University, Social Science Libraries & Information Services); Marc G. Maynard (University of Connecticut, Roper Center); Meinhard Moschner (Zentralarchiv fuer Empirische Sozialforschung-ZA); Ron Nakao (Stanford University); Rob O'Reilly (Emory University); Jostein Ryssevik (Nesstar Ltd.); Janet M. Eisenhauer Smith (University of Wisconsin, Madison); Ken Miller (UK Data Archive); Wendy L. Thomas (University of Minnesota, Minnesota Population Center); Mary Vardigan (ICPSR); Joachim Wackerow (Zentrum fuer Umfragen, Methoden und Analysen-ZUMA).

**Also attending as observers:**
Bill Block (University of Minnesota); Dan Gillman (U.S. Bureau of Labor Statistics); Ann Green, Steering Committee (Yale University); Sanda Ionescu (ICPSR); I-Lin Kuo (ICPSR); Walter Piovesan (Simon Fraser University); Richard Rockwell, Steering Committee (Roper Center); Marion Wittenburg (NIWI/Steinmetz Archive).

Annual Report

After introductions and welcoming comments from Hans Jorgen Marker, Vice Chair of the Committee, and Ann Green, IASSIST President, the Committee briefly discussed the FY04 Annual Report for the Alliance. The DDI Alliance formally began operations on July 1, 2003. While the first fiscal year is not yet complete and thus there is no final budget report, the Alliance will finish out the year with a positive balance that can be used for meetings of working groups and other activities in the upcoming year. There were 25 members in FY04, with the prospect of some new members in FY05.

Upcoming Meetings

In discussing a possible fall 2004 meeting of the Alliance, the possibility of having separate European and American meetings was raised, but this was deemed too difficult to execute. It was decided that the Alliance should attempt to support fall meetings of the Working Groups whose activities are most central at the time. There has already been an informal request from the Structural Reform Working Group to meet in the fall. The Comparative Data group may also be meeting in conjunction with the MetaDater initiative meeting in Cologne in the fall.

The date for the next meeting of the full Expert Committee was scheduled for Sunday, May 22, 2005, in Edinburgh, in advance of the IASSIST conference to be held that week.

Communication Mechanisms

It was generally agreed that ezboard was not working as a general communications vehicle for the Alliance and that we should not invest in the bulletin board technology any longer. Most members prefer to communicate via the email lists that have been established for the Working Groups and for the Expert Committee as a whole. These lists have archives that can be consulted for message threads. It was also agreed that we should continue using the Expert Committee Web page of the DDI Alliance site for file sharing. This site can be password-protected if files of a sensitive nature are to be added. Telephone conference calls have been working fairly well, although this mode of communication can be difficult for participants who are not native English speakers.

Reports of Working Groups

Complex Files Working Group

The Complex Files Working Group had prepared a proposal, which was brought to the Expert Committee for discussion. Pat Doyle described the proposal in detail and walked the Committee through the examples. This proposal is intended to document a system of files as opposed to one file at a time. The proposal can document, for instance, a time series or a structure of person-level and household-level files used in tandem. Another example would be a three-wave longitudinal study in which data are collected at various points in time and then accumulate into a group of related files. The proposal is purely to document these structures independent of applications.

The Complex Files proposal recommends the addition of a new section of the DDI (Section 6) called File Group (similar in concept to Variable Group and Category Group). File Group would be repeatable and would provide the functionality to:

- Identify the set of files that make up the system
- Indicate how they can be used together through a linkage or a join (optional)
- Identify the files generated as a result of the join

A new Section 6 was developed rather than using Section 3 (File Description) in order to provide for an added DDI layer that points to other files. The proposal is mechanistically based and would fit into the modular structure for the new Version 3.0 that is being envisioned. We do, however, need to clean up Section 3 as we move forward.

A comment was made that this is really a revolutionary proposal in that it creates a structure that sits above other DDI instances. We need to take relevant examples and see if they work with this proposal. This is basically a relational system that is being described, and it needs to be generic enough to work in any situation. It is possible that we may need some semantic relationship controlled vocabularies.

We need to decide if the proposal covers comparative data and all types of longitudinal files also. In general, we need to make the distinction that the Complex Files proposal is intended to address the matching of cases, while the Comparative Data area refers instead to the matching and harmonization of variables. Both levels need to be coordinated in the DDI.

Julie Linden volunteered to test the Complex Files proposal with aggregate/tabular data, and Jostein Ryssevik will test it with Nesstar. Pat will act as Architect of the proposal and in that capacity will work with the SRG to populate a spreadsheet detailing the relationships inherent in the proposal elements and attributes. [Note: Pat Doyle died shortly after the meeting. Janet Eisenhauer Smith subsequently agreed to act as Architect of the proposal. -MBV]

Comparative Data/Families of Datasets Working Group

Meinhard Moschner provided a summary of the work of the nine-member group to date. The group distributed an initial brainstorming paper outlining the scope of their substantive concern, which is linking elements across studies or instances and over space and time. The group needs to look not only at the projects designed as comparative studies but also at potential families of studies that may not exist physically but only logically.

Indicating comparability is not easy because methods or measures are not always equivalent. Deviations are unavoidable and also necessary sometimes. We need to document the deviations at the study and variable levels, comment on the reasons for the deviation and on harmonization procedures, and provide for potential trend analysis. Harmonization can sometimes lead to the loss of information -- for example, if one has to collapse categories - and this needs to be documented. In the case of potential trends, we need to be able to link variables and questions across studies and to provide enough information to show that harmonization is possible. This may require a new DDI level - a collection or family level or something like a collection variable group to describe loose trends. The group will prepare a common data model for discussion.

The MetaDater Project, also concerned with comparative data, is in the process of preparing a metadata model, which will be designed to be compatible with the DDI model. There is now an internal vision of the

MetaDater model, and the project is aiming to have a more complete model for an expert workshop in the fall.

It was pointed out that W3C standards to integrate references (XInclude or XLink) may be useful in the data model to describe comparative data. We also need some means of formal description to construct new variables for purposes of harmonization. This could perhaps be a subset of MathML.

ISO-11179 establishes comparability even if there is no actual physical data collection. It lifts variables up to a higher level of abstraction and links to concepts. If we atomize variables into their component parts, as in the ISO-11179 model, do we lose the study context? Keeping links to studies that gave rise to the variables is vitally important. Also, comparability is hard to establish. The DDI needs to keep its information simple and descriptive and provide the information on which researchers can base comparability decisions.

We need to make sure that the DDI enables what comparative researchers really want to do and is useful for them. We might involve researchers from the Comparative Study of Electoral Systems or the Luxembourg Income Study to ensure that we are meeting the needs of the community. Comparative research is affected by methodology, sample design, and many other factors.

Focusing on comparative data issues allows us to compare potentially comparative variables after which an application can capture these new variable groups or relationships into a knowledge base. However, in the process of harmonization, we need to ensure a strict division of labor between the DDI and applications.

It is possible that we should treat harmonization and comparability separately. Should we take the stance that the elements we provide in the DDI are what the researcher needs to know, or should we come up with a measure or index of comparability?

In the Madiera project, the goal is for researchers to make the actual decisions about comparibility but to identify the factors that influence their comparability rankings. It's also important for the researcher to feed back into a system to say that he or she performed a certain harmonization.

We need to have tags to indicate "these measures were designed to be comparable". The results of harmonization itself are in effect a new dataset.

We also need to know that we are measuring the same concept, which is where ISO 11179 again becomes relevant. The DDI already has a concept element, which can point to a vocabulary outside of the DDI; this could be an ISO-11179 repository.

The DDI may need more controlled vocabularies, but the Committee was cautioned not to embed controlled vocabularies into the specification and to keep the DDI XML independent of vocabularies.

Structural Reform Working Group

Wendy Thomas reported for the SRG, which works in parallel with the other working groups and is tasked with maintaining consistency in design across the proposals of the substantive content groups. To this end, the group created a diagram of the data life cycle with a modular structure for review by the Committee to ensure that there was agreement on what the DDI is designed to document. It was noted that there is a chapter of a MetaNet report on the life cycle of statistical data that we should also consult. The SRG also did a mapping of the current DDI tags to the life-cycle model.

This life-cycle model helps to determine what is in and out of scope for the DDI. The diagram starts at the study design stage and continues on to the archiving of a dataset and beyond, with DDI embedded in the process throughout.

The modules of the life cycle are:

- Study/Survey Design - With Concept sitting in a tier above
- Data Collection - With Data Collection Process above
- Data Processing - With Physical Encoding and Logical Encoding above
- Data Dissemination - With Archiving above
- Data Discovery
- Data Analysis

Running left to right through the life cycle are actions that are part of Data Use: Study Discovery, Detailed Discovery, and Data Access.

The current DDI specification is comprehensive enough for a single survey, but in general it represents the tail end of the life cycle.

The sense of the Committee was that the new life-cycle model was appropriate for the DDI. It was noted also that having a model that spans the life cycle of statistical information fits into the new vision of SPSS. SPSS created SPSS Dimensions, which was geared toward data collection and was principally a tool for market research. But now the goal is to have a robust and sophisticated suite of tools that span the data life cyle and that handle large datasets.

It was pointed out that the word "survey" is ambiguous and that the DDI could encapsulate a number of instruments. A question was raised regarding what exactly the digital object being described was - a dataset or a study or some other entity. We need to be clear about our definitions of these terms.

Related to this, there is a preservation metadata model that maps to the OAIS model. Could another model, such as METS, be used to "wrap" the DDI?

In October through December, the SRG will be working on the data model for Version 3.0 with the goal of having people comment by the end of January 2005. This data model will either be in the form of a spreadsheet or a UML model.

A question was raised about whether the DDI is intended to document both the conceptual and the physical. The DDI started with a physical object - a social science codebook -- that was documented and would be preserved. However, we have now separated the physical and the logical, which moves us away from the traditional codebook structure where we started. We need to be able to preserve the conceptual structure archivally without worrying about the physical form.

What we have with this life-cycle model is a set of modules and the DDI instance is a way of combining the modules from different places. We need to also think about versioning across the life cycle.

Aggregate Data, Time, and Geography Working Group

Ilona Einowski reported for this group, which is in the process of obtaining background information to move forward. Wendy Thomas has sent information on the current aggregate model and will also send information on what the NHGIS project has learned in using the nCubes model for that project.

We need to think about how aggregate and tabular data are different. The cube specification needs to be improved in Version 3.0. We want to be able to say that a dimension in one cube is the same as in another cube, and currently there is no way to do that. Right now we have to artificially locate them under the same study, but they could come from different sources.

In terms of time and geography, the group is working on identifying problems with the current specification. The Madiera project completed a review of the existing geographic elements. Atle Alvheim will determine whether the final report can be circulated.

We need suggestions for Version 3.0 in terms of geography, and we need to determine how the DDI relates and should relate to other geographic standards. Ilona will fill in as committee chair for Margaret Low while she is on leave. Julie Linden will look at the mapping from DDI to FGDC.

Instrument Documentation Working Group

It is still an open question how much emphasis the DDI should place on documenting survey instruments. Should this be separate or a part of the DDI? We need to compile a list of potential tags related to instrument documentation that are not currently in the DDI. This can be partially facilitated through ICPSR's collaboration with the Survey Research Operations group at the University of Michigan's Institute for Social Research. This group has created a Blaise documentation program that produces an XML codebook, and ICPSR is currently mapping the XML tags to the DDI. ICPSR will also compile a list of what is missing from the DDI in its current form.

Usability and Outreach Working Group

The ICPSR Web site (www.ddialliance.org) provides a lot of information on how to use the DDI in specific situations, given various sources of information. We need to continue to solicit information on what others are doing and describe the different projects on the site. We also need to find out what people who are not using the DDI need to know in order to understand the value of using it. Providing good examples is extremely important. We could also use outreach materials directed to different audiences.

A useful resource would be to show how to create a Dublin Core record using the DDI. We are currently working on a DDI to MARC conversion.

Outreach to grad students is a potentially promising area now that data integrity issues are so prominent. We need to convince students to document their data to protect themselves.

The view was expressed that the DDI is basically selling itself now and is in fact an easy sell. It is expanding outside the original committee. Transport for London, for example, is going to use it. We need to reach funding agencies as well. We could try to stipulate that projects need to use the DDI to get their funding.

Potential New Working Groups

Other Working Groups have been proposed to address issues of:

- Qualitative Data
- Longitudinal Data
- Historical Data
- Language

We will investigate further and determine the need for separate groups.

Persistent URLs

Joachim Wackerow followed up on his email discussion regarding this issue and suggested that the Alliance may want to make a recommendation on how to identify codebooks in a unique and persistent way, perhaps using URNs, which could be mapped to URLs through file resolvers. However, there is currently no widespread accepted resolution system.

Structured URNs are becoming more common. We should look at OASIS. A question was raised about whether the persistent identifiers should be at the study or the codebook level and whether there should be a central registry. For uniform naming conventions we don't need a registry. We should have a publicly available naming scheme to which we all adhere.

We still have the problem of duplicate holdings. If the Internet domain is the first part of a structured URN, this helps to solve the problem. The SRG considers this issue in its purview and will look into this further. We need to include all archives in such a discussion.

Open Access Protocol

Joachim suggested that the DDI might benefit from having a central interface and repository for search and retrieval of DDI files and a standardized transmission protocol to exchange files.

The industry standard at this time appears to be Simple Open Access Protocol, or SOAP, which is being developed by the W3C as part of Web Services. However, the archives often align themselves with the library community, which uses the Open Archives Initiative Protocol for Metadata Harvesting, or OAI-PMH. SOAP is not part of the OAI specification. If we are interested in developing a registry, we might look at SDMX, which is already going that route. Developing and maintaining a registry has high overhead, so piggybacking on an existing structure would ease that burden.

The UK Data Archive is looking at becoming either OAI or Z39.50 compliant to meet the requirements of its funders. A drawback of OAI is that metadata harvesters can present the results as their own.

We need to investigate the SOAP and OAI protocols further before making a decision.

Procedures Manual

The Manual prepared by the SRG clarifies the process for changing the specification that is outlined in the DDI Alliance Bylaws and distinguishes between proposals for major and minor changes. It sets out a process involving a spreadsheet, which details the relationships between elements and attributes in a proposal; this makes things easier for the Working Groups, who do not then have to write XML. A Working Group can either have a member fill out the spreadsheet or can work with the SRG to build the document.

We are using the Complex Files proposal as a prototype and will be following it through the processes stipulated in the Manual. The proposal currently does not have a corresponding spreadsheet, which would complete Part 1 of the process, but Pat Doyle as the Architect of the proposal will work with the SRG to develop one.

For substantive content groups with overlapping interests, there should be broad discussion of developing proposal before the proposals are formalized and before the Expert Committee has to vote. This will ensure that we don't work at cross-purposes or take radically different strategies. Working Groups should feel free to start to work with the SRG as early in the process as possible. This can be done informally.

Timeline

The main point about the Timeline, which is now published on the DDI site, is that the final date for proposals to be submitted to become part of Version 3.0, which is planned for January 1, 2006, is March 1, 2005. This deadline is intended to provide adequate time for a proposal to make its way through the specified channels.

WORKING DRAFT

# DDI VERSION 3.0 CONCEPTUAL MODEL

STRUCTURAL REFORM GROUP

10 JUNE 2004

## BACKGROUND OF THE CONCEPTUAL MODEL

The conceptual model of the DDI provides the common structure that the technical implementations refer to and describe through various approaches including DTDs and schema. The conceptual model encompasses the logic of the structure covering why and how parts relate to each other. The technical implementations explain how the conceptual model works within a specific technology.

Originally, the DDI took its model from the codebook which assumed a finalized clean version of the data. It was clear early on that many were expanding that concept to mean something much broader and perhaps more complex than a traditional hard copy codebook. There was discussion about using the DDI to capture information during the creation process, but the focus was still on the final data set. Additionally, the "codebook" approach suffered from the lack of a clear idea of what a codebook encompassed. In fact, even a brief glance at the "codebooks" represented in the ICPSR collection shows that the idea of a "codebook" has never been consistently defined in either content or structure, and that the reality of codebook construction reflects infinite variety.

As development of the DDI has progressed we have begun to move towards looking at the DDI as encompassing the development life cycle of a data set. Version 3.0 reflects this change in scope and no longer makes the assumption of a finalized version of the data. In order to support this broader scope, a modular structure is required. This allows the addition of modules or sub-modules as the data is developed, encoded, preserved, disseminated, and analyzed.

In order for the DDI to provide a structure that supports both programming and archival activities, we need to have a well structured and well understood model. We need this in order to provide consistent application of the standard as well as resolve questions regarding the application of the structure to a specific instance of an XML document.

The movement to a modular design for the model has been developing over time and is not a radical change in direction as much as it is recognition of the emerging consensus. It is needed to provide the flexibility for dealing with specialized data files and data sets as well as the variety of technical environments within which we currently work or are in the process of developing.

## GOALS FOR MODULAR DESIGN

- To capture information on the creation of specific data within a production/technical environment in a way that it can be accurately transferred as the data travels through and outside of that environment
- To organize the modules so that they accurately record information about data and the data creations process AND contain the information on structures and relationships necessary for data discovery, extraction and manipulation
- To have basic modules that will work in all technical implementations (specialized modules may not work in all technical implementations)
- To provide specialized modules for special types of data or storage formats so that all elements in the DDI are used in a consistent way
- To organize the elements within modules so that if your system cannot handle a specific module the other modules will still work (use example of a physical store that is described in a specialized module)

## DESIGN OF MODEL

In many ways the original codebook concept captured a specific sub-set of information about the data at a specific point in time, a static picture. In the Version 3.0 model we wish to capture a more dynamic structure, recognizing what information continues through and what changes as the data moves through time. In short, we wish to capture the dynamic nature of the metadata. To do so we need to understand its life cycle.

The metadata life cycle as described by Ann Green and Jean-Pierre Kent in chapter 2.2 of MetaNet Work Package 1: Methodology and Tools,[1] begins with a production phase model (input, throughput, output) and then expands this to include a conceptual phase prior to the input phase, and a repurposing and preservation phase after the output phase. The conceptual phase concentrates on the design issues that take place before a process starts. This includes the design of the entire process model.

> "While output data are dependent on input data,, and can only be produced after input data have been collected, input metadata are inferred from output metadata. One cannot start thinking about what data to collect and how before one precisely knows what the end product will be and how it will be achieved." (Green & Kent, p.33)

The repurposing phase provides for the documentation of secondary analysis resulting from the original data output. The preservation phase focuses on the processes involved in long-term preservation of the data and metadata. This includes format changes to remove system or software dependencies, cleaning, integration and harmonization, and the development of related materials.

[1] http://www.epros.ed.ac.uk/metanet/deliverables/D4/IST_1999_29093_D4.pdf

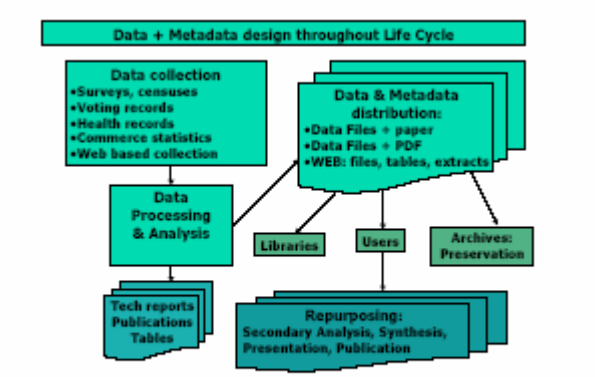



*Figure 1: The Data/Metadata life cycle*

Figure 2 shows an initial working model provided by I-lin Kuo from a programming or data application perspective. I-lin Kuo's model (below) and the Green/Kent life cycle model have parallel features but different areas of detail and emphasis. The *Survey Design* could be seen as analogues to the concept phase described above. *Data Processing* encompasses "Data Processing & Analysis" and "Tech reports" while *Data Distribution* is similar in each. *Data Delivery* may cover parts of "Libraries", "Users", "Archives: Preservation", and "Repurposing", while *Data Analysis* would most likely be contained within the "Repurposing" phase of the Green/Kent model. By adding the module of *Data Discovery*, Kuo bring in the issue of how the information within the structured metadata and data are being used.

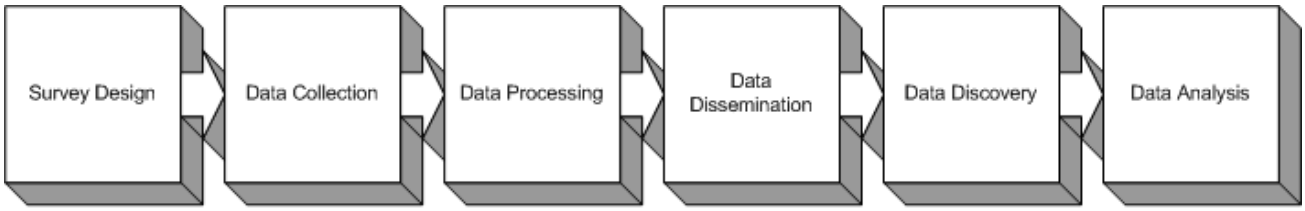



*Figure 2: I-lin Kuo Model*

In Kuo's model the first two boxes focus on specific aspects of the creation of data and documentation, but miss some important details in terms of modeling the DDI, primarily the preservation aspects of archiving and the repurposing of the data discussed in the Green/Kent model. By combining features of the two life cycle models we get a fuller picture.

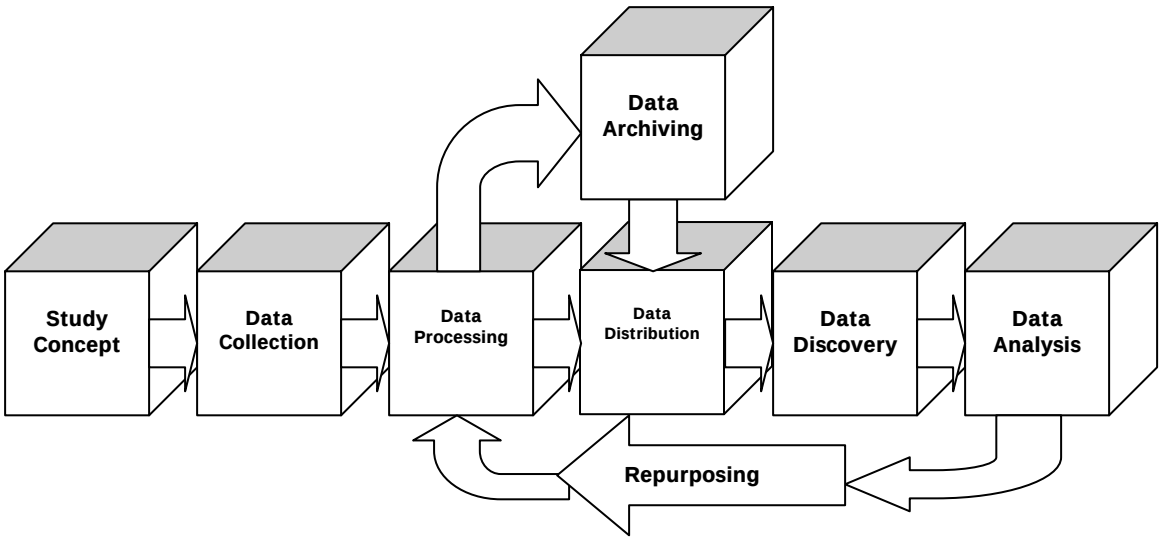



*Figure 3: Combined Life Cycle Model*

The Combined Life Cycle Model incorporates either direct dissemination to users or through data archives and recognizes that data can be reprocessed at a later point in its life cycle, creating an iterative process. However, it is no longer a linear but a circular model. From the concept of time in this model, *Repurposing* follows *Data Analysis* and therefore can't feed back in time. One way to address this is that each circular path becomes a new instance.

We viewed *Repurposing* as being a secondary use of the data from a study. While multiple products could be planned for in the original conceptualization, collection, and processing of the data, *Repurposing* reflected a new conceptual framework. For example this might be a streamlined instructional data set, a specific sampling and restructuring of the data, or combining data from multiple sources to create a new data set (either physically or virtually). The implications of this view include the need for defining the relationships between data products conceived of during the conception process (such as the multiple products of the United States Decennial Census) as well as the ability to define both primary and secondary data sources within the *Data Collection* phase.

## THE BASIC VERSION 3.0 CONCEPTUAL MODEL

The model of metadata below focuses on what is done with the data, in other words the data's life cycle, as opposed to how it is used. Like the model above it contains both a conceptual and data collection module. Rather than a distribution module, further differentiation is given to the description of the output data, dividing it into logical structure and physical structure modules. A combination of the modules to this point (the metadata) and the data are then prepared for discovery and distribution by an entity (here described in the Archiving module) which may further enhance the data or metadata, provide internal identifiers and processing information, and impose rules of access and distribution. In a broad sense it could encompass the "Libraries", "Users", and "Archives: Preservation" phases of the Green/Kent model. It does not address *Data Discovery* or

*Data Analysis* directly, but instead sees these as processes to be supported by the contents of the modules included in the model below.

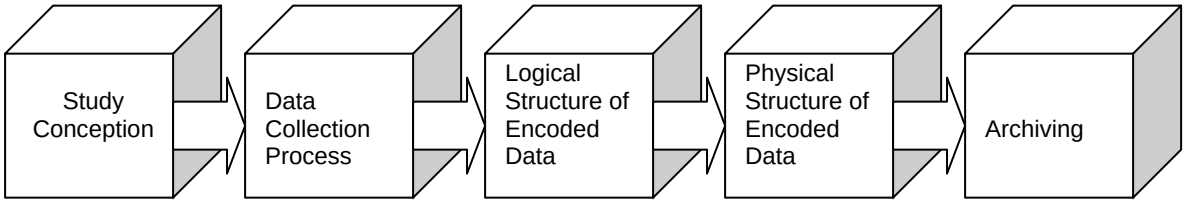



*Figure 4: Modules in the Conceptual Model*

Each of the basic modules can contain multiple sub-modules to reflect the specific descriptive needs of different types of processes or materials. In addition, some modules may be replicated, say if you have two physical instances of the same data in different storage formats (example: ASCII Fixed Format and STATA Data File). In addition, there clearly needs to be a "wrapper" which identifies the modules and module sub-types included in any given instance.

## STRUCTURAL APPROACH FOR CONCEPTUAL MODEL

The color diagram of the DDI Version 3.0 Conceptual Model (attached), shows the inter-relationship of the modules noted in Figure 4 and the those shown in Figure 2. It separates the use of the information in the metadata and data for study level discovery, detailed discovery, and data access. It is important to differentiate various points along this range of usage because of the issues surrounding interoperability with other metadata structures, high level search engines, and systems which are closely tied to a specific technology or use. The following design rules should be used to determine what belongs in each module and whether the module should have a persistent structure or have alternative structures to deal with specialized data files, studies, or technical needs.

### DESIGN RULES

- Persistent sections should be separate from dynamic information
- Information modules should follow through the various lifecycle paths
- Information used for discovery should be in non-specialized modules
- Separation of dynamic materials and non-dynamic materials: What parts change when a data file moves from one "home" to another, or changes something like its physical storage structure? Theoretically those pieces should be modules that can be "swapped" out.
- Information discovery perspective: What information is needed at different levels of discovery/extraction/manipulation and what search engines would be accessing the information at each level? It is beneficial to keep information used by non-social science data specific discover systems together and/or uniformly accessible.

## IMPLICATIONS FOR DATA DISCOVERY

This structure has major implications for data discovery and use. First there are different levels of discovery and different approaches to discovery.

The most basic discovery approach is through the Dublin Core elements and the holding information for a specific archive. In short, what it is and where it is located. The Dublin Core information is persistent while the holding information will change. Dublin Core works for bibliographically based search systems but does not have the content detail or control of other widely used bibliographic systems. In particular it does not carry the geographic detail needed by most geographically based search systems. The use of the field COVERAGE by the Dublin Core does not following the restrictions imposed by geographic search systems and the Dublin Core does not provide options for coordinate based searching.

The next major level of discovery is in-depth identification of study contents at the level of variable or data items labels, category labels, universe statements and other information held in the logical structure material. This level of discover can answer specific questions about what data exists and if it is encoded in a manner that is useful for the individual user. It does not include, but is necessary for, the next level of use, that of data access and/or manipulation.

In order to access data the system needs to be able to identify the following pieces of information:

- Identification of the record and data items required
- Identification of the physical store and the location of the record and data items within that store (a link between the physical description and logical description)
- Logic of record selection
- Ability to process the specific physical data format
- Ability to output, manipulate and/or display the data retrieved from the physical store

## MODULE CONTENTS

**Concept Module**

- 1 – Not a repeatable base module
- Contains identifying information for XML; all Dublin Core elements
- Can be used as a stand-alone bibliographic record (not including holding information)
- Provide information on the context of the study, relationship to series or family of data sets
- Information on the purpose of the study, universe, and coverage in its broad aspects

**Data Collection Process Module**

- Should be able to have multiples of this…multiple questionnaires etc..
- Should this replicate or have different 'swappable' modules to reflect difference processes? Should this be a collection of sub-modules?
- Contain information on the development of the data collection tool and implementation
- Question text should go here as well as intent of questions and interviewer or recorder instruction (this allows questions to act as the source for multiple data items described in one or more logical encodings of the raw data.
- Should contain sub-modules for related resources used for processing or collecting data such as coding schemes, sampling software, etc.
- Should contain sub-modules to include or describe output materials related to the Data Collection and Processing

**Logical Encoding Module**

- Swappable and repeatable to reflect different types of logical data structures derived from the initial data collection
    - Raw-Microdata
        - Simple survey
        - Multiple file survey
        - Time series
        - Dynamic
        - ..
    - Aggregate
    - Others
- Describes the logical encoding of the raw collected data
- Links to questions where appropriate and includes encoding instructions

**Physical Encoding Module**

- 1+ swappable and repeatable
- (Need to allow for multiple stores of the same data set without replicating everything)
- Different modules for different format types
    - One dimensional
        - rectangular
        - hierarchical
    - Two dimensional
        - rows and column
    - Three dimensional
        - rows/columns and layers
    - Proprietary database
        - oracle
        - sass

    - spss
- Links to logical encoded description of data items
- Provides gross file structure description and relationship information between files and/or records

**Archiving Module**

- 1 holding archive information
- Sub-modules could be swappable regarding restrictions, access criteria….how much do these differ? This is one area where individuality by archive is not a problem unless they want to share this access information. There could be a generic base module with swappable sub-modules

## POPULATING THE CONCEPTUAL MODEL

The attached color coded listing of elements from DDI Version 2.0 provides a preliminary suggestion of which module the current elements would fall into. Some of these elements are preceded by colored stars (*) indicating that their "home" is unclear. This may be due to a number of reasons. Sometimes the element is not well differentiated and may contain a mixture of information. At other times the element is used in more than one way and its "home" depends on how it is used in a specific instance. These will be clarified as the conceptual model takes on more structure.

### MAJOR QUESTIONS REMAINING

The following questions will need to be answered before the completion of the conceptual model. They have been raised a number of times in discussions both inside and outside of the DDI Alliance. The include

- How do we define an instance? What constitutes a new instance?
- What modules are required and what are the basic requirements of each module?
- How do we handle versions, editions and copies of the instance as a whole as well as the individual modules?
- How do we leverage the perspective of ISO 11-179 CMR to take advantage of the rich descriptive framework of the concepts underlying the data?
- How will the topics being discussed in the Substantive Working Groups be incorporated? We need to clarify where these will fit conceptually in the life cycle.

## CONCLUSION

This working draft sets out the basic conceptual model proposed for Version 3.0 of the DDI. We have tried to provide sufficient background information for others to understand our perspective and intent for the overall structure of Version 3.0. It is definitely still a work in progress and needs to be further defined, particularly in terms of the contents of the basic modules. However, we also need to confirm that the basic structure is sound and

comprehensive. In order to do that, in-put is required from the Expert Users Group. We propose the following process for accomplishing the development of the conceptual model and production of the first technical implementation by the end of 2004. The timeframe is suggestive and will be adjusted to fit the needs of group, but our goal is to have the conceptual model clearly defined by early October so that work can begin on the technical implementations.

| Time Period | Activity |
|---|---|
| Mid-June to Mid-July | • Get feedback on the broad model<br>• Answer questions if needed<br>• Finalize the basic modules |
| July | • Clarify contents of the basic modules<br>• Populate the structure with the current fields from Version 2.0 |
| August | • Get feedback on distribution of Version 2.0 items<br>• Clarify unclear relationships |
| September | • Expert Committee review of final version |

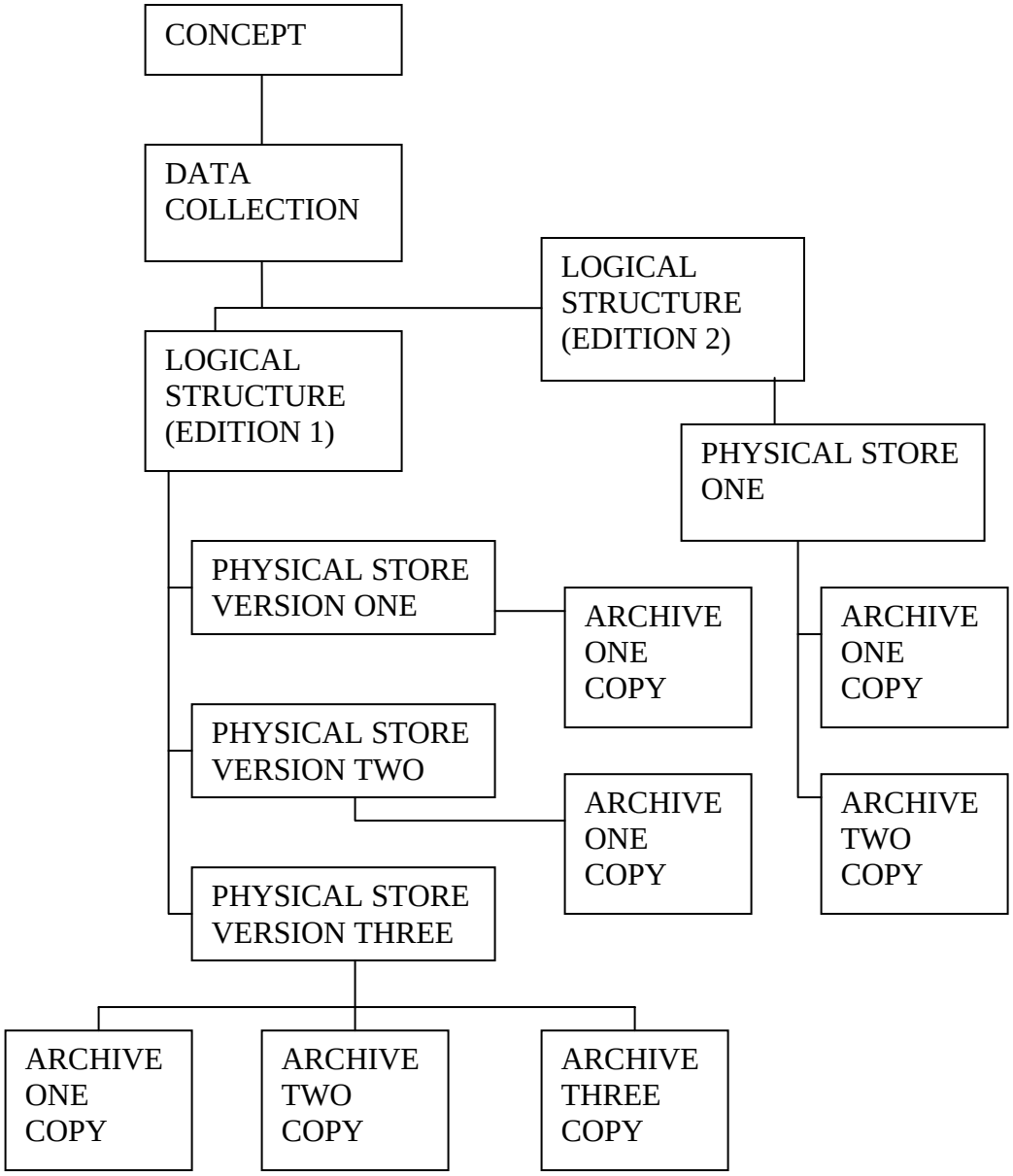
Please direct all comments and questions to the Expert Group Listserve <alliance-experts@icpsr.umich.edu>

## APPENDIX:

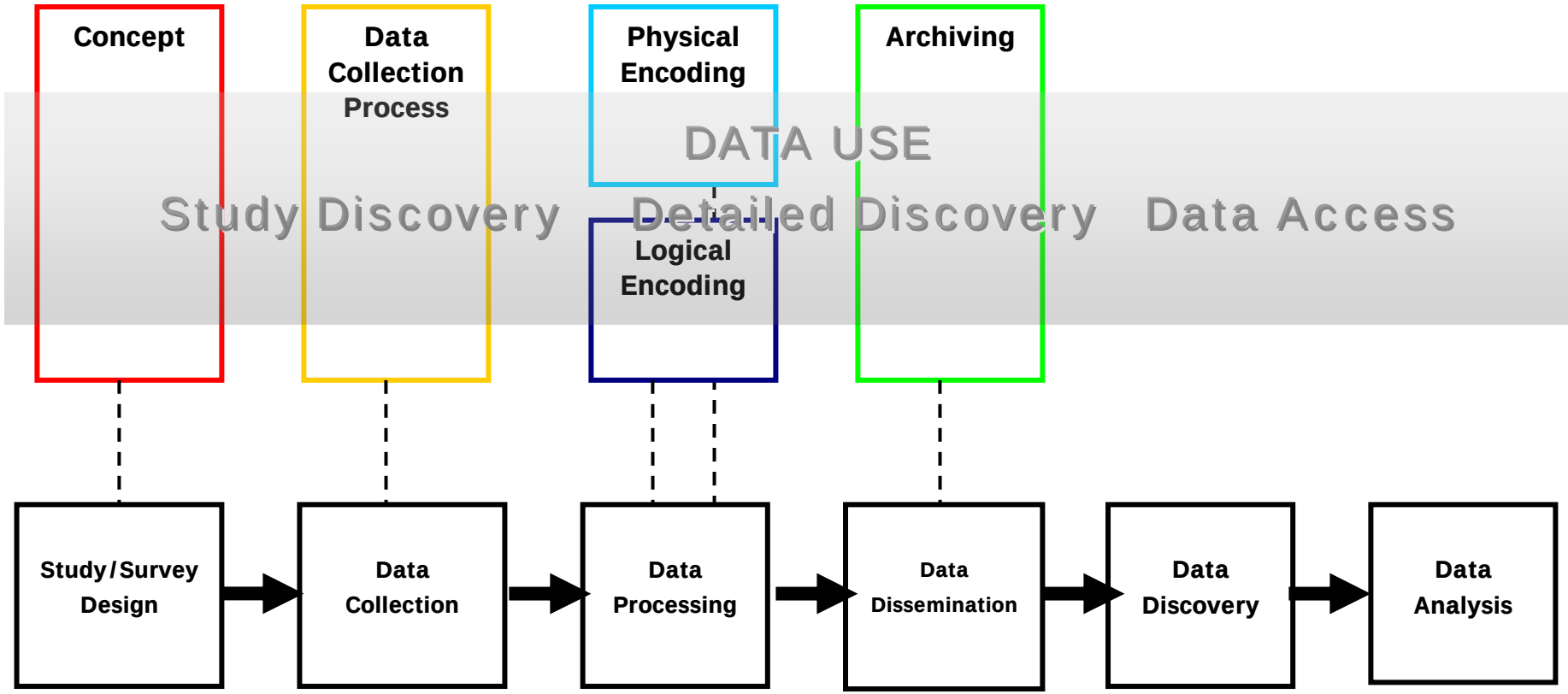
### MULTIPLE USE MODULES, VERSIONS, EDITIONS, AND COPIES

The same data set could also have multiple versions of the same logical structure stored in different physical formats (similar to a translation of text). Finally the same original data set could have multiple editions where information in the logical structure has been corrected, limited in the coverage of elements or records (subsets) or contain added items such as recodes or alternative geographic codes. The graphic on the following page shows a tree of data sets resulting from a single study.

In this hierarchy the study had two editions of the logical structure. Edition 1 has three versions of the data set (for example, fixed format ASCII, a comma delimited file, and a proprietary data base file). Archive 1 hold a copy of all physical stores of each edition of the study. Archive 2 has a copy of one physical store for each edition, and Archive 3 has only a single physical store of edition 1.


CONCEPT
DATA COLLECTION
LOGICAL STRUCTURE (EDITION 1)
LOGICAL STRUCTURE (EDITION 2)
PHYSICAL STORE ONE
PHYSICAL STORE VERSION ONE
PHYSICAL STORE VERSION TWO
PHYSICAL STORE VERSION THREE
ARCHIVE ONE COPY
ARCHIVE ONE COPY
ARCHIVE TWO COPY
ARCHIVE ONE COPY
ARCHIVE TWO COPY
ARCHIVE THREE COPY

# DDI V.3.0 - Conceptual Model 24 May 2004



DDI Codebook Outline -- Version 2.0

WORKING DRAFT

* == ELEMENT IS OPTIONAL & REPEATABLE
+ == ELEMENT IS MANDATORY & REPEATABLE
? == ELEMENT IS OPTIONAL & NON-REPEATABLE
  == ELEMENT IS MANDATORY & NON-REPEATABLE

CONCEPT MODULE
DATA COLLECTION MODULE
PHYSICAL ENCODING MODULE
LOGICAL ENCODING MODULE
ARCHIVING MODULE
OBTAINING DATA MODULE
* COULD BE SECTION DENOTED BY
  COLOR OF STAR DEPENDENT ON USE

0.0 codeBook
1.0 docDscr*
1.1 citation?
1.1.1 titlStmt
1.1.1.1 titl
1.1.1.2 subTitl*
1.1.1.3 altTitl*
1.1.1.4 parTitl*
1.1.1.5 IDNo*
1.1.2 rspStmt?
1.1.2.1 AuthEnty*
1.1.2.2 othId*
1.1.3 prodStmt?
1.1.3.1 producer*
1.1.3.2 copyright?
1.1.3.3 prodDate*
1.1.3.4 prodPlac*
1.1.3.5 software*
1.1.3.6 fundAg*
1.1.3.7 grantNo*
1.1.4 distStmt?
1.1.4.1 distrbtr*
1.1.4.2 contact*
1.1.4.3 depositr*
1.1.4.4 depDate*
1.1.4.5 distDate?
1.1.5 serStmt?
1.1.5.1 serName*
1.1.5.2 serInfo*
1.1.6 verStmt*
1.1.6.1 version?
1.1.6.2 verResp?
1.1.6.3 notes*
1.1.7 biblCit?
1.1.8 holdings*
1.1.9 notes*
* 1.2 guide?
1.3 docStatus?
1.4 docSrc*
1.4.1 titlStmt
1.4.1.1 titl
1.4.1.2 subTitl*
1.4.1.3 altTitl*
1.4.1.4 parTitl*
1.4.1.5 IDNo*
1.4.2 rspStmt?
1.4.2.1 AuthEnty*
1.4.2.2 othId*
1.4.3 prodStmt?
1.4.3.1 producer*
1.4.3.2 copyright?
1.4.3.3 prodDate*
1.4.3.4 prodPlac*
1.4.3.5 software*
1.4.3.6 fundAg*
1.4.3.7 grantNo*
1.4.4 distStmt?
1.4.4.1 distrbtr*
1.4.4.2 contact*
1.4.4.3 depositr*
1.4.4.4 depDate*
1.4.4.5 distDate?
1.4.5 serStmt?
1.4.5.1 serName*
1.4.5.2 serInfo*
1.4.6 verStmt*
1.4.6.1 version?
1.4.6.2 verResp?
1.4.6.3 notes*
1.4.7 biblCit?
1.4.8 holdings*
1.4.9 notes*
1.5 notes*

2.0 stdyDscr+
2.1 citation+
2.1.1 titlStmt
2.1.1.1 titl
2.1.1.2 subTitl*
2.1.1.3 altTitl*
2.1.1.4 parTitl*
2.1.1.5 IDNo*
2.1.2 rspStmt?
2.1.2.1 AuthEnty*
2.1.2.2 othId*
2.1.3 prodStmt?
2.1.3.1 producer*
2.1.3.2 copyright?
2.1.3.3 prodDate*
2.1.3.4 prodPlac*
2.1.3.5 software*
2.1.3.6 fundAg*
2.1.3.7 grantNo*
2.1.4 distStmt?
2.1.4.1 distrbtr*
2.1.4.2 contact*
2.1.4.3 depositr*
2.1.4.4 depDate*
2.1.4.5 distDate?
2.1.5 serStmt?
2.1.5.1 serName*
2.1.5.2 serInfo*
2.1.6 verStmt*
2.1.6.1 version?
2.1.6.2 verResp?
2.1.6.3 notes*
2.1.7 biblCit?
2.1.8 holdings*
2.1.9 notes*
2.2 stdyInfo*
2.2.1 subject?
2.2.1.1 keyword*
2.2.1.2 topcClas*
2.2.2 abstract*
2.2.3 sumDscr*
2.2.3.1 timePrd*
2.2.3.2 collDate*
2.2.3.3 nation*
2.2.3.4 geogCover*
2.2.3.5 geogUnit*
2.2.3.6 geoBndBox?
2.2.3.6.1 westBL
2.2.3.6.2 eastBL
2.2.3.6.3 southBL
2.2.3.6.4 northBL
2.2.3.7 boundPoly?
2.2.3.7.1 polygon+
2.2.3.7.1.1 point+
2.2.3.7.1.1.1 gringLat
2.2.3.7.1.1.2 gringLon
* 2.2.3.8 anlyUnit*
* 2.2.3.9 universe*
* 2.2.3.10 dataKind*
2.2.4 notes*
2.3 method*
2.3.1 dataColl*
2.3.1.1 timeMeth*
2.3.1.2 dataCollector*
2.3.1.3 frequenc*
2.3.1.4 sampProc*

DDI Codebook Outline -- Version 2.0

WORKING DRAFT

* == ELEMENT IS OPTIONAL & REPEATABLE
+ == ELEMENT IS MANDATORY & REPEATABLE
? == ELEMENT IS OPTIONAL & NON-REPEATABLE
  == ELEMENT IS MANDATORY & NON-REPEATABLE

CONCEPT MODULE
DATA COLLECTION MODULE
PHYSICAL ENCODING MODULE
LOGICAL ENCODING MODULE
ARCHIVING MODULE
OBTAINING DATA MODULE
* COULD BE SECTION DENOTED BY
  COLOR OF STAR DEPENDENT ON USE

2.3.1.5 deviat*
2.3.1.6 collMode*
2.3.1.7 resInstru*
2.3.1.8 sources?
2.3.1.8.1 dataSrc*
2.3.1.8.2 srcOrig*
2.3.1.8.3 srcChar*
2.3.1.8.4 srcDocu*
2.3.1.8.5 sources*
2.3.1.9 collSitu*
2.3.1.10 actMin*
2.3.1.11 ConOps*
2.3.1.12 weight*
2.3.1.13 cleanOps*
2.3.2 notes*
2.3.3 anlyInfo?
2.3.3.1 respRate*
2.3.3.2 EstSmpErr*
2.3.3.3 dataAppr*
2.3.4 stdyClas?
2.4 dataAccs*
2.4.1 setAvail*
2.4.1.1 accsPlac*
2.4.1.2 origArch?
2.4.1.3 avlStatus*
2.4.1.4 collSize?
2.4.1.5 complete?
2.4.1.6 fileQnty?
2.4.1.7 notes*
2.4.2 useStmt*
2.4.2.1 confDec?
2.4.2.2 specPerm?
2.4.2.3 restrctn?
2.4.2.4 contact*
2.4.2.5 citReq?
2.4.2.6 deposReq?
2.4.2.7 conditions?
2.4.2.8 disclaimer?
2.4.3 notes*
** 2.5 othrStdyMat*
** 2.5.1 relMat*
** 2.5.1.1 citation*
** 2.5.2 relStdy*
** 2.5.2.1 citation*
** 2.5.3 relPubl*
** 2.5.3.1 citation*
** 2.5.4 othRefs*
** 2.5.4.1 citation*
2.6 notes*

3.0 fileDscr*
3.1 fileTxt*
3.1.1 fileName?
3.1.2 fileCont?
3.1.3 fileStrc?
3.1.3.1 recGrp*
3.1.3.1.1 labl*
3.1.3.1.2 recDimnsn?
3.1.3.1.2.1 varQnty?
3.1.3.1.2.2 caseQnty?
3.1.3.1.2.3 logRecL?
3.1.3.2 notes*
3.1.4 dimensns?
3.1.4.1 caseQnty*
3.1.4.2 varQnty*
3.1.4.3 logRecL*
3.1.4.4 recPrCas*
3.1.4.5 recNumTot*
3.1.5 fileType?
3.1.6 format?
3.1.7 filePlac?
3.1.8 dataChck*
3.1.9 ProcStat?
3.1.10 dataMsng?
3.1.11 software*
3.1.12 verStmt?
3.1.12.1 version?
3.1.12.2 verResp?
3.1.12.3 notes*
3.2 locMap?
3.2.1 dataItem*
3.2.1.1 CubeCoord*
3.2.1.2 physLoc*
3.3 notes*

4.0 dataDscr*
4.1 varGrp*
4.1.1 labl*
4.1.2 txt*
4.1.3 concept*
4.1.4 defntn?
4.1.5 universe?
4.1.6 notes*
4.2 nCubeGrp*
4.2.1 labl*
4.2.2 txt*
4.2.3 concept*
4.2.4 defntn?
4.2.5 universe?
4.2.6 notes*
4.3 var*
4.3.1 location*
4.3.2 labl*
4.3.3 imputation?
4.3.4 security?
4.3.5 embargo?
4.3.6 respUnit?
4.3.7 anlysUnit?
4.3.8 qstn*
4.3.8.1 preQTxt*
4.3.8.2 qstnLit*
4.3.8.3 postQTxt*
4.3.8.4 forward*
4.3.8.5 backward*
4.3.8.6 ivuInstr*
4.3.9 valrng*
4.3.9.1 range*
4.3.9.2 item*
4.3.9.3 key?
4.3.9.4 notes*
4.3.10 invalrng*
4.3.10.1 range*
4.3.10.2 item*
4.3.10.3 key?
4.3.10.4 notes*
4.3.11 undocCod*
4.3.12 universe*
4.3.13 TotlResp?
4.3.14 sumStat*
4.3.15 txt*
4.3.16 stdCatgry*
4.3.17 catgryGrp*
4.3.17.1 labl*
4.3.17.2 catStat*
4.3.17.3 txt*
4.3.18 catgry*
4.3.18.1 catValu?
4.3.18.2 labl*
4.3.18.3 txt*
4.3.18.4 catStat*

DDI Codebook Outline -- Version 2.0

* == ELEMENT IS OPTIONAL & REPEATABLE
+ == ELEMENT IS MANDATORY & REPEATABLE
? == ELEMENT IS OPTIONAL & NON-REPEATABLE
  == ELEMENT IS MANDATORY & NON-REPEATABLE

WORKING DRAFT

CONCEPT MODULE
DATA COLLECTION MODULE
PHYSICAL ENCODING MODULE
LOGICAL ENCODING MODULE
ARCHIVING MODULE
OBTAINING DATA MODULE
* COULD BE SECTION DENOTED BY
  COLOR OF STAR DEPENDENT ON USE

4.3.18.5 mrow?
4.3.18.5.1 mi*
4.3.19 codInstr*
4.3.20 verStmt*
4.3.20.1 version?
4.3.20.2 verResp?
4.3.20.3 notes*
4.3.21 concept*
4.3.22 derivation?
4.3.22.1 drvdesc?
4.3.22.2 drvcmd?
4.3.23 varFormat?
4.3.24 geoMap*
4.3.25 notes*
4.4 nCube*
4.4.1 location*
4.4.2 labl*
4.4.3 txt*
4.4.4 universe*
4.4.5 imputation?
4.4.6 security?
4.4.7 embargo?
4.4.8 respUnit?
4.4.9 anlysUnit?
4.4.10 verStmt*
4.4.10.1 version?
4.4.10.2 verResp?
4.4.10.3 notes*
4.4.11 purpose?
4.4.12 dmns*
4.4.12.1 cohort*
4.4.12.1.1 range*
4.4.13 measure*
4.4.14 notes*
4.5 notes*

** 5.0 otherMat*
** 5.1 labl* )
** 5.2 txt?
** 5.3 notes*
** 5.4 table*
** 5.5 citation?
** 5.6 otherMat*