Alliance for the Data Documentation Initiative

Strategic Plan 2004-2006

The Data Documentation Initiative is a story of great achievement, considerable challenge, and radical change on the horizon. With a shared vision, stakeholders in the social science research enterprise came together and accomplished an estimable goal: to produce a stable metadata standard for describing, finding, and using the survey datasets that underlie much of social science research. The importance of that accomplishment and its potential impact on the conduct of research cannot be overstated.

Since 1995, the initiative has strived to develop a generalized standard that can simultaneously serve the needs of data producers, data archives, and data users. The DDI standard provides a rich and structured set of metadata elements and attributes that not only fully informs a potential data analyst about a given dataset but also facilitates computer processing of the data. Moreover, data producers will find that by adopting the DDI standard they can produce better and more complete documentation as a natural step in designing and fielding computer-assisted interviewing. Acceptance of the standard by the community has been rapid as compared with other standards, and the new Alliance membership model has generated a great deal of interest and enthusiasm. In many senses, the work of the DDI would appear to be finished, subject to minor adjustments over the years as errors or omissions are found or as new needs arise.

However, the Alliance Steering Committee, in its role of providing overall guidance for the new DDI Alliance, believes that there is a great deal of work yet to be done. While much of what was learned over the past seven years will apply in the next three years, the DDI needs additional development to ensure its relevance and value to the social science research community.. What is required is increased responsiveness to the realities of the new world of social science data, data which are much more complex even than they were when the DDI began. The 25 founding members of the Alliance confront new challenges as they continue to shape the DDI specification for use by future generations of social scientists.

This document is intended to lay out some goals and guiding principles for the new Alliance to inform the effort as it moves forward over the next three years. We identify five broad strategic goals in the areas of organization, funding, standards, outreach, and technical development with detail on why they are crucial to the success of the DDI. These goals are not listed in order of priority but rather in a rough order of the chronology of action. The foundation for the needed organizational work has been laid and can be readily built upon. Seeking funding for reaching the remaining goals is actively underway and is essential.

Organizational Goal

When the DDI Alliance begins operations on July 1, 2003, it will have a core of approximately 25 members, with most of the European national archives represented as well as the major archives in the United States, along with some leading research universities.

How should we grow the membership, and what is an appropriate membership goal as we look ahead to 2006? The Steering Committee suggests that the Alliance attempt to add at least five members in each of years 2004-2006, with a goal of having 40 members at the end of the three years. Further, it recommends that some of the growth take place in the sectors of developing countries, smaller colleges and universities, and government and private agencies that produce data. To encourage the first category of members, the Alliance needs to consider a mechanism to help subsidize the membership fees.

While growing the membership will result in added revenues, that is not the primary objective in expanding the Alliance. Rather, the objective is to add members in a strategic way, to ensure that representation in the Alliance is adequate to meet the needs of the various DDI constituencies and that the Alliance is positioned properly in the community so as to have the broadest possible impact and spread the acceptance of the DDI standard.

Funding Goal

The single most important strategic move that the Data Documentation Initiative could have made at the end of seven years was to form the Alliance for the DDI. When 25 or more institutions demonstrate their commitment to an idea by putting up their own money to support its infrastructure, the message could not be clearer to funding agencies: this idea has significant backing. The additional commitment of significant direct support from the Inter-university Consortium for Political and Social Research is further evidence of this backing, as are the smaller in-kind contributions from the member universities and institutions.

Nevertheless, the Alliance for the DDI will not be able to carry out the program outlined in this Strategic Plan without significant external support. It is time to look at funders with larger resources. Members of the Steering Committee are active in seeking such funding, and encourage Alliance members also to aggressively pursue external funding for the initiative. The Steering Committee will develop a means for formally recognizing proposals from Alliance members as Contributing Proposals that will advance the cause of the DDI and its adoption.

Standards Goal

The Steering Committee recommends that the Alliance launch the DDI specification on the path to becoming a recognized standard sanctioned by the International Standards Organization, ISO. Such an action is likely to have a profound impact in terms of garnering wide acceptance of and confidence in the DDI and increasing its use. There is already an effort under way to make the DDI compatible with the ISO metadata exchange standard 11179, “Specification and Standardization of Data Elements,“ which can serve as a springboard for this larger effort. Working to harmonize DDI with other metadata standards such as Dublin Core, GILS, MARC, OAI, etc., is also an effective mechanism for raising the profile of the initiative.

Another component of this standards goal is a long-term objective: housing the DDI standard within an institutionalized standards body that will be responsible for maintenance and revision of the standard into the future.

Outreach Goal

Another critical strategic goal is to enlist the cooperation of data producers in the DDI enterprise. For the initiative to succeed fully (so that the archives become filled with studies with DDI documentation), adoption by the producers of data will be key. The Alliance needs to continue to demonstrate to them that there would be serious advantages to using the DDI *and* that those advantages could be achieved without excessive additional cost (and preferably with a savings). This means that the next generation of the DDI must take the needs of data producers more fully into account. In general, the DDI needs to meet the needs of data producers for a metadata standard that can capture the complexities of today’s social science data collection methods and simplify and improve their work at the same time.

Encouraging interested individuals to write articles about the DDI for submission to peer-reviewed journals is another excellent way to stimulate interest in the standard and to enhance its credibility. Publication confers a sense of legitimacy in the academic world that perhaps no other activity can accomplish. Further, being able to reference a body of literature about the DDI and particularly about its usefulness with respect to the research process would bolster efforts to raise external support from funding agencies.

Finally, the Alliance Charter recommends that the Alliance hold symposia for general discussions of metadata, metadata standards, and the DDI standard. This type of outreach and exposure could draw in new constituencies and serve to promulgate the standard to a wider audience.

Technical Goal

The technical goal has three components: structural reform, substantive content, and usability.

Structural Reform

The Steering Committee believes that the highest-priority task facing the Alliance is the need to establish a semantic data model describing and documenting the elements, attributes, and relationships inherent in the DDI specification. A data model serves to ensure that the data producer and the data user share a semantic domain: what the data producer intended to convey to the end user is accurately and completely conveyed to the data user and to the data archive, if the data pass through such an institution. The data model must be sustainable; that is, it must continue to convey fully and accurately the meaning of the data even when those who produced the data are long gone.

Having such a model will enable us to express the DDI in a number of different formats. It will enable extensibility and modularity in a way that the current expression of the DDI -- the Document Type Definition, or DTD -- cannot. Moreover, a data model will ensure internal consistency and completeness and permit the DDI to move forward along the path to becoming a formal ISO standard.

An important question before the Alliance, as it produces a semantic data model, is how far beyond survey data we wish or need to go. Some of the most important work in the social sciences has involved multi-level analyses that incorporate multiple kinds of information, ranging from individual and household data to community data to national data, little of which may be derived from surveys. We need to decide exactly what types of information we should be modeling and what is rationally outside our purview or could be handled as permissible extensions of the DDI.

Substantive Content

As stated above, the Steering Committee is convinced that constructing a data model is the first major task that the new Alliance should undertake. In addition to structural reform, several issues of content and substance must also be addressed to make the DDI as complete and useful as possible. Based on advice from the original DDI Committee, the Steering Committee recommends that new or revised content in the following areas be considered:

*Aggregate/tabular data, with related dimensions of geography and temporal coverage***.** While considerable time and effort have already gone into the creation of an aggregate/tabular extension to the existing DDI specification, there is concern that the aggregate model may be overly complex. The Alliance needs to take a fresh look at this issue, taking geography and temporal coverage into account. Note that the previous working group on aggregate data recommended that the subject of time receive additional attention in any new or revised specification.

Complex files. While the Committee has dealt with this issue and plans to incorporate a revised extension for handling complex files in the upcoming new version of the DTD, the current specification contains vestiges of other attempts to document complex hierarchical and relational files. The current specification has never been tested against the full range of complex data files that are already being used by data producers and is thus of unknown utility. Moreover, perhaps nothing in the world of data has changed more rapidly and more thoroughly than the database structures employed by data producers. Considerable work needs to be done to anticipate and identify the needed elements, attributes, and linkages and to remove those that are extraneous.

Comparative data/Families of datasets. Real social science data are often complex in ways that are not captured adequately by the DDI. A prime example is that the DDI needs a better mechanism to document comparative research. It is not sufficient to produce a DDI document for each of the studies that one wishes to compare and then to use those DDI instances in tandem. Most studies being compared will be alike in some ways and different in other ways, and the DDI must find a way of coherently expressing this. It must also deal with language differences and with coding differences (e.g., studies rarely use identical coding for party identification across countries). The Alliance needs to work in cooperation with groups such as MetaDater in Europe that are focusing intensively on comparative and cross-sectional data.

We also need to find a solution to the larger problem of which comparative research is a subset: describing datasets that are members of "families" of studies, either across countries (or other populations) or across time. Longitudinal data such as repeated cross-sectional surveys also merit special attention. One of the partners in the Alliance, the Roper Center for Public Opinion Research, has identified repeated cross-sectional surveys that provide many hundreds of data points resulting from questions with identical or nearly identical questions that have been asked over the decades since polling began. This produces a complexity that the current DDI cannot fully capture.

Instrument documentation. The DDI needs to come to terms with the issue of documenting computer assisted interviewing (CAI) survey instruments and how far it intends to go in this direction. While much of the complexity of CAI instruments can be captured in the current DDI specification, what cannot be fully captured is the fact that with CAI techniques, no respondent may have taken precisely the same questionnaire as did any other respondent. Among other things, this means that question order effects may be obscured to the researcher and even to the questionnaire designer. Another issue that arises in documenting survey instruments is that the relationship between an original question and the resulting variables may be difficult to define or describe fully. The DDI must evaluate its capacity to define the universe for a specific question and how that universe was reached in the interview.

Usability

To ensure wide uptake of the emerging standard, we need to expedite as much as possible the use of the DDI, particularly for novice users. This can be accomplished in several ways:

Training. While training has been provided at IASSIST and CESSDA for several years, we need to broaden our training efforts with additional workshops and develop a wider array of useful training materials. Formal training courses need to be created and widely offered, perhaps through the vehicles of the ICPSR and Essex Summer Programs, as well as elsewhere in Europe and the Americas. Many have pointed out that it is difficult to get started in using the DDI and that there is a steep learning curve. The Alliance Web site can play an important role in providing access to tools and other materials, such as step-by-step instructions in getting started with the DDI.

Schools of library or information science may elect to develop course modules on social science data and its documentation, producing a new generation of data librarians who are formally trained in the use of the DDI. Similarly, the eventual end users of the data – such as graduate students in the social sciences – could be taught the use of DDI documentation in the course of introducing them to data analysis. Some of these students will go on to become data producers, further spreading the adoption of the DDI.

Best practice/Controlled vocabularies. With a standard as complicated as the DDI, there will naturally be different ways to mark up a document, none of them "wrong" *per se*, just different paths to the same goal. We need to identify best practice to guide people in the effective and consistent use of elements and attributes and to develop useful examples to follow. Recommendations regarding content and further examples of DDI instances will help increase consistency and improve interoperability among DDI authors.

Tools development. Robust tools for DDI mark-up and conversion are a necessity if the DDI is to be adopted widely, and the DDI Committee has encouraged the development of tools from the outset of the project. The most frequent complaint of new users of the DDI (and among most authors creating XML documents) is that the available mark up software is clumsy and difficult to use. Consideration needs to be given to whether or not to produce DDI-specific software if software manufacturers seem unlikely to produce what is needed. Further, we need to reach statistical software manufacturers and persuade them to incorporate the standard into statistical and other applications as an acceptable data description format.

The DDI standard has always been endangered by the opposition of two pressures: make the standard as comprehensive as possible and make it as simple to use as possible. Tool development may be key to attaining comprehensiveness while enhancing simplicity: a good tool would enable the user of the DDI to employ as little or as much of the standard as is desired, appropriate to the context specified by the user, while concealing until needed other elements of the standard.

Conclusion

The DDI is poised to take a new direction as it reshapes itself into a self-sustaining membership Alliance and confronts structural reform in a world of evolving technologies. Creating a data model would seem to be a good starting point for future action and would permit the DDI to remain flexible and ready to adapt to changing technological environments. In constructing such a data model, DDI participants will need to address issues of substance and content and determine how best to represent social and behavioral data in the model.

Once the data model is in place, the Alliance needs to set the specification on the track to becoming a formally recognized standard and begin to make provisions for an ultimate home for the standard with a mechanism for maintenance and revision.

Concurrently, the DDI project needs to expand the membership, begin a serious outreach effort to data producers and software firms, and pursue a broad range of activities to promote expedited use of the DDI in the social science research community and raise the visibility of the effort.

While there is much to be done, the Steering Committee believes that the new DDI Alliance is optimally structured and well positioned to undertake this next phase of work. We look forward to a productive and successful collaboration with the Alliance partners.

[27-JUNE-2003]