Applications in the Real World: The Counting California Experience with the DDI

*by Patricia Cruse, Ilona Einowski, and Juri Stratford**

Counting California - Background
Government data serve a variety of clientele, ranging from businesses and private citizens to some of the most prominent educational and research institutions in the world. With a population and economy larger than many European countries, the State of California's need for continuous and uniform access to government-produced data has never been more urgent.

While digital technologies have revolutionized data distribution, they have also created new problems. What was once a stable system of print materials has given way to a diffuse, constantly changing array of electronic media, using different formats and access methods. The current climate leaves many would-be users frustrated or bewildered; each new upgrade of software and web browsers only exacerbates the problem.

The preservation and consolidation of historical, or time-series, data are similarly at risk. Government agency web sites often mount new information, but may follow no systematic plan to preserve older, historical data as each update supersedes the last. The lack of cumulative time-series data can effectively cripple any attempt to discern long-term trends and changes.

The need for uninterrupted data access and the preservation of historical data are two of the biggest problems facing government data distribution today. Counting California (http://countingcalifornia.cdlib.org) is a new initiative committed to enhancing California citizens' access to the growing range of social science and economic data produced by government agencies by enabling users access to data compiled by federal, state, and local agencies through a single interface.

Counting California's goals respond directly to these challenges:

- To provide flexible, user-friendly access that meets the diverse needs of the California citizenry;
- To insure uniform, continuous access to both current and historical government data; and
- To foster the ability to share data and work collaboratively between government agencies and other members of the data community.

Development of Counting California began in June 2000. The project team for Counting California includes Patricia Cruse (program leader), Ilona Einowski, Fred Gey, andJuri Stratford (data specialists), Brian Tingle (WWW manager), Margaret Low (Sybase programmer), and Marsha Fanshier (SAS programmer). The project team's programmers initially had to rely on a variety of sources developed by the data specialists to describe the data. However, the project is now using XML exclusively to develop the data descriptions and interfaces to Counting California's data. One of the project team's objectives was to have the metadata drive all the functionality from data discovery to data display: including the creation of citations and related materials, labeling, and the thesaurus, as well as enabling searching. The Data Documentation Initiative (DDI) appeared to be the appropriate set of guidelines to develop the XML and achieve this goal.¡

The Data Documentation Initiative
For the past several years the social science data community has been working on the DDI to produce a metadata standard for social science data resources. The DDI "is an effort to establish an international criterion and methodology for the content, presentation, transportation, and preservation of 'metadata' about the datasets in the social and behavioral sciences." The DDI committee has produced a Document Type Definition (DTD) as the foundation of its data discovery and display system.

The DTD provides the rules for applying XML to the DDI metadata. XML, a dialect of the more general SGML markup language, is used for documents containing structured information. In recent years there has been explosive growth in the use of XML and web-based XML tools including databases, search engines, and editors. As we build additional functionality into Counting California, we will support other XML tools. In keeping with the project's goals, we intend to make the DDI metadata available to the scholarly community. Links to actual XML will be available from the web site. We will encourage

others to build upon our work in creating alternate applications that can also be shared.

The DDI is currently a set of guidelines. Throughout the development of *Counting California*, the project team has relied heavily on the expertise of the data community. DDI committee members Ann Green (Yale University), Wendy Thomas (University of Minnesota), and Cavan Capps (Bureau of the Census) provided guidance in the early stages of the project by answering our many questions about the DTD. Many of our early questions focused on the problem of representing tables of aggregate data using the DDI. These questions lead to a meeting in September 2000 with these DDI committee members. As a result of this meeting, we incorporated Wendy Thomas' proposed extensions to the DDI, which allows for the description of aggregate/tabular data, into the development of the metadata. At present, there are few concrete examples demonstrating how the DDI is to be applied. Mary Vardigan (Inter-University Consortium for Political and Social Research) furnished reference support as we struggled to collect examples.

Project Implementation

To set the stage, let us begin by giving a brief overview of the contents of the elements in the DDI DTD. The Data Documentation Initiative DTD consists of 5 sections:

- Section 1.0: Document Description (Codebook Header) consisting of bibliographic information describing the DDI-compliant document being created.

- Section 2.0: Study Description consisting of information about the data collection, study, or compilation that the DDI-compliant documentation file describes.

- Section 3.0: Data Files Description consisting of information about the particular data file(s) containing numeric and/or numeric plus textual information that the DDI-compliant file describes.

- Section 4.0: Variable Description consisting of information about each of the individual variables/ observations that the DDI-compliant documentation file describes.

- Section 5.0: Other Study-Related Material provides a section for the inclusion of other materials that are related to the study as identified and labeled by the DTD users (encoders). Other Study-Related Materials may include: questionnaires, coding notes, SPSS/SAS/ STATA setups (and others), user manuals, continuity guides, sample computer software programs, glossaries of terms, interviewer/project instructions, maps, database schema, data dictionaries, show cards, coding information, interview schedules, missing values information, frequency files, variable maps, etc.

There are a total of 178 elements in the DDI including Wendy Thomas' extensions, which describe aggregate/ tabular data tables. In order to facilitate the creation of the XML, we carefully examined the full list of elements, and selected only those elements that directly impacted the functionality of the system. The project developed its own internal standards incorporating a subset of the DTD needed to access data and to present displays. As the project evolved, we saw the need for additional elements that we then incorporated. Since we were working with a manageable number of data titles, and the entire system was designed as a set of interacting modules, it was not difficult to incorporate these additional elements. We are currently using only thirty-five of the original 178 elements.

The project used the XMetal software to both develop the XML and to validate the integrity of the XML's adherence to the DTD. As we mentioned before, it is the intent of the project to share the XML with data community. The project is using the DDI to document aggregate data sets at the variable level. This will allow users to go directly from the variable descriptions to all related tables in any data file.

The backbone of *Counting California's* data delivery capacities is SAS/IntraNet, which allows for the integration of SAS and the World Wide Web. Specifically, SAS allows for data extraction via the metadata database and provides the ability to format the data in tables, charts, maps, and graphs. In the future we hope to take full advantage of SAS/IntraNet's capabilities as we add functionality to the system.

Some of the files included in *Counting California* were available from the producers only as Excel spreadsheets. For these files, we used DDI sections 1 and 2 to drive the data discovery. The data specialists created these sections, and the Sybase programmer used the metadata to retrieve the corresponding spreadsheet. We hope to offer PDF versions of the spreadsheets as a future enhancement. We also hope to receive the actual data files, used to create the spreadsheets, directly from the producers thereby allowing the project more flexibility in presenting and combining data.

One of the many advantages of basing the system on a fully loaded DDI is that we eliminate many problems of inconsistency. By conforming to the DDI, all titles, bibliographic citations, concepts, etc. will always appear in the same format. As it currently exists, *Counting California* is not fully populated with DDI-generated material. Some of the "early" screens were generated by hand. This was done to expedite the creation of certain screens when we were still experimenting with the XML. Given all the other tasks on our plates, we have not gone back and reworked those screens. It's difficult to bring yourself to dismantle something that is currently functioning adequately, but it will be done.

Project Functionality
Counting California integrates data from a variety of sources, and provides a standard display regardless of the original format of data. The system provides a number of entry points into the data including subject, geography, study title, or issuing agency. *Counting California* allows users to search the table titles, studies, and geographic areas as free text. Users can search all datasets and tables simultaneously or search tables within an individual dataset. The system is designed to ensure that there are no "data dead-ends": the system never returns zero results.

The metadata drives functionality for *Counting California* including the development of the thesaurus, the creation of citations and references to related materials, searching and the labeling of tables and variables. We have found that using the DDI driven version allows a modular approach, is less labor intensive when incorporating changes, and encourages sharing. At present this data documentation process is labor intensive. The data documentation process must be simplified and to some extent automated to encourage California State agencies to contribute new data sets to *Counting California*.

At present, *Counting California* has only a limited set of features. More features will be added as we learn more about what users want, which will allow us to prioritize development activities. However, our experience has provided a real world application of the DDI standards. Our team members now feel qualified to engage in a productive exchange with the DDI committee on future enhancements of the DDI, including additional elements and expanded documentation.

While we are pleased with the performance of the DDI in providing metadata to power *Counting California*, we are aware that the standards for the content of the elements are not fully developed. We found that we could populate the metadata database with DDI elements that would provide the functionality we desired as long as we were consistent in the application of the DDI within the project. The continued development of both the *Counting California* project and the DDI is an iterative process.

Acknowledgements:
Counting California is a collaborative project funded by the California Digital Library and the Library of California. A portion of the funding for Phase I of *Counting California* comes via special arrangement with the Library of California. This interagency agreement stipulates that the Library of California provides funding, and the CDL implements the project. Some of the funding for Phase II comes through a grant from the Library Services and Technology Act (LSTA), administered by the California State Library. Current strategic partners in data acquisition include the U.S. Bureau of the Census and the California Department of Finance. We are currently negotiating similar arrangements with numerous other state agencies.

* Paper presented at the IASSIST Conference, Amsterdam, May 2001. Patricia Cruse, California Digital Library, Academic Initiatives, University of California, Patricia.Cruse@ucop.edu Ilona M. Einowsky, Assistant Director, UC Data/SRC, University of California, Berkeley, ilona@ucdata.berkeley.edu; Juri Stratford, Government Information and Maps, Shields Library, University of California, Davis, jtstratford@ucdavis.edu.

Footnotes
* Note: activities described in this article do not reflect the current status of Counting California, but rather the status of the project in May 2001.