Metadata Driven Survey Design

A Colectica White Paper

Introduction

In current survey practice, the creation of a data collection instrument involves two distinct steps. The first is survey design, in which a researcher defines the questions and flow of a survey. The second is survey implementation, in which a researcher or programmer turns the design into an electronic or paper survey instrument.

The Colectica platform enables an alternative approach to survey development. Metadata-driven survey design means that the actions taken to define a survey are the same actions that create the survey instrument. Using a metadata-driven approach to survey design provides several benefits, including:

- Less redundant work
- Better, easier data documentation
- Reusability of key survey components
- Increased potential for data harmonization
- Greater research integrity

Metadata in Survey Design

At the core of a survey, of course, are the questions. What is a question? Besides the words that ask for data, a few things are needed to fully represent a question.

In order to know what type of data a question is meant to collect, response options should be specified. Possibilities include numbers, text, lists of categories, scales, dates, or a mixture of these. In some cases, the researcher may wish for the order of response options to be randomized to control for ordering bias.

A researcher might also wish to specify the concept behind a question and to record the intent of a question.

By specifying these metadata at the time the question is created and while the information is most accurate, there is no need to perform separate documentation efforts later. Future users of the collected data will have a full understanding of the researcher's methodology.

Once questions are defined, the survey designer can create the flow of a survey. A survey's flow could be as simple as a series of questions in a particular order.

Many surveys have a more complicated structure that includes loops, sampling, and conditional branching. For example, a certain section of a survey may only be administered to a random subsample of half the respondents in the full sample, or additional questions may be asked of respondents who provide a particular response to a question.

A computer-assisted instrument might also have specialized actions, such as computing dynamic text values or starting a voice recorder at certain times.

In current practice, a researcher may specify all this information in text files, Microsoft Word documents, or hand-drawn flowcharts. These design documents are rarely published. By following a metadata-driven approach to survey design, researchers can ensure that the metadata created during this vital stage of the data collection process are not lost.

Management of the Metadata

The Data Documentation Initiative (DDI) is an open, XML-based standard that provides a structured way to store and exchange these metadata. By using software that implements DDI 3, such as Colectica, researchers can specify the design of their surveys in much the same way as the past: by writing the question text, interviewer instructions, lists of response categories, and other elements that define a survey.

Colectica enables a researcher to define the metadata in a structured way. Although at first the software may not be familiar to researchers, its use can facilitate survey development since it provides standard fields for the required elements, and can quickly point out missing information and conflicts.

Traditional word processing systems or paper cannot provide this instant feedback, and so problems must be communicated from survey developers to the survey designers, changes must be made and resubmitted, and the process repeated several times. This is labor-intensive and error-prone.

A metadata-driven approach to survey design does not remove the need for instrument testing. On the contrary, by specifying a survey in a structured way, software can perform automated testing at an earlier stage and with less effort than traditional methods allow.

Colectica's DDI 3 based model holds enough information to automatically generate a survey instrument in several forms: source code for computer-assisted interviewing systems such as Blaise ®, CASES, web-based survey systems, or in-house data entry systems; paper instruments to be printed; customized Web surveys using X-Forms or other Web technologies; or other forms of survey instruments.

Better, Easier Data Documentation

Besides improving the survey design process, a metadata-driven approach to survey design also increases the quality of a project's data documentation. Instead of the documentation process being a separate step of the data collection phase of a project, it is a byproduct of the design process itself.

Software that generates publishable documentation from the metadata defined by a researcher reduces a great amount of redundant work. This type of work is often a lower priority than processing and analyzing the collected data, but it is vitally important for the usability of the data by researchers who were not involved in the study's design.

By using a metadata-driven approach to survey design, the documentation work is done at a project's outset. This is the point when it is likely to be most accurate.

Enabling Data Harmonization

Colectica stores its structured metadata in a repository. A DDI 3 repository provides computer programs and Web sites a standard way to register, version, search for, and retrieve the metadata created by researchers.

For example, if one researcher creates a set of questions asking about a respondent's socioeconomic status and stores these questions in a repository, another researcher creating a survey can simply search for existing questions about the topic and include those same questions in her own study. This workflow results in a question bank that can be populated and accessed directly by researchers.

A DDI 3-based repository is much more than a question bank, however. In the same way that the researcher who uses a repository-based software tool stores questions in a repository, he stores all other elements of a survey's specification: classification schemes, concepts, universe definitions, survey instruments, and other metadata.

By storing these elements as identifiable, versionable, reusable objects, the potential for cross-study data harmonization is greatly increased. For example, researchers can quickly find all questions in a repository that are related to the "socioeconomic status" concept.

For a given question, researchers can find all surveys that include the question, along with all analysis datasets with variables that are derived from the question.

By connecting multiple repositories into a web of linked data, harmonization is enabled even further.

Colectica Repository can be configured so metadata are publicly available, or things can be restricted to certain individuals, groups, or organizations. Metadata elements managed in a repository can be licensed so that researchers who use another researcher's content agree to certain terms. For example, a researcher may wish to require that use of her content is attributed, or that it is not used for commercial purposes.

Greater Research Integrity

A key tenet of scientific research is that researchers should be able to independently verify the results of published research. Certain aspects of social science research, such as the need to protect respondent confidentiality, make this inherently difficult. However, even in cases where an independent researcher obtains permission to access the appropriate data, validation is often impossible because the original statistical code and relevant documentation are not available.

A metadata-driven approach to survey design, documentation, and analysis can prevent this problem.

In the same way that metadata elements related to data collection can be managed and published as part of a repository, so can metadata about analysis datasets. Datasets, variables, and other published data can be linked back to the source questions and concepts from which they were derived.

This full view of a dataset's life cycle greatly increases the likelihood that a study will be independently verified. This is the key to rigorous, data-based research.

Conclusion

A metadata-driven approach to survey design can lead to better data documentation with less duplicated work.

The DDI standard provides a framework for this metadata-driven workflow. With DDI-based metadata repositories, researchers can reuse metadata elements and create harmonized questionnaires and datasets.

By disseminating data along with documentation that covers the entire life cycle of a study, researchers can allow their findings to be validated.

Contact

The Colectica platform is a software tool for performing metadata-driven survey design and data management.

Colectica's developers have been involved in efforts to standardize survey research and data management practices since 2004. They have taught survey and data management workshops to survey researchers, data archivists, and statisticians from over ten countries.

For additional information about Colectica, the Data Documentation Initiative, metadata-driven survey design, and consulting services, please contact:

Jeremy Iverson
jeremy@algenta.com

Dan Smith
dan@algenta.com

http://www.colectica.com/