Creating Rich, Structured Metadata: Lessons Learned in the Metadata Portal Project

by Mary Vardigan[1], Darrell Donakowski[2], Pascal Heus[3], Sanda Ionescu[4], and Julia Rotondo[5]

Abstract

With support from the National Science Foundation, two long-running social science studies – the American National Election Study and the General Social Survey – partnered with the Inter-university Consortium for Political and Social Research (ICPSR) and NORC at the University of Chicago to improve their metadata and build demonstration tools to illustrate the value of structured, machine-actionable metadata. The partnership also involved evaluating the studies' data collection workflows to determine where in the data life cycle metadata could be captured at source to avoid metadata loss and costly procedures to recreate the metadata later. This article reports on the experience and knowledge gained over the course of the project and also includes recommendations for others undertaking similar work.

Keywords: Metadata, documentation, Data Documentation Initiative (DDI), data life cycle, data dissemination, tools, workflows

Background

The Metadata Portal Project, a collaboration among the General Social Survey at NORC at the University of Chicago, the American National Election Study at the University of Michigan, and the Inter-university Consortium for Political and Social Research, with technical support provided by Metadata Technology North America, was funded by the National Science Foundation (Collaborative Research: Metadata Portal for the Social Sciences, SES-1229957) under the Metadata for Long-standing Large-Scale Social Science Surveys (META-SSS) project to meet the following objectives:

- To develop rich, structured metadata compliant with the Data Documentation Initiative (DDI) standard for two premier time series studies in the social sciences — the GSS and the ANES
- To showcase tools that can be built upon the foundation of rich metadata
- To analyze and improve the projects' workflows
- To capture more metadata at the source

The two-year project resulted in enhanced study- and variable-level DDI markup for both data series as well as a portal linking to prototypes of several useful tools, including a robust search, a variable bank and shopping cart to generate subsets, a cross-study concordance and concept tagging tool, and

> Both groups maintain a "master" or "canonical" version of the questionnaire...

a tool that displays routing paths through a survey. Agreement was also reached to transition both data series to new workflows that enable the export of documentation in DDI format from computer-assisted interviewing systems.

About the Studies and Their Distribution

There are 58 separate studies comprising the ANES: the traditional biennial time series studies (with pre- and post-election surveys in years of Presidential elections)

going back to 1948; pilot studies; panel studies; and special studies of different types. Topics cover voting behavior and the elections, together with questions on public opinion and attitudes of the electorate. ICPSR and ANES are co-distributors of most of the ANES studies.

There is one cumulative file for the GSS that spans the years 1972 to the most recent wave (currently 2012). GSS content encompasses a standard core of demographic, behavioral, and attitudinal questions, plus topics of special interest. The GSS is distributed by NORC, the Roper Center, and ICPSR.

Project Activities

1. File Inventory

The first major phase of the project involved inventorying the relevant files held by all of the partners to ensure a shared understanding of the files in scope for markup and distribution. Some interesting findings were noted during the inventory:

- ICPSR distributed the data in more formats than ANES and had some existing DDI files
- For a few years of the series, ICPSR had grouped multiple data files into single studies, while ANES had kept them separate
- There were a few files that only the ANES was distributing
- ANES distributed in general more documentation than ICPSR did – e.g., they had additional documents on methodology
- Codebook content was basically the same across the distributing organizations, but the formats sometimes differed. Both ICPSR and ANES distributed PDF files, but other .txt files were sometimes available.
- It was noted that while the cumulative file had been the standard GSS product since 1977, ICPSR was still disseminating single-year GSS files for 1972-1977. It was recommended that ICPSR rethink its practice of distributing these single-year files as they had all been subsumed into the cumulative file and may have been revised. At the least, ICPSR was advised to add a note to the metadata records for these files to indicate that the data may have changed and that the cumulative file was the authoritative data source for those years.
- It was noted that the GSS cumulative file had long data records (5000 variables), which raised the issue of whether the file should be reshaped or subsetted to facilitate analysis

These findings and the differences noted across distributors led to a need to define which versions of the files to consider the "authoritative" versions going forward.

2. File Comparison and Defining Canonical Versions

To fully understand differences across the holdings of the distributing partners and to manage project content, a central source for the metadata was required; to that end, the project partners established an iRODS (integrated Rule-Oriented Data System[6]) based file repository. This provided a flexible central system for the file comparison and conversion work, with the capabilities to add metadata and rules to files, sort, search, get notifications of newly deposited files, etc. Organization and logic of the iRODS repository were critical, so Metadata Technology worked with the partners to structure the repository optimally; they also suggested file naming conventions for data and documentation that included the name of the series, type of file (e.g., pilot), and year.

To complement iRODS functionality, a central spreadsheet for all project studies was set up with identifiers and the capability to enter study-level metadata. Based on a model description provided, ICPSR's DDI XML metadata records were imported as a batch into the spreadsheet, and study-level metadata from the ANES and GSS websites were added as well. Metadata from these sources were later integrated.

Using Metadata Technology tools -- such as SledgeHammer and Caelum[7] -- as well as shell scripts, files were parsed, analyzed, and compared, with the following findings noted for ANES:

- File sizes differed across the ANES and ICPSR holdings, but this was to be expected -- in general the number of variables and frequencies agreed across the two organizations
- Study titles differed across ANES and ICPSR -- it was decided to use the ANES titles but to retain ICPSR titles tagged as alternative titles
- Differences, albeit small, were discovered in variable names, labels, category labels, etc. There were many variations in SAS code, most likely having to do with the way ICPSR and ANES produced the SAS scripts. In some instances the ANES version of the ASCII file for study was not fixed but delimited.
- For one study, ICPSR was distributing an older version of the data than ANES
- The ANES 2012 Time Series had just been released, and it was decided to include it in the time series covered by the project

Based on the file comparisons, the project settled on using the ANES version of the ASCII file and corresponding SAS syntax file as the master data/metadata to build the core DDI, updating the variable-level metadata from other sources for substantive differences, especially for the value labels. A decision was made to investigate only major/substantive differences and not typos, minor differences in labels, or file locations and widths, which would not be relevant since the project was designed only for metadata.

3. The DDI Markup Process

The goal of the project was to generate a complete library of DDI markup for all of the ANES and GSS with study-level metadata and variable-level metadata including basic variable descriptions, categories with values and labels, and frequencies, with additional variable-level information to be added when possible. To convert files to DDI format, Metadata Technology used its parsing and extraction tools to produce the core DDI markup in an automated way. Extensive effort also went into developing custom parsers for extracting metadata from legacy text files available on the ANES website, and combining various metadata sources into a final DDI XML document for each study.

Which DDI specification to use

Before the markup process could begin, a decision had to be made about which version of DDI to use, even though the tools available were agnostic as to the DDI version.

The DDI Alliance distributes two main product lines. DDI Codebook (DDI-C) is designed to include all of the elements of a typical social science codebook needed to facilitate effective data analysis. DDI Lifecycle (DDI-L) has a broader focus: To document and manage data across the entire life cycle, from conceptualization to data publication and analysis and beyond. In the end the project settled on DDI Codebook Version 2.5 for three main reasons. First, ICPSR had been using DDI-C for many years and already had existing study descriptions and some variable-level information in DDI-C. Second, most of what the project aimed to accomplish with the DDI metadata library could be done using the simpler of the standards. A final rationale was that Version 2.5, which had recently

been released, was seen as a bridge to DDI-L if a conversion to the more complex standard were required later. Other projects making similar decisions might consider these three factors as they determine which DDI product line will best suit their needs.

Automating markup

As noted, Metadata Technology used parsing tools, custom development, and scripts to produce the markup in an automated way as much as possible. This was not always straightforward, however. ANES OSIRIS-like codebooks had a lot of rich detail at the variable level that the project wanted to capture, including interviewer instructions, lead-ins to questions, forward and backward question flow, etc., but automating the conversion of these different variable components was difficult as the formatting was not always uniform. PDF format was another barrier to markup as each file had to be converted to editable text to extract the needed metadata. Identifying patterns in the text so that "families" of study documentation could be parsed together in a more efficient way was an effective solution for some of the heterogeneity encountered. In the end most of the markup was done programmatically with manual markup performed when necessary.

In terms of content to include at the variable level, the project ultimately used the following elements:

- Variable name
- Variable ID
- Variable label – short
- Variable label – long
- Variable group
- Literal question
- Summary statistics
- Category label
- Category value
- Category frequencies
- Notes (substantive notes relevant for data analysis)
- SHA1 hash for question text and value labels

A major innovation was that MTNA computed a hash for variable classifications based on a string composed of all codes and categories. This hash would enable reuse of identical classifications and also facilitate conversion from DDI Codebook to DDI Lifecycle.

4. Building Tools

Database and search

At the end of the processes described above, new DDI metadata were produced for 58 ANES surveys (79,521 variables) and the cumulative GSS 1972-2012 dataset (5,558 variables). HTML reports were also produced for each study. These metadata were loaded in a BaseX[8] database for querying and retrieval and indexed with Apache Solr[9] to facilitate full and faceted searches. Both systems are available over a public REST API. A web-based application leveraging these two services was built as a proof of concept. The tool allows users to search and select variables and collect them in a shopping basket, which can in turn be used for generating data subsetting scripts for SPSS/SAS/Stata and producing customized codebooks in HTML/PDF.

Visualization of question routing

The project included an exploratory effort to capture/document question flow through an older survey. The process for marking up this information involved selecting a study to use as a test case, tagging the documentation in DDI, and running an ICPSR-created tool called RUG (Reverse Universe Generator) to capture the flow. The markup to highlight "system missing" had to be done manually in conjunction with the codebook as the available syntax did not carry this information.

Input to the tool was an ASCII data file and DDI 2.5 variable descriptions. The DDI variable-level metadata had system missing values flagged at the category level. The DDI "missing type" attribute on the category element was used. The assumption was that for each relevant variable a unique code was assigned to system missing values, and no other types of missings (DK, NA, etc.) on the same variable carried the same code. The system missing flagging was done manually to create a working input for the tool. (This could be automated if the same code, and preferably the same label, were consistently assigned to system missing values across a dataset.) The weight variables were also flagged using the attribute "weight" on the variable element. This markup was also done manually, but was not so onerous, as there were a limited number of weight variables in a typical dataset.

The RUG tool identified the system missing code on each given variable and then regressed that variable on all of the other variables to find perfect matches between the system missing code and other codes on the searched variables. It only looked for single-variable dependencies and nested variables. It did not check for complex universe logic based on multiple variables.

RUG generated variable-level universe information (DDI-C universe element, a child of the variable element) pointing to the source of the dependency as well as the relevant categories involved.

It assumed that "correlation equals causation" and automatically created universe metadata when a correlation was found between the system missing values of a particular variable and one or more categories of another variable.

In some instances, the independent variables were not found. A closer analysis of the data and original documentation (used in creating the DDI) showed that on quite a few variables the system missing category included cases that were not truly system missing, but represented responses like 'no comment', 'no second mention', 'no pro or con', etc., from respondents who were actually asked the question. This was an important finding that directly impacted the expected tool output: for a satisfactory output, the input data and documentation need to contain accurate system missing information, assigning unique codes for system missing values and documenting them in an unambiguous way.

The overall assessment of the RUG tool effort was that it was an interesting experiment but that RUG would be of limited use as a production tool, given the constraints related to the data and metadata input. Legacy datasets would not be good candidates for the tool, as they would require case-by-case evaluation as well as metadata editing and perhaps even data reprocessing, adding to the time spent. The RUG experiment highlighted the importance of high quality, complete, and accurate data documentation and clean datasets. The major lesson learned was that question routing markup should ideally be part of the documentation deposited with archives to avoid this costly retrofit work.

Concept comparisons

The project also investigated how ANES and GSS operationalize some important concepts by building an integrated crosswalk of concepts. The concept comparison was generated by first using the ANES cumulative file and then looking at the GSS for comparable concepts. Since the ANES cumulative file groups variables in just a few broad topics, the work was expanded to

include the ANES Core Utility, which offers more granularity although it only covers recent time series going back to the 1990s. Related to this work on concepts, a prototype concept tagging tool was built. This permits the individual user to tag variables by concept and then build a crosswalk to compare variables over time or across different studies. It is also possible to create public lists so that an organization can apply its own authoritative tagging to its content and make it publicly available.

5. Re-envisioning the Process: Markup at the Source

The process to mark up legacy documentation chronicled above was laborious and time-intensive, involving a large team of people, specialized tools, and manual work. Much of the work performed, e.g., adding question text to variables, was in essence restoring information to its original state as found in the CAI interview environment.

A logical solution to avoiding this scenario in the future is to capture the metadata at the source – from the original CAI instrument -- and export to XML when the data are exported from the interview software. This would eliminate costly work on legacy materials as metadata would be harvested once at the source. To explore this idea further, the grant included funding to hold a workshop for all three partners to explore changes in the workflow of data collection and dissemination and how the projects might transition to capturing metadata at the source.

The focus of the meeting, held April 24-25, 2014, in Ann Arbor, Michigan, was for ANES and GSS to share their processes and compare them, applying what had been learned about creating metadata over the life of the project in order to capture more and better metadata – ideally, with less work. ANES and GSS staff made presentations about their surveys describing workflow processes, opportunities for capturing metadata and paradata, and challenges they faced in creating a harmonized process for the ANES and GSS surveys.

The ANES discussion focused on several of the new tools created to enhance researcher and staff experience with the survey. Specifically, the Questionnaire Development Tool sparked discussion of the workflows involved in creating the Time Series questionnaires of the ANES. This tool permits collaboration on the questionnaire: users of the tool can draw from a pool of questions previously used and construct new questions. They can specify question provenance, randomization, timings, etc. The group also discussed the two main survey databases used by ANES staff – the questionnaire database and the variables database – to facilitate the creation of the questionnaire and the ultimate codebook. The questionnaire database feeds into the variable database, as do observations and notes from the field. ANES staff want to include as much information as possible in these databases in order to generate a subset for a well-constructed codebook, but often the metadata information they need from the data collectors is not automatically provided – ANES staff must push for its release. It was noted that mappings between DDI and these two internal ANES databases would be worthwhile.

Though the type of paradata desired by ANES staff may be consistent – timing of questions, timing of mailings, etc. – the systems and formats used to document these events are not consistent across vendors, nor is the functionality always immediately usable. For example, data collectors can provide timestamp information on every key stroke but that doesn't necessarily help researchers understand how long each question took to administer and answer. The ANES process during data collection involves very close monitoring by the ANES staff, and the staff works on the documentation continuously.

The GSS discussion focused first on a high-level overview of the GSS workflow process over the three-year cycle, describing areas of overlap in the pre-production cycle of the next round and the post-production cycle of the current round. Discussion then moved to NORC's early attempts to map the GSS to the GSBPM (Generic Statistical Business Process Model), a reference model developed by national statistical organizations from around the world to standardize the production of official statistics. This mapping work involved interviewing NORC GSS staff, doing an environmental scan of GSS dissemination sites to examine the types of metadata available to researchers, and creating a visual flow of the GSS work process.

During the discussion on creating post-production documentation, the group found similarities between the ANES and the GSS, though the processes were different. Both groups maintain a "master" or "canonical" version of the questionnaire, which is then used to check data collected from the field. Actions are taken when deviations are found – either by adding notes in the databases for the codebooks in ANES or by doing post-production checks against the CAPI in GSS to see if the discrepancy was due to a mechanical failure that can be resolved easily or not.

Metadata capture – and specifically where in the process it is captured – was a theme that arose frequently during the workshop. For both the GSS and ANES it appeared that most metadata were entered into codebooks and databases manually rather than through any automated process. Both surveys recorded metadata during data collection, but this information was often ad hoc observations and did not always follow specified pre-planned processes. ANES discussed a previous round where they had created rigid procedures for call note documentation for the data collectors. They found that with a strong process in place, they got very useable data – but this was the result of a lesson learned from a previous request for call note documentation that led to unusable information (not consistent, incomplete, etc.) being delivered.

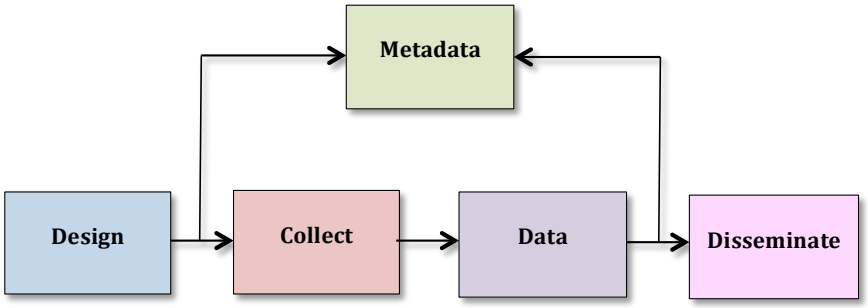

Figure 1: Current workflow

Three draft high-level harmonized workflows were created during the workshop to better understand the existing processes and to re-envision new processes. The standard case with metadata as a final input but not really incorporated into the system until the dissemination phase appears in Figure 1:

Then an ideal workflow with a data and metadata done in sync was brainstormed:

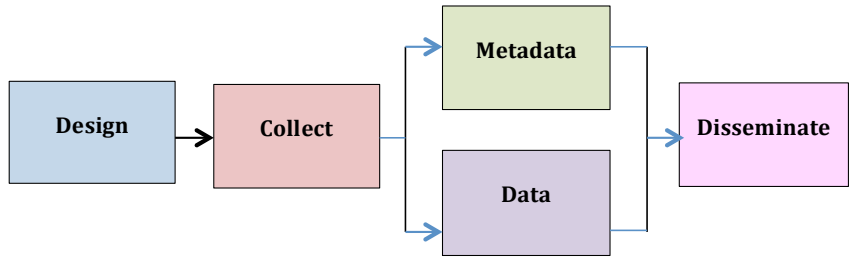

Figure 2: Future workflow

The group discussed potential delays in the design phase, benefits from having data and metadata being generated simultaneously though remaining conceptually separate, and how useful the workflow shown in Figure 2 would be to the end-user – for example, if ANES specified the type of metadata they wanted from their data collector, it might not be formatted in a useful way for the researcher.

A third version of the workflow was offered that brought paradata and auxiliary data into the process:

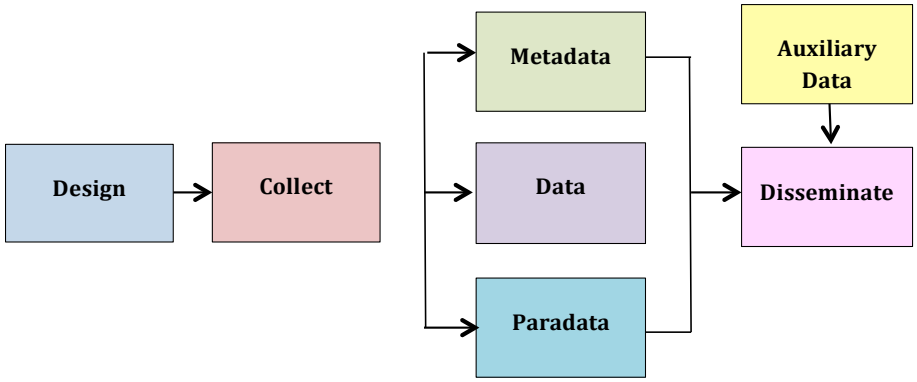

Figure 3: Future workflow with enhancements

6. Capturing data transformations

Under the auspices of the Metadata Portal project, a second meeting was convened to discuss the creation of a tool that would update DDI XML metadata when the associated data file changed. Such a tool would be necessary if the above future workflow were implemented. Projects often make changes to the data after they are exported from CAI software, and there is a need to keep the metadata synchronized as data are transformed. Also important is recording the provenance and history of changes to the data in the documentation so that users are informed of data transformations over time.

The meeting was comprised of technologists and developers, most of whom had already created tools to work with DDI XML. A consensus emerged at the meeting that the group should develop a Standardized Data Transformation Language (SDTL) that would be neutral in terms of statistical software packages. This would enable the creation of a tool to capture data changes and publish them in the synchronized documentation. The tool could be invoked at any point in the data life cycle when a codebook was needed.

Good progress was made on the SDTL during the meeting and plans to pursue funding for development of the tool were communicated and discussed.

Conclusion and Next Steps

Project participants learned a lot during the course of the project about the challenges of converting legacy documentation to machine-actionable form. This is labor-intensive work that should only be done once. The goal for the future is to capture machine-actionable metadata from the source and to have this marked-up documentation deposited in archives. We also hope to encourage others to leverage the XML documentation produced to create new tools beyond those created for the Metadata Portal. Over the project period, the partners identified several next steps, described below, to carry on the work begun by the collaborative research project.

Exporting DDI XML from CATI-CAPI programs

As shown in the graphics above, the ideal workflow involves exporting DDI XML along with the data from the CAI system and then maintaining them in parallel through to dissemination. One way to accomplish this goal is to require export from the CAI systems when commissioning a survey. ANES goes through a bidding process to identify the data collection firm for each wave of the survey, so they could stipulate in the request for bids that the data be delivered along with DDI XML documentation. GSS uses an internal system at NORC for data collection but could also work on an XML export. Both partners expressed an interest in transitioning to this kind of process, starting with the next data collection cycle.

Interestingly, after the project ended a related effort -- the "Survey Metadata: Barriers and Opportunities" Meeting held June 26, 2014, in London -- resulted in a published DDI profile for questionnaire documentation as well as a collaborative statement calling upon the survey design, production, and archiving communities to take leadership in facilitating survey metadata exchange through adoption of shared metadata standards for questionnaire and data description. The profile and statement for endorsement are available at http://www.ddialliance.org/survey-metadata-reusability-and-exchange. This kind of best practice supports and

validates the vision for the future coming out of the Metadata Portal project.

Documentation data transformations

Work on a Standard Data Transformation Language (SDTL) was started but is not yet complete. There is a commitment from the participants in the initial meeting on this topic to continue to develop the SDTL as it is an essential foundation for tools to capture provenance and data transformations across statistical packages. A meeting to continue the work will likely take place in 2015

Exporting DDI XML from ANES databases

A mapping for the ANES codebook database has been completed, and the next step would be to use the mapping to export DDI XML from the database. This could also be done with the ANES Questionnaire Development Tool. If the questionnaire were marked up in DDI, this could serve as input to the CAI process.

Capturing paradata

The workshop involving the project partners had a strong focus on paradata – which paradata items the ANES and GSS capture, how they use paradata internally, and which types of paradata they make available or would like to make available. It was decided that this is a fruitful area for further exploration and collaboration.

Resolving versioning issues

One of the key lessons learned on the Metadata Portal project was the importance of establishing canonical versions of the data for the two data series. ANES and ICPSR are co-distributors of the ANES data, while NORC, the Roper Center, and ICPSR all distribute the GSS. This situation with multiple distributors means that the data can easily get out of sync.

As we learned during the workshop held to analyze the workflows and business processes of the ANES and the GSS, the GSS workflow is particularly vulnerable to versions being out of sync. NORC sends its final file to Roper, which makes some changes to the data (related to missing values) to integrate it with their iPOLL database. ICPSR then gets the data from Roper. NORC may make changes and issue errata, but these changes are not pushed out to the other distributors. While we are not aware that this situation has resulted in divergent analytic results because of different files being used, it is possible that this is occurring and we simply do not know about it.

This is a wider problem that affects any dataset with multiple distribution points. For example, ICPSR and its counterparts in Europe co-distribute several datasets, giving rise to the potential for different versions in circulation. With open access to data becoming the norm around the world, this problem is likely to escalate.

The participants on the Metadata Portal project considered a range of solutions to address this issue. A simple fix is for the GSS to push out notifications when there is an updated GSS file available. ANES integrates version numbers into its data files as variables, which is another simple solution. Also needed are tighter versioning controls and rules for these series and persistent identifiers to the data to uniquely identify them. This should be coupled with online access to all versions of the data for replication purposes.
To complement these fixes, creating a checksum registry to hold the authoritative version of the data could be useful. Other distributors could compare their versions with the authoritative version to ensure that they have identical files. One possibility is to explore the use of the Universal Numeric Fingerprint[10] as the checksum. This is a solution that might have wider adoption. Ultimately, the best solution is for the co-distributors to send users to a central source for data downloads, but this will take some time and cultural and technological changes to implement for ANES and GSS.

Resolving differences between portal metadata and ANES/ICPSR metadata

While, as noted above, a central authoritative version of all files is the end goal, during the transition to that state study- and variable-level metadata available on the Metadata Portal will differ from what ANES and ICPSR currently provide on their websites. We will need to address this issue with the goal of minimizing the number of different versions and confusion for users. It is likely that ANES and ICPSR will need to update their collections to incorporate metadata enhancements that resulted from the project.

References

American National Election Studies. http://www.electionstudies.org/. Accessed October 1, 2014.

General Social Survey. http://www3.norc.org/GSS+Website/. Accessed October 1, 2014.

Generic Statistical Business Process Model. http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model. Accessed October 1, 2014

Inter-university Consortium for Political and Social Research. http://www.icpsr.umich.edu/. Accessed October 1, 2014.

Meta-SSS site. http://metasss.mtna.us/. Accessed October 1, 2014.

Metadata Technology North America. http://www.mtna.us/. Accessed October 1, 2014.

Roper Center for Public Opinion Research. http://www.ropercenter.uconn.edu/. Accessed October 1, 2014.

Notes

1. Mary Vardigan is an Assistant Director at ICPSR. Email: vardigan@umich.edu
2. Darrell Donakowski is Director of Studies for the ANES. Email: dwdonako@umich.edu
3. Pascal Heus is Vice President of MTNA. Email: pascal.heus@metadatatechnology.com
4. Sanda Ionescu is a Documentation Specialist at ICPSR. Email: sandai@umich.edu
5. Julia Rotondo is a Senior Research Analyst at NORC. Email: Rotondo-Julia@norc.org
6. http://www.irods.org
7. http://www.openmetadata.org/site/?page_id=362
8. http://www.basex.org
9. http://lucene.apache.org/solr/
10. http://thedata.org/book/universal-numerical-fingerprint