Mapping the General Social Survey to the Generic Statistical Business Process Model: NORC's Experience

by Scot Ausborn, Julia Rotondo and Tim Mulcahy[1]

Abstract

As a part of the Metadata Portal Project, with support from the National Science Foundation, NORC mapped the General Social Survey workflow to the Generic Statistical Business Process Model (GSBPM) to determine where in the survey cycle DDI-based metadata could be more effectively captured. Lessons learned from the process include a better understanding of utilizing the flexibility of the GSBPM model and a recommendation to collect paradata in a collaborative, facilitated workshop rather than mapping responses from individual staff. Information gained from the mapping has proven useful in identifying areas of metadata and paradata collection enhancements.

Keywords:
Metadata, paradata, Data Documentation Initiative (DDI), Generic Statistical Business Process Model (GSBPM), workflows

Background

The Metadata Portal Project, funded by the National Science Foundation as part of the Metadata for Long-Standing Social Science Surveys (META-SSS, SES-1229957) initiative, is a collaborative effort 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 assistance from Metadata Technologies North America (MTNA). The project's objectives are:

- 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 primary deliverable of the project is a web-based portal leveraging DDI-compliant metadata to provide a range of new tools for researchers working with GSS and ANES data. The portal incorporates an enhanced search engine for both datasets, comprehensive variable and concept banks, and a subsetting feature for generating custom datasets. One of the project tasks supporting the creation of the metadata portal – and intended to sustain it going forward – is an analysis of the business processes surrounding the production of survey data to determine where in the survey cycle DDI-based metadata may be captured to avoid having to generate it retroactively. By taking this initial step

> The GSBPM is a schema for parsing statistical production workflow that consists of nine high-level processes

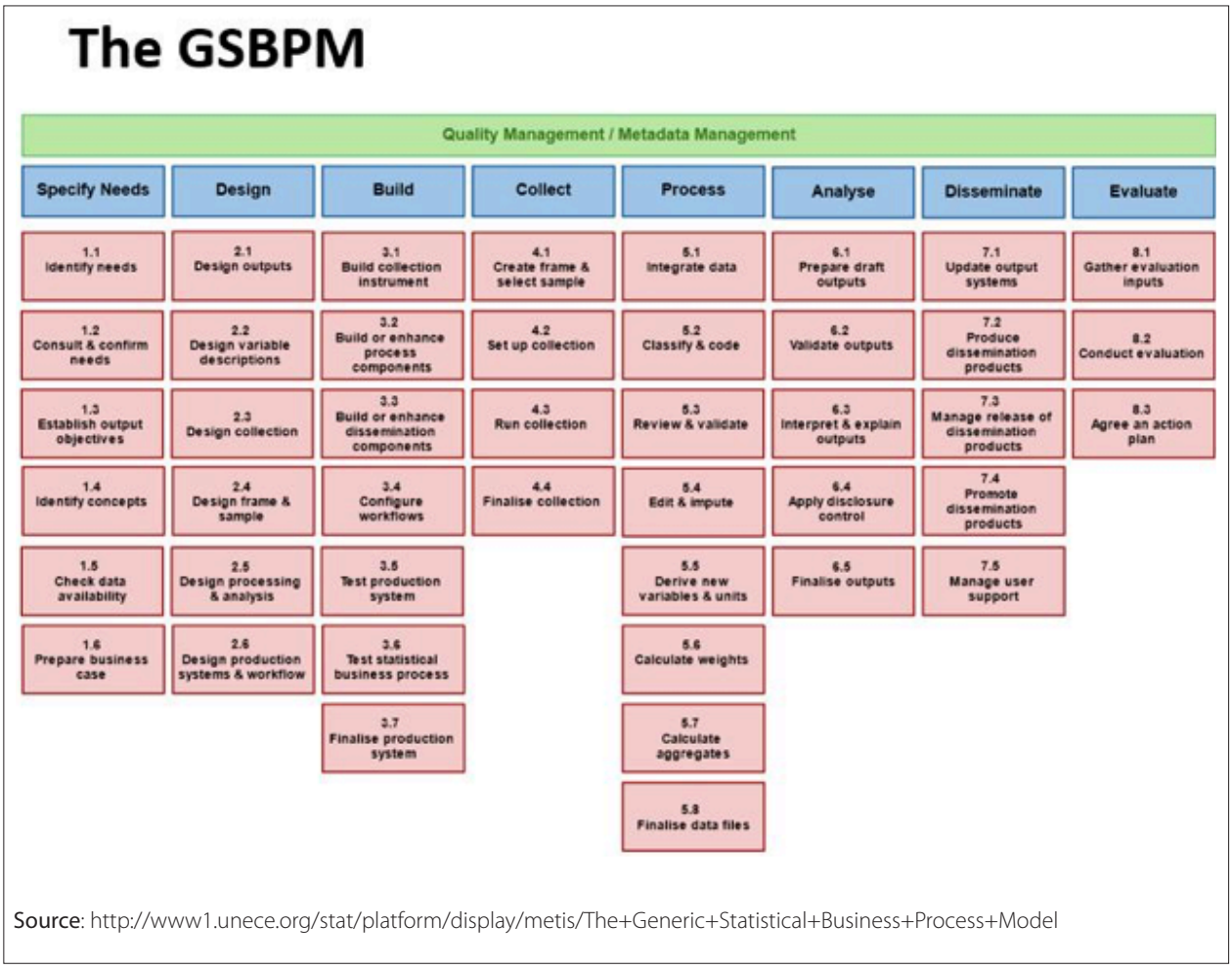
toward a DDI-based workflow, the goal is to enhance the metadata available to researchers in the portal and to realize greater efficiencies in the survey cycle itself by identifying redundant processes, such as duplicate data transformation, that could be remediated with a metadata-based approach.

The GSBPM

In order to better understand the workflow processes associated with the production of the General Social Survey (GSS), NORC conducted a survey of internal GSS staff asking them to explicate their respective roles on the survey in terms of the Generic Statistical Business Process Model (GSBPM)[4] . The GSBPM is a schema for parsing statistical production workflow that consists of nine high-level processes and several sub-processes under these.:

sub-processes. As a result, the questionnaire followed a highly structured format. Respondents were asked to provide detailed information regarding the facets (inputs, outputs, actions, tools, etc.) of each sub-process in addition to a brief overview of the sub-process itself.

One of the challenges of using the language provided by the GSBPM is that it is highly abstract, requiring some deduction to understand how the process being described corresponded with internal GSS processes. Thus in creating the questionnaire, additional explanatory text was required to help tailor GSBPM language to GSS specific processes. A GSS staff member with experience in several different aspects of the survey workflow was essential in order to create the additional explanatory text. Similarly,



Source: http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model

Correlating aspects of the GSS workflow to elements of the GSBPM allowed NORC to gain a comprehensive and integrative view of the individual efforts that together produce the survey. Additionally, gathering the GSS paradata in this manner also facilitated the identification of processes in the workflow where metadata relevant to dissemination and discovery of the survey data is potentially being lost or ineffectively captured. By identifying and remediating these points, it is intended that the survey be produced more efficiently while better meeting the needs of researchers analyzing the data

Questionnaire design

The questionnaire distributed to internal GSS staff (hosted online at http://dataenclave.org/gss) was adapted directly from the language used in GSBPM descriptions of individual processes and

the volume of GSBPM sub-processes required the selection of appropriate respondents for a particular sub-process in order to prevent survey fatigue.

On the technical side, it was determined that web-based dissemination of survey questions would best facilitate data collection and analysis. To implement this NORC used a standard LAMP-stack design, with the webpage coded in standard HTML/CSS and data stored in a MySQL database using PHP. The database schema used a separate table for each sub-process, with the respective columns storing the respondent's ID, an overview of the process from their perspective, and the different facets of that process.

Survey execution

Prior to the survey link being distributed to the respondents, an email from the Senior Vice President of the department producing the GSS was sent to reinforce the value of the survey and the expectations of completing it. Once the survey link was distributed, respondents were given one week to enter the information for the specific sub-processes that were assigned to them. Respondents were allowed to quit the survey and return later; however, they were unable to retrieve previously saved answers at a later time. In retrospect, allowing respondents to login and retrieve saved answers would have been helpful for respondents, but needed to be balanced against time to develop and implement. Another possibility for achieving this functionality would have been to set a cookie in the respondent's browser.

Given the direct request from senior management to respondents, the survey garnered a high response rate. However, the responses indicated that some respondents may not have been targeted well, with a few stating that the sub-processes they had been assigned to provide information for were not part of their work with the GSS
.

Collecting, compiling, and cleaning internal survey responses

Once the survey collection period had ended, the survey responses were downloaded from the MySQL database to a CSV file. From there, the file was opened in Excel and cleaned to ensure that within a sub-process each respondent's answers were only captured once. Some respondents had encountered technical difficulties with the form, mostly due to browser compatibility issues, and ended up submitting the same answers in excess of five times for the same sub-process. Responses that did not add new information to the survey (e.g., a respondent entering "skip" or "N/A" into the comment box) were deleted from the file as well. Finally, formatting was introduced to improve legibility of the responses for analysis.

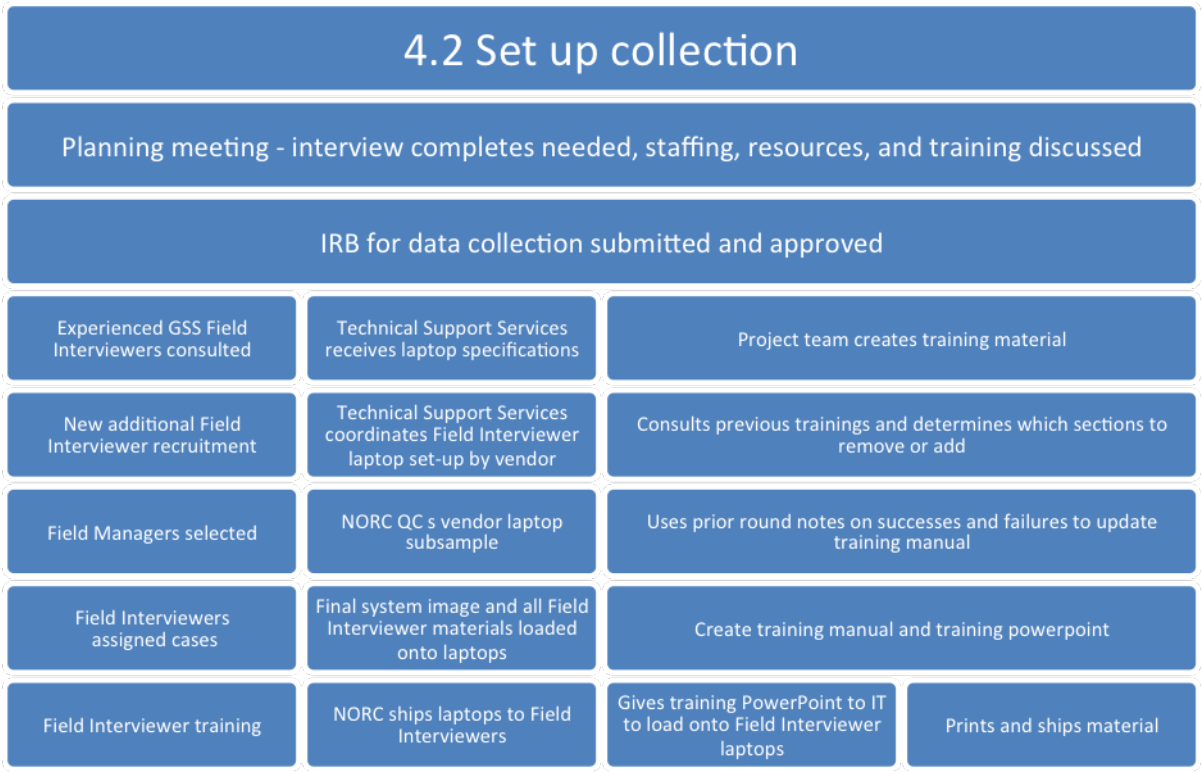
Mapping to the GSBPM

After the responses had been cleaned, the NORC team examined the responses given for each of the GSBPM sub-processes and attempted to create a comprehensive overview of the process. Challenges became immediately apparent in the process, including:

- Determining if responses truly belonged in the sub-process they were placed in by respondents
- Determining what happened in cases where no responses were given for sub-processes
- Within the GSS, it was often the case that multiple sub-processes were happening simultaneously while in the model they occurred linearly

When consolidating the survey responses into an overview, research analysts noted that respondents would often reply to one sub-process with information that might better fit in another. For example, some responses were submitted under the 2.5 Design Statistical Processing Methodology that upon review seemed to be a better fit for 2.2 Design Data Collection Methodology. In instances such as these, the responses were moved to the new section, but annotated so that the team could track how responses had moved. Responses were moved because while the respondents were the experts in the GSS, they were not as familiar with the GSBPM as the research analysts working on mapping the GSS to the GSBPM. Therefore, while the team worked to ensure that all information submitted to the GSBPM was included in the combined workflow overview, if the responses given to a particular section seemed a better fit to another section, the team decided to move the response.

Other challenges included having no responses for parts of a sub-process or entire processes. For example, NORC staff did not submit any responses for many of the sub-processes within the 1.0 Specify Needs section of the GSBPM, including 1.3 Establish Output Objectives and 1.4 Identify Concepts. Because there



were no responses, the team did not include these sub-processes in the overview.

Finally, a difficulty was that one sub-process within the GSBPM was sometimes too broad to clearly show multiple processes working simultaneously. For example, within the GSBPM sub-process 4.2 Set up collection, many distinct action items are performed by different GSS teams to accomplish this task. The responses revealed that while certain common steps occurred within that sub-process, it actually contained three separate team processes – each with their own steps, paperwork, inputs, outcomes, and purposes. The team kept all the responses together within the sub-process narrative to maintain cohesion within the GSBPM, but as can be seen below by the diagram of the sub-process, it was not a natural fit.

Lessons Learned

Overall, the challenges faced by the research team in mapping the GSS workflow to the GSBPM can be traced to a rigid adherence to the GSBPM model. The research team began the process of collecting paradata with the model and then asked GSS staff to discuss their processes within its framework, leading to poorly fitting sub-processes – where some sub-processes are empty while others are so full that they lose clarity. In retrospect, a better process might have been to start by asking GSS staff to detail their process, map out the steps, and then see how that process compared to the model. In that respect, the NORC team did not fully exploit the main benefit of the GSBPM – namely that the tool is meant to be a customizable starting point rather than a rigid endpoint.

Going forward, if NORC were to conduct this study again, we would hold a workshop in which GSS staff would be able to engage with one another and discuss the workflow processes rather than having each person provide his or her input in isolation. Furthermore, it would be highly beneficial to have an expert in GSBPM (or perhaps the complementary Generic Statistical Information Model) to conduct the mapping of workflow rather than asking staff members to conceptualize their work in terms of the abstract language provided by the GSBPM model. By doing this, staff would simply describe what they do, rather than reacting to a question or sub-process that might be interpreted as having no relevance to their work.

Nevertheless, while the method of gathering GSS paradata had its difficulties, the information gleaned from this study has proven useful in terms of identifying points in the workflow where metadata might be enhanced as well as how the collection of paradata using a GSBPM or similar model might be improved upon. It is the desire of the NORC team that this experience should prove instructive for other institutions wishing to conduct a workflow study of their own statistical production processes.

References

Andritsos, Periklis and Keilty, Patrick (2014) Level-Wise Exploration of Link

Notes

1. Scot Ausborn is a Systems Engineer for the Data Enclave at NORC. Email: Ausborn-Scot@norc.org
2. Julia Rotondo is a Senior Research Analyst at NORC. Email: Rotondo-Julia@norc.org
3. Tim Mulcahy is the Program Area Director of the NORC Data Enclave. Email: Mulcahy-Tim@norc.org
4. http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model