GENERIC LONGITUDINAL BUSINESS PROCESS MODEL

By Ingo Barkow, William Block, Jay Greenfield, Arofan Gregory, Marcel Hebing, Larry Hoyle, Wolfgang Zenk-Möltgen

3/15/2013

DDI Working Paper Series – Longitudinal Best Practices, No. 5

This paper is part of a series that focuses on DDI usage and how the metadata specification should be applied in a variety of settings by a variety of organizations and individuals. Support for this working paper series was provided by the authors' home institutions; by GESIS - Leibniz Institute for the Social Sciences; by Schloss Dagstuhl - Leibniz Center for Informatics; and by the DDI Alliance.

Generic Longitudinal Business Process Model

DDI – DOCUMENTING THE HELIX

FORWARD

This paper is the product of one of the three working groups at Dagstuhl event 11382. The group was charged with producing a reference model for the process of longitudinal data production and use, with an emphasis on the specification and management of the supporting metadata. This model is designed to be useful for the gamut of study types where data are collected across time, including panel studies and repeated cross-sectional studies. It should also be useful for single cross-section studies.

INTRODUCTION

The intention of this document is to provide a generic model that can serve as the basis for informing discussions across organizations conducting longitudinal data collections, and other data collections repeated across time. The model is not intended to drive implementation directly. Rather, it is primarily intended to serve as a reference model against which implemented processes are mapped, for the purposes of determining where they may be similar to or different from other processes in other organizations. It may also prove useful to those designing new longitudinal studies, providing reminders of steps which may need to be planned.

This is a reference model of the process of longitudinal and repeat cross-sectional data collection, describing the activities undertaken and mapping these to their typical inputs and outputs, which would then be described using DDI Lifecycle.

With early roots in the social sciences, this model is grounded in human science. Elements such as *anonymizing data* (step 5.8 in Figure 5) and *managing disclosure risk* (step 8.6) relate directly to research on people, whether a biomedical study or a study on political attitudes. The model was developed with longitudinal surveys being the archetypal study type so many of the examples in this paper relate to surveys. Nevertheless, the model described here is intended to be applicable to a wider range of study types. This model should be just as applicable to a longitudinal series of experiments as a survey (see Block et al. 2011).

This model is not intended to be comprehensive. It is intended to be descriptive of a generalized view of longitudinal data collection. This model may be extended or specialized to describe specific processes within an organization. Appendix A provides one example of extending this model by incorporating elements from another process model.

Relationship to Previous Work

This document was produced at the 2nd (2011) Dagstuhl Workshop on Longitudinal Data. It builds upon the materials produced at the 2010 workshop on the same subject among which was a high level process model describing the relationships between different waves of an ongoing data collection (see Hoyle et al. 2011). These were depicted in the diagram shown below (see Figure 1).

In this diagram we see a high level process depicting concept, collection, processing, analysis, distribution and discovery activities. This paper elaborates upon that high level view, adding a deeper level of detail. There is a cyclic aspect to the high level model, as data collection waves are conducted. Earlier waves will impact succeeding waves, and we have attempted to show where these interactions take place in terms of the metadata associated with collection activities.

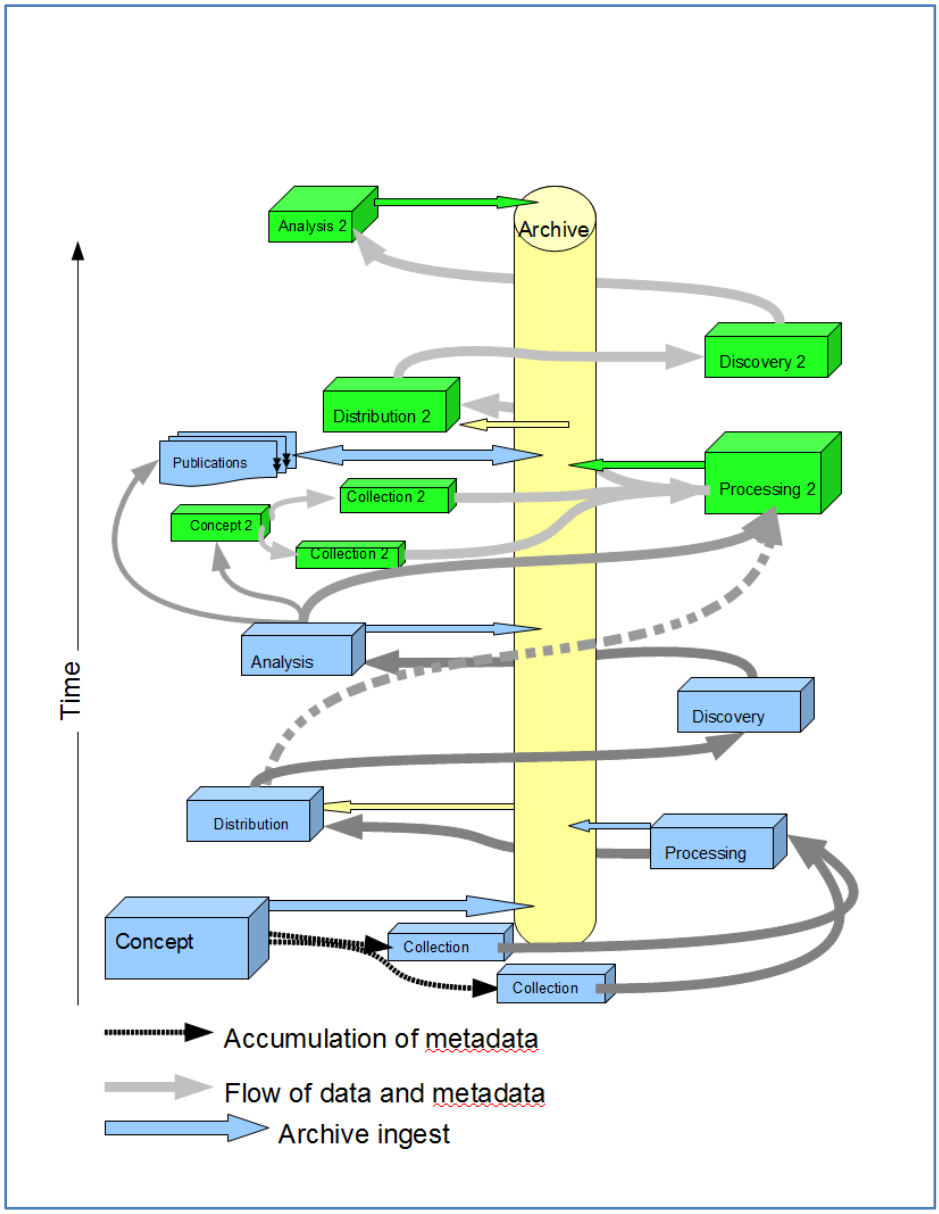

Figure 1. A helical view of the data and metadata lifecycle

Relationship to the DDI Lifecycle model and the GSBPM

DDI Lifecyle contains a high-level model of the data life cycle (see Figure 2). This high-level model was used in developing the Generic Statistical Business Process Model (GSBPM)[1]. The GSBPM is a model developed for statistics agencies to allow them to define and compare processes within and between organizations. It is a much more detailed reference model describing statistical production within the official statistics domain. The GSBPM uses a non-linear style which was found to be appropriate for modeling longitudinal data collection (see UNECE Secretariat 2009). Our model, the GLBPM, takes the approach of having a non-linear path through a matrix of alternatives, as in figure 6 below, directly from the GSBPM.

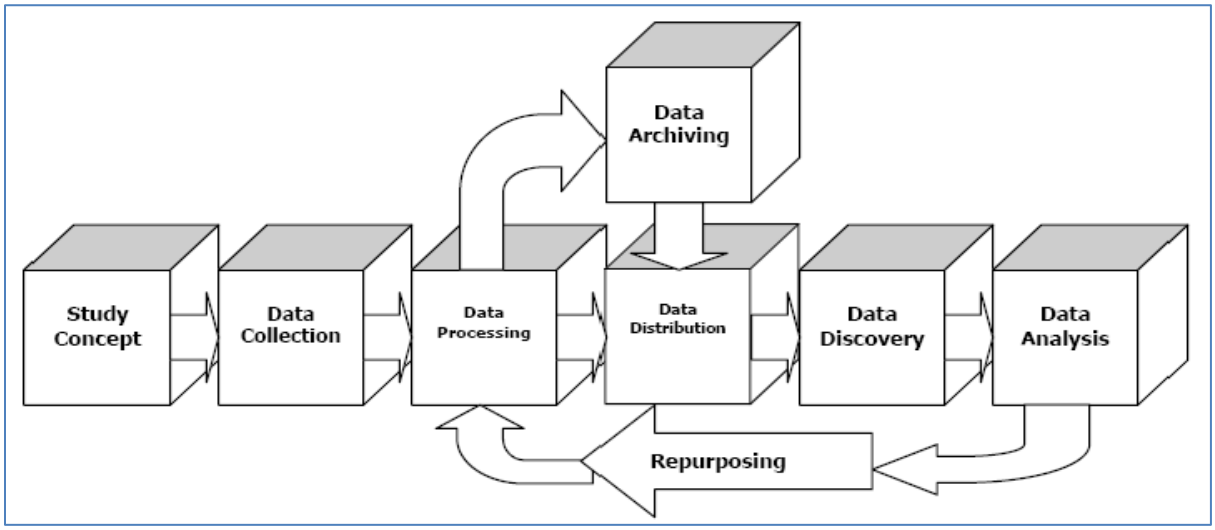

Figure 2. The DDI combined lifecycle model (taken from Data Documentation Initiative (DDI) Technical Specification, Part I: Overview Version 3.1 October 2009)

The model presented in this document has many similarities to the GSBPM, although it differs in some specific activities as a result of different practice in the real world and also as a result of the different terminology used in the social sciences. Furthermore, in social science research, data are commonly collected with specific analyses in mind. These must be integrated into early planning stages. Because GSBPM is based on the DDI Lifecycle model and because the model presented here is based on GSBPM, it is possible to map it against the DDI Lifecycle model. This is shown in the figure below.

[1] For more information about the GSBPM see:
http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model

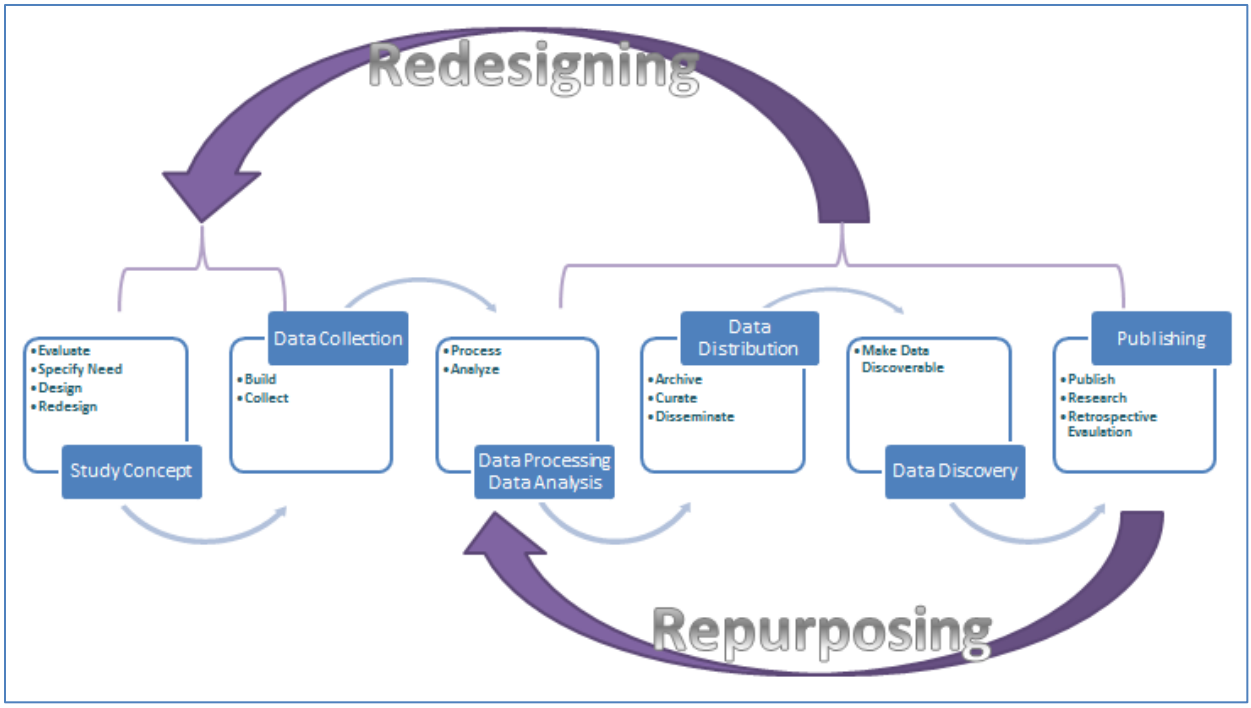

Figure 3. Mapping of the GLBPM against the DDI combined lifecycle model

HIGH LEVEL VIEW OF THE MODEL

The highest level of the model (Figure 4) presents nine steps roughly organized around time. These depict a single wave of data collection within a repeated study. These steps are used to conceptually organize more detailed sub-steps (Figure 5), which are not necessarily organized in a linear (or uni-directional) fashion when describing an actual process. In other words, many different paths through the model in Figure 5 are possible (see Figure 6). The nine high level steps are presented in the diagram below.

Beneath the nine steps are other processes that are significant to the data collection process and which occur throughout. These processes are not the focus of the GLBPM.

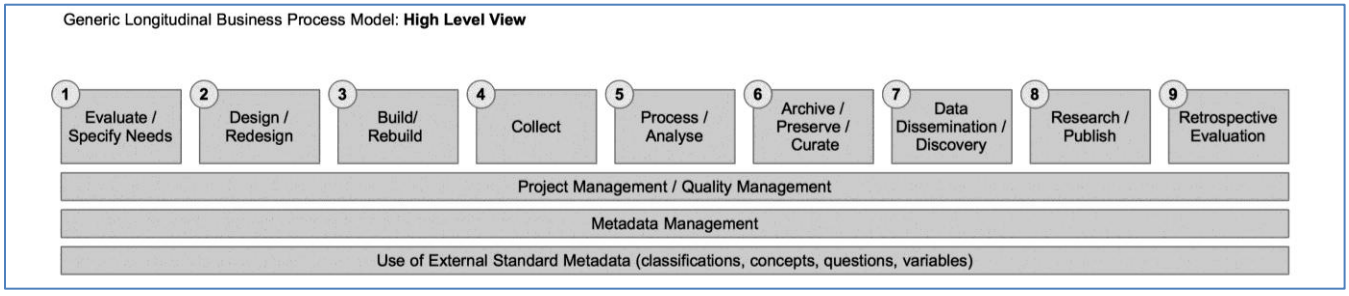

Figure 4. Generic Longitudinal Business Process Model – High level view

In general terms, the products of each high level step can be summarized as follows:

Step 1: Study design

Step 2: Methodological design

Step 3: Instrument and documentation

Step 4: Raw data/metadata

Step 5: Processed data/metadata and logical data products

Step 6: Published[2] and migrated versions of Step 5 products

Step 7: Physical data products

Step 8: Citations and publication

Step 9: Assessment report, modification plans

THE MODEL OVERVIEW

The following diagram presents an overview of the GLBPM. This is a non-linear model. The high level view (the boxes across the top, numbered 1-9) represents a series of steps that are organized across time in a general fashion. The sub-steps (organized in columns below the high level steps) represent possible activities within the high level steps. This presentation is intended to allow a specific process to be mapped against these steps in whatever order they would actually occur. This may mean moving between the numbered high level steps to identify specific sub-steps representing the activities being described.

[2] In the context of DDI, publication occurs when access to metadata is given to anyone outside of the internal group responsible for creating the metadata.

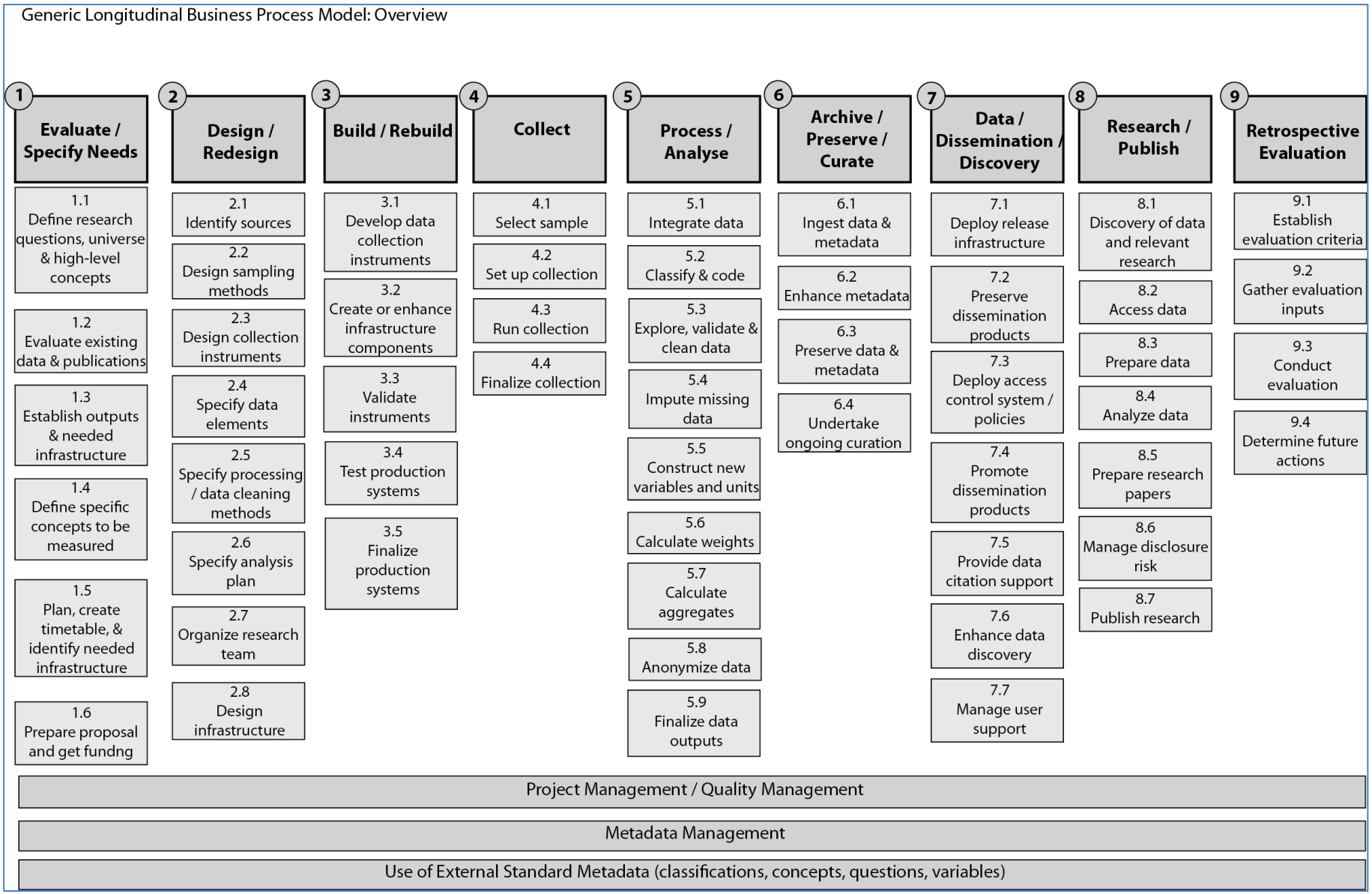

Figure 5. Generic Longitudinal Business Process Model (GLBPM): overview

Paths through the Model

Figure 6 below shows one possible sequence of steps through the model. This hypothetical study involves a survey administered twice by an already established team to a single panel using the same instrument each time. The study begins at step 1.1with initial design work, then moves to step 1.3 and 1.4 with decisions on the summary tables to be produced. Next comes step 9.1 establishing the between-round evaluation criteria. Steps 2.3, 2.4 and 2.5 finalize the design of the collection method and steps 3.1, 3.4 and 3.5 implement it. Collection proceeds through all step 4 sub-steps. Data are cleaned and aggregated in steps 5.3 and 5.7 and the final Round 1 outputs are produced in step 5.9. Round 1 data are preserved on the local file server in step 6.3. Initial analysis occurs in step 8.4. Evaluation of Round 1 occurs in steps 9.2, 9.3, and 9.4.

Round 2, in blue, begins with new data collection using the already selected sample at step 4.2 and follows the same path from there as Round 1 until final data analysis at step 8.4. Final results are published in steps 8.5, 8.6, and 8.7.

This study included no analysis of change across time for individual respondents. If it had, then steps 5.1, 5.2, 5.4, and 5.5 might have come into play. The process also points to a possible lack of a long-term preservation plan, with no steps 6.1 and 6.2.

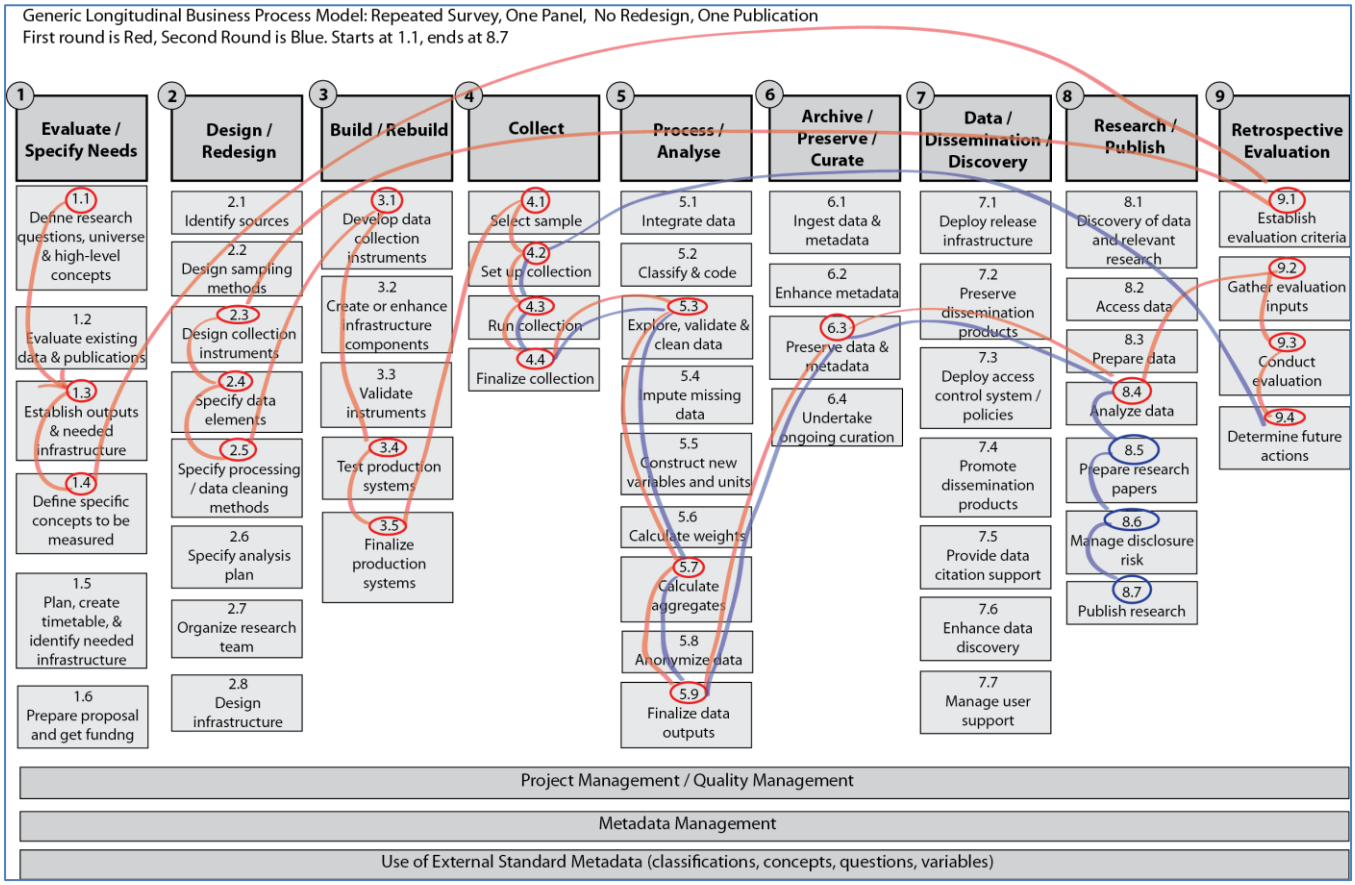

Figure 6. The path through the model for a hypothetical study

Project Management/Quality Management

There are models describing project management and its intersection with quality management, e.g., the CMMI model (Capability Maturity Model Integration developed by the Carnegie Mellon Software Engineering Institute). The model presented here recognizes the importance of these processes to data collection but does not model them directly. They are in operation throughout the data collection process.

Metadata Management

The management of metadata is critical to the process of data collection. When modeling the data collection itself, however, metadata are assumed to exist and be available as inputs and outputs for many of the steps inherent to GLBPM. The process of metadata management is not modeled here. For an example of a study using DDI, see Brislinger et al. 2011.

Use of External Standard Metadata

Some metadata are made available for re-use in data collection but are not produced by the data collector. This type of metadata is typically published by external organizations which specialize in its production. Examples of this include ISCED, a standard classification published by UNESCO for use with international education data. The forthcoming paper from this workshop, "Structuring Metadata for Reuse: Building Foundational Metadata," will address ISCED in more detail.

Locating and selecting external metadata for use in the study being designed will be a part of the activities described here (see steps 1.2, 2.1, 2.3, 2.4 in Figure 5).

RELATIONSHIPS AMONG LONGITUDINAL WAVES

In a longitudinal study, data elements must be traceable both within a single data life cycle iteration and across as many life cycle iterations as there are waves. Otherwise we cannot distinguish measures that are the same from those that are different in the series of measures we take over time. One visualization of these iterations over time presented earlier in Figure 1 is what has come to be called the "tornado". In Figure 7 we look at the tornado from the point of view of its eye and begin the discussion of best practices for assuring the traceability of data elements over time.

In the eye of the tornado we archive, preserve, and curate. Archiving and preserving are the necessary basis for the traceability of data elements. It is, however, curation that connects the dots.

Curation is a function of metadata management. Curators or, again, metadata librarians attend to drawing semantic relationships between study objects at various stages of the data life cycle. At the end of each data life cycle metadata librarians take on a special leading role during the retrospective evaluation.

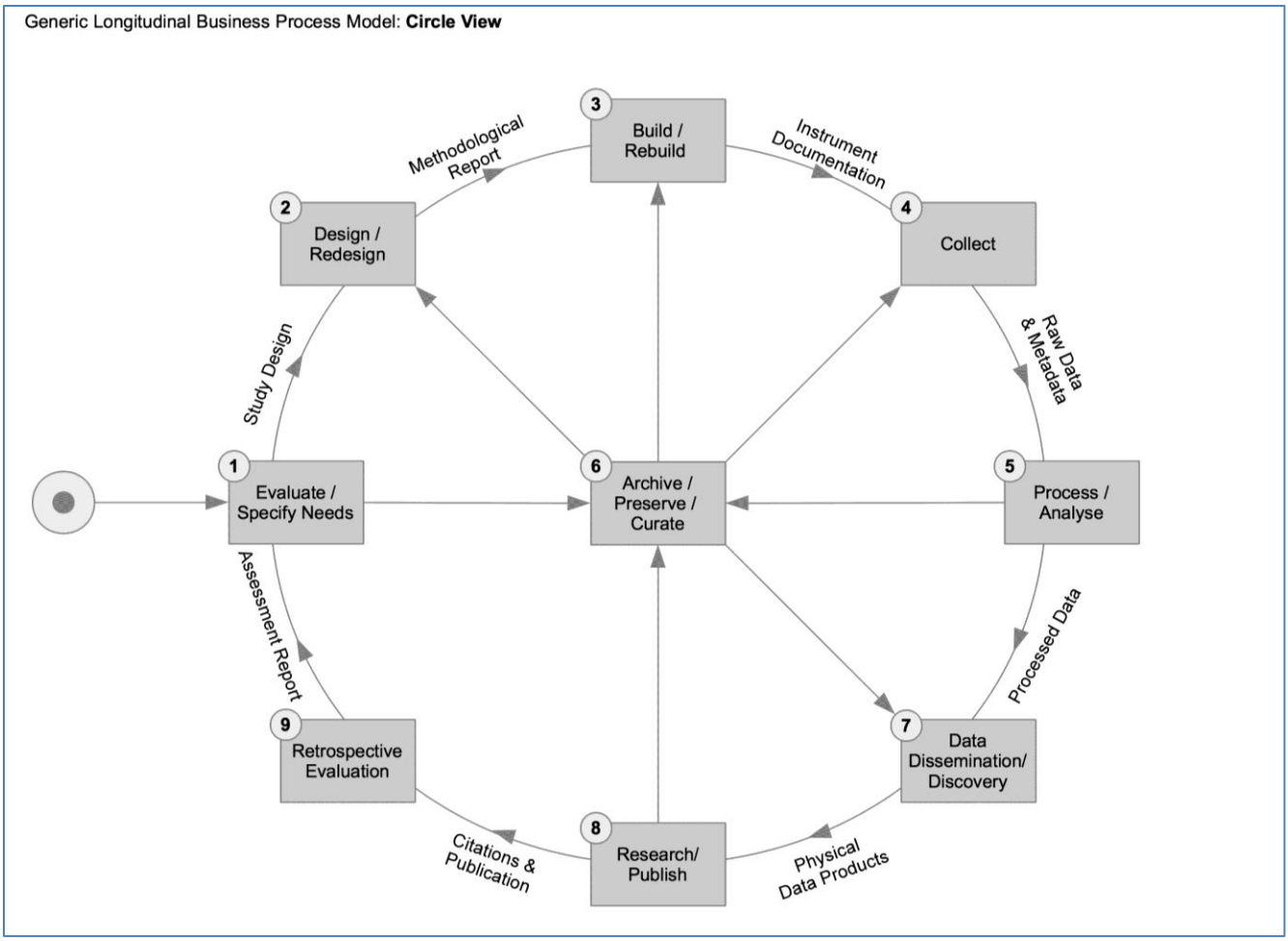

Figure 7. Generic Longitudinal Business Process Model -- Circle view

Retrospective Evaluation and Refactoring

Project and quality management principles require that at the end of each longitudinal wave, regardless of which steps in the General Model a survey or study passes through, a retrospective evaluation occurs. During this evaluation there is refactoring. Refactoring in this context is not software refactoring. It is process (e.g., survey) refactoring. In survey refactoring, study objects are revisited to determine which ones are specific to a wave and which ones can be shared or reused across waves. Very often in a study when we go through the "tornado" the first time, study designers and builders lack this perspective. Instead, no matter their experience, study designers and builders lack at least some foreknowledge as they move through the data life cycle defining, building, and executing the study in line with some but usually not all of the steps in the General Model. Perspective grows with retrospective evaluation.

Beginning in Version 3, DDI began to define and support study objects in such a way as to support survey/study refactoring. This occurred with the introduction of resource packages and groups. What follows are some retrospective evaluation best practices in connection with the General Model and the DDI Data Lifecycle.

Retrospective Evaluation of Selected DDI Elements					
Study Concept	**Data Collection**	**Data Processing Data Analysis**	**Data Distribution**	**Data Discovery**	**Publishing**
Revisit Universe Refreshment Strategy? Replacement Strategy? Adjust Population? Create Resource Packages for ConceptualComponents that are reusable across StudyUnits **Refactor StudyUnits** Move collection schemes from StudyUnits to Group? Add StudyUnit?	**Revisit Collection Strategy** Add Collection Events? Replace questionnaire data elements with direct observation? Modify Concepts to reflect new data sources?	**Refactor LogicalRecords** Move round-specific data elements into their own logical records?	**Revisit Physical Data Products** Include extant data? If yes, prepare resource package(s) and add physical structures by reference Create new physical products upon demand? If yes, subset and recombine LogicalRecords	**Revisit Data Elements** Modify PHI and PII tagging of data elements to facilitate access control? Add dimensions to the knowledge space in which data elements are located? If yes, modify data element tags to locate them in more dimensions	**Revisit OtherMaterials** Add/modify OtherMaterial at all lifecycle stages for publication **Revisit Citations**
Specify Comparisons / Note LifeCycleEvents					

Table 1. Retrospective evaluation of selected DDI elements

DETAILED DESCRIPTION OF STEPS/SUBSTEPS

1 Evaluate/Specify Needs

In a new study there is a requirement to specify the general purpose of a research endeavor. This will often include the research question(s), the reason(s) for the study, goals, outcomes, subjects of the study, and other high level aspects of the work. This stage is often characterized by ongoing discussion among the principal researchers and requires a good deal of investigation into the current state of understanding within the domain being investigated. What data exist? What has been published? What impediments need to be overcome? What data are being asked for?

In a study that repeats, the retrospective evaluation will provide many inputs that do not exist in the early stages of a new study, and which can result in specific changes to the study. These changes may include an extension of the scope, changes to the universe, modified or new research questions, etc.

1.1 DEFINE RESEARCH QUESTIONS, UNIVERSE & HIGH-LEVEL CONCEPTS

As a study begins and as each new wave or phase begins, high-level conceptualization underpins what follows. This work may be based on researcher knowledge and experience, new theoretical work, prior studies (or earlier waves), a preliminary literature review, and more. At the beginning of new phases of a project, the results of retrospective evaluation (steps 9.x) may initiate reconceptualization.

1.2 EVALUATE EXISTING DATA AND PUBLICATIONS

A more rigorous literature review may follow an initial one. Existing data might be evaluated in terms of comparability of universe, concepts, and categories, as well as in terms of access restrictions, cost, and format. This activity highlights the importance of good metadata for the existing data.

1.3 ESTABLISH OUTPUTS & NEEDED INFRASTRUCTURE

This step may include preliminary estimates of needed staff, equipment, space, travel, and so on. Preliminary development of a project team to include people with a variety of skills may occur here, as potential collaborators are recruited.

It may also be desirable to consider the general form of needed analyses and outputs. Will data be published as static tables, dynamic Web sites, or as a Web service? Will there be access control issues? How long will the data need to be preserved? Where might long-term preservation be housed and funded?

1.4 DEFINE SPECIFIC CONCEPTS TO BE MEASURED

Measures for dependent and independent variables, controls, and other classification measures need to be chosen. Measures that repeat across waves can be documented in a DDI *ResourcePackage* and used by reference. This practice can extend to using measures that have been used in other studies – documenting their comparability. The specific universe to be measured may need to be defined.

1.5 PLAN, CREATE TIMETABLE & IDENTIFY NEEDED INFRASTRUCTURE

More detailed planning than in 1.2 above may be necessary either due to the complexity of the project or through the need to develop a formal proposal. A wide range of complexity of the planning process is possible, ranging from a relatively simple document to the development of detailed project management models. This step may also involve identification of potential funding sources and developing at least an informal timetable for the proposal process.

Several funding sources now require a detailed data management plan as part of a proposal. Developing the data management plan may require some preliminary work in several of the "later" steps in the model. A data management plan may require a description of how the data will be captured, which in turn will require at least an initial consideration of sampling, instruments, data elements, and processing from steps 2, 3, and 5. The plan may require some consideration of data and metadata standards and file formats, archival setting, and distribution policies and procedures (steps 5, 6, and 7). Part of developing the data management plan will involve decisions about for how long the data will be preserved and at least preliminary arrangements with the organization which will ultimately preserve and curate the data.

1.6 PREPARE PROPOSAL & GET FUNDING

In many cases involving human subjects research, approval will be required from an institutional review board (IRB), necessitating a formal project proposal. Funding agencies will have their own requirements for the content of a proposal. Increasingly this may require a formal data management plan in addition to a detailed budget proposal. This step may require travel to meet with potential funders.

2 Design/Redesign

2.1 IDENTIFY SOURCES

This step may involve identifying specific sources of existing data and of outside expertise. It may be necessary to identify people or institutions controlling access to potential research subjects or other resources necessary for the project. Some instruments may be proprietary and specific arrangements may have to be made to use them.

2.2 DESIGN SAMPLING METHODS

A simple sampling method may be inadequate to accurately represent a particular universe. Special procedures may be needed to sample difficult to reach portions of the universe. This step should also include estimates of statistical power and consideration of the analysis methods demanded by the chosen sampling method.

2.3 DESIGN COLLECTION INSTRUMENTS

Concepts to be measured may have been chosen, but in this step the specific form of the measurement tool is designed. For surveys this includes specifying categories and codes, the wording of questions, and the flow of the questionnaire. Other concepts may need to be measured through analysis of physical samples, as with biomarkers, or through the action of some mechanical or electronic instrument, as with measuring physical location. Custom sensors may need to be designed.

Data may also be obtained by coding something observed, either in real-time, or from recorded material. Observational settings may be structured, as in interviews or focus groups, or unstructured. Data may also be obtained through mining public online sources. This may involve software design.

Careful consideration should be given to any methods to improve data quality at the collection point rather than in post-processing.

2.4 SPECIFY DATA ELEMENTS

The OECD defines a data element as "a unit of data for which the definition, identification, representation, and permissible values are specified by means of a set of attributes (Source: http://stats.oecd.org/glossary/detail.asp?ID=538)." Defining data elements for the study goes hand-in hand with designing collection instruments. The ultimately desired data elements, though, may be measured indirectly. A rate, for example, may be measured as a distance and a time. Body Mass Index may be measured as a weight and a height. More complex scale scores may need to be calculated. This step may identify what post-processing is necessary to compute the final set of data elements.

Designing data elements to be reused across waves or phases of a longitudinal study will be important. The use of versioned, persistent identifiers for new data elements, as implicit in DDI, is an important part of this step. Choosing data elements that have been used in other studies is also desirable, enhancing the potential for reuse.

2.5 SPECIFY PROCESSING & CLEANING METHODS

This step identifies the exact procedures to generate any derived measures as well as any steps needed to improve data quality. These processes become an important part of the definition of a data element. When they change, the meaning of the data element may change. Careful documentation here is an important part of documenting data quality.

Software may be purchased or developed to facilitate automated cleaning. Use of visualization tools may be planned. Staff training might need to be scheduled.

2.6 SPECIFY ANALYSIS PLAN

It may be desirable, or required to specify an analysis plan in advance of collecting data. This does not preclude deviation from the plan, but may strengthen the impact of certain types of results when a prespecified plan is carried out (see step 8.4).

2.7 ORGANIZE RESEARCH TEAM

More formal arrangements for building the team are made here. Recruitment, hiring, and training may take place. An organizational structure may be developed. Some thought should be given to the appropriate leadership style, and agreement on roles for the project should develop.

2.8 DESIGN INFRASTRUCTURE

This step may involve arranging for space and equipment for the team, communications infrastructure, travel arrangements, and more. Computing infrastructure may need particular attention – e.g., hardware, software, networking, security, and storage. Thought should be given to where bottlenecks might occur during production phases.

Infrastructure needs will extend beyond the collection, analysis, and publication phases to the archival life of the data.

3 Build/Rebuild

3.1 DEVELOP DATA COLLECTION INSTRUMENTS

Here the design developed in step 2.3 is actually implemented. This may involve the use of commercial or open source tools or involve custom programming (or hammering and soldering). Best practice is to use tools that develop the instrument automatically from the metadata developed in earlier steps (see Iverson 2009).

3.2 CREATE OR ENHANCE INFRASTRUCTURE COMPONENTS

Staff may be hired and trained, space occupied, equipment installed, software written and/or installed, and more.

3.3 VALIDATE INSTRUMENTS

Pretesting may reveal the need for redesign or reimplementation of instruments. Questionnaire flow should be thoroughly tested. Simulated or pre-test data may be run through anticipated analysis procedures.

3.4 TEST PRODUCTION SYSTEMS

This may involve testing any computing infrastructure, including tests of security and capacity and any backup facilities to ensure that pretesting scales to the actual data collection. Testing assumptions about staff capacity could be done in this step also -- for example, testing whether each interviewer can really complete interviews in the time estimated.

3.5 FINALIZE PRODUCTION SYSTEMS

Switch over from test systems. Activate production identity management. Adjustments to scheduling may occur here.

4 Collect

4.1 SELECT SAMPLE

The production sample is selected. Any issues with the sampling processes need to be thoroughly documented.

4.2 SET UP COLLECTION

Software and staff are ready, pretest data are cleared out, final access to subjects and settings is arranged, notifications to subjects are sent out, appointments are made, and so on.

4.3 RUN COLLECTION

Data begin coming in. Data security and backup procedures are in place. Metadata and paradata generated during collection are preserved.

4.4 FINALIZE COLLECTION

This step may involve notifications or payments to subjects, rescheduling, debriefings of collection staff, closing down facilities, or making arrangements for future collection phases.

5 Process/Analyze

Once the raw data are collected, the processing and analyzing phase of a longitudinal study forms the heart of the data work. This work is often internal to the project and usually precedes public dissemination or subsequent research with the data. The inputs often consist of raw data collected in Step 4, but can also include other forms of data that are linked, merged, or otherwise used to process, analyze, or improve the data at the core of the project.

5.1 INTEGRATE DATA

This step may involve joining data from parallel collection streams as well as joining collected data with external data. Joining data may involve transformations to harmonize data with different units of measurement or classification and coding schemes. Different streams of data may have been collected at different levels in a hierarchy of units of analysis – e.g., for households and for persons in households. Preserving software code used to integrate data is an important part of documenting the data.

5.2 CLASSIFY & CODE

Some data may need classification and coding. Open-ended questions, for example, may need to be processed by trained coders. Training may need to begin before production data are available. Supervision of coding may need to be set up. A study may involve qualitative data analysis techniques coding audio or video. Automated techniques such as text mining may be applied to classify chosen units of analysis. These actions may result in new variables (see 5.5 below). Parameter settings used to perform automated processing should be documented.

5.3 EXPLORE, VALIDATE AND CLEAN DATA

Almost all data will have some degree of error. Statistical and visualization techniques may be employed to search for problems. These activities can also generate measures of data quality, which should be documented.

5.4 IMPUTE MISSING DATA

Data collection projects almost universally experience instances of missing data or data that are inconsistent, logically impossible, or are otherwise in need of improvement. To overcome this limitation, various methods of data imputation are often used. In brief, data imputation involves the creation of data based on a set of rules that are clearly specified and intended to produce results that are scientifically valid. The creation of an entirely synthetic dataset may be viewed as an extreme form of data imputation, in that all the records in a completely synthetic dataset are imputed.

An exploration of the reasons for missing data can be crucial. If the occurrence of missing values relates to some aspect of the study or subjects, interpretation of an analysis may be impossible.

The methods used for imputation are described in DDI as *GenerateInstructions,* which are referenced from the variable using the *ImputationReference* element. If imputation methods change across the waves of a study, this would be documented in the comparison of the variables (see Ionescu et al. 2010).

5.5 CONSTRUCT NEW VARIABLES AND UNITS

Transformations may be applied to combinations of variables to produce new variables, as in the computation of scales, or the computation of a BMI score from weight and height. Classification and coding (see step 5.2) may have generated new variables. Some variables may need to be transformed to different units of measurement, or perhaps transformed to fit a different distribution.

For certain types of data, automated procedures may construct new variables. Pairwise distance measures, for example, can be processed by a multidimensional scaling program to produce spatial dimension variables. Parameters used in such procedures should be documented.

New units of analysis may need to be generated. Transcripts for persons might need to be transformed to make paragraphs or sentences the unit of measurement for some text mining techniques.

5.6 CALCULATE WEIGHTS

In order for statistics to accurately reflect each universe measured, one or more weights may need to be calculated for observations.

5.7 CALCULATE AGGREGATES

Aggregate measures, e.g., counts, mean, median, low, high, and so on, may be calculated across combinations of classification variables. Sets of aggregates and associated classification variables may be output as data cubes. Where there are confidentiality constraints with the data, care may need to be taken to not produce aggregates which can reveal data about individual observations, for example the one person over 100 years old in a particular geographic unit.

5.8 ANONYMIZE DATA

Anonymization is a complex process potentially involving legal issues which vary across political boundaries. Techniques may involve removing or recoding variables deemed to be personal identifiers (government-issued ID numbers, specific geographic location, etc.). Other computational techniques may involve suppressing data in cells with small counts in data cubes (see step 5.7). Statistical techniques which may add noise to the data may be employed to prevent disclosure (see, for example, Eurostat, *Statistical Disclosure Control*).

Different sets of data may be produced at different levels of anonymization, each having different access policies.

5.9 FINALIZE DATA OUTPUTS

Data outputs may need to be transformed from the form convenient to the tools used for the steps above to forms appropriate for distribution. Some analysts may need the raw or cleaned raw data. Others may be able to use data with just direct identifiers removed. Public datasets may have more thorough anonymization applied. Summary datasets or graphics may be produced for publication. Data and metadata may have been maintained in a relational database for processing and analysis and need to be exported into more open formats for distribution, a DDI XML file, for example. Plans may also be made to make data available through an online service, or through the Semantic Web (see the Wikipedia article "Linked data").

6 Archive/Preserve/Curate

6.1 INGEST DATA & METADATA

In some ways, the activity of archiving data, whether or not to a formal archive, is central to a longitudinal study (see Figures 1 and 6). Throughout the course of the whole study, data and associated metadata will need to be preserved, if only for use in later phases of the study.

As publications are generated, it will be good practice to be able to reproduce the exact set of data that were used in analyses for the publications. If the data are to be preserved at an archive (organization), something like an OAIS Submission Information Package SIP (http://public.ccsds.org/publications/archive/650x0b1.pdf) will need to be produced for submission. These activities may involve converting data and metadata from some in-house format to a more generally accessible format for the long term.

The Producer-Archive Interface – Methodology Abstract Standard (PAIMAS) (ISO 20652: 2006) does seem relevant here with four phases (page 11 of Principles and Good Practice for Preserving Data): "These phases make explicit the steps an organization must take when archiving data."

6.11 Preliminary – Define the information to be archived
6.12 Formal Definition – Develop agreement
6.13 Transfer – Actual transfer of the objects
6.14 Validation – Validate the transferred objects

6.2 ENHANCE METADATA

Metadata from the data collection and analysis phases of the study may be enhanced in multiple ways on an ongoing basis. They may be integrated into some widely searchable system (see, for example, the Wikipedia article "Linked data"). Metadata may also accumulate as data are cited and reused. Having links from the data to reuses and from those reuses to the data will enhance the value of the data.

An important issue for enhancing is the connection between scholarly publications and data. These connections may be made by establishing persistent identifiers (e.g., like DOIs) for datasets that can be published later (see 7.5). Another enhancement may be to translate metadata and documentation into other languages to make it understandable for other communities.

6.3 PRESERVE DATA & METADATA

Data preservation requires ongoing activity. Storage media decay or become obsolete, new formats become necessary, metadata accrue with ongoing access and use, and desirable access methods evolve. Most projects have funding for a limited period. In many cases the need for preservation will outlive the original funding. Arrangements for ongoing preservation may need to be made with local institutional repositories or more global archives. These arrangements, made toward the beginning of a project may generate requirements for ingestion activities described in step 6.1 above. Access control policies may also need to be established.

6.4 UNDERTAKE ONGOING CURATION

Once the data are in an archive additional curation activities may generate metadata which can be recorded as DDI LifeCycleEvents. Data may be migrated to new formats, or replicated to multiple sites. Legal contracts for access to confidential data may be drawn. Assessment of disclosure risk will be an ongoing activity for the life of the data, as other external data and procedures with the potential for allowing disclosure emerge.

7 Data Dissemination/Discovery

7.1 DEPLOY RELEASE INFRASTRUCTURE

Data dissemination may be handled through a repository or an archive. An ongoing project may also take a more hands-on approach to dissemination – with staff and/or a Web site as a contact point. The latter will require development of an infrastructure.

7.2 PREPARE DISSEMINATION PRODUCTS

A variety of dissemination products may be generated at multiple points in a project. These may include raw or processed data, summarized data, tables, graphics, and datasets and scripts for various statistical packages. The latter may involve restructuring both data and metadata to fit the underlying data models of the target packages. A variety of dynamic applications may be produced, including online or standalone visualization or analysis tools. Application programming interfaces (APIs) may be developed to allow external software to directly access the data. Access control and licensing policies may apply in the preparation of these products.

7.3 DEPLOY ACCESS CONTROL SYSTEM & POLICIES

Each of the dissemination products may need different access control polices and systems to apply them. Raw data might need to be accessible only under strict confidentiality terms. Summary data or graphics might have more lenient terms. Applications might need to have the policies built in and thoroughly tested.

7.4 PROMOTE DISSEMINATION PRODUCTS

Use of the data depends on people finding them. Citation of the data in publications is one traditional method of promotion. Ensuring that detailed, well-structured metadata are available through search mechanisms is another. Creation of persistent Digital Object Identifiers (DOIs) will enhance the ability to locate the data.

7.5 PROVIDE DATA CITATION SUPPORT

Persistent identifiers linked to the current source of the data will ensure that the data can be cited, that statistics about citation can be computed, and that the data can be found from the citation. The DataCite organization (http://datacite.org) provides one such mechanism.

7.6 ENHANCE DATA DISCOVERY

In order to make variables and questions more discoverable, they may be tagged with metadata. This tagging may occur prior to a wave. Retrospective analysis may also reveal the need to refactor, leading to changes in the way variables and questions are grouped.

The organization curating the data may undertake some of these activities. Archives may create metadata such as catalog records for searching, index those records with subject terms, and prepare metadata for variable level search.

As datasets grow larger, it may not be possible to transfer them easily – or even at all. Online tools to extract, summarize, analyze, and visualize data may be required.

7.7 MANAGE USER SUPPORT

Complex data, or simple data about complex topics, may require providing support to those trying to reuse the data. In a large study those users may be part of the project. A support infrastructure may be needed. Sophisticated analysis methods employed in a study may require specialized expertise.

8 Research/Publish

8.1 DISCOVERY OF DATA & RELEVANT RESEARCH

Within a longitudinal study, once data are ready for analysis an additional search for existing data and research, extending a search that might have been done at step 1.2, might be desirable. Longitudinal studies involving meta-analyses in which no new data are collected are also possible.

8.2 ACCESS DATA

Data from an ongoing study may need to be extracted, possibly from multiple sources. Access controls and identity management may come into play, particularly in multi-institutional studies.

8.3 PREPARE DATA

Data may need to be refactored from their form in production systems to forms usable by data analysts. Different analyses may require new variables. A spatial analysis, for example, may require geocoding the data. Data structures may have to be changed. To fit the requirements of analysis tools, a “tall skinny” form of a table containing one row for each combination of subject and time period might need to be transformed into a “short wide” table with one row for each subject and separate columns for each time period.

8.4 ANALYZE DATA

Analysis of the data may involve running a predetermined set of programs on the data, or might involve a complex iterative process. In either case, good practice would involve the specification of an analysis plan before the data collection begins and recording deviations from that plan as analysis proceeds. This is particularly true in cases where there are many possible statistical significance tests on a set of data. Finding

a small set of prespecified significant results is interpreted very differently than searching for and choosing any statistically significant results out of a large set of possibilities.

Replication, the "gold standard" in science, requires a detailed description of the analysis process. Best practice might involve including analysis computer scripts along with related metadata (version of the software used to run the script, operating system and hardware used, etc.) in the extended metadata to be archived.

8.5 PREPARE RESEARCH PAPERS

In addition to text, publications may include tables and graphics. Methods used to produce those figures should be included in the metadata associated with the data used for the publication. In cases where the data are analyzed from an extract from an evolving production system, good practice would involve either preserving the extract or some method to recreate exactly that set of data. Best practice might involve including in a publication a persistent identifier usable for locating the data (see, for example, The DOI® System) as well as a reproducible identification measure for the **content** of a set of data, such as the Universal Numeric Fingerprint (see Altman and King 2007) to ensure that attempts at replication are really using the same data even if represented in a different software format.

8.6 MANAGE DISCLOSURE RISK

When there are confidentiality constraints on the underlying data, tables and graphics derived from those data must be evaluated for disclosure risk. Detailed statistical models may also require evaluation.

8.7 PUBLISH RESEARCH

Publication of research results may require negotiation of arrangements for archiving the data and metadata used for the publication. In some cases, a publisher might require a copy of the data for its archive. Careful negotiation of access rights would be prudent. Metadata for the dataset used in the publication should be updated to include a citation of the publication.

9 Retrospective Evaluation

Any project that takes place over an extended time period will experience change. A retrospective evaluation will be important in both evaluating the impact of unplanned change and in determining the need for planned change. The evaluation may include inputs from the quality and project management processes as well as metadata accumulated through the project. Outside evaluators may be useful or even required.

A retrospective evaluation may assess the degree to which goals were met. It may also consider changes in the project environment - administrative, physical, and intellectual (e.g., has someone invented better measurement methods?). Choosing to make mid-course improvements will always have to be balanced against the possibility of confounding study results. For a more detailed discussion see Greenfield et al.

9.1 ESTABLISH EVALUATION CRITERIA

Criteria will need to be established for determining which planned or unplanned changes are significant, whether external events need to be documented, and whether established goals are being met. All of the lifecycle elements are candidates for scrutiny. Design, sampling, data collection, data processing, analysis, and the retrospective evaluation process itself all should be evaluated. Comparison of expenditures against the budget and the rate of progress against the planned timeline may also be important.

9.2 GATHER EVALUATION INPUTS

The criteria for evaluation will guide the selection of inputs for the evaluation process. Some inputs may be collected as the earlier phases of the study progresses.

9.3 CONDUCT EVALUATION

Some components of the evaluation may occur as the study progresses. Some information about the data collection process will be available as a round of collection finishes. More may be revealed as data processing and analysis proceed. Analysis of patterns of missing data, for example, may uncover flaws in the collection process.

9.4 DETERMINE FUTURE ACTIONS

Evaluation may suggest changes to the project. Once again, the decision as to whether to implement those changes will entail a careful consideration of the impact on the ability to achieve projects goals. Some changes may have to wait for a future study.

Choosing to make some changes may initiate a revisiting of some of the earlier steps in the process model. A change to a variable may necessitate redesign of an instrument and so on.

OUTPUTS

Each step in the model may generate outputs: reports, data, and metadata. A particular step may generate output multiple times throughout the course of a project. During the proposal phase, the creation of a data management plan in step 1.5 may necessitate initial work in data element selection, instrument design, team building and more. These steps would then be revisited as the project got under way.

Metadata outputs may take many forms including sets of concepts, categories and codes, description of data collection instruments, narrative, and programming scripts. Accumulating these outputs in a formal structure will serve to make them more searchable and reusable. In a longitudinal study DDI can also facilitate the documentation of additions and changes across time with explicit comparisons. An important part of project management will include the selection of tools to manage this metadata structure.

REFERENCES

Altman, Micah, & King, Gary. (2007). A Proposed Standard for the Scholarly Citation of Quantitative Data. *D-Lib Magazine*, 13(3/4).

Block, William C., Andersen, Christian Bilde, Bontempo, Daniel E., Gregory, Arofan, Howald, Stan, Kieweg, Douglas, & Radler, Barry T. (2011). Documenting a Wider Variety of Data Using the Data Documentation Initiative 3.1. In Mary Vardigan, Stefan Kramer, & Larry Hoyle (Eds.), DDI Working Paper Series – Longitudinal Best Practice. doi: http://dx.doi.org/10.3886/DDILongitudinal01

Brislinger, Evelyn, deNijsBik, Emile, Harzenetter, Karoline, Hauser, Kristina, Kampmann, Jara, Kurti, Dafina, Luijkx, Ruud, Ortmanns, Verena, Rokven, Josja, Sieben, Inge, Ros, Ivet Solanes, Stam, Kirsten, van de Weijer, Steve, van Vlimmeren, Eva, Zenk-Möltgen, Wolfgang. (2011). European Values Survey EVS 2008 Project and Data Management GESIS-Technical Reports (2011 | 12 ed.): GESIS – Leibniz Institute for the Social Sciences.

Carnegie Mellon Software Engineering Institute. Capability Maturity Model Integration (CMMI) Overview. http://www.sei.cmu.edu/cmmi/

Chapman, Pete, Clinton, Julian, Kerber, Randy, Khabaza, Thomas, Reinartz, Thomas, Shearer, Colin, & Wirth, Rüdiger. (1999, 2000). CRISP-DM 1.0 Step-by-step data mining guide. CRISP-DM consortium: NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringen en Bank Groep B.V (The Netherlands). http://www.the-modeling-agency.com/crisp-dm.pdf

Data Documentation Initiative Alliance. (2009). Data Documentation Initiative (DDI) Technical Specification, Part I: Overview Version 3.1

Data Documentation Initiative Alliance. (2009). DDI 3.1 XML Schema Documentation (2009-10-18), from http://www.ddialliance.org/Specification/DDI-Lifecycle/3.1/XMLSchema/FieldLevelDocumentation/

European Commission/EUROSTAT. Statistical Disclosure Control. http://epp.eurostat.ec.europa.eu/portal/page/portal/research_methodology/methodology/statistical_disclosure_control

International DOI Foundation. (2011). The Digital Object Identifier System. http://www.doi.org/

Greenfield, Jay, Conrey, Riki, Kaun, Sophia, Shilonskaya, Alexandra, & Smith, Daniela. (2011). Retrospective Evaluation, Maintaining Common Data Elements and a Common Model in Longitudinal Studies Version 1.1 (unpublished manuscript).

Hoyle, Larry, Castillo, Fortunato, Clark, Benjamin, Kashyap, Neeraj, Perpich, Denise, Wackerow, Joachim, & Wenzig, Knut. (2011). Metadata for the Longitudinal Data Life Cycle. In Mary Vardigan, Stefan Kramer, & Larry Hoyle (Eds.), DDI Working Paper Series – Longitudinal Best Practice. doi: http://dx.doi.org/10.3886/DDILongitudinal03

Ionescu, Sanda, , with Larry Hoyle; Mari Kleemola; Martin Mechtel; Olof Olsson; and Wendy Thomas. (2010). Using DDI 3 For Comparison. In Larry Hoyle, Mary Vardigan, & Michelle Edwards (Eds.), DDI Working Paper Series -- Use Cases. doi: http://dx.doi.org/10.3886/DDIUseCases03

Iverson, Jeremy. (2009). Metadata-Driven Survey Design. IASSIST Quarterly 2009 (Spring/Summer). http://www.iassistdata.org/iq/metadata-driven-survey-design

Leaper, Nicole. A Visual Guide to CRISP-DM Methodology. http://exde.files.wordpress.com/2009/03/crisp_visualguide.pdf

UNESCO Institute for Statistics. ISCED: International Standard Classification of Education. http://www.uis.unesco.org/Education/Pages/international-standard-classification-of-education.aspx

UNECE Secretariat (prepared by Steven Vale). (2009). Generic Statistical Business Process Model (Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS). http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model

Wikipedia contributors. Linked data, *Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=Linked_data&oldid=470715927

APPENDIX A: EXAMPLE OF MAPPING CRISP-DM

It has been discussed how not every study follows all the steps in the General Longitudinal Business Process Model (GLBPM). It is also the case that there may be more specific steps that a survey or study follows that are not represented in the General Model. In this event projects can harmonize the more general project independent reference model with a project specific one.

An example that comes to mind is a project that conducts internal analyses on the data prior to publication. Such a project might be interested in determining whether it is in fact feasible to assemble, for example, a predictive model for retention using the constructs it currently plans to release. In this instance a study might go about harmonizing the GLBPM and CRISP-DM. CRISP-DM is the Cross Industry Standard Process for Data Mining (Chapman et al., 1999, 2000). In the figure below steps from the CRISP-DM model are added to the GLBPM (see Figure 8).

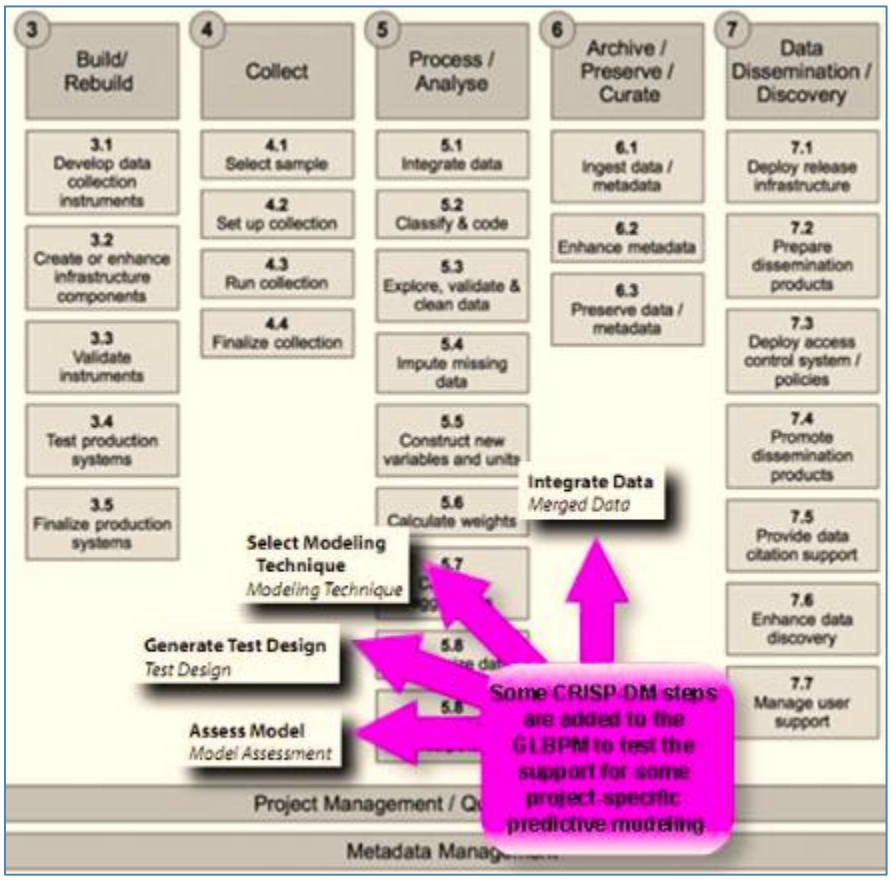

Figure 8. Steps from CRISP-DM added to GLBPM

CRISP-DM is a data mining process model that got under way as a European Union project in 1996. The first version of the methodology was released in 1999. Contributors included ISL (acquired by SPSS), NCR Corporation (which spawned the Teradata warehouse), and Daimler-Benz. In 2006 the consortium began work on a second version of the process model shown below (see Figure 9).

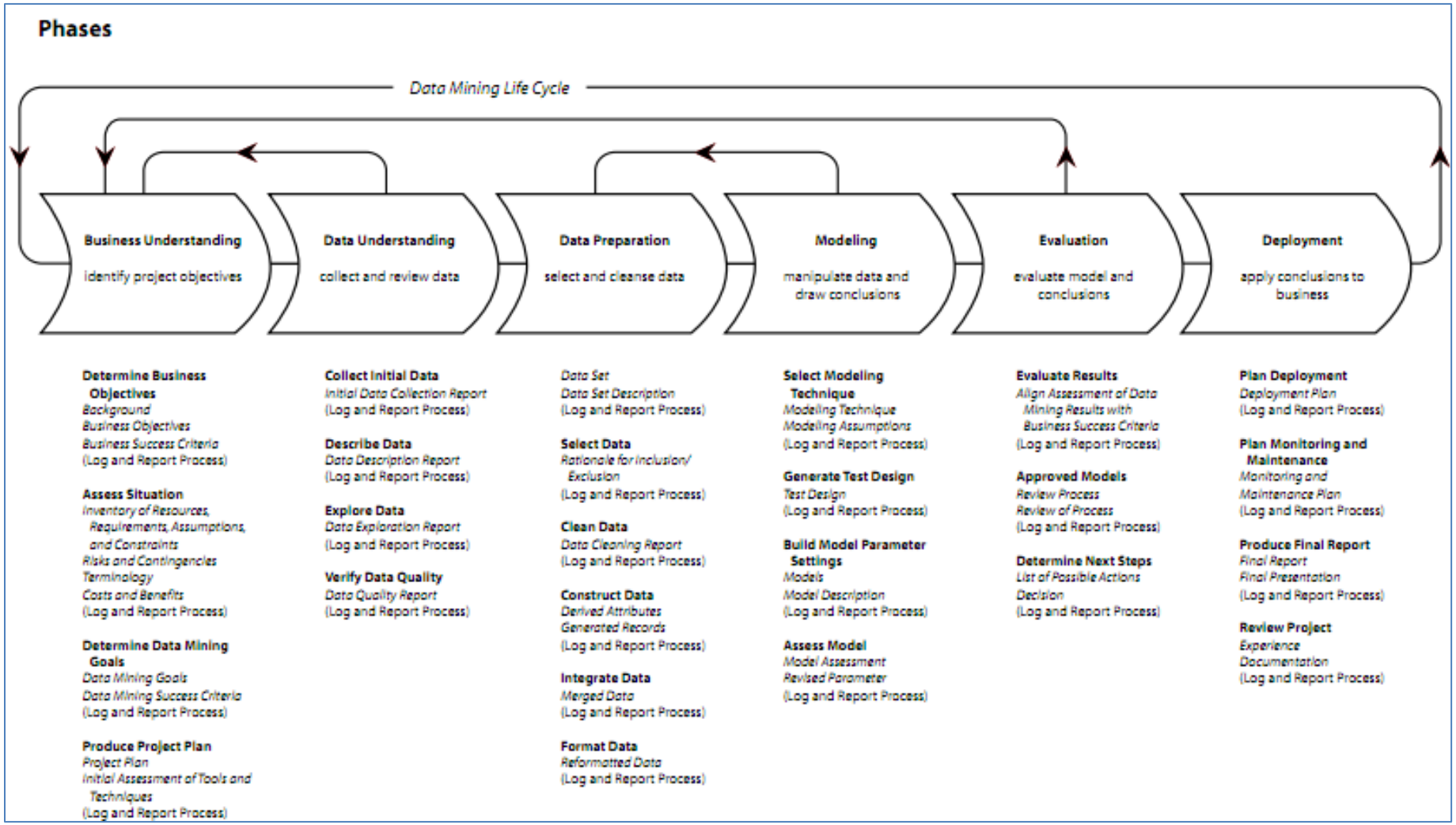

Figure 9. The CRISP-DM model (Leaper)

APPENDIX B

Acknowledgments

This paper is one of several papers that are the outcome of a workshop held at Schloss Dagstuhl - Leibniz Center for Informatics in Wadern, Germany (Dagstuhl Event 11382), on September 17-23, 2011. Papers were edited by Alerk Amin, William Block, Jeremy Iverson, Joachim Wackerow, and Marion Wittenberg. Larry Hoyle was editor for the workshop as a whole. Mary Vardigan (Inter-university Consortium for Political and Social Research [ICPSR], University of Michigan, USA) is the editor of the DDI Working Paper Series (http://www.ddialliance.org/resources/publications/working).

Workshop Title:
DDI: Managing Metadata for Longitudinal Data – Best Practices II

Link: http://www.dagstuhl.de/11382

Organizers:
Arofan Gregory (Open Data Foundation - Tucson, US)
Wendy Thomas (Population Center, University of Minnesota, US)
Joachim Wackerow (GESIS - Leibniz Institute for the Social Sciences, DE)

Participants in the workshop:

- Alerk Amin — CentERdata
- Ingo Barkow — DIPF – Educational Research and Educational Information
- William (Bill) Block — Cornell Institute for Social and Economic Research (CISER), Cornell University
- Joan Corbett — Scottish Centre for Social Research (ScotCen)
- Johan Fihn — Swedish National Data Service (SND)
- Jay Greenfield — Booz Allen Hamilton
- Arofan Gregory — Open Data Foundation (ODaF)
- Marcel Hebing — German Socio-Economic Panel Study (SOEP), DIW - Berlin - German Institute for Economic Research
- Larry Hoyle — Institute for Policy & Social Research, University of Kansas
- Jeremy Iverson — Colectica
- Merja Karjalainen — Swedish National Data Service (SND)
- Sophia Kuan — Booz Allen Hamilton
- Hilde Orten — Norwegian Social Science Data Services (NSD)
- Abdul Rahim — Metadata Technology Inc., North America
- Bodil Stenvig — Danish Data Archive (DDA)
- Wendy Thomas — Minnesota Population Center (MPC)
- Joachim Wackerow — GESIS - Leibniz Institute for the Social Sciences
- Marion Wittenberg — Data Archiving and Networked Services (DANS)
- Wolfgang Zenk-Möltgen — GESIS - Leibniz Institute for the Social Sciences

APPENDIX C