

The **IPUMS Business Process Model (IPUMS BPM)** should be applied and interpreted flexibly. It's not a rigid framework in which all steps must be followed in a strict order. Instead it's used to identify the possible steps throughout a project – acquiring, transforming, and documenting data. A single project may have several paths through the model depending on specific needs, repetition involved in updates, versions, or iterations of the project.

## **Evaluate / Specify Needs**

This is the beginning phase of formulating a research project. It's the time for brain-storming and high-level concept formulation. The basic framework of the proposed project comes through this phase.

### *1.1 Define research needs, coverage & high-level concepts*

Define the initial research idea, identify where new research is needed and formulate the idea for a proposal. This step would also look at intended users and their needs.

### *1.2 Evaluate existing data & publications*

Look at what already exists for similar research projects. Existing data and publications need to be taken into consideration to see what added value this proposed project may have to researchers and the public. Also, this would take into consideration what data may already exist and if they cover the proposed project or could be useful for said project.

### *1.3 Establish outputs & needed infrastructure*

Identify the outputs needed to make data useful to the targeted users (as defined in 1.1). This step also defines the infrastructure for the outputs, such as databases and search engines.

### *1.4 Identify specific concepts to be harmonized*

Specify the conceptual coverage of the project, identifying topical areas of research that the product will support.

### *1.5 Plan, create timetable, & identify needed infrastructure*

Develop a framework for the project including a timetable and goals. How will it be completed? What steps are needed? What infrastructure needs to be created to support the project?

### *1.6 Identify partners*

Identify who will partner on this project and groups or individuals needed for successful completion. This includes identifying sources for the raw data as well as partners for creating infrastructure and potential external members of the project team. Partners should be identified as acting partners through appropriate metadata.

### *1.7 Prepare proposal and get funding*

Write the formal proposal and have it be reviewed by managers to be submitted for funding. This may be an iterative process.

## **Design / Redesign**

This phase describes the development and design activities. It expands the initial proposal into an actionable plan.

### 2.1 *Identify sources*

Identify the data sources to be used for the project and how they will be procured. Provenance metadata should be captured to track the lifecycle of the data throughout the project. Identify the required provenance metadata.

### 2.2 *Design sampling methods*

Determine the sampling methods required for the procured data. Create the representative subset needed for the project used for the analysis of the data. This process needs to be documented in the metadata identifying what took place in the sampling process to get the sub-set used for the project.

### 2.3 *Design capture process*

Determine the appropriate method of capture for the needed data. The actual activities may vary according to the source of the data and may include electronic transfer, harvesting, delivery on media or other appropriate method. This step also involves the design of formal agreements with data providers and should specify the requirements and responsibilities of both the project and the data provider. This sub-process also includes the design of project-specific data management tools used for capture.

### 2.4 *Specify data elements and related metadata*

Identify the data elements required for the project and the metadata needed to identify and accurately manage the data. This will include identifying metadata for the data set (collection process, purpose, source, etc.) as well as for individual elements.

### 2.5 *Specify processing / data cleaning methods*

Specify the steps to be taken in “processing” and “cleaning” the data. This can include specifications of routines for coding, editing, inputting, validating, and finalizing data sets.

### 2.6 *Specify evaluation plan*

Determine what standards will be used to evaluate the project through all its stages. Evaluation may be done by the project team itself as well as managers. This may include the evaluation of data processes. Metadata should be captured as it relates to data uses.

### 2.7 *Organize research team*

Identify the team for the proposed project both in terms of required skills and individuals. Specify all staff necessary for the project – managers, researchers, technical staff, etc.

### 2.8 *Design infrastructure*

Design the necessary tools for data capture, cleaning, analysis, organization, etc. The design could include databases for storage, search algorithms for use in research, etc..

## **Build / Rebuild**

This phase describes the development and building of the protocols, processes, software, and tools used by the project.

### 3.1 *Develop data capture processes*

Determine the protocols for capturing data. This may include instrument development. This should include consideration of the metadata that should be collected during the capture process as well as

the relationships between the metadata associated with data capture and the metadata directly related to the definition of the data captured.

### *3.2 Create or enhance infrastructure components*

Build new and/or enhance existing components and services needed for the “Process/Analyze” phase. Services may include ingest tools, workflow frameworks, metadata management services, etc.

### *3.3 Validate processes and tools*

Processes and tools can be validated through testing and/or through research regarding the validity of a specific protocol or tool.

### *3.4 Test production systems*

The act of testing the assembled services and related workflows. This step includes technical testing and sign-off of new systems and routines, as well as confirmation that existing routines are suitable for use in this project.

### *3.5 Finalize production systems*

Verify that the activities, processes and services work well as a unified system, including modified and newly-created processes.

## **Collect**

This phase describes the acts of selecting, negotiating for, and collecting data and metadata for the project.

### *4.1 Select sources*

Determine the sources of data desired for use in the planned project.

### *4.2 Negotiate access and distribution rights*

Clarify and codify access and distribution rights with the agency providing data to the project.

### *4.3 Capture data*

May include the delivery of data from the producer, selective downloading of data from a data provider, harvesting, or primary data capture.

### *4.4 Obtain metadata*

Metadata includes descriptive metadata as found in a data dictionary, source information on how the data was originally captured (including questionnaires, protocols, instructions, sampling processes), and the who, what, where, and when of the original data capture.

### *4.5 Create sample*

Implement a sampling procedure on a data set, if necessary.

## **Process / Analyze**

This phase describes the common processes used by the project when working with data from project input to output. Metadata captured during these activities results in a provenance trail for the project, dataset, and possibly data items.

### *5.1 Validate data against metadata*

Verify that the metadata accurately reflects the data in terms of location, definition, documentation of codes, ranges, etc.

### *5.2 Select and restructure data*

Identify required data and resulting extract. This may include recoding, the creation of new data items, and/or reorganization.

### *5.3 Clean and anonymize data*

Includes correction required due to documentation from the source (published corrections), dealing with undocumented and/or out of range codes, and/or implementation of confidentiality protocols.

### *5.4 Impute missing data*

Where data are considered incorrect, missing, or unreliable, new data may be inserted during this sub-process.

### *5.5 Harmonize selected data*

The planning and processing involved in design, implementing, and documenting harmonization of data.

### *5.6 Calculate weights*

Includes determining weights and application of weights, generation of weight variables, and documentation of the process of creating and use of weights.

### *5.7 Calculate aggregates*

Includes determining aggregates, processing, documentation, and storage of aggregates.

### *5.8 Validate processed data*

Validation of processed data may include the use of check sums or digital fingerprints, reverse engineering, use of external aggregates to validate sampling results, and implementation of quality standards.

### *5.9 Finalize data outputs*

Includes verification of overall packages of data such as SIP, AIP, and/or DIP packages to ensure both accuracy and completeness.

## **Archive / Preserve / Curate**

Describes the activities undertaken to preserve data and metadata in a manner that is persistent, uniquely identifiable, and usable by future researchers.

### *6.1 Ingest data & metadata*

Establish protocols regarding selection and scheduling for ingesting data and metadata for preservation purposes.

### *6.2 Enhance metadata*

Identify required or preferred metadata, seek additional metadata sources, and add process metadata relating to the process of archiving/preservation.

### *6.3 Capture process/provenance metadata*

Identify what process/provenance metadata is required for archiving (may exceed that needed by an end user). Specify where metadata should be captured, type of capture (documentary, actionable), and means of capture.

### *6.4 Preserve data and metadata*

Covers the technical process of preservation including both content and relationships (internal to an AIP and to the collection of a whole).

### *6.5 Undertake ongoing curation*

Conduct on-going management of archival content, access, and persistence.

### **Data Dissemination / Discovery**

Activities intended to make data discoverable and available to end users.

#### *7.1 Deploy release infrastructure*

Implement and maintain infrastructure used to support discovery and dissemination.

#### *7.2 Preserve dissemination products*

Determine what should be preserved regarding DIPs and how this is accomplished.

#### *7.3 Deploy access control system / policies*

Determine and document policies regarding access and the systems to support these policies.

#### *7.4 Promote dissemination products*

Market and promote the product and sub-products to identified communities of use.

#### *7.5 Provide data citation support*

Establish policies regarding citation and provide citation information through various media (web site, DIP, DOI).

#### *7.6 Enhance data discovery*

A broad category which can cover user testing of systems, response to user feedback, workshops, promotion, publications, exploration of other systems, etc.

#### *7.7 Manage user support*

Handle questions from users and document feedback to inform system development.

### **Research / Publish**

Activities focused on supporting the analysis and use of data through research and publication.

#### *8.1 Obtain listing of publications based on the data product*

Determine what level of publication coverage is needed for user interest, funder feedback, or internal development and how to obtain and maintain that information.

#### *8.2 Maintain publication database*

Processes for adding to, maintaining, and disseminating a publication database.

#### *8.3 Manage versioning*

Establish and document versioning rules and how mid-versioning change information is logged and disseminated.

#### *8.4 Deposit metadata in related systems*

Documents how repositories are selected, the process of providing metadata to these repositories, and integration into the overall metadata tracking system.

#### *8.5 Manage disclosure risk*

Determine, measure, and handle disclosure risk.

### **Retrospective Evaluation**

Focuses on the evaluation of the processes and activities as they relate the goals of the project. Results of evaluation may be used to improve future iterations of the same project, a future project, or similar processes used by other projects.

*9.1 Establish evaluation criteria*

Document the goals, targets, and audience for evaluation. Specify how evaluation is fed back into system improvements.

*9.2 Gather evaluation inputs*

Collect data and information required for evaluation and organize evaluation responses.

*9.3 Conduct evaluation*

Describes the act of evaluation and the metadata related to that process.

*9.4 Determine future actions*

Specifies how evaluation impacts future development of this project or other projects. Clarifies who is involved in the process and how it may be implemented.