IPUMS Workflow

November 2021

Intent

This document describes the interrelated processes of IPUMS activities and maps these processes to the requirements of related archival models. IPUMS integration and documentation of census and survey data makes it easy to study change, conduct comparative research, merge information across data types, and analyze individuals within family and community context. IPUMS includes multiple collections that deliver and preserve individual-level microdata and small-area tabular data. Individual projects focus on either the United States or a broader international coverage. Each collection covers multiple years, supporting research over time; time periods vary by collection, but many provide data covering multiple decades.

Data Producer: IPUMS projects obtain data produced by national statistical agencies, data archives and other data producers, then harmonize it to support analysis over time and space. The results of this work cover most of the data submitted to the IPUMS Archive and to the IPUMS live data store.

Archival Preservation: Snapshot versions of IPUMS collections are preserved and made available through archival access. As part of the preservation snapshot process large data files are divided into relevant subsets reflecting the overall coverage of the data in the original file and the standard usage patterns as determined by the project management. Some original data files and documentation are archived under agreement with the originating agency for the purpose of preservation, but not dissemination. The archive also maintains copies of source data and documents as provided by the project for informational and business continuity purposes.

Dissemination: The IPUMS live data store provides sophisticated online exploration, a tool to create data subsets customized to users' research questions, and online analysis of data. Archived data and documents are available as standard files for download through a separate archive access system.

IPUMS disseminates and archives data. In these roles we turn to two major standards: Open Archival Information System (OAIS) and Generic Statistical Business Process Model (GSBPM). This document addresses the requirements of the OAIS list of Mandatory Responsibilities and is organized around the OAIS Functional Entities, which reflects the flow of information packages through the management entities of the organization. (see Appendix A) The detailed work of the data producing projects, archive management, and delivery activities are organized by an IPUMS Business Process Model based on the GSBPM of the United Nations Economic Commission for Europe. (see Appendix B)

Definition:

- **Dataset:** A set of data made accessible through a designated interface within the IPUMS Live Data Store (access system for current versions of all datasets) and subsequently through the Past Version Archive system. Datasets receive a DOI for each published version of the dataset.
- **Data Series:** A versioned series of datasets. Each dataset DOI references the previous version and immediate next version of the dataset in the series using the tags isNewVersion of and isPreviousVersionOf references.

- **Collection:** An assemblage of related datasets, encompassing one or more data series. Collections are assigned a single DOI and refer to all included datasets and their versions over time.
- **Project:** A funded work group within IPUMS that produces a collection or collections of datasets to support the goals of the project.

Project Workflow

The primary goal of individual IPUMS projects' workflows is to create integrated datasets from data collected by external data producers. This process feeds into the OAIS framework at several points ensuring that input data and metadata, as well as the published collections of the project, are well documented and preserved. The IPUMS Business Process Model, modeled on the Generic Statistical Business Process Model (GSBPM) assists us in identifying common tasks while allowing individual projects to follow the workflows required by their specific needs. Descriptions of the implementation of the OAIS framework and the IPUMS Business Process Model are provided in Appendix B. An important feature of the IPUMS Business Process Model is its use in identifying points at which we capture pieces of metadata required to meet the mandatory responsibilities of providing adequate reference, contextual, provenance, fixity, and access information.

Preservation Description Information

The OAIS Magenta Book (pg.4-30) lists information required for preservation compliance:

- **Reference Information**: taxonomic, reference, and registration information used to identify and/or describe Content Information
- **Context Information**: relates the Content Information on a data object to its environment, including production information
- **Provenance Information**: documents the history of the Content Information, providing some assurance of the likely reliability of the Content Information
- **Fixity Information**: provides the Data integrity checks or validation/verification keys used to ensure that the particular content Information has not been altered
- **Access Rights Information**: identifies access restrictions pertaining to the Content Information, including the legal framework, licensing terms, and access control

IPUMS Activity

Table 1 identifies the Mandatory Responsibilities that an organization must discharge in order to operate an OAIS Archive and the steps IPUMS has taken to meet these responsibilities (information in gray indicates actions that are underway but not complete):

Table 1: Organizational Responsibilities

Mandatory Responsibility	IPUMS Actions
Negotiate for and accept appropriate information from the information Producers.	External Producers: – If data are not publicly available, memorandum of understanding (MOU) outlines data agreement allowing IPUMS to disseminate data to third parties – We maintain provenance and contextual information on data from external producers IPUMS: – Complete provenance and context for each data object – Process information for each data object [currently this information is gathered in several ways in varying levels of detail; policy to support capturing this information exists; consistent process for achieving this is a goal and is in progress]
Obtain sufficient control of the information provided to the level needed to ensure Long Term Preservation.	– Data and information are maintained in formats for long term preservation (ASCII, UTF8, PDF-A) – MOU's include preservation and distribution rights agreements
Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided, thereby defining its Knowledge Base.	– Each project has a clearly defined Designated Community, but in general it is academic and policy researchers engaged in comparative research in the social and behavioral sciences
Ensure that the information to be preserved is **Independently Understandable** to the Designated Community. In particular, the Designated Community should be able to understand the information without needing special resources such as the assistance of the experts who produced the information.	– IPUMS develops and maintains extensive and detailed documentation on content and contextual relationship information, which is available on each collection's website. – Documentation accompanying preservation copies (snapshot versions described below) does not currently contain all information on the website. [Progress is being made on ensuring that information currently available

	through IPUMS search systems and/or in related physical archives is clearly identified and linked to the collections and individual data objects where appropriate]
Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, including the demise of the Archive, ensuring that it is never deleted unless allowed as part of an approved strategy. There should be no ad-hoc deletions.	– Policies are in place to support long term preservation from a management perspective including an institutional commitment and agreements with other OAIS archives through Data Preservation Alliance for the Social Sciences (Data-PASS) – Policies are in place to preserve information following versioning changes – Processes to support clear versioning and preservation of earlier content [in progress]
Make the preserved information available to the Designated Community and enable the information to be disseminated as copies of, or as traceable to, the original submitted Data Objects with evidence supporting its Authenticity.	– Registered as a maintenance organization with DataCite – DOI obtained for all IPUMS products – Source data retained and described in documentation – Working on improving documentation of provenance chain and data transformations [in process of recording source information in a uniform manner; identifying specific production information to retain in provenance chain]

Activity flow of an IPUMS Project

Figure 1 illustrates a subset of the IPUMS OAIS implementation model, focusing on the workflow from data producer to IPUMS user and the creation of Archive Information Packages. The full model is illustrated and described in Appendix B. Note that the input to the archive (blue boxes) are structured Submission Information Packages that arise out of the project production work. Table 2 provides detail of the metadata content captured at each step that are used to inform and run the IPUMS Live Data Access system and populate both the Submission Information Package (SIP) content for the Archive and Dissemination Information Package (DIP) content for the customer of the IPUMS Live Data Access System. The archive creates the Archive Information Package (AIP) from the content of the SIP plus administrative and processing information from the archive.

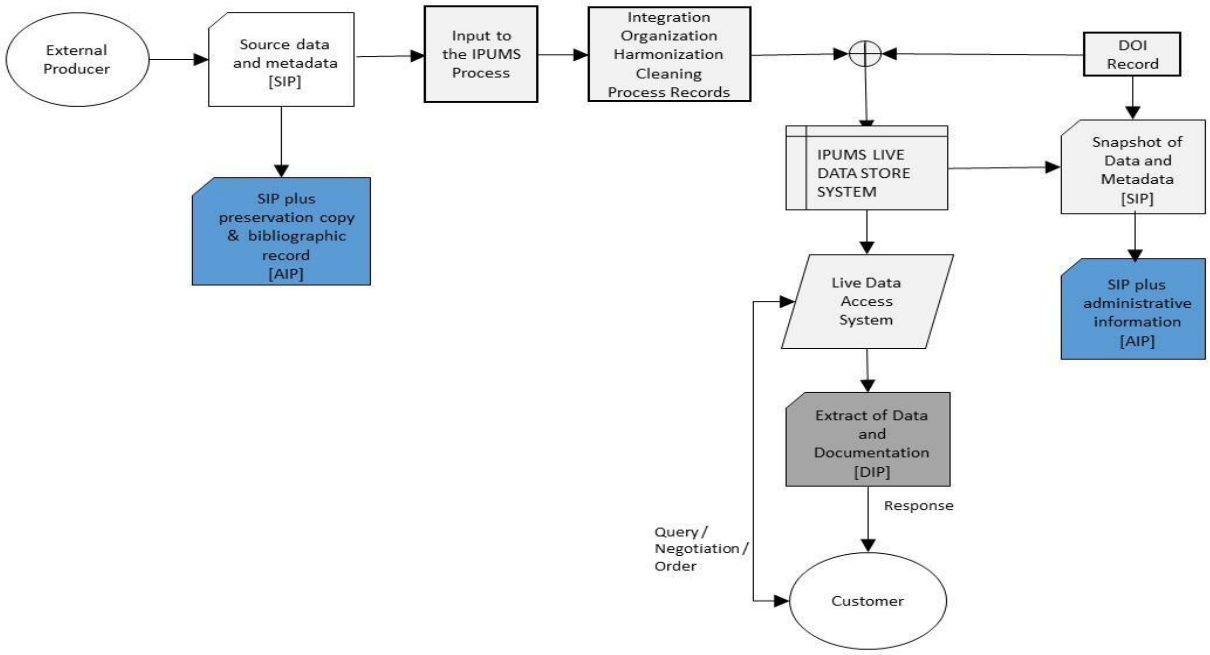


Figure 1: IPUMS Project Workflow (subset of Figure B1)

Table 2: IPUMS - Workflow Description

Activity	**Information Captured**
External Producer	Producer information, alternate source
Source data / metadata	Data; format documentation; variable/datum level conceptual definitions appropriate for the data type, source information, and methodology information; legal requirements and MOU for redistribution rights; and processing instructions, code lists, and geographic definitions. For microdata this generally includes variable definitions, variable data source, data collection form, collection instructions, sampling information. For aggregate data, the table and dimension descriptions, data source, universe, geographic definitions, and imputation information.
SIP – source content	Data and documentation selected by the project for preservation and support of reference, context, and provenance understanding. Source information. Date of acquisition. MOU.
AIP – Source	See Table 3. Archive - Workflow Description (AIP – external producer)
Input to IPUMS process	Results of: Verification of receipt of required documents. Verification of data against metadata (layout, undocumented codes, etc.). Standard confidentiality checks (need for top or bottom coding, small n for response value, etc.). Translation of proprietary data structures.
Integration, Organization, Harmonization	Examples of content include: Software to translate from source data to integrated variable definitions; comparability information; definition of IPUMS specific variables; harmonized data content; geographic integration/harmonization methodology; time series aggregation methods; algorithms for determining means, ranges, etc.

DOI Record	DOI, publication date, citation, contributors, geographic locations, funders, abstract
IPUMS LIVE DATA STORE	Data, data specific metadata, linking tables, DOI number, citation, geographic locations
SIP - Snapshot and DOI	Upon publication, a snapshot of predefined extracts covering all data in the dataset, DOI information, formatted as standard DIP plus separate set of DDI-Codebooks for each extract and dataset as whole.
AIP – IPUMS	See Table 3. Archive - Workflow Description (AIP – internal producer)
Live Data Access System	Information: Collection level metadata; citation and terms of use; data and metadata required to support discovery, access, and linkage; data description and source information as appropriate (Microdata example: variable description, source, comparability, values, availability by sample, count by sample, methodology information, related resources, sample descriptions, sources, contributors, funders, and usage guides. Aggregate example: table and dimension descriptions, universe, geographic availability, related resources, sources, contributors, funders, and usage guides) Note that changes to the data and metadata that do not trigger a version change are recorded with a date stamp in a data series change log. This allows updates and corrections to be made as needed during the life of a major version.
DIP – Extract negotiated by and delivered to Customer	Microdata extract defined by the user through negotiation with Live Data Access System. DDI-Codebook with extract-specific content (contains variable level source, definition, and comparability information). Simple codebook (name, value, physical location, extract filter). Set-up files normally provided (currently SAS, SPSS, Stata). Aggregate data includes a full geographic link to spatial files in the data set and provides a data summary, data dictionary, and citation and use statement in a simple codebook structure. Related spatial files can be included in the download on request. Work continues to expand the coverage of the DDI-Codebook to include more extensive metadata from the Live Data Access System. This would include more source question text and comparative information. This would replace references back to the Live Data Access System. This information would be included in the SIP for the archive.

DOIs

IPUMS is organized into separate collections with unique branding identities. A collection may have one or more individually identified data series. For example, IPUMS CPS consists of a single data series covering each published version of the Current Population Survey (CPS) dataset. IPUMS Global Health consists of two data series, IPUMS DHS and IPUMS PMA. A DOI is created and filed for each version of an IPUMS dataset (e.g., IPUMS CPS, IPUMS DHS, and IPUMS PMA). A collection DOI record is used to connect the related datasets within the data series of the collection. For example, the IPUMS Global Health collection lists all datasets published over time for the two data series found in the collection, IPUMS DHS and IPUMS PMA. The IPUMS CPS collection lists the datasets within the single data series of IPUMS CPS. A DOI is created and filed for each published version, and the collection record is updated

with a new hasPart relationship to designate the new version (Figure 2). Each project defines what constitutes a published version change. All changes to a published collection are recorded in the date-stamped change log of the collection. The DOI record contains Title, Creator(s), Contributors, Funders, Publication Date, Version, Abstract, Geographic Locations (country coverage), and relationships (isPartof, isNewVersionOf, and isPreviousVersionOf). Relationships are used to provide the provenance chain between versions and membership within a specific collection

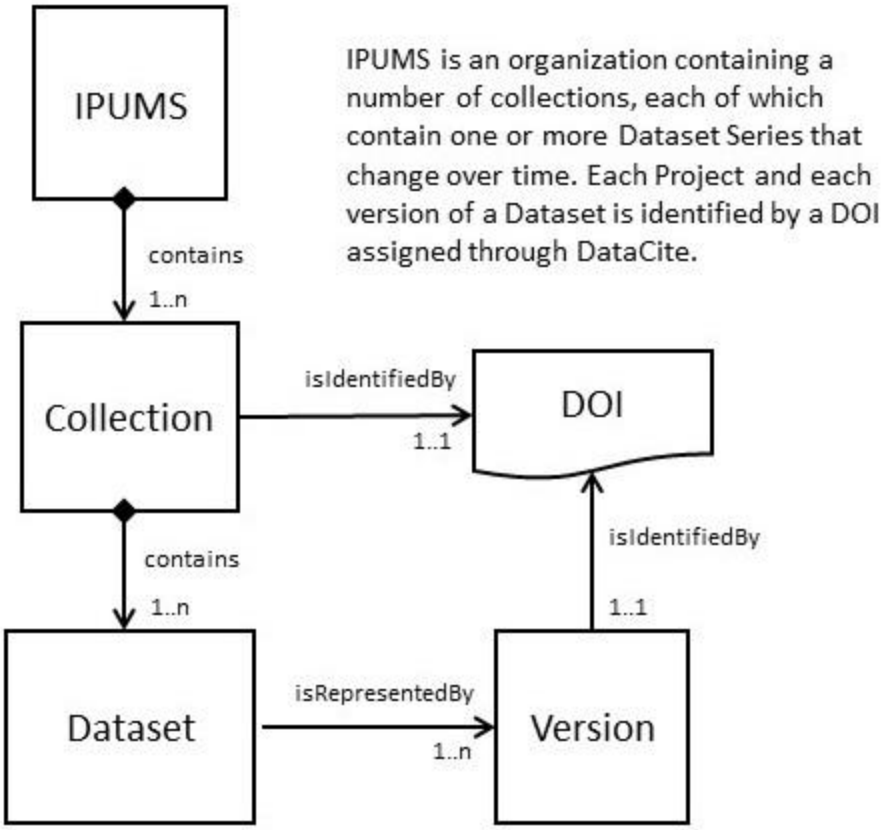


Figure 2: Product organization

Activity Flow of the IPUMS Archive

The IPUMS Archive accepts two types of SIP profiles (Figure 3). The first is from an external producer commonly obtained as part of the material submitted to a project. These may include data files and/or supporting documents obtained for the purpose of understanding and processing the data for inclusion in the project. The project determines what objects are submitted and provides information on the source of the material. Occasionally IPUMS will accept archival material directly from an external source for the purpose of supporting current or planned IPUMS projects. For example, a collection of census materials on Oceania obtained from the EastWest Center on their closure or a detailed collection on the redistricting process in Detroit from a retiring consultant.

The second SIP profile covers preservation copies of each version of the datasets published by the project. These are obtained from the IPUMS Live Data Store when initially published. These include a snapshot of the new version of the dataset plus DOI record content. Content coverage reflects the data

and documentation made available through the Live Data Access System to the customer using predetermined subsets of data that will facilitate future access to each major version of the collection.

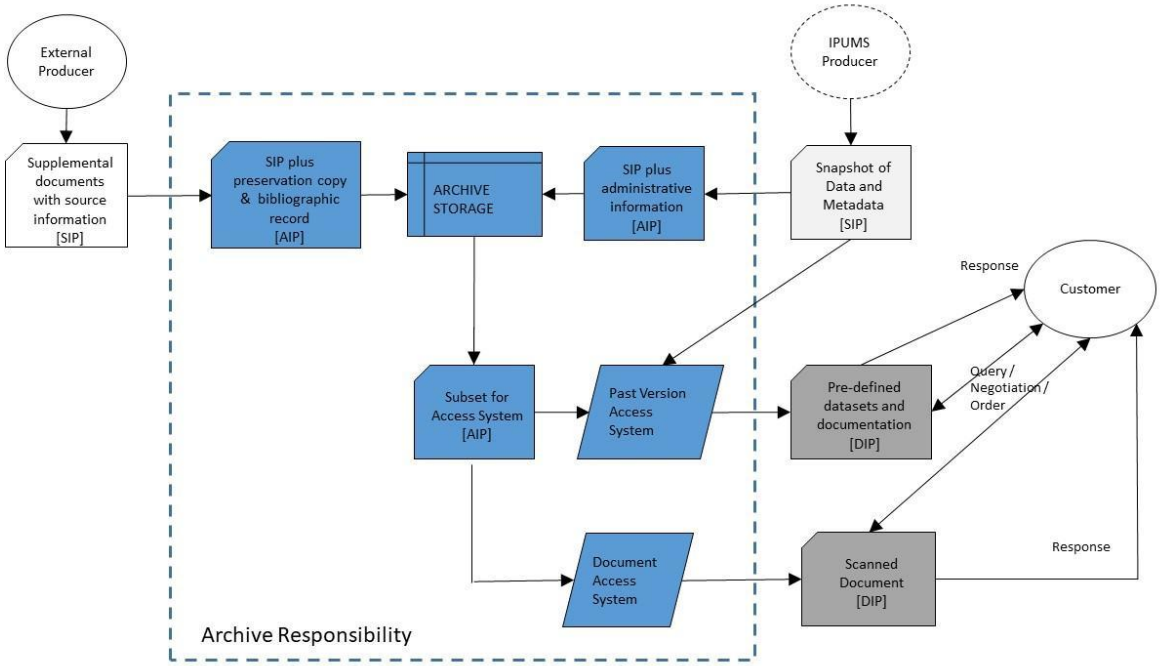


Figure 3: Archive Workflow (subset of Figure B1)

Table 3 identifies each activity area in Figure 3 and indicates the content preserved or captured to ensure access to information on reference, context, provenance, fixity, and access requirements.

Table 3: Archive - Workflow Description

Activity	Information
SIP – external producer	Document and name of provider; permission for distribution via a MOU with external data provider, if not publicly available data (providers include government agencies, intermediate archives, and genealogy organizations); identification of those documents which may be preserved but NOT redistributed (ex. Original data files).
AIP – external producer	Assignment of system unique identifier; creation of bibliographic record including Title, Creator, Producer, Date of production, relationship to broader collection, related Census or Survey date, country, language, descriptors (general and internal for project support), source, link to electronic copies (full and/or partial), and description where clarification is needed. Creation of partial or full-image (PDF-A or TIF) based on current archive policy for preservation and discovery purposes. Note that focus is on clarity of information content rather than physical image representation. Cleaning rules cover straightening, cleaning, and reconstruction of content where needed for clarity.

	Data are replicated in an archival format with layout documentation (ASCII fixed or delimited files).
SIP – internal producer	All datasets are subset into pre-defined extracts to facilitate future use. DDI-Codebook extract-specific content. Simple codebook (name, value, physical location, extract filter). Set-up files normally provided (currently SAS, SPSS, Stata). DDI-Codebook for complete dataset. DOI record. DDI-Codebook contains variable level source, definition, and comparability information.
AIP – internal producer	Expanded DDI-Codebook integrating additional collection level information and additional source question text for selected collections. Snapshot date. Verification of DDI-Codebook.
Archive Storage	Consists of Snapshot output, electronic documents, bibliographic records, DOI records, original data.
Subset for Access System	Bibliographic records, documents with redistribution rights, DOI records, Snapshot output, access rights. Change log for each collection.
Past Version Access System	DOI linked landing page. Citation and abstract from DOI record. Link to change log with applicable date to identify revisions made during the valid period for the version. Link to access requirements. Access to Snapshot content. Links to current and other versions of the dataset as well as the datasets within other data series that are a part of the collection (for example: https://ipums.org/projects/ipums-international/d020.v7.1).
Document Access System	Bibliographic record. Document images. [under development]
DIP – Past Version	Snapshot content for selected predefined sub-set. Change log access.
DIP - Document	PDF of selected document

Appendix A – Implemented Standards

OAIS – Open Archival Information System

OAIS Section 3.1 - MANDATORY RESPONSIBILITIES

This subsection establishes mandatory responsibilities that an organization must discharge in order to operate an OAIS archive.

The OAIS must:

– Negotiate for and accept appropriate information from information Producers.

– Obtain sufficient control of the information provided to the level needed to ensure Long-Term Preservation.

– Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided.

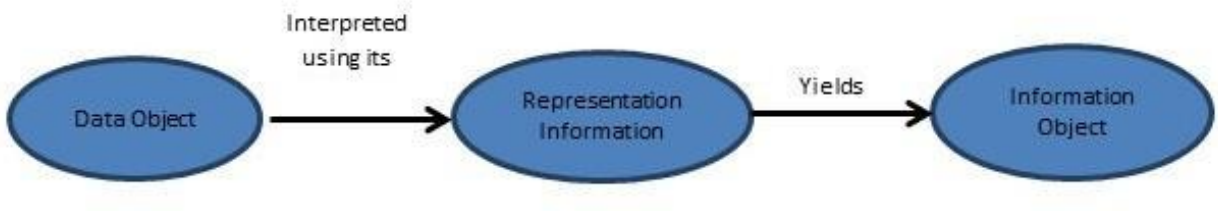
– Ensure that the information to be preserved is Independently Understandable to the Designated Community. In other words, the community should be able to understand the information without needing the assistance of the experts who produced the information.

– Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, and which enable the information to be disseminated as authenticated copies of the original, or as traceable to the original.

– Make the preserved information available to the Designated Community.

Definition of Conformance

This conformance summary is based on the *CCSDS Recommended Practice for an OAIS Reference Model* (June 2012)[1]. Conformance to the OAIS Model is defined as a system which supports the following model and the list of Mandatory Responsibilities (pg. 3-1).



Obtaining Information from Data (pg. 2-4, Figure 2-2)

For IPUMS, Data Objects may be obtained from an external producer or produced internally. The Representation Information we receive from external producers may be limited, but for internally

[1] http://public.ccsds.org/publications/archive/650x0m2.pdf

produced Data Objects, our goal is complete and detailed Representation Information for the purpose of yielding quality Information Objects. Representation Information includes the following (pg.2-6, 2-7):

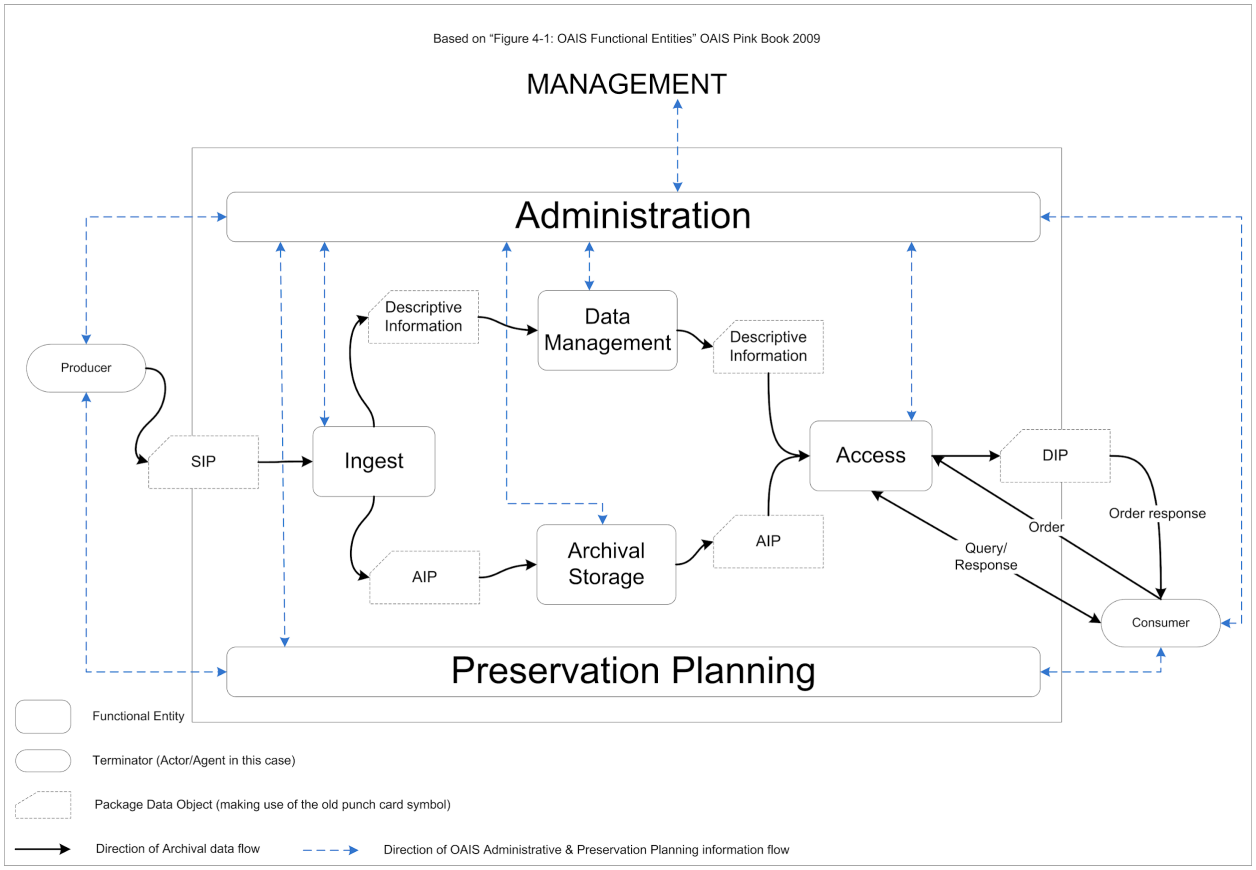
- Provenance: The custody and process history of the Data Object
- Context: The relationship of the Data Object to other Data Objects
- Reference: A unique identifier
- Fixity: Protection from undocumented alteration (i.e. checksum)
- Access Rights: Information covering the access, preservation, distribution, and usage of the Data Object

OAIS Framework

OAIS provides a framework from which to review the processes in IPUMS that manage the flow of data and metadata through the system. IPUMS is not strictly an archive, and as such, the OAIS model does not reflect the full range of its activities. IPUMS does not engage in primary data collection nor does it serve solely as an archive that ingests, preserves, and provides access to a depositor's data. The primary activities of IPUMS focus on acquiring data from an external producer, processing the data and related metadata to integrate it for the purposes of comparative research, providing a means of access to facilitate that research, and then delivering customized packages of data and metadata to the consumer. OAIS focuses on the movement of three data package objects through an archival system:

- SIP: Submission Information Package
- AIP: Archival Information Package
- DIP: Delivery Information Package

These packages move through a system containing the activities of Ingest, Data Management, Archival Storage, and Access, informed by an Administration layer and Preservation Planning.



2

Submission Interactions – SIPs

IPUMS serves as the producer for most collections, selecting data and metadata for input, then integrating and harmonizing data for the content of the final collection. Except for a limited number of input datasets, IPUMS is responsible only for archiving and delivering data and metadata from its final collections. IPUMS maintains other input data files for the purpose of provenance.

Archival Storage – AIPs

Upon publication, each new collection or dataset receives a unique identifier through DataCite. IPUMS creates a snapshot of each new dataset and the associated metadata for preservation purposes. With each subsequent version, IPUMS supplements the AIP of the previous version with a change log of activity taking place between the two versions, providing clear provenance trails between versions.

Consumer Interactions – DIPs

IPUMS currently delivers DIPS through two types of consumer interactions, Ad-hoc and Event Based. Ad-hoc interactions are supported by our dataset-based search systems (available only for the current version of the dataset within a collection) which provide information on the content of the DIP, how it is delivered, and the consumer's rights in terms of usage of the data. The data are free. Representation information not delivered directly to the consumer (e.g., detailed comparison or methodology information) is available on-line within the search system itself.

[2] Image provided by Herve LHours, UKDA

Event Based consumer interactions occur primarily with related registries and repositories. The DIP generally contains only metadata but occasionally covers the complete Data Object and Representation Information package (e.g., Geospatial files in DataONE). For each customer IPUMS has a clear agreement on the overall content of the DIP and the minimum level of information provided.

Business Process Models

GSBPM – Generic Statistical Business Process Model[3]

"The Generic Statistical Business Process Model (GSBPM) describes and defines the set of business processes needed to produce official statistics. It provides a standard framework and harmonised terminology to help statistical organisations to modernise their statistical production processes, as well as to share methods and components. The GSBPM can also be used for integrating data and metadata standards, as a template for process documentation, for harmonising statistical computing infrastructures, and to provide a framework for process quality assessment and improvement. These and other purposes for which the GSBPM can be used are elaborated further in Section VII. This version of the GSBPM is aligned with version 1.2 of the Generic Statistical Information Model (GSIM) and version 1.2 of the Generic Activity Model for Statistical Organisations (GAMSO).[4]

GLBPM – Generic Longitudinal Business Process Model

The GLBPM is a modification of the Generic Statistical Business Process Model that focuses on the longitudinal survey process as employed in longitudinal data gathering by academic, governmental, and private research organizations. [5] IPUMS used this as a source of information on the longitudinal process as well as an approach for modification of the GSBPM. A brief representation of GLBPM is found at http://ddionrails.org/glbpm/

[3] https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1
[4] https://statswiki.unece.org/display/GSBPM/I.+Introduction
[5] https://ddialliance.org/sites/default/files/GenericLongitudinalBusinessProcessModel.pdf

Appendix B – IPUMS Implementation of Standards – General Approach

OAIS Implementation

General approach to implementation of OAIS

IPUMS projects define the negotiation for and acceptance of information from producers (SIP content) where the data are not publicly available. We have a well-defined listing of preferred documentation for census data, and similar profiles exist for projects using other data sources. SIPs created from internal datasets (those held in the IPUMS Live Data Access system) are like those for externally produced data and may have tighter requirements.

The context of the data collected by IPUMS is defined by the individual projects within the overall mission and vision of IPUMS.

Unique identifiers are provided of all versions of our products and the content is maintained in common non-proprietary standard formats to support continued access and use over time.

Our contractual arrangements include rights to retain and distribute (within confidentiality constraints) all the data and metadata we obtain. Our processes create archival quality copies of the SIP content (ASCII data files and PDF versions of print documentation). DIP content is provided in readily transferable formats (ASCII data files, shapefiles, geoTIFFS, ASCII readable setup files for common analysis software, DDI structured codebooks, and simple text data dictionaries).

IPUMS Activities in the OAIS Framework Context

Figure B1 illustrates the full range of IPUMS activities and clarifies their relationship to the OAIS Framework where the activities of the archive, as outlined by OAIS, are within the blue box. The white objects in the diagram represent actors and activities external to IPUMS. The light gray boxes are activities related to the IPUMS projects and are managed by those projects and technical staff. The dark gray objects represent the packages of content (DIPs) provided to the customer following a query and order activity. The blue objects are the responsibility of the archive management. Materials are received through SIPs which conform to requirements based on their source. They are modified at ingest to the Archive Storage system (AIP) and in preparation for presentation (if required) by the archival access system. The customer receives a package of data conforming to their final request (DIP).

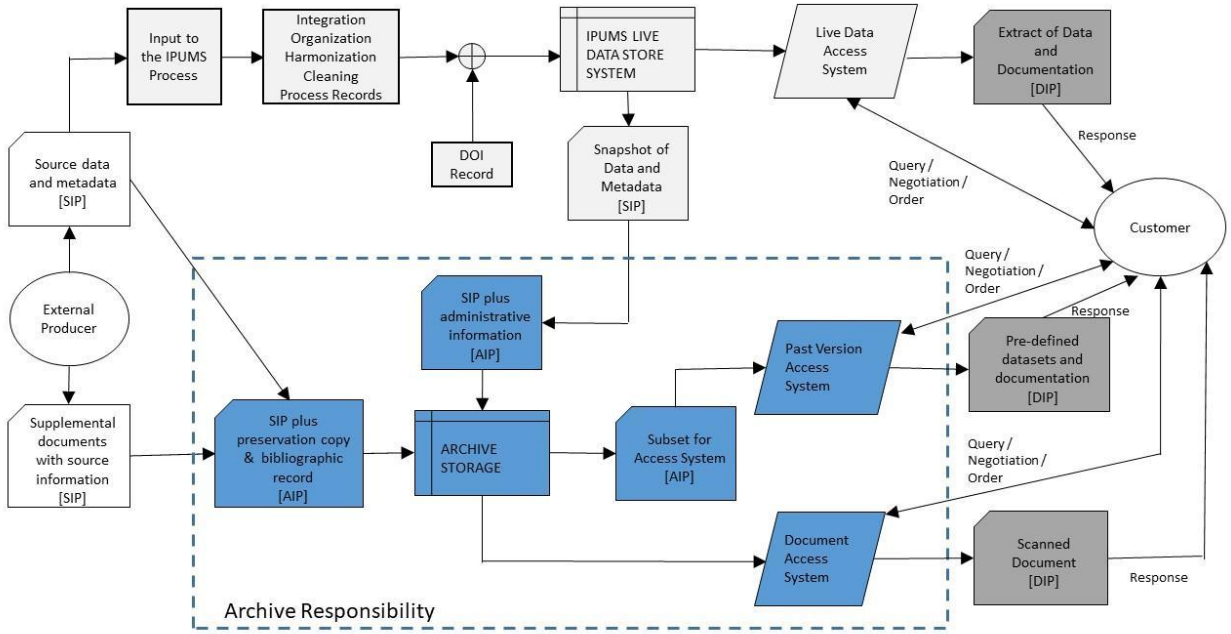


Figure B1: *IPUMS Activities in OAIS Context*

Data Submitted Through SIP

External data are only obtained within the context of a project. The scope, source, and required documentation is determined by the project and obtained from the external producer. Rights and permissions for the use and redistribution of the information is obtained at this point. If the data are to be retained for preservation purposes, the information goes to the project and is separately submitted as an information package for ingestion to the archive. The SIP contains the data; documentation of the content, capture, and processing of the data (as much as available); and permissions for long-term management of the data.

Additional documents related to data provided to a project may also be submitted to the IPUMS archive. These documents may relate to censuses in general or other major areas of data collection for IPUMS. For example, we have an extensive collection of census related documents from the United Nations Statistical Division, United States Census Bureau International Collection, and smaller archives such as the East West Center, the Latin American and Caribbean Demographic Centre (CELADE), and private collections.

Upon publication of a dataset by a project in the IPUMS Live Data Store, a snapshot of the dataset and selected metadata are created and submitted to the Archive for storage and eventual access on the Past Version Access System.

AIPs for Storage and Access

All SIPs are processed to ensure the use of archival formats for content and to provide additional information related to processing and identification. This AIP contains the original SIP, archival copies and additional processing information. Prior to entry to the Past Version Access System, the content of dataset snapshot SIPs is repackaged and organized to provide manageable sub-sets for user exploration and access. The contents of the AIPs are non-proprietary formats of both the data and metadata. The

documentation (including analytic setup files) is provided in multiple formats to facilitate future use. The AIP also includes separate copies of DDI documentation that may be set up for query purposes.

Access Systems and DIP

IPUMS supports two primary means of access to its live data (current version of an IPUMS dataset) and access to archived content (AIPs of past versions of IPUMS datasets and related documents):

- IPUMS provides sophisticated access to both data and metadata of the IPUMS live data store. The consumer explores the metadata online, via web-based access systems or via application programming interfaces (APIs), and specifies the content of the required sub-set from a dataset . The DIP from this system includes the customized extract in non-proprietary format, set-up files for commonly used statistical analysis packages, a simple text codebook, and a DDI (XML) codebook for use in other DDI aware tools or through a transformation to a viewable presentation of the customized codebook.
- Archived content (in AIPs) is accessed through a search and download system which enforces access limitations pertaining to individual AIPs. Customization of DIPs is limited to selection of pre-determined subsets of data and metadata (for example, country samples by census year). The Document Access System (under development) will provide a DIP consisting of the selected scanned document.

IPUMS Business Process Model

In developing a business process model, the goal has been to facilitate communications within IPUMS and with our designated communities, including OAIS archives, statistical agencies, geographic data community, as well as science and research data communities. Our original systems were developed around specifications for access. Our plan is to expand this system to support capturing SIP content and relevant processing stages between SIP and DIP to ensure a clear trail from one to the other.

IPUMS draws on two models: the Generic Statistical Business Process Model (GSBPM)[6] developed by United Nations Economic Commision for Europe (UNECE) and the Generic Longitudinal Business Process Model (GLBPM),[7]a related derivative in the research world. IPUMS is in the process of developing a similar model for the business flow of the IPUMS projects. Each project and iteration of a project varies in specifics and order of the steps followed, but there is a commonality in the goal of each step and often in the process used to accomplish the task.

The IPUMS Business Process Model (BPM) is a customization of the GSBPM and GLBPM, primarily in the areas of Collect and Process/Analyze. These changes reflect the use of secondary data sources and the work of harmonization and integration to create a data infrastructure that supports research across time and space. The model displays the top level (headers) and first level of breakdowns. Additional lists are being developed and reviewed to specify specific sub-activities and tasks where the use of a uniform approach may be implemented. The next steps are to map input and output from activities for the various IPUMS projects, specifying where current processes have been automated or structured in a way to provide uniform output.

[6] https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1
[7] https://www.ddialliance.org/sites/default/files/GenericLongitudinalBusinessProcessModel.pdf

The goal of this model internally is to identify common areas of activity as well as the actual production processes of new and continuing projects within IPUMS. Identification of common activities helps focus IT and organizational approaches that, while beginning in one project, may be applicable to others. This is important to IPUMS in its effort to provide a unified platform for the delivery of data to researchers. The model is also used to identify IPUMS activities where metadata may be generated and/or captured. This approach is being used to focus work on improving the capture and preservation of provenance metadata at the sub-project level.

In addition to capturing harmonization and integration activities, IPUMS BPM also takes into consideration the following IPUMS features:

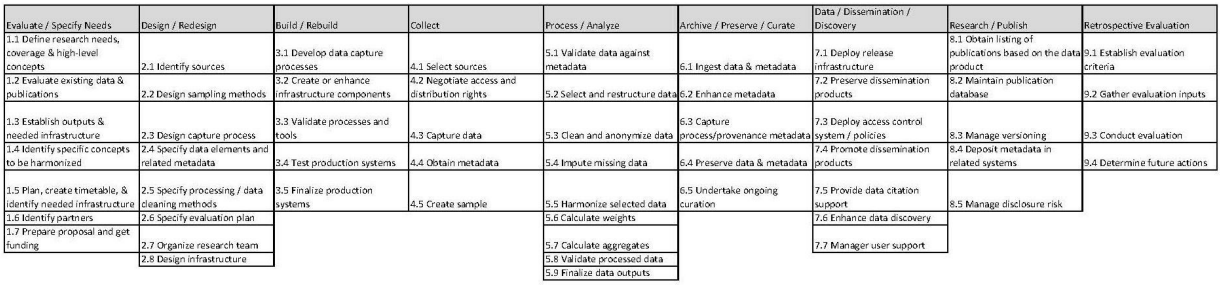
- Funded projects from peer-reviewed competitive funders
- Identification of activities that generate metadata needed for preserving provenance
- Focus on data reuse rather than primary data capture
- Development of a cross-project technical system to support data/metadata management, preservation, and delivery
- Providing a system to support on-going expansion of collections as new iterations of data become available

This chart in Figure B2 is intended to clarify the overall work processes employed within the IPUMS projects. IPUMS requires this type of mapping as the individual project processes must reflect the needs and constraints of their data sources and goals. The value of this approach is that it lets project staff discuss specific activities in a common language, technical staff understand where common tools may be of value, and administrative staff identify process developments in one project that could benefit others. Most importantly, it allows individual projects to identify their own path through the process activities for each iteration and identify specific inputs and outputs in terms of their own needs.

As in the GSBPM, general process areas have been divided into sub-steps which can then be further expanded. The current model has 9 process areas broken down into 4-9 sub-steps. The document IPUMSBPM-Listing extends the detail of these sub-steps in the areas of Evaluate/Specify Needs, Collect, Process/Analyse, Archive/Preserve/Curate, and Research/Publish.

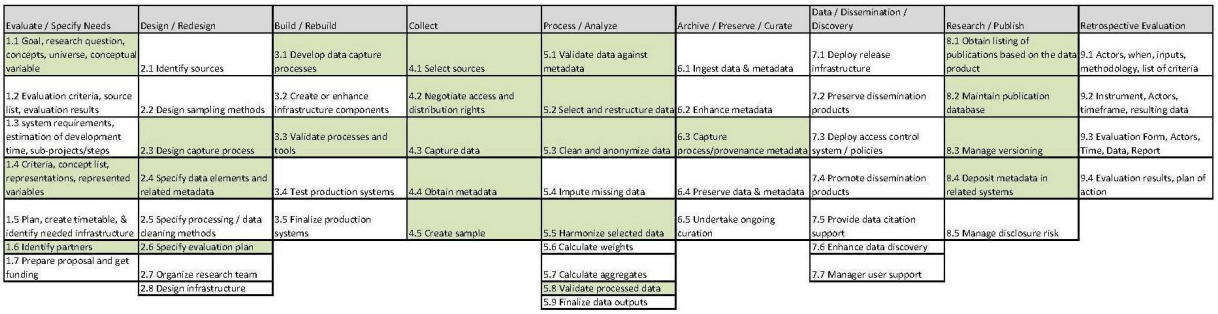
The figures on the following pages provide the First Level Process Model which covers the general activity areas (first row) and general processes within each activity area. Further details are found in a separate document. The second image (Figure B3) shows the use of the model identifying where the IPUMS process may produce metadata that would be required to track the provenance of data through the projects. The third image (Figure B4) shows the relation of several processes that are related to the content and construction of the SIP, AIP, and DIP objects in the OAIS model.

Figure B2: First Level Process



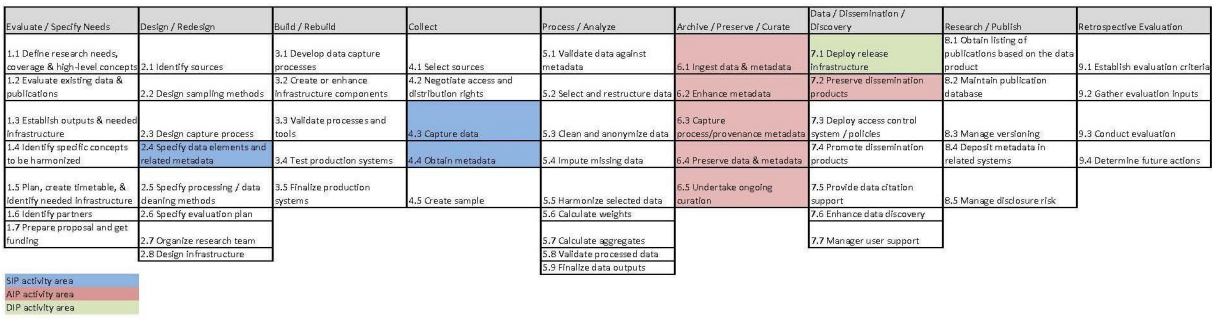
Evaluate / Specify Needs	Design / Redesign	Build / Rebuild	Collect	Process / Analyze	Archive / Preserve / Curate	Data / Dissemination / Discovery	Research / Publish	Retrospective Evaluation
1.1 Define research needs, coverage & high-level concepts	2.1 Identify sources	3.1 Develop data capture processes	4.1 Select sources	5.1 Validate data against metadata	6.1 Ingest data & metadata	7.1 Deploy release infrastructure	8.1 Obtain listing of publications based on the data product	9.1 Establish evaluation criteria
1.2 Evaluate existing data & publications	2.2 Design sampling methods	3.2 Create or enhance infrastructure components	4.2 Negotiate access and distribution rights	5.2 Select and restructure data	6.2 Enhance metadata	7.2 Preserve dissemination products	8.2 Maintain publication database	9.2 Gather evaluation inputs
1.3 Establish outputs & needed infrastructure	2.3 Design capture process	3.3 Validate processes and tools	4.3 Capture data	5.3 Clean and anonymize data	6.3 Capture process/provenance metadata	7.3 Deploy access control system / policies	8.3 Manage versioning	9.3 Conduct evaluation
1.4 Identify specific concepts to be harmonized	2.4 Specify data elements and related metadata	3.4 Test production systems	4.4 Obtain metadata	5.4 Impute missing data	6.4 Preserve data & metadata	7.4 Promote dissemination products	8.4 Deposit metadata in related systems	9.4 Determine future actions
1.5 Plan, create timetable, & identify needed infrastructure	2.5 Specify processing / data cleaning methods	3.5 Finalize production systems	4.5 Create sample	5.5 Harmonize selected data	6.5 Undertake ongoing curation	7.5 Provide data citation support	8.5 Manage disclosure risk	
1.6 Identify partners	2.6 Specify evaluation plan			5.6 Calculate weights		7.6 Enhance data discovery		
1.7 Prepare proposal and get funding	2.7 Organize research team			5.7 Calculate aggregates		7.7 Manager user support		
	2.8 Design infrastructure			5.8 Validate processed data				
				5.9 Finalize data outputs				

Figure B3: Metadata Creation Points



Evaluate / Specify Needs	Design / Redesign	Build / Rebuild	Collect	Process / Analyze	Archive / Preserve / Curate	Data / Dissemination / Discovery	Research / Publish	Retrospective Evaluation
1.1 Goal, research question, concepts, universe, conceptual variable	2.1 Identify sources	3.1 Develop data capture processes	4.1 Select sources	5.1 Validate data against metadata	6.1 Ingest data & metadata	7.1 Deploy release infrastructure	8.1 Obtain listing of publications based on the data product	9.1 Actors, when, inputs, methodology, list of criteria
1.2 Evaluation criteria, source list, evaluation results	2.2 Design sampling methods	3.2 Create or enhance infrastructure components	4.2 Negotiate access and distribution rights	5.2 Select and restructure data	6.2 Enhance metadata	7.2 Preserve dissemination products	8.2 Maintain publication database	9.2 Instrument, Actors, timeframe, resulting data
1.3 system requirements, estimation of development time, sub-projects/steps	2.3 Design capture process	3.3 Validate processes and tools	4.3 Capture data	5.3 Clean and anonymize data	6.3 Capture process/provenance metadata	7.3 Deploy access control system / policies	8.3 Manage versioning	9.3 Evaluation Form, Actors, Time, Data, Report
1.4 Criteria, concept list, representations, represented variables	2.4 Specify data elements and related metadata	3.4 Test production systems	4.4 Obtain metadata	5.4 Impute missing data	6.4 Preserve data & metadata	7.4 Promote dissemination products	8.4 Deposit metadata in related systems	9.4 Evaluation results, plan of action
1.5 Plan, create timetable, & identify needed infrastructure	2.5 Specify processing / data cleaning methods	3.5 Finalize production systems	4.5 Create sample	5.5 Harmonize selected data	6.5 Undertake ongoing curation	7.5 Provide data citation support	8.5 Manage disclosure risk	
1.6 Identify partners	2.6 Specify evaluation plan			5.6 Calculate weights		7.6 Enhance data discovery		
1.7 Prepare proposal and get funding	2.7 Organize research team			5.7 Calculate aggregates		7.7 Manager user support		
	2.8 Design infrastructure			5.8 Validate processed data				
				5.9 Finalize data outputs				

Figure B4: OAIS Information Package activities



Evaluate / Specify Needs	Design / Redesign	Build / Rebuild	Collect	Process / Analyze	Archive / Preserve / Curate	Data / Dissemination / Discovery	Research / Publish	Retrospective Evaluation
1.1 Define research needs, coverage & high-level concepts	2.1 Identify sources	3.1 Develop data capture processes	4.1 Select sources	5.1 Validate data against metadata	6.1 Ingest data & metadata	7.1 Deploy release infrastructure	8.1 Obtain listing of publications based on the data product	9.1 Establish evaluation criteria
1.2 Evaluate existing data & publications	2.2 Design sampling methods	3.2 Create or enhance infrastructure components	4.2 Negotiate access and distribution rights	5.2 Select and restructure data	6.2 Enhance metadata	7.2 Preserve dissemination products	8.2 Maintain publication database	9.2 Gather evaluation inputs
1.3 Establish outputs & needed infrastructure	2.3 Design capture process	3.3 Validate processes and tools	4.3 Capture data	5.3 Clean and anonymize data	6.3 Capture process/provenance metadata	7.3 Deploy access control system / policies	8.3 Manage versioning	9.3 Conduct evaluation
1.4 Identify specific concepts to be harmonized	2.4 Specify data elements and related metadata	3.4 Test production systems	4.4 Obtain metadata	5.4 Impute missing data	6.4 Preserve data & metadata	7.4 Promote dissemination products	8.4 Deposit metadata in related systems	9.4 Determine future actions
1.5 Plan, create timetable, & identify needed infrastructure	2.5 Specify processing / data cleaning methods	3.5 Finalize production systems	4.5 Create sample	5.5 Harmonize selected data	6.5 Undertake ongoing curation	7.5 Provide data citation support	8.5 Manage disclosure risk	
1.6 Identify partners	2.6 Specify evaluation plan			5.6 Calculate weights		7.6 Enhance data discovery		
1.7 Prepare proposal and get funding	2.7 Organize research team			5.7 Calculate aggregates		7.7 Manager user support		
	2.8 Design infrastructure			5.8 Validate processed data				
				5.9 Finalize data outputs				

SIP activity area
AIP activity area
DIP activity area