IPUMS

Working Papers

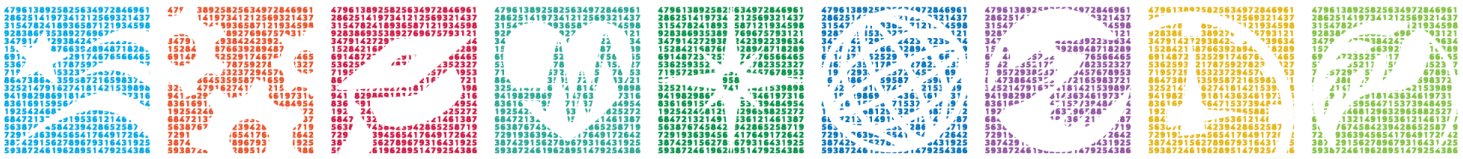
The IPUMS Business Process Model: Instituting a Workflow Mapping Strategy to Support Archival Processes

Diana L. Magnuson†
University of Minnesota

August 2024

Working Paper No. 2024-06
DOI: https://doi.org/10.18128/IPUMS2024-06

†Address correspondence to Diana Magnuson: magn0031@umn.edu.



The IPUMS Business Process Model: Instituting a Workflow Mapping Strategy to Support Archival Processes

Diana L. Magnuson[1]

Abstract

The IPUMS Preservation Archive is instituting a workflow mapping strategy to further identify IPUMS process and metadata capture points to expand its holdings in the data archive. Drawing on two business process models, the Generic Statistical Business Process Model (GSBPM) and the Generic Longitudinal Business Process Model (GLBPM), archival staff have created an IPUMS Business Process Model (IPUMS BPM). The IPUMS BPM reflects 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. Internally, the IPUMS BPM provides a clear visualization of the IPUMS workflow from external submission of data, harmonization process, documentation, extraction systems, and archival preservation of metadata. The challenge for archival staff is furthering the understanding and adoption of the IPUMS BPM within the IPUMS project groups, and to identify metadata production points that require the intervention of the archive for provenance and preservation purposes. This paper identifies the value of instituting this mapping approach to gain a clearer understanding of the role of the archive within project work cycles, points where production and preservaton activities intersect, and opportunities to expand archival holdings.

Keywords

business process model, archive, metadata, preservation

Introduction

IPUMS at the University of Minnesota has created the world's largest accessible database of census and survey microdata, and geographic summary tables. The primary work of IPUMS is data harmonization—making census and survey data compatible across time and space. IPUMS integration and documentation makes it easy for researchers to study change, conduct comparative research, merge information across data types, and analyze individuals within family and community contexts. As of this writing, the IPUMS suite of products contains nine harmonized data collections. International

data comes from over one hundred national and regional statistical organizations. All data and documentation are freely available to the global public.

The IPUMS Preservation Archive has historically functioned as a secondary unit to the main IPUMS data harmonization product line. As such, our data curation work was often operating in "stealth" mode from the perspective of IPUMS data product managers (Magnuson 2015a). The first IPUMS product, what is now known as IPUMS USA, was launched in 1993 (Magnuson and Ruggles 2022). IPUMS International followed in 1999, and formal agreements with partner international statistical organizations committed the Minnesota Population Center (now the Institute for Social Research and Data Innovation) to curating and preserving tens of thousands of ancillary materials used in support of IPUMS International data harmonization work (Ruggles et al. 2003). In 2001 the Minnesota Population Center hired data curator Wendy Thomas and under her guidance a nascent manuscript curation workflow began to take shape, culminating in the launch of a public-facing document access system in 2023 (Magnuson 2024).

2016 was a watershed year for IPUMS archival preservation work. The Minnesota Population Center (MPC) reorganized into the Institute for Social Research and Data Innovation (ISRDI), with IPUMS becoming its own center within the new structure. At this point in time IPUMS "clarified its mission in terms of data harmonization, access, curation, and preservation;" and developed institutional guidelines around data product versioning, preservation, and assigning DOIs (Magnuson and Thomas 2023). The role of the archive during this realignment was enhanced but still outside the immediate vision of most internal IPUMS stakeholders.

The effort to achieve Core Trust Seal certification, motivated by our institutional realignment and internal commitment to establishing clear versioning rules around our data products and registering DOIs, began in 2018 and culminated successfully in 2023. Core Trust Seal (CTS) certification required us to pull together policy documentation that was scattered, incomplete, out of date, or simply non-existent for practices already in place. Building IPUMS policy documentation to complete the CTS application, especially around our preservation practices, clarified our institutional strengths

and identified areas to refine. For the IPUMS Preservation Archive, this process documentation provided the blueprint for future action.

This paper will demonstrate the power of a clear business process model for developing archival goals in an organizational setting in which the archive function is vital but secondary to the main product.

Expanding IPUMS Preservation Archive work

Four areas of institutional maturation across thirty years of IPUMS institutional history encouraged the administrative decision to expand the work and holdings of the IPUMS Preservation Archive.

First, the steady growth of the number, size, and complexity of IPUMS projects over its thirty-year history necessitated the clarification, expansion, and reconfiguration of IPUMS archival work. With funding from both the University of Minnesota and the National Science Foundation (NSF), the Social History Research Laboratory at the University of Minnesota converted existing public use samples of the U.S. census from 1880 to 1980 (excluding 1890 which was destroyed by fire in 1921) into a single coherent series with extensive documentation. The resulting IPUMS data series was first disseminated through an anonymous FTP site, and the first dataset was downloaded on November 19, 1993 (Magnuson and Ruggles 2022). Over time, the IPUMS data integration project expanded to include other U.S. microdata sources: the Current Population Survey (IPUMS CPS), the American Community Survey (in IPUMS USA), the National Health Interview Survey and Medical Expenditure Panel Survey (IPUMS Health Surveys), and survey data on scientists and engineers (IPUMS Higher Ed). IPUMS International (first data release in 2002) harmonizes and disseminates census and survey data from around the world, presently partnering with 103 countries. Global health data are harmonized and disseminated as IPUMS DHS, IPUMS MICS and IPUMS PMA. The data collection incorporates historical full-count international census data of the North Atlantic Population Project (NAAP, first released in 2001; now included in IPUMS International) and IPUMS USA Full Count data (1790-1950).

Aggregate data are harmonized and disseminated through the National Historical Geographic Information System (IPUMS NHGIS), IPUMS Terra (combining global population, land use, and environmental data; now decommissioned), and the International Historical Geographic Information System (IPUMS IHGIS). Another aggregate data collection, the Contextual Determinants of Health (CDOH) provides access to measures of disparities, policies, and counts, at the state and national level, for historical marginalized populations. IPUMS Time Use has harmonized data from time diary surveys: ATUS, AHTUS, and MTUS.

Second, increasing expectations of funding organizations regarding IPUMS' preservation practices motivated internal effort to align with external standards. IPUMS is currently supported by a variety of funding agencies and foundations including the National Institutes of Health (NIH, NIA), the National Science Foundation (NSF) and the Bill and Melinda Gates Foundation.[2] IPUMS' commitment to preservation, discoverability, documentation, and dissemination goes beyond our grant promises. We recognize that the unique resources IPUMS has created and maintains need to be accessible long into the future for new research, scholarly replication, and for unanticipated creative uses of the data. The added impetus of evolving funder expectations around preservation increased the institutional visibility of the archive and its professional concerns.

Third, the developing expertise of IPUMS IT teams encouraged and supported the decision to expand the IPUMS Preservation Archive. Across the thirty years of IPUMS institutional history, a unique partnership between researchers and technologists grew in which systems and tools were collaboratively developed to support IPUMS data and metadata harmonization, documentation, and dissemination (Ruggles et al. 2023, Ruggles et al. 2003a, Ruggles et al. 2003b, Esteve and Sobek 2003, Fitch and Ruggles 2003, Block and Thomas 2003, Hall et al. 1999, Ruggles et al. 1996). The IPUMS IT team grew from graduate student support in the 1990s, its first full-time hire in 2000, to (currently) seventeen full-time staff of software developers, UX/UI specialists, data engineers, operations staff, and managers (Fabrizio 2023, Magnuson 2015b). The IPUMS IT team of designers, developers, and system administrators "collectively build all of the systems required to produce and disseminate the

many IPUMS data products."[3] The longevity of IPUMS IT staff (twice the industry average) is testimony to the vitality of this dynamic and productive partnership between researchers and technologists (Fabrizio 2023).

Lastly, the rigorous CTS application process requires applicants to assess ongoing compliance and compliance stretch goals with respect to "the characteristics required to be a trustworthy repository for digital data and metadata."[4] This assessment involved a significant investment of time to develop a sustainable metadata culture within ISRDI, including: articulating the archive role within the organization; creating an archival workflow that makes sense in the unique IPUMS environment;[5] producing and leveraging documentation;[6] and working to comply with recognized international preservation standards.[7] A valuable byproduct of all this work was illuminating the enormous intellectual investment IPUMS product teams contributed to collecting, harmonizing, organizing, cleaning, documenting, and disseminating IPUMS' unique data collections.

Maturation in these four areas brought the IPUMS organization to the point where the will and capacity coexisted to expand the archival preservation workflow to include additional metadata produced by IPUMS data products, as well as key input artifacts, such as source datasets for which IPUMS is the primary holder. The challenge for archival staff was to leverage this opportunity to communicate to IPUMS administrators, project managers, and IT staff the benefits of adopting the IPUMS BPM to support the expansion of the IPUMS Preservation Archive.

The IPUMS Business Process Model (IPUMS BPM)

The IPUMS Business Process Model (IPUMS BPM) was fleshed out by data curator Wendy Thomas as part of the CTS application process. Drawing on two business process models, the Generic Statistical Business Process Model (GSBPM) and the Generic Longitudinal Business Process Model (GLBPM), Thomas created an IPUMS Business Process Model (IPUMS BPM).[8] Developing an organization model that clearly and accurately reflected the workflow of IPUMS products and archival processes was a crucial step in developing CTS application materials and a key document clarifying

"the role of the archive within the organization and provid[ing] the archive with the means to clearly present that role and identify specific touchpoints to the IPUMS project workflows" (Magnuson and Thomas 2023). The key to the IPUMS BPM is its flexibility: commonalities between the processes of individual IPUMS projects can be identified in the model while allowing for differences in the selection and ordering of tasks within each project over time. IPUMS project managers are not "locked in" to one path through the IPUMS BPM, thus preserving their project specific workflow and our institutional standards.

The current iteration of our business process model contains nine general process areas with sub-levels that are tailored to IPUMS specific tasks (Table 1). These nine process "activities" and their respective sub-levels are presented in outline, descriptive, and "map" formats, and described in detail in the IPUMS Archive Workflow documentation.[9]

Table 1. IPUMS Business Process Model (IPUMS BPM) outline

Evaluate/Specify Needs
1.1 Define research needs, coverage and high-level concepts
1.2 Evaluate existing data and publications
1.3 Establish outputs and needed infrastructure
1.4 Identify specific concepts to be harmonized
1.5 Plan, create timetable, and identify needed infrastructure
1.6 Identify partners
1.7 Prepare proposal and obtain funding
Design/Redesign
2.1 Define research needs, coverage and high-level concepts
2.2 Evaluate existing data and publications
2.3 Design capture process
2.4 Specify data elements and related metadata
2.5 Specify processing/data cleaning methods
2.6 Specify evaluation plan
2.7 Organize research team
2.8 Design infrastructure
Build/Rebuild
3.1 Develop data capture processes
3.2 Create or enhance infrastructure components

3.3 Validate processes and tools 3.4 Test production systems 3.5 Finalize production systems
Collect
4.1 Select sources 4.2 Negotiate access and distribution rights 4.3 Capture data 4.4 Obtain metadata 4.5 Create sample
Process/Analyze
5.1 Validate data against metadata 5.2 Select and restructure data 5.3 Clean and anonymize data 5.4 Impute missing data 5.5 Harmonize selected data 5.6 Calculate weights 5.7 Calculate aggregates 5.8 Validate processed data 5.9 Finalize data outputs
Archive/Preserve/Curate
6.1 Ingest data and metadata 6.2 Enhance metadata 6.3 Capture process/provenance metadata 6.4 Preserve data and metadata 6.5 Undertake ongoing curation
Data/Dissemination/Discovery
7.1 Deploy release infrastructure 7.2 Preserve dissemination products 7.3 Deploy access control system/policies 7.4 Promote dissemination products 7.5 Provide data citation support 7.6 Enhance data discovery 7.7 Manage user support
Research/Publish
8.1 Obtain listing of publications based on the data product 8.2 Maintain publication database 8.3 Manage versioning 8.4 Deposit metadata in related systems

8.5 Manage disclosure work
Retrospective Evaluation
9.1 Establish evaluation criteria 9.2 Gather evaluation inputs 9.3 Conduct evaluation 9.4 Determine future actions

Armed with the IPUMS BPM and administrative encouragement, archival staff organized an initial meeting with each of the nine IPUMS project management teams to examine the IPUMS BPM and begin a conversation about how archival staff was positioning to support IPUMS project teams more effectively and expand the IPUMS Preservation Archive work. After several meetings, it became clear that utilizing the activity "map" visualization of the IPUMS BPM (Figure 1) facilitated an immediate grasp of project specific IPUMS workflows and archival touchpoints from external submission of data, harmonization process, extraction systems, documentation, and archival preservation of metadata.

Figure 1. IPUMS Business Process Model (IPUMS BPM) as an activity "map"

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				

From an archival perspective, the purpose of the meetings was threefold. First, to use the IPUMS BPM as a locus to identify preservation priorities for each project. Second, to expand project managers' thinking about preserving intellectual contribution relating to process and methodology of developing IPUMS data collections. Historically IPUMS data collection teams have primarily been (rightly) concerned with preserving the data that is disseminated to users and less intentionally attentive to preserving the pieces of intellectual activity that contributed to the data harmonization

process. Preserving all these elements is important not only to IPUMS' institutional history narrowly, but a significant contribution to social science infrastructure more broadly. Lastly, the meetings persuaded IPUMS project managers of the future utility both internally for staff and externally for data users, of preserving their enormous investment collecting, harmonizing, documenting, and disseminating IPUMS' unique data collections. Once project teams began talking about the intellectual activity underlying the processes and metadata they were creating, they eagerly identified potential areas for long term preservation (Table 2).

Table 2. Preservation priorities of IPUMS project teams (denoted by "x")

Key below*	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**
Annual Reports to Funders			x	x					
Git Hub				x		x			
Grant Proposals	x	x	x	x	x	x	x	x	x
Process Documentation	x	x	x	x	x	x	x	x	
Producer Documentation			x	x		x			
Software Tools (IPUMS IT)				x		x			
Syntax			x	x	x				
Training Materials (internal)	x	x	x	x	x	x	x		
Training Materials (external)	x	x	x	x	x	x	x	x	x
Translation Tables			x						

Webpages	x	x	x	x	x	x	x	x	x
Wiki	x	x	x	x	x	x	x	x	x

*(1) USA; (2) CPS; (3) International; (4) Global Health; (5) NHGIS; (6) IHGIS; (7) Time Use; (8) Health Surveys; (9) HigherEd; Terra (now decommissioned); CDOH; and Historical Census Projects

After these productive information gathering meetings, archival staff created five clear action points. First, for each IPUMS project the archive will identify metadata creation points using the IPUMS BPM. Second, in consultation with project managers and IPUMS IT, archival staff will differentiate between business continuity preservation (short term) and archive preservation (permanent) practices for the content and metadata identified for preservation. Next, project metadata capture requests will be prioritized. Fourth, archival staff will collaborate with IPUMS IT to create tools for archivist led metadata capture and preservation. Finally, a metadata capture schedule will be reviewed with project stakeholders and implemented in collaboration with IPUMS IT.

Advantages of implementing the IPUMS BPM

As we have previously noted, implementing the IPUMS BPM has advantages for the projects, the administrative team, the IT team, and the IPUMS Preservation Archive. First, common vocabulary is used across all organizational entities, streamlining and enhancing archival related communication. Second, the IPUMS IT team can better develop tools for use across projects, which in turn advances efficiencies and economies of scale around all aspects of the IPUMS data acquisition, harmonization, documentation, dissemination, and preservation workflow. At the administrative level, process and tool developments can be identified for use across projects and for future grant development purposes. Third, the IPUMS BPM is flexible enough to provide both institutional continuity and individualized project workflows. Lastly, the IPUMS BPM flags the areas of project metadata production that require the attention of the archive for provenance and preservation purposes (Magnuson and Thomas 2023).

Strategic use of the IPUMS BPM will support and further four IPUMS Preservation Archive objectives. First, to provide access to current and previous versions of IPUMS data. Second, to retain internal-use data and metadata for the purpose of provenance and quality assurance. Third, to meet archival requirements of external funding agencies, demonstrating compliance with data, metadata, and archival standards, including disciplinary standards of data users and the digital preservation community. Finally, to maintain a digital preservation program that is nimbly responsive to the ever-changing technological environment.

For the IPUMS Preservation Archive staff in particular, use of the IPUMS BPM has continued to clarify the role of archival function within the organization and provide us "with the means to clearly present that role and identify specific touchpoints to IPUMS project workflows" (Magnuson and Thomas 2023). Further, use of the IPUMS BPM in conversation with IPUMS project managers has expanded communication regarding the vital importance of the enormous investment project teams have contributed over the history of the IPUMS data harmonization production work. Preservation of this important intellectual work in developing process, systems, and tools for data harmonization, documentation, and dissemination is vital for the intelligent use and analysis of the data by contemporary and future researchers.

Conclusion

Is your archival function situated in an organizational setting in which your preservation work is vital but secondary to the main product? Are you playing preservation "catch-up" with a fast growing "main product" within your organization? The IPUMS BPM, a workflow mapping strategy employed by the IPUMS Preservation Archive, is applicable to other data archive contexts, especially those in which preservation work is secondary to the primary product of the institution. Actionable steps to move forward with, develop, and expand archival goals include:

- Create a business process model (BPM) that reflects the workflow and metadata creation points for your organization.

- Invest in using the BPM to effectively communicate with organization stakeholders.
- Align metadata preservation, discoverability, documentation, and accessibility with international standards.
- Leverage collaborative work with IT staff to develop long- and short-term goals. Identify strengths and weaknesses of your current archival data management plan and how IT involvement can help you reach your preservation goals.
- Maintain a digital preservation program that is responsive to the changing technological environment.
- Do not waiver from the position that preservation, discoverability, documentation, and accessibility of archival material in all its forms is valuable to the main product of your organization.

Investing in creating and implementing a business process model will benefit your archival workflow, support your short- and long-term preservation goals, highlight where production and preservation activities intersect, and thus enhance archival support of your organization's main product.

References

Block, W. and Thomas, W. (2003) "Implementing the Data Documentation Initiative at the Minnesota Population Center," Historical Methods: A Journal of Quantitative and Interdisciplinary History, Volume 36, No. 2.

Esteve, A. and Sobek, M. (2003) "Challenges and Methods of International Census Harmonization," Historical Methods: A Journal of Quantitative and Interdisciplinary History, Volume 36, No. 2.

Fabrizio, F. (2023) "IPUMS Technology: 30 Years of Innovation, A Unique Partnership of Researchers and Technologists," Data-Intensive Research Conference, Minneapolis, Minnesota.

Fitch, C.A. and Ruggles, S. (2003) "Building the National Historical Geographic Information System," Historical Methods: A Journal of Quantitative and Interdisciplinary History, Volume 32, No. 1.

Hall, P.K, Fitch, C., Canaday, M., Ebeltoft-Kraske, L., Ronnander, C., and Thomas. K.M. (1999) "IPUMS Metadata: Documenting 150 Years of Census Microdata," Historical Methods: A Journal of Quantitative and Interdisciplinary History, Volume 32, No. 3.

Magnuson, D.L. (2024) "Stewarding our resources: Building a sustainable IPUMS archival document access system," IASSIST Quarterly, Volume 48 Number 1.
https://iassistquarterly.com/index.php/iassist/article/view/1095/1032

Magnuson, D.L. (2015a) Wendy Thomas interview, University of Minnesota, March 24, 2015.

Magnuson, D.L. (2015b) Todd Gardner interview, University of Minnesota, May 27, 2015.

Magnuson, D.L. and Ruggles, S. (2022) “Challenges of Large-Scale Data Processing in the 1990s: The IPUMS Experience,” IEEE Annals of the History of Computing, pp. 71-83. https://ieeexplore.ieee.org/abstract/document/9972862

Magnuson, D.L. and Thomas, W. L. (2023) ”Expanding our perspective: Building a sustainable metadata culture,” IASSIST Quarterly, Volume 42 Number 2. https://iassistquarterly.com/index.php/iassist/article/view/1046

Ruggles, S., Cleveland, L., and Sobek, M. (2023) ”Harmonizing Global Census Microdata: IPUMS International,” in Irina Thomescu-Bubrow, Christof Wolf, Kazimierz M. Slomeczynski, and J. Craig Jenkins (eds) Survey Data Harmonization in the Social Sciences. New York: Wiley, pp. 207-226. DOI:10.1002/9781119712206

Ruggles, S., Sobek, M., King, M.L., Liebler, C., and Fitch, C.A. (2003a) ”IPUMS Redesign,” Historical Methods: A Journal of Quantitative and Interdisciplinary History, Volume 32, No. 1.

Ruggles, S., King, M.L., Levison, D., McCaa, R., and Sobek, M. (2003b) ”IPUMS International,” Historical Methods: A Journal of Quantitative and Interdisciplinary History, Volume 32, No. 2.

Ruggles, S., Sobek, M., and Gardner, T. (1996) ”Disseminating Historical Census Data on the World Wide Web,” IASSIST Quarterly, Volume 20 Number 3. https://iassistquarterly.com/index.php/iassist/issue/view/15

Endnotes

[1] Diana L. Magnuson is Curator and Historian at the Institute for Social Research and Data Innovation, University of Minnesota (magn0031@umn.edu).
[2] https://www.ipums.org/about/funding
[3] https://tech.popdata.org/about/about-isrdi-it
[4] https://www.coretrustseal.org/why-certification/requirements/
[5] https://www.ipums.org/workflows
[6] https://assets.ipums.org/_files/ipums/Digital_Preservation_Framework_May2022.pdf
[7] https://dataverse.nl/dataset.xhtml?persistentId=doi:10.34894/CRANSO
[8] https://assets.ipums.org/_files/ipums/workflows/IPUMS_Archive_Workflow_Nov2021.pdf
[9] https://assets.ipums.org/_files/ipums/workflows/IPUMS_BPM_Outline_Nov2021.pdf