IPUMS
Working Papers

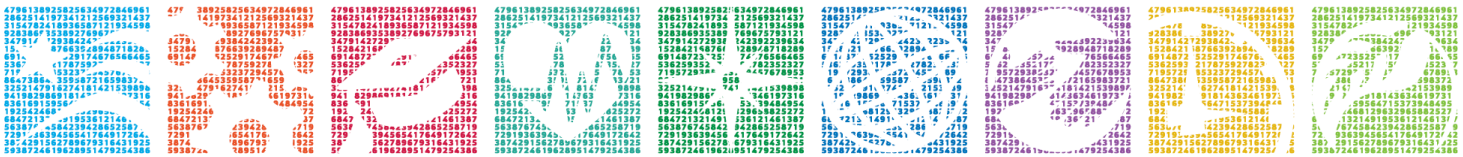
Stewarding Our Resources: Building a Sustainable IPUMS Archival Document Access System

Diana L. Magnuson†
University of Minnesota

October 2023

Working Paper No. 2023-03
DOI: https://doi.org/10.18128/IPUMS2023-03

†Address correspondence to Diana Magnuson: magn0031@umn.edu. Support for this research was funded by the National Science Foundation: IPUMS International grant NSF SES0851414.



Abstract

IPUMS at the University of Minnesota has created the world's largest accessible database of census and survey microdata. The IPUMS suite of products contains nine harmonized data products. The largest of these projects, IPUMS International (IPUMS-I) has supported the curation and preservation of ancillary materials received during data acquisition efforts. Archival staff have preserved thousands of unique pieces of census and survey documentation, creating bibliographic records using an extended Dublin Core profile that supports the use of controlled vocabularies to enhance findability. The goal of this curation work was to create a searchable, and downloadable document access system for our internal use and to support IPUMS researchers. This paper describes our experience constructing a tool that supports exploration and dissemination of these archived materials. During this development, we gained valuable insight about stewarding our resources that are applicable to research organizations responsible for curating, preserving, and disseminating similar archival materials.

Introduction

Over the last thirty years, IPUMS at the University of Minnesota has created the world's largest accessible database of census and survey microdata (Magnuson and Ruggles 2022). 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. Data come from over 100 national and regional statistical organizations.

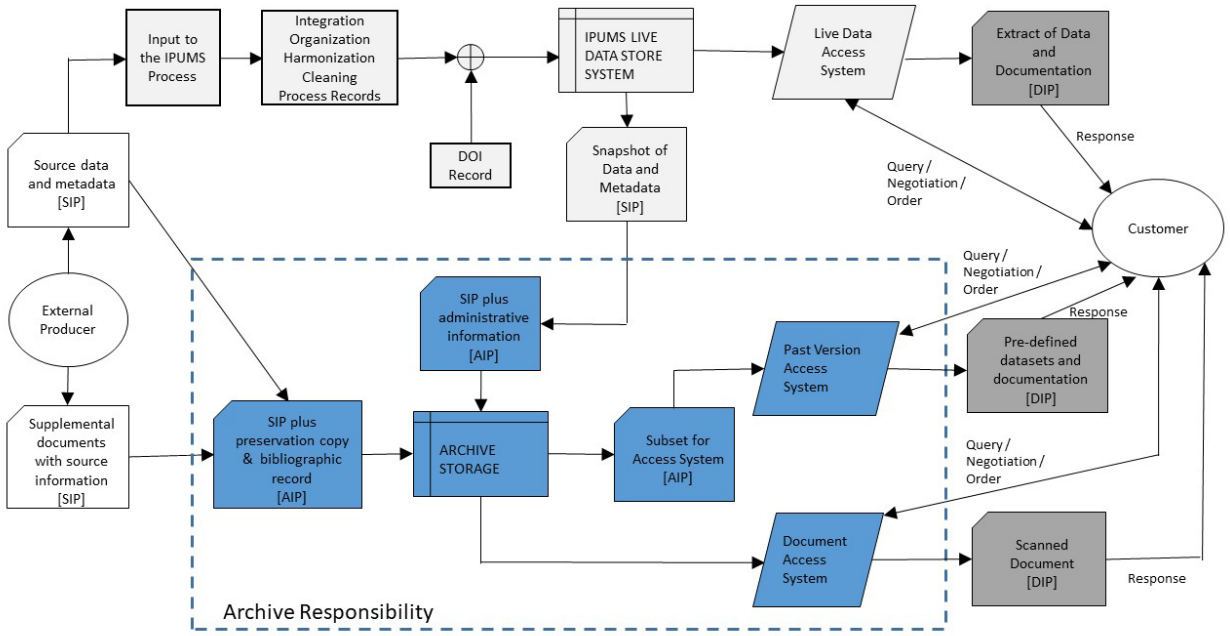
The context of this paper is specific to the IPUMS International (IPUMS-I) data project. Beginning in 1999, with a social science infrastructure grant from the National Science Foundation (NSF), IPUMS-I had a simple yet audaciously ambitious goal: preserve the world's microdata resources and democratize access to those resources. Twenty-four years later, the project goals continue to be: collecting and preserving census and survey data and documentation; harmonizing those data; and disseminating the harmonized data free of charge (Ruggles and McCaa et al. 1999-2004; McCaa and Ruggles 2000; Ruggles et al. 2003). IPUMS-I data are coded and documented consistently across countries and over time to facilitate robust comparative research.

Over more than two decades, IPUMS-I amassed tens of thousands of ancillary materials in support of its data harmonization work. These materials came from the United States Census Bureau (USCB), United Nations Statistical Division (UNSD), Latin American and Caribbean Demographic Center (CELADE), East West Center, Centre Population et Dévelopement (CEPED), and over one hundred national statistical agencies. Examples of this material include correspondence, maps, enumerator instructions, supervisor instructions, training materials, codebooks, publicity, reports, newspaper clippings, unpublished papers, census timetables, data processing materials, and technical manuals. The ancillary materials in our

collection attest to the varied technical, business, social, and economic aspects of conducting censuses and surveys across time and space.

Preservation and dissemination of data products is already part of the IPUMS workflow (Figure 1). For the purposes of this paper, archival workflows are highlighted in blue. We have discussed elsewhere the role of the archive in the IPUMS workflow (Magnuson and Thomas 2003). This case study will focus on the development of the document access system within the archive workflow (lower right).

Figure 1. IPUMS workflow



The IPUMS-I grant has funded the curation and preservation of the ancillary materials acquired by the project. For over two decades, archival staff have preserved thousands of unique pieces of census and survey documentation, creating bibliographic records using an extended Dublin Core profile that supports the use of controlled vocabularies to enhance findability. The goal of this work was the creation of a simple, searchable, and downloadable document access system.

An early iteration of the effort to make available basic international census documentation was the creation of a census forms webpage. The goal of the census forms page was primarily to demonstrate to potential international data partners the scope and capacity of IPUMS-I. The World Population Census Forms webpage offered the minimal functionality of a hyperlink list to access this single class of documents (Ruggles et al. 2003).[1] While useful for organizing, identifying, and accessing standard international census forms, this static utilitarian webpage challenged us to consider how we could better steward the thousands of additional materials entrusted to us. We needed a more sophisticated tool to provide access to our archival resources.

Building the IPUMS document access system

Building the IPUMS document access system required short- and long-term plans, building a warehouse of flexible metadata, navigating the priorities and schedule of the Institute for Social Research and Data Innovation (ISRDI), and working with the ISRDI Information Technology IT Product Team to develop a web application that is simple, sustainable, discoverable, and capable of disseminating basic metadata and pdfs of digitized archival materials.

Vision

Beginning in 1999, ancillary documents began streaming into the archive as a product of IPUMS-I acquisition efforts. It is very likely that some of these materials are uniquely held by IPUMS. Some of these materials arrived with the express understanding that they would be made publicly available through some means after their lifecycle in the IPUMS microdata harmonization work had concluded. Other acquired materials were broader than the IPUMS data collection and covered countries, censuses, and surveys, for which IPUMS does not yet disseminate microdata. The fragility of some documents created a race against time to carefully preserve them for future use. Taken together, this collection reflects an exciting diversity of archival materials. We cannot predict the uses to which future researchers might put these materials, but it is our obligation to preserve them and make their innovative future use possible.

Data curator Wendy Thomas recognized the significance of all these materials and anticipated the day when an archival document access system would be a reality. Thomas' experience providing research support for social science data users and her technical expertise curating data and metadata positioned her well for envisioning an IPUMS document access system (Magnuson 2015). Thomas immediately took steps to assemble the architecture necessary to build a flexible metadata warehouse. From the beginning, Thomas advocated for a simple online tool that would provide basic metadata and downloadable pdfs of digitized archival materials. The tool was envisioned to be sustainable long-term by the archivist, with minimal intervention of IT staff beyond the development of the initial web application.

Building a metadata warehouse

Building a metadata warehouse for the IPUMS-I archival document access system has been years in the making. First, Thomas built up the scaffolding for logical intake and workflow for digital and manuscript materials. This was no small task. Documents did not arrive on a fixed schedule or in uniform packaging, but rather, as agreements were made with supporting statistical entities and in whatever form they happened to exist (McCaa and Ruggles 2000; Ruggles et al. 2003). Thousands of physical and digital documents needed to come under archival control before they were subject to the metadata creation stage.

Next, Thomas developed a process for creating structured bibliographic records for each piece of archival material. Using an extended Dublin Core profile, Thomas hand-tailored a controlled vocabulary to the archival dimensions of IPUMS-I ancillary materials. The decision to use Dublin Core was a practical one. Dublin Core is recognized worldwide as a basic core of bibliographic records and cataloging systems that do not use Dublin Core invariably provide information on mapping to it (Thomas 2023).[2] Prior to the creation of the web application, archive staff used Thomas' controlled vocabulary to search for and organize materials, using simple text string searches employing Perl scripts.

Accessing the archive of processed metadata has consistently relied on use of controlled vocabularies to enhance findability and flexibility in the short run for project staff and in the long run for future external researchers. Along the way, archivists developed and refined training and reference materials to support undergraduate student workers who created bibliographic records for each unique piece of archival material. An unintended but happy result of the development and refinement of these instructional and reference materials over time was the curation of institutional history documenting the work, growth, and development of the IPUMS archival processes.

Institutional priorities

As noted above, census and survey data harmonization and dissemination are the central activities of IPUMS at the University of Minnesota. In this fast-paced environment, it has been incumbent upon archival staff to nurture an appreciation for expanding our institutional archival curation activities. IPUMS administrators, principal investigators, project managers, and research data scientists all recognize the importance and utility of an archival data access system, but quite naturally their priorities have historically focused on grant writing, data acquisition, data harmonization, and data dissemination.

The key for archival planning and productivity, then, has been identifying IPUMS priorities that naturally connect with, and enhance the short- and long-term goals of the archive. First, the bold mission of IPUMS is to democratize access to the world's social and economic data for current and future generations.[3] Second, a "central goal" of IPUMS-I from the beginning was "to create an inventory of surviving census microdata and documentation," including "enumerator instructions, census forms, codebooks, studies of data quality, and any other ancillary documentation we can locate for all countries that will allow us access to this documentation" (Ruggles et al. 2003). Third, the decision to begin assigning DOIs to IPUMS data products was triggered by the increasing requirements of external funding organizations to conform to "standard archival practice using the open archival information system (OAIS) model and digital object identifiers (DOI)" (Magnuson and Thomas 2023). Assigning DOIs forced broader

internal discussions around access, curation, and preservation-- all concerns of the archive. Fourth, documenting these internal discussions became a steppingstone to building a successful Core Trust Seal application, which in turn highlighted the preservation work of the archive (Magnuson and Thomas 2023).[4] These priorities all have strong ties to IPUMS archive concerns. Intentionally communicating and nurturing those connections to internal IPUMS stakeholders has been essential to moving archival goals forward.

The steady growth of the IPUMS-I data project, increasing expectations of external funding organizations regarding IPUMS preservation practices, and a transition in IPUMS archive leadership due to a retirement, eventually led to concrete steps toward building an online document access system. The IPUMS IT Product Team, the group of developers responsible for the IPUMS web dissemination system and related components, added that task to their quarterly project calendar (Ruggles et al. 2015; Magnuson and Thomas 2023). The persistent commitment to building a warehouse of flexible metadata positioned the archive well to take advantage of this opportunity to fulfill a long-standing grant deliverable. In addition, the IT Product Team was looking for a discrete project to train a new hire and introduce them to project workflows. The evolutionary moment had finally arrived for IPUMS to undertake building an archival document access system.

The archive was invited to propose a small defined project to the IT Product Team and the timing was ideal. Archive personnel had already submitted a presentation proposal to the 2023 International Association for Social Science Information Service & Technology Conference promising to document the effort to build a document access system. The acceptance in February 2023 for an IASSIST presentation in early June 2023 fit the IPUMS IT Product Team timeline and meant the goal of the presentation would not merely be documentation but an actual demonstration of the functional access tool.

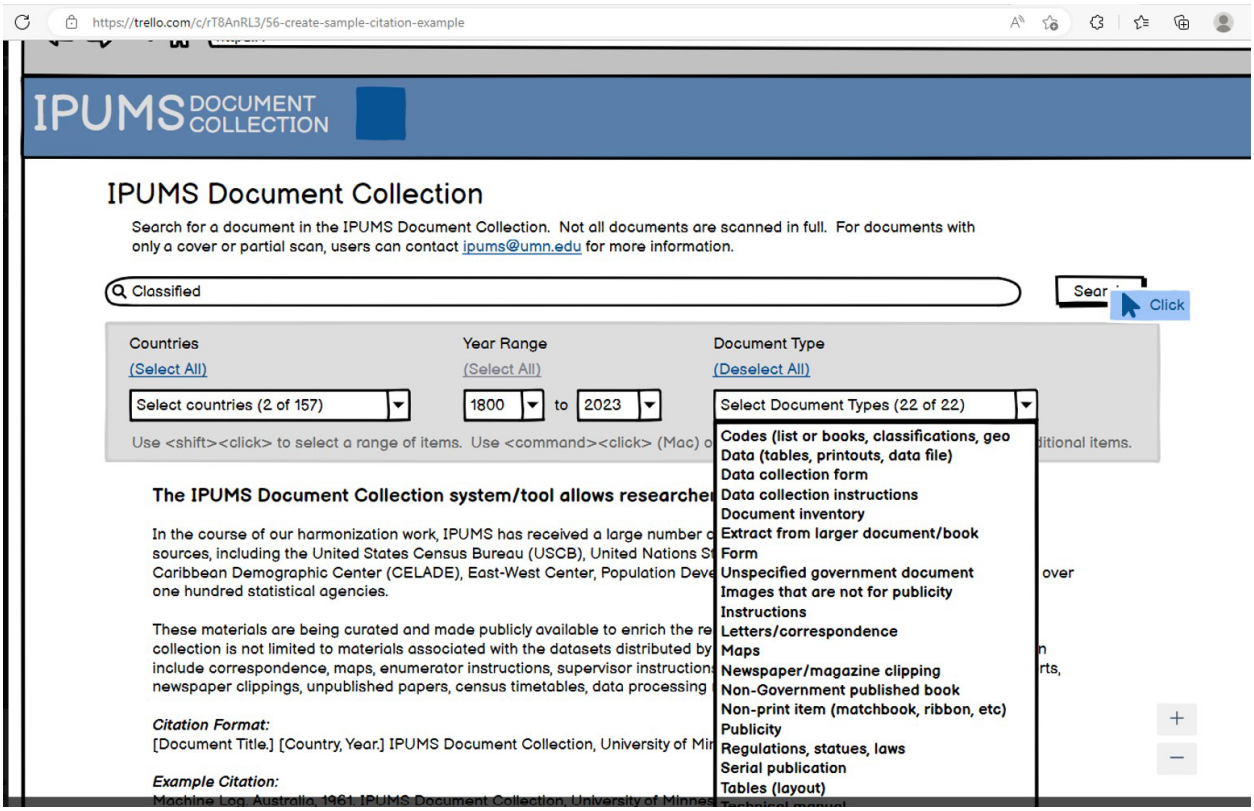
IPUMS IT Product Team

The process of working with the IPUMS IT Product Team to develop a web interface was highly stimulating and collaborative. One throughline Product Team member (the new hire) was consistently assigned to the project, while several Product Team members stepped in and out of the workflow according to the need for their area of expertise. The first formal step in the collaboration was completing an IT Project Data Sheet (PDS) to record information from the archive covering four areas: 1) objective statement, project motivation, project context, primary stakeholder; 2) major milestones and deadlines, project scope; 3) Minimum Viable Product (MVP), most important quality attributes, trade-off matrix; 4) known business issues and risks, and how success will be measured. Working through the PDS to its completion was a worthwhile intellectual exercise for archival stakeholders. It forced us to problematize the project in productive directions, identifying "must have," "valuable to have," and aspirational (future) developments to the system functionality and user interfaces.

Once the project was approved by the Product Team and underway, there was regular written (email) and oral (Zoom) dialogue clarifying project goals and developing a workflow. Tasks were tracked via the web-based management tool Trello. For five months leading up to the project "sprint" (final two weeks of the project), periodic check-ins to track progress and clarify action points were conducted over Zoom every few weeks, primarily with the throughline Product Team member. The archive was presented with a proof of concept and shortly thereafter a prototype of the user interface by the software developer. Two weeks before the project deadline the IT Product Team launched their sprint and communication increased to daily brief "standup" meetings. During the standup, IT and archive contributors assessed and assigned tasks, asked questions, and held everyone accountable to the specifications of the PDS. During

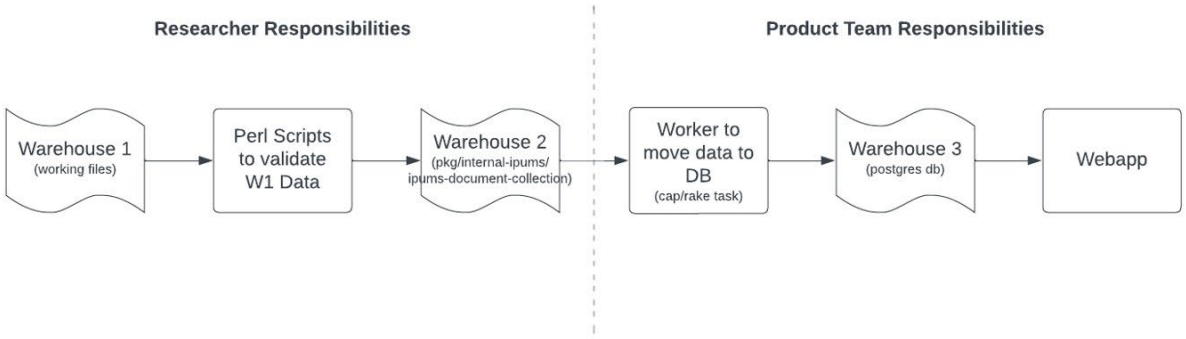
the sprint period, the IT team created “wireframes,” illustrating the proposed user interface content, functionalities, and intended behaviors for archival review and feedback (Figure2).

Figure 2. Example of IPUMS Document Collection wireframe



The Product Team visualized a product workflow containing three dynamic data “warehouses” (Figure 3). Warehouse 1 contains pdfs of the scanned documents and the XML metadata working files created by undergraduate student workers. Files in Warehouse 1 are subject to archivist (“Researcher”) validation against internal archive standards and cleaning. After completing validation and cleaning the archivist moves the metadata to Warehouse 2, where it is ready to be processed by IT. The Product Team then takes the PDF and XML metadata files from Warehouse 2, subjecting these files to IT computational validation and moves the files to Warehouse 3, to be deployed to the web interface.

Figure 3. Metadata workflow responsibilities for researcher and product team



The collaboration between the IT Product Team and archive personnel produced valuable learning in both directions. Archive personnel gained an appreciation and understanding of the elements necessary to build a scalable user interface: including consistency of content, reliable search filters, clear expectations regarding cardinality of individual elements, and exceptions to standard rules. In turn, the role and concerns of the archive are now in the IT Product Team's line of sight. This is especially important for the archive, as its work is often viewed as being conducted on the periphery of IPUMS product workflows.

IPUMS Document Collection

The IPUMS Document Collection was the minimum viable product (MVP) resulting from the collaboration between archive staff and the IPUMS IT Product Team.[5] The launch of the IPUMS Document Collection tool in early June 2023 triggered a new phase in product development.

PDF documents and metadata from the region of Oceania served as a test set to represent the quality and content of the metadata records and to evaluate the ability of these records to support the functionality of the web interface. The Oceania test data set is relatively small (9 countries, 626 document records, 29 collection records, and 653 pdf files), but it represents most of the special cases found in the

overall collection. Thus, the Oceania test data set served as an essential diagnostic at several points in the development of the web interface.

First, the test data set made clear for IPUMS IT staff what files were going to be provided to them by the archive. Second, archive staff divided responsibility for validation based on where in the workflow was most effective to test the metadata. Third, communication during the refinement of the MVP clarified record content in terms of immediate and future requirements that will support consistently applied additional search options. Going forward, the IPUMS Document Collection will be built out as ancillary materials from new regions are processed through Warehouses 1, 2 and 3.

Conclusion

Adopting a long-range *mise en place* strategy was essential to successfully launching the IPUMS Document Collection. Experienced French chefs practice the skill of preparation — *mise en place* or — "everything in its place" to successfully create a tasty dish. The IPUMS archival *mise en place* strategy gathered, organized, and prepared the anticipated requisite elements necessary for web and user interfaces. This effort took place over time, in expectation of an eventual funding and/or institutional opportunity. When the institutional opportunity presented itself, archival staff were ready.

As noted, 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 to study change, conduct comparative research, merge information across data types, and analyze individuals within family and community contexts. All IPUMS resources are free of charge. In this organizational context, the archive has a supporting but vital role. Stewarding and expanding access to our archival resources in support of IPUMS involved:

- Articulating what areas of the organization our archival functions are related to. It was incumbent upon archive staff to clearly and consistently identify where the work of the archive touched the main product of the organization and how archival stewardship enhanced the main product.

- Knowing the priorities of our organization. Framing archival goals within the context of IPUMS deliverables required archival staff to plan metadata creation in terms of how it would be efficiently conveyed to users of IPUMS data products.
- Investing in building flexible and scalable metadata processes and infrastructure.
- Leveraging work with the IPUMS IT Product team. Through constructive and deliberate communication, archival staff developed short-and long-term goals, identified strengths and areas for refinement of archival data management, and educated IPUMS IT staff on the vital role of the archive within the organization.
- Strategizing scalable deliverables. In the IPUMS-I context, we focused on developing the scaffolding for intake of digital and manuscript materials, tailoring a controlled vocabulary using an extended Dublin Core profile, and building a simple, but extensible, web interface.

The ancillary materials in the IPUMS document collection attest to the technical, business, social, and economic aspects of conducting censuses and surveys across time and space. In addition, ancillary materials contextualize contemporary uses of and responses to historic censuses and surveys. Expanding and deepening our search and delivery system will provide findability and accessibility to a rich set of supporting archival documentation that will illuminate census development and implementation processes across time and space. For example, access to materials documenting the development of enumeration forms and procedures over time supports researchers' understanding of how statistical entities responded to the challenges of collecting demographic data on difficult to enumerate populations. Creating a sustainable, discoverable, and searchable access system for a broad range of archival census and survey materials will support the IPUMS mission of democratizing access to the world's social and economic data and enable transformative scholarship. We believe curation and public

availability of these materials enrich IPUMS products but also the innovative research of IPUMS data users.

References

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

Magnuson, D.L. and Ruggles, S. (2022) "Challenges of Large-Scale Data Processing in the 1990s: The IPUMS Experience," *IEEE Annals of History and 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 47, No. 2. https://iassistquarterly.com/index.php/iassist/article/view/1046

McCaa, R. and Ruggles, S. (2000) "IPUMS-International: A Global Project to Preserve Machine-Readable Census Microdata and Make Them Useable," *Handbook of International Historical Microdata for Population Research*, edited by Patricia Kelly Hall, Robert McCaa and Gunnar Thorvaldsen, pp. 335-346. https://international.ipums.org/international/microdata_handbook.shtml

Ruggles, S., King, M.L., Levison, D., McCaa, R and Sobek, M. (2003) "IPUMS International," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, Volume 32, No. 2. https://www.tandfonline.com/doi/abs/10.1080/01615440309601215

Ruggles, S., McCaa, R., Sobek, M. and Cleveland, L. (2015) "The IPUMS Collaboration: Integrating and Disseminating the World's Population Microdata," *Journal of Demographic Economics*, Volume 81. DOI:10.1017/dem.2014.6

Ruggles, S., McCaa, R., Levison, D., Gardner, T. and Sobek, M. (1999-2004) "International Integrated Microdata Access System." SBR9908380, Methodology, Measurement, and Statistics Program, NSF.

Thomas, W. (2023) "Why Dublin Core," General Information, ISRDI Archive.

Notes

[1] https://international.ipums.org/international/census_forms.shtml

[2] https://www.dublincore.org/specifications/dublin-core/profile-guidelines/

[3] https://www.ipums.org/mission-purpose

[4] https://www.coretrustseal.org/

[5] https://documents.ipums.org