

FAIR Vocabularies in Population Research: report of the IUSSP-CODATA Working Group on FAIR Vocabularies

George Alter, Arofan Gregory, Steven Mceachern, Darren S Bell, Derek Burke, Robert Chen, Alessio Cardacino, Nada Chaya, David Barraclough, Rowan Brownlee, et al.

▶ **To cite this version:**

George Alter, Arofan Gregory, Steven Mceachern, Darren S Bell, Derek Burke, et al.. FAIR Vocabularies in Population Research: report of the IUSSP-CODATA Working Group on FAIR Vocabularies. IUSSP; CODATA. 2023. hal-04096418

HAL Id: hal-04096418
https://hal.science/hal-04096418

Submitted on 12 May 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

FAIR VOCABULARIES IN POPULATION RESEARCH

Report of the IUSSP-CODATA Working Group on FAIR Vocabularies



Members of the Working Group:

George Alter, Co-chair (University of Michigan, IUSSP)

Arofan Gregory, Co-chair (DDI Alliance, CODATA)

Steven McEachern, Co-chair (Australian National University, DDI Alliance, CODATA)

Darren S Bell (UK Data Archive)

Derek Burk (University of Minnesota, Institute for Social Research and Data Innovation (ISRDI), IPUMS)

Robert Chen (Center for International Earth Science Information Network (CIESIN), Columbia University)

Alessio Cardacino (Italian National Statistical Institute)

Nada Chaya (Arab Council for the Social Sciences)

David Barraclough (OECD and SDMX)

Rowan Brownlee (Australia Research Data Commons)

Tom Emery (Erasmus University Rotterdam)

Patrick Gerland (United Nations)

Cristina Giudici (Sapienza University of Rome)

Abdulla Gozalov (Statistics Division, United Nations)

Edgardo Greising (International Labour Organization)

Sanda Ionescu (ICPSR)

Taina Jääskeläinen (Finnish Social Science Data Archive and CESSDA Vocabulary Service)

Chifundo Kanjala (ALPHA Network)

Vladimira Kantorova (United Nations)

Joseph Larmarange (CEPED and Demopaedia)

Pablo Lattes (United Nations)

Jared Lyle (ICPSR)

Diana Magnuson (University of Minnesota, Institute for Social Research and Data Innovation (ISRDI), IPUMS)

Melissa Meinhart (Equality Insights)

Santosh Kumar Mishra (S.N.D.T. Women's University)

Romesh Silva (United Nations, UNFPA)

Thomas Spoorenberg (United Nations)

Philipp Ueffing (United Nations)

Jay Winkler (ICPSR)

The Working Group gratefully acknowledges advice and assistance from:

Susana Adamo (Center for International Earth Science Information Network (CIESIN), Columbia University)

Franck Cotton (INSEE)

Simon Cox (Commonwealth Scientific and Industrial Research Organisation)

Alejandra Gonzalez Beltran (UKRI Science and Technology Facilities Council)

Edith Gray (IUSSP and Australian National University)

John Graybeal (BioPortal, Stanford University)

Simon Hodson (CODATA)

Thomas LeGrand (IUSSP and Université de Montréal)

Barbara Magagna (Environmental Agency Austria)

Paul Monet (IUSSP)

Mark Musen (BioPortal, Stanford University)

John Scialdone (Center for International Earth Science Information Network (CIESIN))

Alex de Sherbinin (Center for International Earth Science Information Network (CIESIN))

Mary Ellen Zuppan (IUSSP)

Suggested citation:

IUSSP-CODATA Working Group on FAIR Vocabularies. (2023). "FAIR Vocabularies in Population Research." Paris: IUSSP and CODATA. https://doi.org/10.5281/zenodo.7818157



This work is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted (images).



LightField Studios. *Group of multiethnic senior friends embracing in bar*. Shutterstock. Retrieved 28 April 2023.

TABLE OF CONTENTS

1.	**EXECUTIVE SUMMARY**	**07**
1.1	Summary	07
1.2	Key findings	08
2.	**OBJECTIVES OF THE WORKING GROUP**	**10**
3.	**DEFINING DEMOGRAPHY**	**12**
3.1.	IUSSP and the Language of Demography	12
3.2.	Demopædia	13
3.3.	DemoVoc	13
4.	**FAIR VOCABULARIES**	**15**
4.1.	Levels of controlled vocabularies	15
4.2.	FAIR vocabularies	16
5.	**USE CASES**	**20**
5.1.	Data discovery	20
5.2.	Data integration	21
5.3.	Data harmonization	21
5.4.	Data production	22
6.	**STANDARDS AND TECHNOLOGIES FOR DATA DISCOVERY AND EXCHANGE**	**24**
6.1.	Metadata standards	24
6.1.1.	DDI	24
6.1.2.	SDMX	24
6.2.	Semantic web standards	25

7. CASE STUDIES 29
7.1. European Language Social Science Thesaurus (ELSST) 29
7.2. SDMX Global Registry 30
7.3. IPUMS 30

8. REPRESENTATION OF DEMOGRAPHIC TERMS AND CONCEPTS IN EXISTING ONLINE VOCABULARIES 33

9. OPPORTUNITIES AND NEEDS 38
9.1. Concept Vocabularies 38
9.2. Operational and Technical Vocabularies 39
9.3. Code Lists and Code Mappings 39

10. RECOMMENDATIONS 41
10.1. Recommendations for IUSSP 41
10.1.1. FAIR Vocabulary of Demography 41
10.1.2. Development of a FAIR Vocabulary of Demography 41
10.1.3. Outreach to the SDMX and DDI communities 42
10.1.4. Promoting Data Harmonization 42
10.2. Recommendations for SDMX 42
10.3. Recommendations for the DDI Alliance 43
10.4. Recommendations for IPUMS 43
10.5. Recommendations for CODATA 44
10.6. Recommendation for IUSSP and CODATA: Harmonizing Vocabularies to Harmonize Data 44

Appendix: Governance of a FAIR Vocabulary of Demography 45
Glossary 46
References 58



ra2 studio. *Demographic related charts, diagrams and graphs hovering over young hand.* Shutterstock. Retrieved 28 April 2023.

1 EXECUTIVE SUMMARY

1.1 SUMMARY

This report describes the role of controlled vocabularies in the documentation and dissemination of demographic data in the light of the FAIR principles that all data should be "Findable, Accessible, Interoperable, and Reusable" by both humans and machines (Wilkinson et al., 2016). Population research is an empirically focused field with a long tradition of widely shared, easily accessible, data collections. The FAIR Principles point to ways that this tradition can be enhanced by taking advantage of emerging standards and technologies. Our work builds on the "Ten Simple Rules for making a vocabulary FAIR" (Cox et al., 2021), prepared by a group formed at a workshop convened by CODATA and DDI to describe how a FAIR vocabulary will work with international standards for documenting and sharing social science data.

Controlled vocabularies play a central role in data sharing by associating data with concepts and by defining which categories or codes may be applied. FAIR vocabularies specify globally accessible persistent identifiers to distinguish data items that are the same from those that are different. Consider the most basic variable in demographic analysis: age. The Organization for Economic Cooperation and Development (OECD) has a list of 643 age categories, while the UN Population Division copes with more than 1100 age groups. If the meanings of variables in a dataset are only available through human-readable documentation, like a pdf, harmonizing data from two providers will remain a tedious manual process. However, if the age categories are linked to persistent identifiers in machine actionable metadata, software can be programmed to harmonize age groupings. If these operations are performed across dozens of variables in hundreds of data sources, enormous amounts of human time will be saved.

Construction of the infrastructure for FAIR data has begun. Demographic concepts are already included in vocabularies developed by other disciplines, like medicine, with definitions that conflict with usage in population research. **Therefore, there is a need for a FAIR vocabulary of demographic concepts endorsed by an authoritative institution in the field of population science.**

IUSSP has a long history of working with the UN and other agencies to define demographic concepts (International Union for the Scientific Study of Population, 1954; Vincent, 1953). Those efforts currently exist in electronic forms (Demopædia and Demovoc) that provide a base for a multilingual FAIR Vocabulary of Demography. We argue that a FAIR Vocabulary of Demography will have important benefits for the population research community represented by IUSSP, and we conclude with recommendations for IUSSP and other important organizations.

In addition to summarizing the activities of the Working Group, this report is intended to serve as an introduction to the standards and infrastructure used to share social science data. Most demographers have never heard of URIs, SDMX, or DDI, even though they use services from the UN, ILO, OECD, CESSDA, IPUMS, and other organizations that depend on these standards. Understanding key features of the international data infrastructure will help IUSSP leadership to influence its development.

1.2 KEY FINDINGS

Our primary recommendation is that IUSSP should create a standing committee to create a multilingual FAIR Vocabulary of Demography.

- A FAIR Vocabulary of Demography will enable software to improve discovery and automate access to demographic data.
- Online vocabularies including demographic terms already exist, and most of them define terms in ways incompatible with demography. Population research will be at a disadvantage without an authoritative FAIR vocabulary of its own.
- IUSSP has a long history of defining demographic concepts and terms in multiple languages, and a FAIR Vocabulary of Demography can be built existing online resources Demopædia and DemoVoc. IUSSP is indebted to the Institut National d'Études Démographiques (INED) for its leadership in compiling the multilingual dictionaries of demography and for the preservation of both Demopædia and DemoVoc.

The infrastructure that creates and shares demographic data rests on standards and technologies, such as SDMX and DDI. IUSSP and CODATA should work with other organizations to assure that this infrastructure will take advantage of FAIR vocabularies. Accordingly,

- The FAIR Vocabulary of Demography should reach out to other stakeholders by providing services in SDMX, DDI, and other metadata standards.
- The SDMX and DDI communities should implement registries for concepts, variables, and codelists that conform to FAIR principles.
- By virtue of its global leadership in data harmonization, IPUMS can influence other organizations by making its library of variable definitions, code schemas, and mappings available with persistent identifiers.
- IUSSP and CODATA should partner with others to promote research and development of tools that can compare and link items (concepts, variables, and codes) in different registries and repositories.

JANUARY 2023



BearFotos. *Adult man answers questions of the interviewer at home*. Shutterstock. Retrieved 28 April 2023.

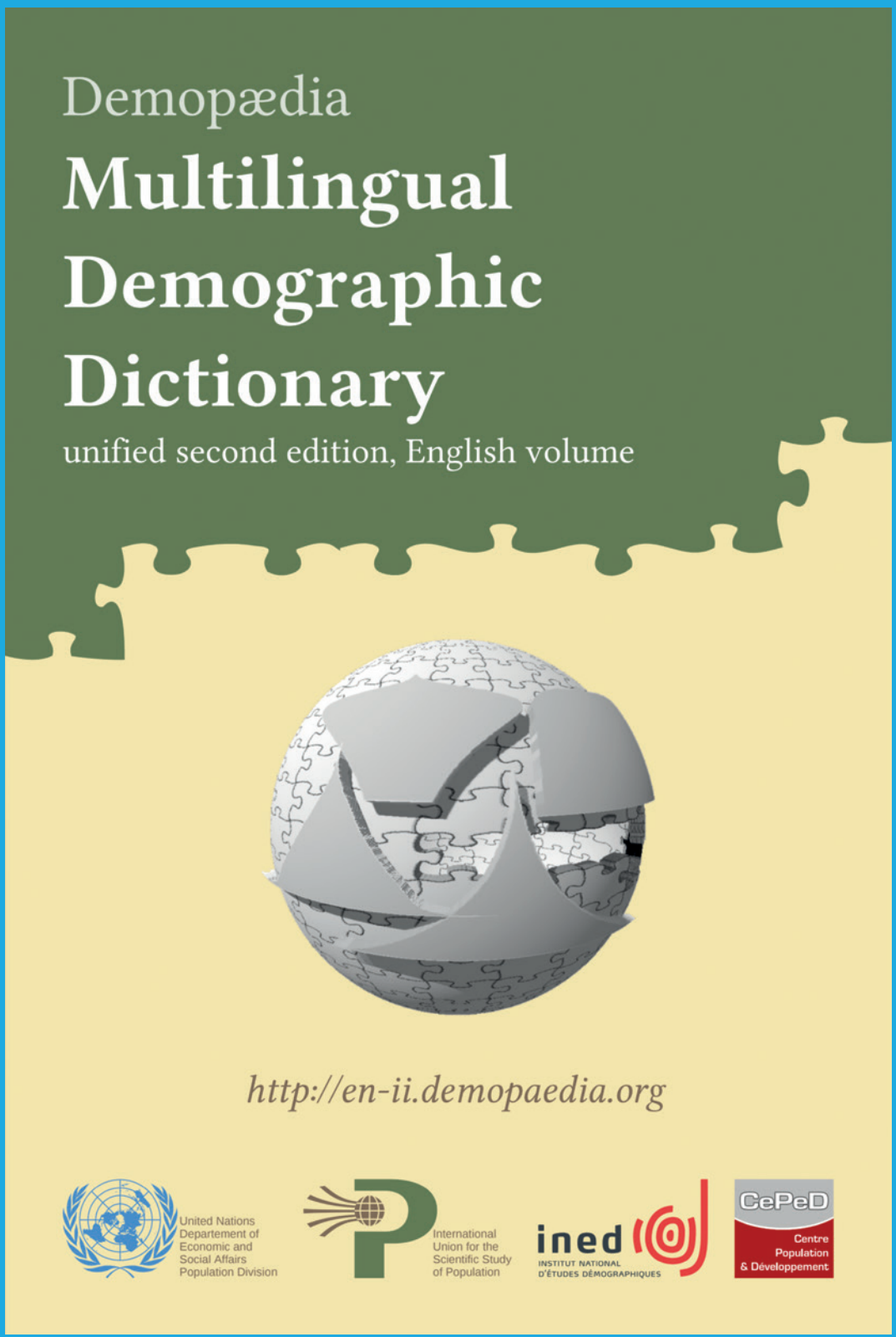
2 OBJECTIVES OF THE WORKING GROUP

The FAIR Vocabularies for Population Research Working Group is a joint initiative between IUSSP and CODATA, the Committee on Data of the International Science Council, in response to the growing movement to make data "Findable, Accessible, Interoperable, and Reusable" (FAIR). Population research is an empirically focused field with a long tradition of widely shared, easily accessible data collections. The FAIR Principles point to ways this tradition can be enhanced by taking advantage of existing and emerging standards and technologies. The Working Group focussed on the development of FAIR Vocabularies for population data, an essential step in making data reusable and interoperable. Our work builds upon "Ten Simple Rules for making a vocabulary FAIR" (Cox et al., 2021).

FAIR vocabularies yield benefits when data from different sources must be combined. Consider the most basic variable in demographic analysis: age. OECD has a list of 643 age categories, and the UN Population Division copes with more than 1100 age groups. The handling of special age groups (e.g., totals, unknown age, open age groups, etc.) require special provisions. If the meanings of variables in a dataset are only available through human-readable documentation, like a pdf, harmonizing data from two providers will remain a tedious and time consuming manual process. However, if the age categories are linked to persistent identifiers in machine actionable metadata, software can be coded to harmonize age groupings. If these operations are performed across dozens of variables in hundreds of data sources, enormous amounts of human time will be saved. Thus, the ultimate goal of this initiative is to make demographic data more interoperable by publishing controlled vocabularies that can be discovered and acted upon by software.

The Working Group began meeting regularly in May 2021 with more than 25 scientists from academia and international agencies. Our meetings explored both the principles and mechanisms of FAIR vocabularies and the existing infrastructure for documenting and disseminating data in the social sciences. The group has discussed the FAIR principles as they apply to controlled vocabularies and related technologies used to make them accessible. Our case studies included the European Language Social Science Thesaurus (ELLST), a FAIR vocabulary of social science concepts developed by the Consortium of European Social Science Data Archives and the IPUMS Project, which creates and distributes harmonized data from international censuses and other sources. We also delved into two widely used standards used to document social science data: the Data Documentation Initiative (DDI) used for microdata, and Statistical Data and Metadata eXchange (SDMX), developed by international organizations for the exchange of statistical data. This report also describes IUSSP's historical contribution to vocabularies of demographic terms and two current online versions of those vocabularies: Demopædia and DemoVoc.

Our recommendations propose ways that IUSSP can bring the benefits of FAIR data to the population research community. We explain how IUSSP can build on existing vocabularies to create a multilingual FAIR Vocabulary of Demography. We also make recommendations for moving the existing social science data infrastructure (SDMX, DDI, IPUMS) toward FAIR principles. We believe that IUSSP and CODATA can influence the organizations that support that infrastructure in ways that will benefit demography and the social sciences in general.



International Union for the Scientific Study of Population. (2017). *Demopædia. Multilingual Demographic Dictionary.* Second edition, English Volume.

3 DEFINING DEMOGRAPHY

3.1. IUSSP AND THE LANGUAGE OF DEMOGRAPHY

IUSSP has a long history of documenting and translating demographic terminology across multiple languages. As demography developed into a self-conscious discipline, it developed its own unique terminology. Demographers have invented new words, like "fecundability", and they have assigned very specific meanings to seemingly general terms, like "natural fertility" and "intrinsic rate of growth" (Petersen, 1983; Pressat, 1980). In 1949, the Population Division of the United Nations was assigned the task of providing a "Multilingual Dictionary of Demography," and they turned to IUSSP for help with this project (Gubry, 2007; Vincent, 1953). The first multilingual dictionary was published in 1954 (International Union for the Scientific Study of Population, 1954), and revised editions were issued until 1987.

The Multilingual Demographic Dictionary follows a model described by Paul Vincent (1953) as "la collection de volumes unilingues à double index." Each volume consists of a set of short essays in which several related terms are defined. Although essays in each language follow the same model, authors have flexibility to explain particular features of usage in each language. The terms used in each essay are assigned reference numbers that are consistent across languages. An alphabetic index in each volume links a term to its reference number and the essay in which it is defined. A numeric index in each volume allows users to find a specific term in each language. Thus, one can translate a term from English to French by looking up its reference number in the alphabetic index of the English volume and then finding the French term in the numeric index of the French volume.

In 1973, the UN Population Division was asked to employ computer technology to create a database that could be used for indexing documents (Gubry, 2007; Hankinson, 1993). This project was undertaken by the Committee for International Co-operation in National Research in Demography (CICRED). The first Population Multilingual Thesaurus was released in 1979, and versions in several languages have been published (Hankinson, 1993). In 1981, a working group for maintaining the thesaurus was established by the Population Information Network (POPIN), and subsequent versions are known as the "POPIN Thesaurus." The thesaurus was maintained until 1993, and it was the base for DemoVoc.

The POPIN Thesaurus was intended for indexing and retrieving documents held by libraries and other organizations. It lacks much of the explanatory text found in dictionaries of demography. However, it does provide structured links ("Narrower term," "Broader term," "Related Term") to help users find related documents.

Although the Multilingual Demographic Dictionary and the POPIN Thesaurus have not been updated in many years, they still have legacies on the Internet in the form of Demopædia and DemoVoc, which we describe here. These efforts are important, because they can serve as the core of a new FAIR Vocabulary of Demography as we recommend below.

The Working Group is pleased to recognize the important role played by the Institut National d'Études Démographiques (INED) in these efforts. From Paul Vincent, who edited the first Multilingual Demographic Dictionary, to Nicolas Brouard, who led the creation of Demopædia, INED has provided leadership and support for defining demographic terms in multiple languages.

3.2. DEMOPÆDIA

Demopædia (Demopædia, 2013) was initiated in 2005 by the French National Committee of the IUSSP to make accessible on the web the Multilingual Demographic Dictionaries produced by the United Nations and IUSSP between 1953 and 1987. The IUSSP Council approved Demopædia, as an official IUSSP scientific activity in 2012. The current version is also supported by the Population Division of the Department of Economic and Social Affairs at the United Nations. Demopædia is maintained by the Institut national d'études démographiques (INED).

Demopædia was conceived as an Open Population Encyclopaedia, and IUSSP members have been encouraged to contribute. The project currently covers 18 different languages. The Demopædia team developed procedures and tools for harmonizing content across languages, which had become unsynchronized in the latest published volumes of the Multilingual Demographic Dictionaries. To encourage community participation, Demopædia is built on Wiki technology, which provides tools for discussion as well as tracking contributions.

Entries in Demopædia are in the form of brief articles defining several related terms. Terms within each article are indexed and hyperlinked to other articles. Parallel articles are available in each supported language, and a translation table shows terms in other languages. An SQL database of terms and definitions has been built from the Wiki.

3.3. DEMOVOC

DemoVoc (Thésaurus DemoVoc, 2015) is an online thesaurus created by the INED Library from the POPIN Thesaurus database. DemoVoc includes about 2,000 concepts and 2,500 terms in both French and English. Librarians and researchers at INED updated the thesaurus until 2015, and the thesaurus is currently used by the INED archives. In 2019, the Library of INED was merged into the unified library and documentation system for the new Campus Condorcet, which is dedicated to the humanities and social sciences.

DemoVoc is a FAIR vocabulary in the sense described below. It is built on an open-source technology, Skosmos (National Library of Finland, n.d.), which provides indexing, searching, and browsing as well as direct communication with computers via an application programming interface (API). The structured links used in the POPIN Thesaurus ("Narrower term," "Broader term," "Related Term") have been translated into SKOS (Simple Knowledge Organization System), a reference system used on the "semantic web" (see below).



Media Lens King. *African grandfather reads a book with his grandson*. Shutterstock. Retrieved 28 April 2023.

4 FAIR VOCABULARIES

4.1. LEVELS OF CONTROLLED VOCABULARIES

A controlled vocabulary is a standardized set of words and phrases used to organize and describe knowledge. Controlled vocabularies are fundamental for data sharing, because they simplify communication, discovery, and reuse of data. As an empirical and data-driven discipline, controlled vocabularies are pervasive in population research. For example, occupations reported in censuses and surveys must be coded into a standard set of categories, such as the International Standard Classification of Occupations (ISCO) (International Labour Office, 2012), to remove alternative spellings and enable groupings by industry, social status, etc. Some controlled vocabularies are maintained by institutions, like ISCO, which is a product of the International Labour Organization (ILO), while others are created by data producers and researchers as the need arises.

Figure 1 illustrates three types of controlled vocabularies relevant to demographic data.[1] A Concept is an abstraction that refers to a type of information. For example, age is a general term referring to the length of time since the birth of a subject. A Concept is turned into data when it is the target of an instrument, such as the question "How old were you on your last birthday?" We will refer to the results of an instrument that generates data as a Variable or Measure. Most Concepts can be measured in different ways, and an abridged version of age, such as a set of age groups, is also a Variable representing the concept of age.

The difference between "age at last birthday" and "age group" is apparent when we look at the responses for each of these variables. "Age at last birthday" produces a set of integer values beginning with 0 and very rarely exceeding 120. An "Age group" combines these values into categories, such as the five-year age groups shown in Figure 1. We will call the set of responses to a variable a Value Domain or Code List. Note that a five-year age group, such as "0 to 4", is a category not a number, even though it may be represented by a number in the data.

Figure 1. Levels of Controlled Vocabularies

Level	Example	
Concept	Age	
Variable / Measure	Age at last birthday	Age group
Value Domain / Code List	(0, 1, 2, …)	("0 to 4", "5 to 9", "10 to 14", … "90 or older") ("0-14", "15-64", "65 and older")

1 Figure 1 is a simplified version of the DDI "variable cascade", which can be found in the "DDI-Cross Domain Integration: Detailed Model" (DDI Alliance, 2020).

The levels in Figure 1 describe a hierarchy that becomes broader as we move from Concept to Variable to Value Domain. Many Variables may be used to measure Age, and a Variable like age group can be categorized in many different ways. The one-to-many relationships of Concepts to Variables and Variables to Value Domains creates both opportunities and difficulties.

Data catalogs use the relationships between Concepts and Variables to help users to navigate their collections. A researcher may begin with "mortality" and then choose among more specific types of mortality (infant mortality, maternal mortality, occupational mortality, etc.) as a path to a relevant data set.

Controlled vocabularies are important for showing differences as well as similarities, especially for Value Domains. If a researcher is trying to combine data from several sources, differences in Value Domains can prevent comparisons. For example, the classification of marital status has become much more complicated than it was a decade ago. Until recently, marital status was usually classified into four categories: single, married, divorced, widowed. The Eurostat SDMX registry[2] divides marital status into 29 categories with separate codes for consensual unions, registered partnerships, and same/opposite sex unions. How does one compare a data set in which "consensual union" is considered a marital status to one that does not include that category? Harmonizing data sets using different Value Domains is one of the most common and time consuming tasks for researchers who combine data from multiple sources. (See discussion of IPUMS below.) Thus, controlled vocabularies tell us both when things are similar and when they are different.

Controlled vocabularies are a key part of the infrastructure for finding and accessing information on the Internet. As such, they have become an important focus of the FAIR data movement.

4.2. FAIR VOCABULARIES

FAIR stands for Findable, Accessible, Interoperable, and Reusable, first introduced in 2016 by Wilkinson et al. (2016) as the goals of the FAIR Guiding Principles (Figure 2). FAIR has been enthusiastically received by an existing community committed to expanding access to scientific data, and the FAIR principles are endorsed by CODATA, the Research Data Alliance, and World Data System, as well as UNESCO, OECD, and research funding agencies in Europe and the U.S. (European Research Council, 2018; National Institutes of Health, 2021; National Science Foundation, 2022; OECD, 2006; UNESCO, 2021).

2 Obtained from the Eurostat SDMX registry API with query "https://ec.europa.eu/eurostat/api/dissemination/sdmx/2.1/codelist/ESTAT/MARSTA?detail=referencestubs&completestub=true"

Figure 2. FAIR Guiding Principles from (Wilkinson et al., 2016)

The FAIR Guiding Principles
To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifer of the data it describes
F4. (meta)data are registered or indexed in a searchable resource
To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available
To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data
To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

It is important to recognize that FAIR aims to make data available to machines as well as humans. FAIR data will speed scientific discovery by automating many time-consuming data processing tasks. Computer programs will be able to find relevant data and perform data preparation and analysis steps. The benefits of automated data analysis are already apparent in population research, where comparative studies of standardized data series, like the Demographic and Health Surveys, often combine hundreds of data sets from dozens of countries. FAIR promises to enable automation of data processing for studies that integrate many different data sources.

All of the FAIR Guiding Principles are fundamentally about metadata, and achieving machine actionable data requires machine-readable metadata. Metadata must be in formats that machines can read and process. Humans can read and understand lightly formatted text documents, but computers cannot. Computers require metadata in structured formats that allow programs to identify the role of each item (e.g., author, title, variable name, variable label, etc.). Standards for structured metadata files describing social science data have developed over the last two decades. The two leading standards are Statistical Data and Metadata eXchange (SDMX) and the Data Documentation Initiative (DDI), (discussed below).

In addition, FAIR metadata uses persistent identifiers to recognize and locate digital objects. Persistent identifiers (PIDs) are codes that can be resolved to a web page, web service, or document accessible on the Internet. Persistent identifiers that conform to established patterns are often described as URIs (Uniform Resource Identifiers) or IRI (Internationalized Resource Identifiers). Persistent identifiers may be globally unique (URNs), resolvable to locations on the Internet (URLs),

or both. Since persistent identifiers are expected to be stable and long-lasting, they are usually associated with registry systems that assign and translate them into URLs. For example, journal articles are assigned Digital Object Identifiers (DOIs), one kind of persistent identifier, which are resolved to a publisher's web page by a registry service called Crossref. Similarly, most data repositories assign DOIs to data sets through an organization named DataCite. Implementation of FAIR involves a dramatic increase in the use of persistent identifiers, which will be used to tag not only data files but attributes of data files, such as concepts, variables, value schemas, and values.

Persistent identifiers allow computers to answer questions about complex digital objects. For example, how do we know when variables in two different data sets are the same? For a human, the answer is contained in a codebook or user guide, which may have an image of the survey questionnaire. Programming a computer to handle all of the possible permutations in a codebook is extremely difficult. But computers can easily process structured metadata files (e.g., SDMX or DDI), which can include persistent identifiers describing variables and their value schemas. If variables in two data sets have the same persistent identifiers, they are the same.

Persistent identifiers can also show computers relationships between digital objects. For example, suppose that we are interested in data about fertility. Assume that an online thesaurus of demographic terminology is available in which "parity progression rate" and "crude birth rate" are tagged as measures of the concept "fertility" through their persistent identifiers. A computer using this thesaurus to discover measures of fertility will return data sets including either the "parity progression rate" or the "crude birth rate." Unlike an Internet search engine, searching by persistent identifier will not return unrelated measures, like soil fertility or the fertility of dolphins.

A group formed at a CODATA-DDI workshop has published advice for making an existing vocabulary FAIR (Cox et al., 2021), which has guided the preparation of this report. These "ten simple rules" (Figure 3) emphasize the importance of governance, which is mostly implicit in the original FAIR guidelines (Wilkinson et al., 2016). Since a FAIR vocabulary is intended to be an authoritative and reliable reference, it must be supported by well-established institutions that can assure its long-term stability and availability. Moreover, a FAIR vocabulary must have procedures to assure that it continues to represent best practices in its scientific community. It is thus very appropriate for IUSSP to undertake this role for population science.

Figure 3. Ten simple rules for making a vocabulary FAIR

Ten simple rules for making a vocabulary FAIR	
Rule 1.	Determine the governance arrangements and custodian of the legacy vocabulary
Rule 2.	Verify that the legacy-vocabulary license allows repurposing, and agree on the license for the FAIR vocabulary
Rule 3.	Check term and definition completeness and consistency in the legacy vocabulary
Rule 4.	Establish a traceable maintenance-environment for the FAIR vocabulary content
Rule 5.	Assign a unique and persistent identifier to (a) the vocabulary and (b) each term in the vocabulary
Rule 6.	Create machine readable representations of the vocabulary terms
Rule 7.	Add vocabulary metadata
Rule 8.	Register the vocabulary
Rule 9.	Make the vocabulary accessible for humans and machines
Rule 10.	Implement a process for publishing revisions of the FAIR vocabulary

Source: (Cox et al., 2021)



Wheaton, Ashley. (2009). *A female surveyor interviews a female participant in her home.* Wikimedia Commons. Retrieved 28 April 2023.

5 USE CASES

This section describes potential applications of FAIR vocabularies in population research. We focus on cases that enable automation of tasks that currently require human intervention.

5.1. DATA DISCOVERY

FAIR vocabularies will enhance data discovery tools that already exist by adding more precision and clarifying relationships between concepts and variables. Researchers searching for data are likely to use two technologies. First, they can rely on standard search engines, like Google and Bing, which process all of the content on the Internet. Search engines are very good at finding data published in online tables. They are less able to find variables within data sets, but there are initiatives, like schema.org, that allow webmasters to provide details about digital objects, including data sets, in ways that search engines can access. Second, researchers can go directly to the catalogs of data repositories (e.g., CESSDA and Data-PASS) and data providers, like statistical agencies. These catalogs often have functionality that search engines cannot provide, like variable-level searching and online analysis tools. Both of these approaches are designed for humans, and it can be difficult to program machines to use them. As everyone knows, searching for the wrong words can produce a flood of irrelevant results. When searching is based on persistent identifiers found in FAIR vocabularies, the outcomes are more consistent and precise.

In a world where FAIR vocabularies have been harmonized and integrated, it will be possible to navigate from concepts to measures to data across all three levels described in Figure 1. A user who selects a concept from a thesaurus will be directed to measures of that concept, such as links from “fertility” to “parity progression ratio” and “crude birth rate.” Data catalogs will use the same identifiers as the thesaurus, allowing a researcher to quickly find data sets containing each measure. It will also be possible to filter results by value schemas, such as geographic areas or alternative ways of grouping ages. Since FAIR vocabularies are machine actionable, the researcher may write a program to harvest all estimates of parity progression ratios by social status and year for a specific country.

5.2 DATA INTEGRATION

We use data integration to refer to the process of combining data from multiple sources. This can be a very time-consuming task, because it requires ensuring that each source is measuring the same concepts in the same ways. This ambiguity is removed when data are accompanied by machine-actionable metadata with persistent identifiers for variables and value schemas. The child-woman ratio, a common measure of fertility, is a ratio of children to women in the childbearing ages in a census or survey. The most common ratio is between children aged 0-4 and women aged 15-49, but other age groups may be used, only ever-married women may be counted, or only children living with their mothers included. If each variant of the child-woman ratio has its own persistent identifier, a computer program can merge data from a variety of sources.

As other scientific domains adopt FAIR principles, the potential for combining data from multiple sources increases. For example, demographic data can be augmented with a wide variety of contextual variables ranging from soil quality maps to weather histories to distances to health care facilities.

5.3. DATA HARMONIZATION

Differences between data sets are often overcome through a process of data harmonization. The world leader in data harmonization is IPUMS (see below). The IPUMS International project, one of nine IPUMS collections, distributes microdata from more than five hundred censuses and surveys spanning over one hundred countries. The IPUMS International "harmonization table" for marital status has four main categories (single, married, separated/divorced, widowed), which appear in more than fifty variants. For example, a person who is separated/divorced might have been recorded as "divorced," "divorced, unregistered", "annulled," "separated," "married living separately," "separated from consensual union," or "spouse absent." IPUMS International has assigned unique codes to each of these categories and provides data on marital status grouped into categories that are comparable across time and place. In addition to the "harmonization table" for marital status, the IPUMS International website provides the program code used for harmonization.

If data sets included persistent identifiers showing how marital status is coded, a computer program could apply IPUMS harmonization procedures. For example, if the desired outcome is a variable showing age in five-year groups, a computer can be programmed to read the persistent identifier for the value schema of an age variable and select an appropriate harmonization table to produce five-year age groups.

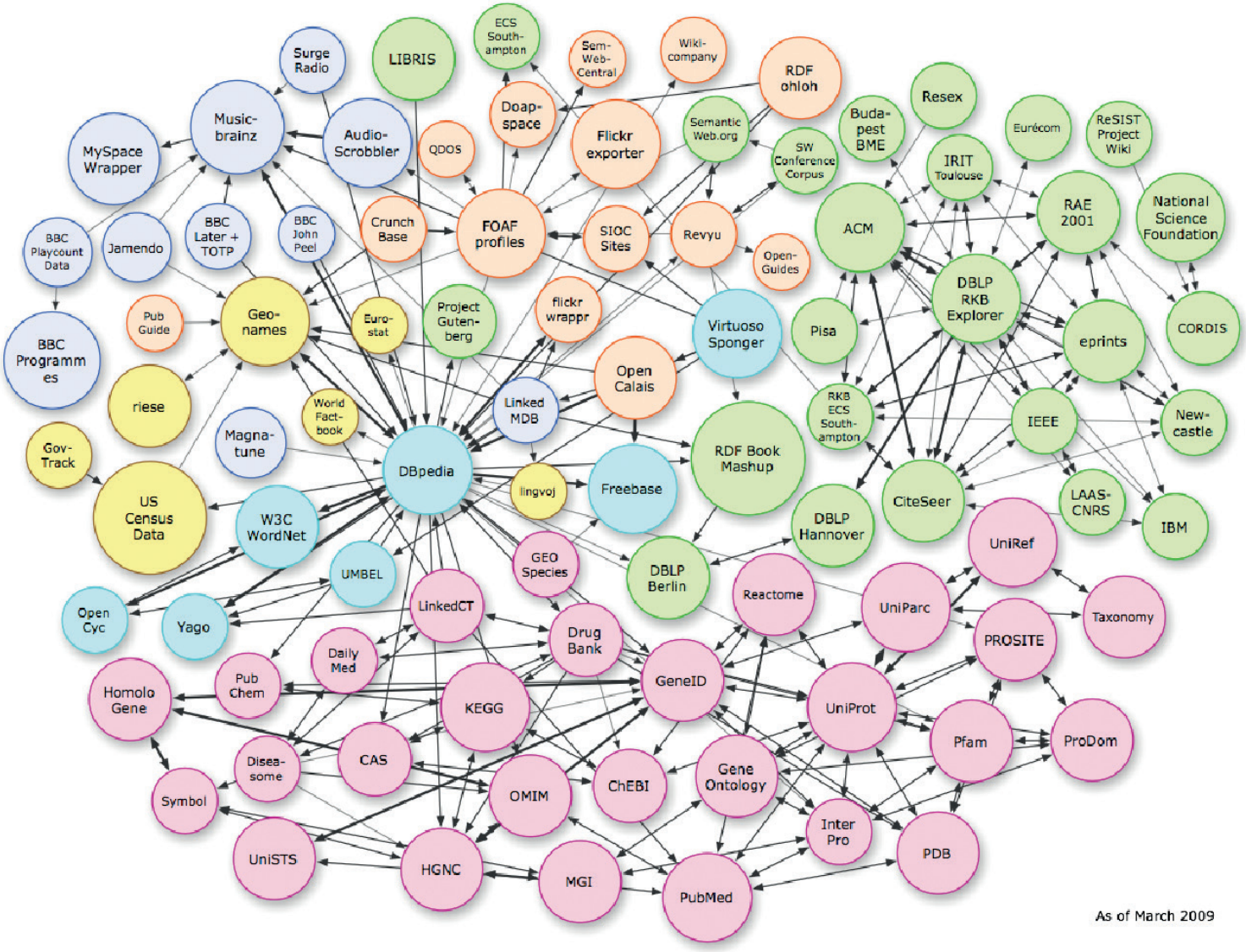
5.4. DATA PRODUCTION

FAIR vocabularies can be integrated in data production processes in ways that reduce costs and make data easier to share. The benefits of standardization in data collection and processing have long been apparent to producers and distributors of national statistics. International agencies, such as the UN, OECD, and ILO, have agreements with national statistical offices to ensure that data used in official reports are consistent and comparable. These concerns resulted in Statistical Data and Metadata eXchange (SDMX), an international initiative for standardizing and facilitating the exchange of statistical data. FAIR vocabularies are very consistent with the mission and approach of SDMX. Embedding persistent identifiers in SDMX metadata provides assurance that data are consistent with international standards and agreements. This can save time and effort for both statistical agencies and data aggregators like the UN. As discussed below, much of the infrastructure for implementing FAIR vocabularies already exists.

Even data collection projects not designed for international exchange can benefit from FAIR vocabularies. We outline here how a new survey can be designed using a metadata driven workflow.

> The first stage of a new project involves translating research concepts into instruments. Researchers can benefit from earlier research by consulting a question bank, such as the CESSDA Euro Question Bank (EQB), that includes high quality questions in multiple languages (CESSDA, 2022b). Items in the question bank will be linked to data sets, which can be used to compare and evaluate them. If the question bank is tied to a thesaurus, like the European Language Social Science Thesaurus (ELSST), searching for concepts will produce a list of relevant instruments (CESSDA, 2022a). When a list of survey questions has been selected, all of the metadata (question text, response values, response codes, etc.) can be downloaded from the question bank. This information can be fed directly into a survey design language to configure a computer-assisted interview (CAI) system. Since data received from the CAI system will be consistent with all of the metadata used to create the survey, documentation, such as codebooks, can be generated automatically. The data and metadata can be transferred to a data repository, which will publish them for discovery on the Internet.

Almost all of the steps in this ideal workflow are already feasible. ELSST already exists, the EQB is under construction, and most of the workflow has been implemented in the Colectica suite of tools (Colectica, 2022). Questionnaires can be assembled from a variables database and translated directly into CAI systems, like Blaise and CASES. Some statistical organizations are also developing concept and classification management systems, such as the Ariā system at Statistics New Zealand (Statistics New Zealand, 2022).



Bizer, C., Heath, T., & Berners-Lee, T. (2009). *Linked Data - The Story So Far. Int. J. Semantic Web Inf. Syst., 5,* Figure 2. Retrieved 28 April 2023.

6 STANDARDS AND TECHNOLOGIES FOR DATA DISCOVERY AND EXCHANGE

> Pure data is barren. For example, the game scores 4 to 3 and 2 to 1 mean almost nothing until you identify the sport, the team names, and when the games were played. The data are 4, 3, 2, and 1. The metadata (information about the data) provided is that these data are game scores. The metadata needed for the data to be useful are the sport, team names, and dates. It would also help if it were explained that these were women's Olympic soccer (football) games.
>
> (Bank for International Settlements et al., 2002, p. 6)

The social sciences are fortunate to have two well-established standards for metadata. The Data Documentation Initiative (DDI) is an international standard for describing data from surveys and other observational methods. Statistical Data and Metadata eXchange (SDMX) is a standard used by statistical agencies to describe and share statistical data. DDI and SDMX are complementary standards used for different purposes, and the organizations that maintain them are committed to consultation and cooperation.

6.1. METADATA STANDARDS

6.1.1. DDI

The Data Documentation Initiative grew out of decades of cooperation among social science data archives in the U.S. and Europe (Vardigan et al., 2008). DDI provides a machine-actionable format for describing digital data derived from a survey, census, or other series of observations. Social science data is distributed to researchers with a "codebook" or "user guide," describing the contents of the data file, such as the meanings and physical locations of variables, as well as information about the authors and methods used to produce the data. Codebooks and data catalogs began as printed documents, but the emergence of direct downloading of data over the Internet made digital representations of metadata much more efficient. DDI can be rendered into digital (e.g., pdf) codebooks, thus eliminating the need for printing, storing, and mailing codebooks. DDI can also be indexed and searched, resulting in very capable online catalogs and data discovery tools.

DDI describes data in a rectangular matrix, where rows are observations and columns are variables. Observations may refer to people, countries, time periods, etc. This is the standard format used by statistical analysis software, such as SPSS and Stata. R and Python refer to this format as a "dataframe."

DDI is primarily used by data archives, but its use is spreading among data producers. Recent developments, such as the Colectica (Colectica, 2022) suite of tools, have created metadata-based workflows that can reduce the costs of producing and documenting new data. For example, Colectica includes a variable database from which data producers can design entire questionnaires and translate their questionnaires into code for widely-used computer-assisted interview systems.

6.1.2. SDMX

SDMX resulted from a partnership of seven major international organizations aimed at improving the efficiency of data exchange and data sharing (Bank for International Settlements et al., 2002). All of these organizations collect data from national statistical agencies, banks, or other sources, which they combine and publish for policymakers and the public. SDMX provides a way to standardize agreements for the exchange of data between international organizations and member states or agencies. SDMX also allows international organizations to provide access to their data more quickly and to develop new interactive tools. SDMX version 3.0, which was released in 2021, makes it easier to reuse existing artifacts, like codelists, and to map equivalences between different artifacts.

SDMX is typically used to describe a multidimensional data structure known as a "data cube." Data cubes describe a quantity, called a "measure," that has been

observed across a number of "dimensions." Data cubes are especially appropriate for aggregate data, such as counts, rates, averages, or sums. There are usually more than three dimensions, which often include geographic and temporal units. For example, a set of unemployment rates broken down by sex, age, year, and country can be described as a four-dimensional matrix with a measured value, the unemployment rate, in each cell. Software allows the user to "slice" a data cube into sub-matrices by filtering the values on each dimension. For example, if the user filters on "Belgium," they will get a two dimensional matrix with dimensions sex and age.

The SDMX community has developed protocols for application programming interfaces (APIs), which allow software applications to communicate with each other. The SDMX APIs provide tools for obtaining both metadata and data, and a number of organizations are implementing SDMX APIs.

Both SDMX and DDI are usually stored in XML (Extensible Markup Language) format, which is both machine- and human-readable. XML is an international standard published by the World Wide Web Consortium (W3C) (Extensible Markup Language (XML) 1.0 (Fifth Edition), 2008). SDMX also supports other formats, such as CSV and JSON.

6.2. SEMANTIC WEB STANDARDS

The Semantic Web, also called Linked Data, is a set of technologies that attach meaning to digital objects on the Internet. Web applications can use these meanings to automate reasoning and perform functions currently requiring human intervention (Berners-Lee et al., 2001).

Resource Description Framework (RDF) is a way of making statements about Web resources. Each statement consists of three parts: subject, predicate, and object. The predicate describes some relationship between the subject and the object (Cyganiak et al., 2014). RDF statements are usually called "triples." Figure 4 shows a simplified example of RDF triples. The first row shows that "mortality" is a "narrower" concept than "demography." The last two rows show two concepts "narrower" than "mortality": "infant mortality" and "occupational mortality." These relationships allow a computer program to infer that "infant mortality" is a concept within "demography."

Figure 4.

Subject	Predicate	Object
demography	narrower	mortality
mortality	narrower	infant mortality
mortality	narrower	occupational mortality

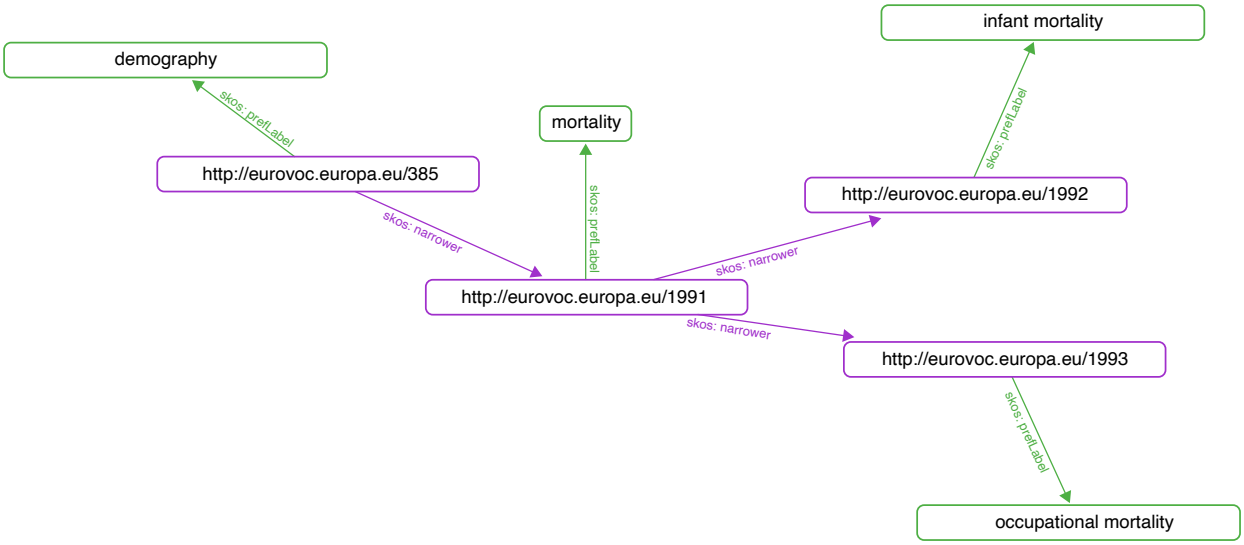
On the Semantic Web concepts (such as ”demography,” “morality,” and “narrower”) are identified by Universal Resource Identifiers (URIs), which link to locations on the Web. URIs remove any ambiguity about terms that may be used differently in different communities. Every URI is unique, and most URIs point to a webpage with explanatory information. Figure 5 illustrates this principle with triples downloaded from Eurostat’s Eurovoc thesaurus (Publications Office of the EU, 2022). Most items in Figure 5 are URIs pointing to pages that have additional information. RDF triples are often described as a graph (Figure 6).

Since the URIs are intended for machines, we have also included triples with the predicate “prefLabel” to provide a human-readable name in Figure 6. We only show English labels here, but Eurovoc includes labels in every language used in the European Union. In RDF, “prefLabel” may be modified by a language identifier, which makes it easy to support as many languages as possible. The use of URIs for identifying concepts that can be expressed in different languages is analogous to the reference numbers used in the Multilingual Demographic Dictionary (International Union for the Scientific Study of Population, 1954) and Demopædia.

Figure 5.

Subject	Predicate	Object
http://eurovoc.europa.eu/385	http://www.w3.org/2004/02/skos/core#prefLabel	“demography”
http://eurovoc.europa.eu/385	http://www.w3.org/2004/02/skos/core#narrower	http://eurovoc.europa.eu/1991
http://eurovoc.europa.eu/1991	http://www.w3.org/2004/02/skos/core#prefLabel	“mortality”
http://eurovoc.europa.eu/1991	http://www.w3.org/2004/02/skos/core#narrower	http://eurovoc.europa.eu/1992
http://eurovoc.europa.eu/1991	http://www.w3.org/2004/02/skos/core#narrower	http://eurovoc.europa.eu/1993
http://eurovoc.europa.eu/1992	http://www.w3.org/2004/02/skos/core#prefLabel	“infant mortality”
http://eurovoc.europa.eu/1993	http://www.w3.org/2004/02/skos/core#prefLabel	“occupational mortality”

Figure 6. Example RDF Graph



The predicates in Figure 5 are defined in the SKOS (Simple Knowledge Organization System) thesaurus, which describes relationships between concepts. All items in the SKOS thesaurus have URIs that resolve to human-readable definitions. XKOS (Extended Knowledge Organization System) extends SKOS to provide better representations of statistical classifications and relationships between concepts in classifications.



Reyes, Fernanda. *Portrait of a Latin family hugging in rural area.* Shutterstock. Retrieved 28 April 2023.

7 CASE STUDIES

This section describes three initiatives that contribute to the goals of FAIR in the social sciences with direct relevance to population research. The European Language Social Science Thesaurus (ELSST) is a new multilingual thesaurus of social science concepts that has been built on FAIR principles. The SDMX Global Registry is an effort by the international statistical community to increase standardization of documentation and facilitate data sharing. IPUMS is the world leader in harmonizing data from censuses and other sources.

7.1. EUROPEAN LANGUAGE SOCIAL SCIENCE THESAURUS (ELSST)

The European Language Social Science Thesaurus (ELSST)[3] is a multilingual thesaurus for the social sciences. ELSST is published by the Consortium of European Social Science Data Archives (CESSDA)—a partnership of data repositories in 22 countries—with thesaurus content and RDF representations managed by the UK Data Service. The thesaurus includes more than three thousand concepts, which are available in 16 languages. CESSDA partners use ELSST to facilitate data discovery by associating concepts with data in their collections. Data producers are beginning to use ELSST to manage their internal workflows. For example, in the Netherlands the Open Data Infrastructure for Social Science and Economic Innovations (ODISSEI) (ODISSEI, 2020) has been created for publishing the metadata of Statistics Netherlands so that it is accessible to researchers wanting to access the data for research. This metadata catalog is large, written in Dutch, and exceptionally complex to navigate. However, this metadata has been annotated using the ELSST vocabulary. In the ODISSEI metadata portal, this then enables researchers to use English language search terms and will return semantically similar results. Now if you search for "fertility", it will understand the conceptual link to "births" and will know the dutch term "Bevalling". This will then allow the relevant files on births in the Netherlands to be shown in the results. In the future, these annotations will also include the vocabularies used in SDMX.

ELSST was designed to be a FAIR vocabulary, and it meets or exceeds all of the criteria described in the "Ten Simple Rules" (Cox et al., 2021). All items in ELSST are assigned linked data URIs and URNs for referencing by DDI metadata used by CESSDA partners to describe their holdings and populate their data catalogs. Designed to be RDF native from the outset, both at the storage and dissemination level, ELSST is structured using SKOS predicates ("broader", "narrower", "related") which can help researchers to perform searches of varying granularity, depending on the precision with which concepts are assigned to e.g. studies, datasets, variables, questions. ELLST offers both human and computer interfaces. Humans can browse the ELSST Thesaurus (https://elsst.cessda.eu/access-elsst) for definitions of concept and translations into other languages. Computers can obtain the same information via APIs. ELSST is maintained and deployed using widely used open source software VocBench (Tor Vergata University of Rome, n.d.) and Skosmos (National Library of Finland, n.d.).

Through their participation in the Working Group, members of ELSST leadership asked for advice on the representation of demographic concepts. We expect that ELSST will be very responsive to suggestions from a future FAIR Vocabulary of Demography.

3 https://thesauri.cessda.eu/elsst

7.2. SDMX GLOBAL REGISTRY

The SDMX Global Registry was created for storing SDMX objects that are widely used in the SDMX community. It is intended to improve data sharing by making it easier for data providers to reuse the common building blocks of SDMX, such as concept schemes, codelists, and data structure definitions. For example, the SDMX Global Registry provides structures for reporting on the UN's Sustainable Development Goals (SDGs). Users can browse the registry to find approved artifacts and download them in SDMX formats. The SDMX Global Registry can also be accessed by computers through APIs. All of the items in the SDMX Global Registry are provided with both URNs and resolvable URLs consistent with FAIR principles.

Objects in the SDMX Global Registry have been vetted by working groups and approved by SDMX sponsor agencies. The codelist for Degree of Urbanization, for example, was studied and tested by six international organizations before its adoption by the UN Statistical Commission in 2020. Since the Global Registry is primarily intended for global and cross-domain artifacts, much more SDMX content will be stored in registries devoted to more specific communities and subjects. The SDMX community is encouraging all of these registries to use the same open source software, and discussions aimed at creating a federation that will dynamically resolve references across multiple SDMX registries are under way.

7.3. IPUMS

When the Integrated Public Use Microdata Series (IPUMS) project was launched at the University of Minnesota in 1991, the project faced a challenging computing environment. Eight public use samples (PUMS) covering the period from 1850-1980 were available for research. Created by different researchers and entities at different times, the PUMS were not interoperable across datasets with the exception of 1960 and 1970 (both created by the U.S. Census Bureau). Documentation of datasets was also non-standardized, incomplete, and inadequate for automated processing. Work to harmonize the extant PUMS resulted in the creation of a structured hierarchical metadata system for harmonizing microdata and the development of a customizable data extract system. These two significant technical innovations were scalable, replicable, and applicable to a broad range of census and survey data (S. Ruggles, 1991; S. Ruggles et al., 1995, 1997)

Thirty years later, the signature activity of IPUMS is harmonizing variable codes and documentation to be fully consistent across datasets within a collection (IPUMS, 2021). IPUMS provides census and survey data from around the world, integrated across time and space. The development of an extensive technical infrastructure to support large-scale data integration through a structured metadata system means that thousands of population datasets are interoperable. Furthermore, IPUMS data and documentation integration makes it easy for researchers to study change, conduct comparative research, merge information across data types, and analyze individuals within family and community context (Kugler & Fitch, 2018). All IPUMS data and services are available through online data access systems free of charge for registered users.

Although IPUMS data and metadata are findable, accessible, interoperable, and reusable in general terms, they are not FAIR in the more technical sense described in the article "Ten Simple Rules for making a vocabulary FAIR" (Cox et al., 2021). IPUMS was created to address the same problem that motivates FAIR – the fact that data "in the wild" are often not findable, accessible, interoperable, or reusable. The IPUMS approach to solving this problem is to pull in select datasets and thoroughly harmonize and document them such that within an IPUMS data collection, samples from different years or geographies are indeed seamlessly interoperable. This approach greatly enhances the value of those select datasets to the research community, but it is time- and labor-intensive. Less attention has been paid to the goal of making IPUMS data play FAIR-ly with datasets outside of IPUMS.

IPUMS offers users rich metadata along with the data it disseminates, but these data and metadata are accessed through the IPUMS website rather than programmatically. IPUMS users browse the metadata and build data extracts using a graphical user interface on the website of a particular IPUMS data collection. Variable-level metadata available on the website include variable labels, descriptions, comparability notes, availability across samples, questionnaire text, and (for some collections) links to the original pre-harmonized "source" variables in the case of variables that have been harmonized across multiple samples. In the case of the IPUMS International project, the crosswalks that were used to recode the original source variables into a harmonized variable are also available on the website. Variables are also organized into topical groups for ease of browsing.

All of the metadata available on the IPUMS website is stored internally in machine-actionable formats, so they are ripe for programmatic access. Currently, IPUMS NHGIS is the only IPUMS collection with a metadata API, but IPUMS plans to add metadata APIs covering all the data collections in the coming years. This will be an important step towards making IPUMS data and metadata FAIR, but it will not be sufficient without other actions. In particular, FAIR principles imply that IPUMS should assign persistent identifiers to all entities, such as variables, codelists, or codelist mappings.



Coman, Lucian. *African mother with baby girl, Botswana*. Shutterstock. Retrieved 28 April 2023.

8 REPRESENTATION OF DEMOGRAPHIC TERMS AND CONCEPTS IN EXISTING ONLINE VOCABULARIES

The process of building the Semantic Web has already begun in the form of thousands of online vocabularies in dozens of scientific domains[4]. Many of the key terms and concepts used in population studies are already defined in these vocabularies, often in ways that are inconsistent with their definitions in demography. Table 1 illustrates this problem by showing the definitions of two important demographic concepts: fertility and fecundity. These terms are a useful test, because other scientific domains do not define them in the same ways.

The meanings of fertility and fecundity among English-speaking demographers were defined in 1933 at one of the earliest meetings of the Population Association of America (PAA) (Fairchild, 1934). In demography, fertility refers to reproductive performance, while fecundity describes the capacity to reproduce. Thus, fertility is measurable, and fecundity is a potential that may or may not be realized. The *Oxford English Dictionary* (https://www-oed-com.proxy.lib.umich.edu/view/Entry/69506?redirectedFrom=fertility#eid) suggests that these meanings had begun to diverge in the nineteenth century. However, Anglophone demographers appear to be alone in using these definitions. Demographers speaking French and Spanish assigned these definitions in the opposite way. Thus, measurable childbearing is *fécondité* in French and *fecundidad* in Spanish, and the capacity to reproduce is *fertilité* or *fertilidad*. Moreover, other scientific disciplines in the English-speaking world do not follow the definitions of fertility and fecundity established in demography.

Table 1 shows definitions of fertility and fecundity in twenty online resources. Panel A gives the definitions adopted by the PAA in 1933 and shown in *Demopædia*, IUSSP's online thesaurus. The seven definitions in Panel B are consistent with demographic usage. Most of these definitions were located through EuroVoc, a set of vocabularies published by the European Union. Most of the EuroVoc vocabularies do not include definitions, but they do link to corresponding terms in other languages. Panel C shows twelve definitions that reverse the meanings of fertility and fecundity used in demography. A majority of these definitions appear in biomedical sources, but three general sources, ELSST, Wordnet and Wikidata, also use non-demographic definitions. At the time of writing this report, the entry for fertility in Wikipedia has a banner requesting help because its current definitions are inconsistent and confusing. Paradoxically, most of the sources using a non-demographic definition of fertility also include the term "fertility rate" under its demographic definition. Since a rate cannot be computed for an unrealized potential, demographers invented the term "fecundability" to describe an empirical measure of fecundity.

The implication of Table 1 is that many researchers searching for data on human reproduction will be misdirected, because their search engine is using a non-demographic definition of fertility. IUSSP can take steps to ameliorate this situation. First, IUSSP can sponsor an authoritative FAIR vocabulary of demographic terms. Second, IUSSP can work with vocabulary providers to distinguish between demographic and non-demographic definitions of terms. A vocabulary can include multiple definitions of the same term by providing the context in which each definition is used, as the APA Dictionary of Psychology does.

4 See https://lov.linkeddata.es/dataset/lov

Table 1. Definitions of Fertility and Fecundity in Online Vocabularies

Vocabulary	Fertility	Fecundity
A. Definitions used in population studies		
Population Association of America 1933 (Fairchild 1934)	Fecundity expressed in **performance** and therefore measurable	Physiological **capacity** to participate in reproduction
Demopædia	**Fertility** and **infertility** refer to reproductive **performance rather than capacity**, and are used according to whether there was actual childbearing or not during the period under review. ES: fecundidad FR: fécondité	The **capacity** of a man, a woman or a couple to produce a live child is called **fecundity**. ES: fertilidad FR: fertilité
B. Definitions consistent with demographic usage		
EuroVoc	ES: fecundidad FR: fécondité	
European Science Vocabulary (EuroSciVoc)	ES: fecundidad FR: fécondité	
ECLAS Central Library of the Commission of the European Communities	ES: fecundidad FR: fécondité	
STW Thesaurus for Economics ZBW - Leibniz Information Centre for Economics	DE: Fertilität	
UNESCO Thesaurus	Human reproduction FR: Fécondité ES: Fecundidad	
UNBIS Thesaurus United Nations Libraries	Used for Human Fertility FR: Fécondité ES: Fecundidad	
APA Dictionary of Psychology American Psychological Association	1. in biology, the potential of an individual to have offspring. Although most frequently applied to females, it may also refer to reproductive capacity in males. 2. in demography, the number of live children born to an individual or within a population.	1. in biology, a measure of the number of offspring produced by an individual organism over a given time. 2. in demography, the general capacity of a human population to have offspring. A below-average capacity is termed **subfecundity**.

Vocabulary	Fertility	Fecundity
C. Definitions not consistent with demographic usage		
ELSST European Language Social Science Thesaurus	THE POTENTIAL FOR HUMAN REPRODUCTION FR: FERTILITÉ "SCOPE NOTE: DO NOT CONFUSE WITH 'FERTILITY RATE' AND 'BIRTH RATE' WHICH REFER TO ACTUAL CHILDBEARING." **Note: ELSST is revising its coverage of demographic terms with advice from demographers on the IUSSP-CODATA Working Group.**	FECUNDITY Prefer using FERTILITY
FAO agrovoc	The capacity to conceive or to induce conception. It may refer to either the male or female. Of humans, animals and plants; EN: Fertility altLabel: Fecundity ES: Fertilidad altLabel: Fecundidad FR: fertilité altLabel: Fécondité	
WIKIDATA	natural capability to produce offspring ES: Fertilidad FR: Fertilité	actual reproductive rate of an organism or population, measured by the number of gametes (eggs), seed set, or asexual propagules. **Fecundity** is similar to fertility, the natural capability to produce offspring
WordNet	S: (n) birthrate, birth rate, fertility, fertility rate, natality (the ratio of live births in an area to the population of that area; expressed per 1000 population per year) S: (n) fertility, fecundity (the state of being fertile; capable of producing offspring)	S: (n) **fecundity**, fruitfulness (the intellectual productivity of a creative imagination) S: (n) **fertility**, fecundity (the state of being fertile; capable of producing offspring) S: (n) **fruitfulness**, fecundity (the quality of something that causes or assists healthy growth)
Ontobee Ontology of Biological Attributes	A reproductive quality inhering in a bearer by virtue of the bearer's initiating, sustaining, or supporting reproduction. [database_cross_reference: PATOC:GVG]	A reproductive quality inhering in an organism or population by virtue of the bearer's potential reproductive capacity as measured by the number of gametes. [database_cross_reference: Wikipedia:http://en.wikipedia.org/wiki/Fecundity]

Vocabulary	Fertility	Fecundity
C. Definitions not consistent with demographic usage, *continued...*		
National Cancer Institute Thesaurus	The ability of an individual to produce offspring.	
Read Codes, Clinical Terms Version 3 (CTV3)	Ability to conceive	
Medical Subject Headings (MESH)	The capacity to conceive or to induce conception. It may refer to either the male or female.	
Computer Retrieval of Information on Scientific Projects Thesaurus (CRISP)	capacity to conceive or to induce conception; may refer to either male or female.	
Phenotypic Quality Ontology (PATO)	A reproductive quality inhering in a bearer by virtue of the bearer's initiating, sustaining, or supporting reproduction.	A reproductive quality inhering in an organism or population by virtue of the bearer's potential reproductive capacity as measured by the number of gametes.
Gender, Sex, and Sexual Orientation Ontology (GSSO)	The ability of an individual to produce offspring.	A reproductive quality inhering in an organism or population by virtue of the bearer's potential reproductive capacity and measured by the number of gametes.
International Classification for Nursing Practice (ICNP)	Capacity to participate in the conception of a live foetus that delivers as a viable child.	



IVASHstudio. *Indian groom dressed in white Sherwani and red hat with stunning bride in red lehenga stand and hold each hands walking outside*. Shutterstock. Retrieved 28 April 2023.

9 OPPORTUNITIES AND NEEDS

The Working Group strongly believes that IUSSP will have an important impact on demography and the social sciences more generally by developing a FAIR Vocabulary of Demography. This Vocabulary will build on IUSSP's long history of supporting dictionaries and thesauruses in multiple languages. We envision a tool that will serve new purposes both within and beyond population research. A rapidly growing field in biomedical research is using controlled vocabularies to find connections among data sets and publications in the search for new drugs and therapies. FAIR vocabularies will open similar opportunities for new forms of demographic research.

Controlled vocabularies already play important roles at all stages of the data life cycle, and different kinds of vocabularies are needed for designing, collecting, cataloging, discovering, and reusing demographic data. To understand the need for controlled vocabularies, it is helpful to consider different types of vocabularies and how they would be used by the demographic community.

9.1 CONCEPT VOCABULARIES

Concept vocabularies are the most commonly discussed vocabularies in social science and are primarily used for cataloging and searching libraries and archives. They are used to annotate data, papers, individuals or any other digital object with social science concepts so that they can be readily identified and retrieved. Concept vocabularies provide a hierarchical overview of a field by organizing concepts into sets and subsets. The resulting graph improves search functionality and returns results more relevant to a specific concept and search term. ELSST is an increasingly used vocabulary in this area and contains an array of demographic concepts such as fertility, marriage, mortality etc. A FAIR Vocabulary of Demography will enrich other vocabularies by providing accurate definitions of demographic concepts and more extensive links among concepts. Demography can be better represented in this area by mapping its own authoritative vocabulary to other resources such as ELSST.

Concept vocabularies improve the findability of data, but they have limited usefulness for interoperability. Knowing that two datasets both contain information on fertility is only useful for human researchers, it is not useful for machine readable interoperability between datasets. Most concepts can be implemented in many different ways, and other tools are needed for automating data merging and harmonization. We anticipate that the FAIR Vocabulary of Demography will begin with simple hierarchical relationships, such as those found in DemoVoc and ELSST. However, it may evolve in the direction of a formal ontology, which can include more complex relationships between terms.

9.2 OPERATIONAL AND TECHNICAL VOCABULARIES

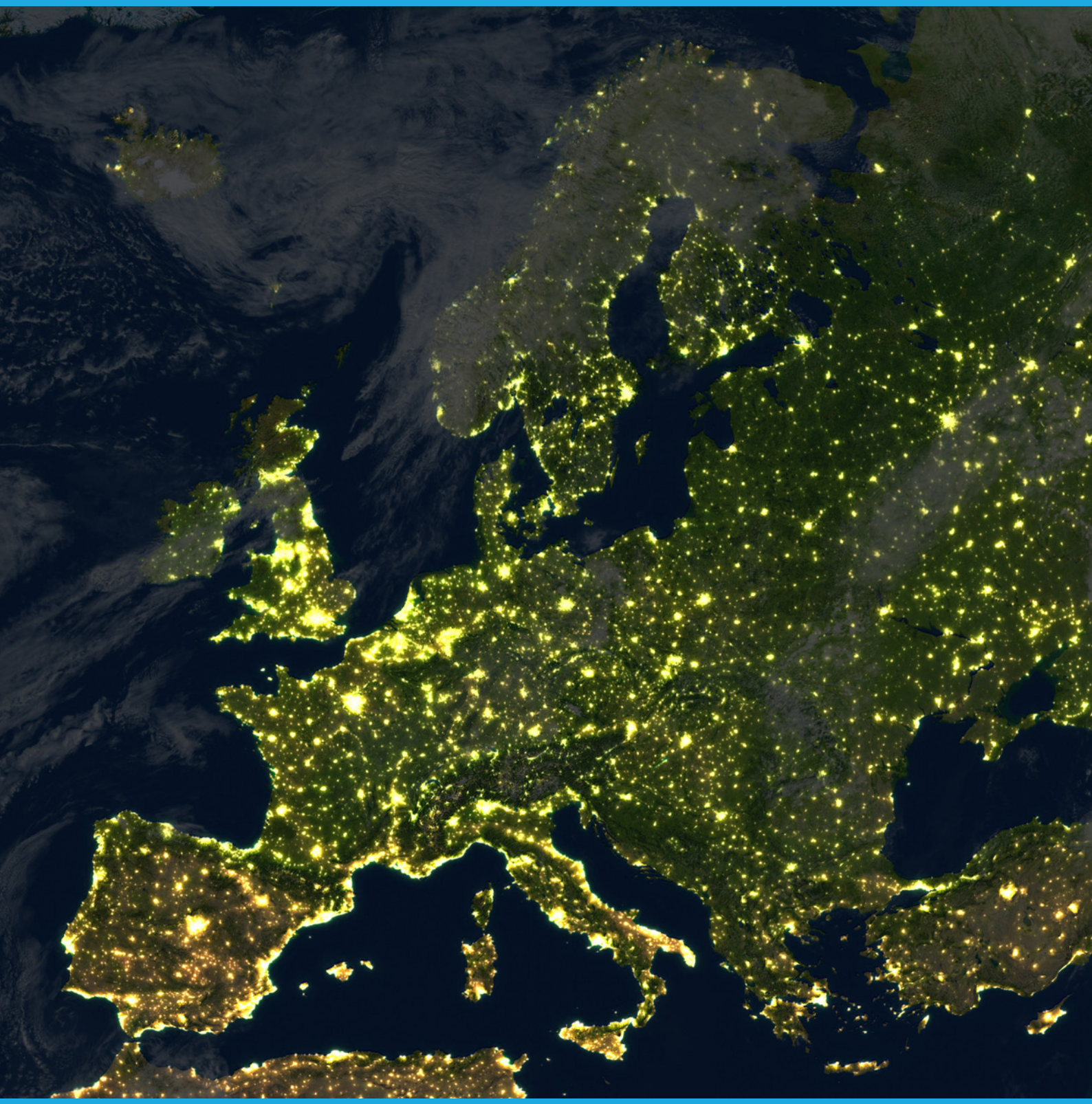
Machine readable interoperability for large scale, semi-automated analysis, such as those conducted in the hard sciences, requires well-defined and precise operational vocabularies. Demography as a field is blessed with such definitions but they are not FAIR. When researchers generate a specific variable within a dataset, there is no way to identify in a machine readable manner that the data they have generated meets specific criteria. For example, the Human Mortality Database (HMD), a hugely successful and well regarded data source in demographic research, standardizes mortality data across space and time. This is a fundamental data source in global demography, which has been vital for understanding the pandemic. Yet, if you want to understand how HMD calculates any measure, you need to consult their methodological protocol in PDF and read the relevant equation. If you have created your own dataset using the same definition/operationalization as HMD, you have no way to declare this. There is no way to tell the world that your data are consistent with data in the HMD.

To solve this problem, an operational vocabulary of demography is required. The FAIR Vocabulary of Demography can assist resources like HMD, the Human Fertility Database (HFD), and others by providing persistent identifiers for measures and methodologies. The definitions and measures of key concepts such as period life expectancy, total fertility rate, birth rate and other life table concepts are well defined and accepted within demography. If the FAIR Vocabulary of Demography codifies and publishes these definitions as FAIR digital objects, the demographic community can cite these standards wherever relevant. In the life sciences, extensive machine readable vocabularies are already used for machine intensive research, such as discovering new drugs. The application of these standards will open new avenues for demographic research.

There is a similar need for technical definitions of data sources in a FAIR vocabulary. Demography as a field delineates itself from many other fields of social science by a keen focus on the underlying data source, precise definition of the statistical universe, and fastidious attention to detail on the quality of the data source. A demographer's contribution to any social science symposium is often simply to ask "what is the denominator?" In collaboration with CESSDA, DDI has begun to incorporate survey methodological applications such as 'sampling method' (CESSDA, 2019). Demography can improve this vocabulary by providing technical definitions of population data systems, such as censuses and vital registration. For example, population registers, which are increasingly used to define sampling frames in Europe, differ from censuses in important ways (Poulain et al., 2013).

9.3 CODE LISTS AND CODE MAPPINGS

The FAIR Vocabulary of Demography should also include Code Lists of allowable categories for demographic variables and Code Mappings translating between alternative Code Lists. Code lists were not part of earlier demographic dictionaries or thesauruses, but they are much more important in a world of FAIR data. As noted above, even basic demographic attributes, like marital status and gender, have become contentious, and the enormous success of the IPUMS Project shows the value of harmonized data to the research community. IUSSP should be proactive in partnering with organizations that deal with these issues (like the UN Population Division) and in marshaling the expertise of the international demography community.



Harvepino. *Europe at night. 3D illustration with detailed planet surface and visible city lights.* Elements of this image furnished by NASA. Shutterstock. Retrieved 28 April 2023.

10 RECOMMENDATIONS

10.1 RECOMMENDATIONS FOR IUSSP

10.1.1. FAIR Vocabulary of Demography

The central recommendation of this report is that IUSSP should develop an authoritative FAIR Vocabulary of Demography. A trustworthy vocabulary for demographic concepts and measures will benefit both population research and the wider scientific community. IUSSP is uniquely placed to develop this vocabulary, because of its global membership and reputation among both academic and non-academic demographers. IUSSP has a long history of defining and standardizing demographic terms, most recently in Demopædia. Demopædia is a remarkable achievement, and the FAIR Vocabulary of Demography will repackage much of that work in a new form.

As discussed above, FAIR vocabularies play a key role in the ongoing development of tools and services known as the semantic web. These technologies depend upon URIs (Uniform Resource Identifiers) in machine actionable thesauri and ontologies to attach meaning to content. A FAIR Vocabulary of Demography can ensure that these services represent demographic terms correctly, and it will enable new forms of data discovery and data access.

IUSSP leadership in developing a FAIR Vocabulary of Demography will have effects far beyond population research. First, members of the Working Group have already advised ELSST about revising its representation of demographic concepts. Secondly, a complete FAIR vocabulary will be welcomed by other services that describe and distribute demographic data. For example, national and international statistical organizations can tag their data products with IUSSP-provided URIs to make them more discoverable and interoperable. Lastly, demography will become a model for other social science disciplines.

10.1.2. Development of a FAIR Vocabulary of Demography

A FAIR Vocabulary of Demography can be developed very rapidly, because it can be built from existing content in DemoVoc and Demopædia. Combining DemoVoc and Demopædia will take advantage of the strengths of each resource. DemoVoc is already FAIR. It assigns URIs to terms, and it uses SKOS to describe relationships among concepts. However, DemoVoc lacks definitions, and extracting these definitions from Demopædia will save a large amount of time. Similarly, DemoVoc currently supports only French and English, but terms in Demopædia have been translated into 16 other languages.

We recommend that IUSSP constitute a standing committee to produce and maintain the FAIR Vocabulary of Demography. The Appendix "Governance of a FAIR Vocabulary of Demography" below sets out a model for managing both the content and technology of a FAIR vocabulary.

IUSSP should seek partnerships with other organizations in support of the FAIR Vocabulary of Demography. In particular, the UN Population Division worked closely with IUSSP and CICRED on earlier versions of the Multilingual Demographic Dictionary and Population Multilingual Thesaurus. Many organizations that produce, distribute, and analyze demographic data will benefit from the Vocabulary.

There will be costs associated with creating a FAIR Vocabulary of Demography, such as blending content from DemoVoc and Dempaedia. Funding agencies in both Europe and the US are promoting FAIR, and the Working Group believes that a grant proposal from IUSSP, CODATA, and partners would be successful.

10.1.3. Outreach to the SDMX and DDI communities

IUSSP can increase adoption by data providers in the SDMX and DDI communities by exporting content from the FAIR Vocabulary of Demography into the formats used by these communities. For SDMX, this means creating Concept Schemes and Codelists in SDMX format. Ideally, these SDMX artifacts will be hosted by the SDMX Global Registry or one of the other SDMX registries. Similarly, entries in the FAIR Vocabulary of Demography can be exported as concepts and variables in DDI format, which will be available to DDI-based data catalogs and data acquisition tools like Colectica. In both cases, the artifact will include URLs pointing to the FAIR Vocabulary of Demography. Both SDMX and DDI are likely to develop federated registries for metadata and data discovery, and IUSSP should be alert to ways of participating in these services. The Working Group believes that reaching out to these communities will make it much easier for data providers to incorporate IUSSP-approved definitions into their products.

10.1.4. Promoting Data Harmonization

The lack of standardization in concepts, measures, and codes across multiple sources is often the most time consuming and costly part of data gathering. The enormous contribution of the IPUMS Project shows the enormous value of harmonized data to the population research community. However, there is often a disconnect between the interests of data users and the problems faced by data producers, who must adapt research instruments to local languages and circumstances. There are no easy solutions to these problems, but IUSSP can advance FAIR data principles in constructive ways.

First, IUSSP can facilitate and publicize standards for concepts, measures, and classifications in demography. Data standardization is only possible when there is agreement within the scientific community. As the international organization of the population research community, IUSSP can build consensus around existing standards and encourage the development of new standards.

Second, consensus building is already implicit in many IUSSP activities, and small steps can make it explicit. For example, IUSSP Scientific Panels can be asked to report on any data-related standards that emerge from their deliberations. New concepts, recommended survey instruments, and standard classification schemes can be added to the FAIR Vocabulary of Demography and disseminated to the SDMX and DDI communities, and Scientific Panels can be encouraged to publish articles explaining these developments.

10.2. RECOMMENDATIONS FOR SDMX

1. The development of SDMX was motivated by the same concerns for sharing data that resulted in the FAIR principles, and the SDMX community will benefit greatly from a few small steps that will make SDMX FAIR. We recommend that the SDMX community develop guidelines for implementing FAIR consistently across all SDMX artifacts and registries.
2. The most important step is the implementation of FAIR identifiers, i.e., persistent and resolvable identifiers for concepts, concept schemes, codelists, and codes. SDMX identifiers are currently text strings (URNs), which could be replaced or supplemented with URLs. SDMX is already encouraging the re-use of identifiers across domains, and the SDMX Global Registry was created for that purpose. We recommend that FAIR identifiers be provided by all SDMX registries and that the scope of the SDMX Global Registry be expanded to serve as a central resource for all vocabulary identifiers.

   We are aware that SDMX agencies may have other reasons for using existing identifiers, such as continuity with existing databases. However, there are ways to add FAIR identifiers to SDMX metadata without removing existing identifiers. For example, URIs linking to external identifiers (e.g., FAIR Vocabulary of Demography) can be included as SDMX Annotations. These annotations could use standard terms like SKOS to describe the relationship between the SDMX item and the external object.
3. SDMX should also move to accommodate semantic web applications by providing metadata and possibly data in RDF. Translation of SDMX concept schemes and codelists into RDF is a good place to start, because it will allow explicit use of relationships specified in SKOS and XKOS, which are more expressive than the SDMX standard.
4. We urge the SDMX community to promote best practices with the aim of simplifying data

dissemination through SDMX APIs and other distribution mechanisms. SDMX is a large and sophisticated standard, and there are often alternative ways to represent the same data. For example, the standard procedure in SDMX is to have a single Measure called OBS_VALUE. However, it is possible to assign more than one quantity to OBS_VALUE by using a "Measure Dimension" to describe the content of OBS_VALUE. Although this is a valid use of SDMX, it requires additional program code for anyone trying to automate data access from multiple SDMX data repositories. Although this example should be resolved by new features in SDMX 3.0, it illustrates differences in applying the SDMX standard that can frustrate users of the SDMX APIs.

10.3. RECOMMENDATIONS FOR THE DDI ALLIANCE

1. We encourage the DDI Alliance to develop guidelines and practices for registries of vocabularies, variables, and value domains similar to the development occurring in the SDMX world. DDI registries should work to encourage reuse of survey instruments and coding schemes in ways that will make it easier to discover and harmonize data. Since the DDI community is very decentralized, we expect that DDI registries will be federated using the DDI Agency Registry to simplify the assignment of persistent identifiers.

   As an XML based standard since its inception, DDI has supported URNs but as new serializations emerge, support should be accelerated for RDF URIs that are natively resolvable to locations on the Web.

2. IUSSP strongly recommends that data producers apply URIs when data is created. It is much more efficient for persistent identifiers to be assigned by data producers, and these identifiers will provide more accurate information than annotations applied by data repositories late in the data life cycle. Tools for assigning persistent identifiers throughout the data lifecycle are already available. Wherever possible, these URIs should point to existing FAIR vocabularies, such as ELSST and the future FAIR Vocabulary of Demography.

3. The DDI Alliance should also take other steps to facilitate harmonization of data sets. The assignment of persistent identifiers described in DDI registries is a pre-condition for reducing the costs of data harmonization, but other steps, such as using tools like SKOS to map variables to concepts are needed. We strongly encourage the DDI Alliance to contribute to research and the development of tools for automating linking identifiers across registries discussed below.

10.4. RECOMMENDATIONS FOR IPUMS

1. IPUMS leadership moving toward compliance with the FAIR principles will have a far-reaching impact on the social sciences. The first step is to make IPUMS metadata programmatically accessible. This will require assignment of URIs to entities such as variables, codelists, and codelist mappings (crosswalks) in IPUMS data collections. Since IPUMS already uses DDI for its codebooks and data extracts, an IPUMS registry federated through the DDI Agency Registry will be very welcome.

2. IPUMS can have a major impact by linking its data descriptions and URIs to other data and concept registries. In particular, much of the IPUMS collection consists of microdata from censuses, and IPUMS users will greatly benefit from links between IPUMS microdata and aggregate data in SDMX data repositories. Links between IPUMS and the SDMX community should also result in new applications of IPUMS codelist mappings. IPUMS is in an ideal position to contribute to research and development toward automated linking of concepts and variables described below.

10.5. RECOMMENDATIONS FOR CODATA

1. This Working Group would be pleased to share its experiences with other domains, especially related fields in the social sciences. We believe that several lessons from our work are worth sharing:
    - We benefited from a wide range of skills. Members of our group had backgrounds in a variety of data related specialties, including data production, data archiving, and metadata development. However, it is essential to include scientists with research experience in the field.
    - We strongly recommend reaching out to all stakeholders and existing data infrastructures. For example, our colleagues from the UN, ILO, and OECD helped us to understand how statistics are shared between national and international organizations and the essential role of the SDMX standard.
    - Scientific domains should understand that if they do not create FAIR vocabularies for their own fields, someone else will do it. Many terms from demography and other social sciences appear in ontologies and thesauruses created for the biomedical community, which sometimes uses very different definitions. Broad efforts, like ELSST, are designed to serve a wide range of fields, and they would welcome expert advice.
2. We encourage CODATA to convene and publish recommendations on the governance and sustainability of FAIR vocabularies. We were fortunate to be in contact with Barbara Magagna and other leaders in this area, but their expertise needs to be widely shared.

10.6. RECOMMENDATION FOR IUSSP AND CODATA: HARMONIZING VOCABULARIES TO HARMONIZE DATA

A central problem moving forward will be the proliferation of persistent identifiers (PIDs) pointing to the same or closely related concepts and variables. The institutional structure of social science data is very decentralized, and many organizations will provide vocabulary registry services. A future IUSSP FAIR Vocabulary will be one among many vocabulary services. This means that common concepts and variables will be assigned different PIDs by multiple registries. The SDMX Global Registry is expected to maintain only a small percentage of the concept and variable descriptions used by national and international statistical agencies. To achieve the goals of FAIR data, the demography community will need to know when PIDs in different vocabularies are describing the same things.

The working group recommends that IUSSP and CODATA encourage research and development using natural language processing (NLP) and machine learning (ML) to compare and link PIDs across registries describing demographic data. A large body of research on this problem has already been done by computer scientists under the headings of "concept normalization" and "entity linking". Development has been especially important in biomedical research, which relies heavily on well-developed ontologies and controlled vocabularies. For example, the National Library of Medicine in the U.S. has developed a Unified Medical Language System (UMLS), which includes a "metathesaurus" linking terms across many vocabularies and software for extracting concepts from text.

APPENDIX: GOVERNANCE OF A FAIR VOCABULARY OF DEMOGRAPHY

The long-term sustainability of a FAIR Vocabulary of Demography requires a commitment by IUSSP to perform ongoing governance and maintenance tasks. As in many other areas, IUSSP will depend heavily on the voluntary participation of members. We describe here a governance model provided by Barbara Magagna, who coordinated the EnvThes thesaurus of Long Term Ecological Research and Experimental Ecology.

Roles:

Vocabulary technical administrator (VTA): The VTA is responsible for the operation of the vocabulary platform. This person should be experienced with the semantic technologies used in the platform, such as SKOS. If IUSSP contracts for a vocabulary platform, the VTA may be part of the organization providing the service.

Vocabulary content administrator (VCA): The VCA is responsible for incorporating changes in the vocabulary. Major changes to the vocabulary will be decided by the Vocabulary Expert Group, but the VCA should be empowered to make minor changes and respond to users of the vocabulary. Depending on the vocabulary platform, the VCA will update the vocabulary directly or work with the VTA to do so. The VCA should be familiar with demographic concepts and measures as well as semantic services. Experience in data production, data archiving, or library science will helpful.

Vocabulary expert group (VEG): The VEG will be constituted as an IUSSP Scientific Panel led by a Chair appointed by IUSSP leadership. The VEG will review the content of Demopædia and decide on all important revisions and extensions. It will be composed of IUSSP members with experience in population research and demographic methods. The group should draw on all of the sub-fields of population studies, and expertise in data acquisition (censuses, surveys, and emerging technologies) as well as data analysis will be useful. The VCA will be an ex officio member of the VEG. Appointments to the VEG will be no longer than three years. The VEG will meet as needed based on consultations between the VEG chair and the VCA.

Language experts: Demopædia is a multilingual thesaurus, and the VEG Chair should be empowered to add experts in selected languages to the VEG as needed.

IUSSP membership: IUSSP members are the ultimate authority on the vocabulary used in population research. The VEG will describe its activities in IUSSP communication channels, and it will periodically solicit suggestions for revising and extending Demopædia.

Tasks:

1. With assistance from the VTA, the VCA will create a github project or other system to track change requests and other tasks. Access to the request tracking system should be prominently displayed on the vocabulary platform.
2. The VEG will meet at least once a year to review content and usage of Demopædia and to resolve any outstanding requests in the ticketing system.
3. The VEG Chair may appoint working groups for specific projects, such as adding or extending support in an under-served language.
4. All changes to Demopædia are implemented by the VCA with help from the VCT as needed.

GLOSSARY

API	Application Programming Interface	An API allows two computer programs to talk to each other. The API specifies how to request a service from a program and the format of the response.
CESSDA	Consortium of European Social Science Data Archives	CESSDA is a partnership of social science data archives across Europe, with the aim of promoting the results of social science research and supporting national and international research and cooperation.
CV	Controlled vocabulary	A controlled vocabulary is a carefully selected list of words and phrases, which are used to tag units of information (document or work) so that they may be more easily compared across applications, including search. Controlled vocabularies solve the problems of homographs, synonyms and polysemes by a bijection between concepts and authorized terms.
DDI	Data Documentation Initiative	DDI is a set of standards for metadata files for describing microdata. DDI is used by most of the data repositories in the social sciences. DDI is maintained by the DDI Alliance:
FAIR	Findable, Accessible, Interoperable, Reusable	FAIR is a set of principles for improving access to data. The FAIR principles emphasize making data accessible to computers without human intervention.
IRI	Internationalized Resource Identifier	An IRI is an 'Internationalized Resource Identifier', which allows for non-ascii characters to be included.
JSON	JavaScript Object Notation	JSON is a simple data exchange format. JSON is a text format that uses tagged fields to identify data elements.
ONT	Ontology	An ontology encompasses a representation, formal naming and definition of the categories, properties and relations between the concepts, data and entities that substantiate one, many, or all domains of discourse. More simply, an ontology is a way of showing the properties of a subject area and how they are related, by defining a set of concepts and categories that represent the subject.
OWL	Web Ontology Language	Semantic Web language designed to represent knowledge about things, groups of things, and relations between things. OWL documents are ontologies based on a computational logic that can be exploited by computer programs.
PID	Persistent Identifier	A PID is a label that uniquely names a person, place or thing. PIDs are managed by organizations that guarantee that they are kept up to date. DOIs (Digital Object Identifiers) are a kind of PID. DOIs for journal articles are managed by Crossref, and DOIs for data are managed by DataCite. See URI.
RDF	Resource Description Framework	RDF is a model used for exchanging data on the Web. RDF consists of 'triples' that describe relationships between things. This creates linking structures which can be represented as graphs. RDF is used for Web-based ontologies with specifications like OWL and SKOS.

SDMX	Statistical Data and Metadata eXchange	SDMX is a metadata standard used by statistical agencies for describing and exchanging data. SDMX is sponsored by a number of international agencies, including the UN, OECD, EUROSTAT, and the World Bank.
SKOS	Simple Knowledge Organization System	SKOS is a common data model for sharing and linking knowledge organization systems, such as thesauri, taxonomies, classification schemes and subject heading systems.
URI	Uniform Resource Identifier	A URI is a text string identifying a specific resource. A URI may be a persistent identifier (URN), a locator (URL), or both.
URL	Uniform Resource Locator	A URL is a kind of URI that includes a location and protocol for accessing a resource on the Web.
URN	Uniform Resource Name	A URN is a globally unique persistent identifier.
XML	Extensible Markup Language	XML is a simple specification for exchanging data in text files. Items in an XML file are identified by tags, which can be designed for any purpose.

REFERENCES

Bank for International Settlements, European Central Bank, International Monetary Fund, Organization for Economic Co-operation and Development, Statistical Office of the European Communities, & United Nations Statistical Division. (2002, March 6). *Common Open standards for the Exchange and Sharing of Socio-economic Data and Metadata: The SDMX Initiative*. Statistical Commission, Joint UNECE/Eurostat Work Session, Luxembourg. https://sdmx.org/wp-content/uploads/wp111.pdf

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web Scientific American, May, 2001. *Scientific American, Inc.*

CESSDA. (2019). *CESSDA Vocabulary Service*. https://vocabularies.cessda.eu/vocabulary/SamplingProcedure?v=1.1

CESSDA. (2022a). *ELSST Thesaurus*. https://www.cessda.eu/Tools/ELSST-Thesaurus

CESSDA. (2022b). *European Question Bank*. https://www.cessda.eu/Tools/EQB

Colectica. (2022). *Colectica*. http://www.colectica.com/

Cox, S. J. D., Gonzalez-Beltran, A. N., Magagna, B., & Marinescu, M.-C. (2021). Ten simple rules for making a vocabulary FAIR. *PLOS Computational Biology, 17*(6), e1009041. https://doi.org/10.1371/journal.pcbi.1009041

Cyganiak, R., Wood, D., & Lanthaler, M. (2014). *RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation, 25 February 2014*. W3C. http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

DDI Alliance. (2020). *DDI-Cross Domain Integration: Detailed Model*.

Demopædia. (2013). http://demopaedia.org/

European Research Council. (2018, February 23). *Open Research Data and Data Management Plans Information for ERC grantees*. ERC: European Research Council. https://erc.europa.eu/content/open-research-data-and-data-management-plans-information-erc-grantees

Extensible *Markup Language (XML) 1.0 (Fifth Edition)*. (2008). https://www.w3.org/TR/2008/REC-xml-20081126/

Fairchild, H. P. (1934). Organization for research in population. *Population, Journal of the International Union for Scientific Investigation of Population Problems*, 79–84.

Gubry, F. (2007). *From Dictionaries to Thesaurus: Lessons for Demopaedia The need for an up-to-date vocabulary in the field of demography*. International Workshop on Demopaedia, Paris. https://www.researchgate.net/profile/Francoise-Gubry/publication/256124812_From_Dictionaries_to_Thesaurus_Lessons_for_Demopaedia_The_need_for_an_up-to-date_vocabulary_in_the_field_of_demography/links/0c960521dbc7b3300a000000/From-Dictionaries-to-Thesaurus-Lessons-for-Demopaedia-The-need-for-an-up-to-date-vocabulary-in-the-field-of-demography.pdf

Hankinson, R. K. (1993). *POPIN Thesaurus: Population Multilingual Thesaurus*. Committee for International Cooperation in National Research in Demography.

International Labour Office. (2012). *International Standard Classification of Occupations 2008 (ISCO-08): Structure, group definitions and correspondence tables*. International Labour Office.

International Union for the Scientific Study of Population. (1954). *Multilingual demographic dictionary. English volume*. United Nations, Department of Economic and Social Affairs.

IPUMS. (2021). *IPUMS*. https://www.ipums.org/

Kugler, T. A., & Fitch, C. A. (2018). Interoperable and accessible census and survey data from IPUMS. *Scientific Data, 5*(1), Article 1. https://doi.org/10.1038/sdata.2018.7

National Institutes of Health. (2021). *Data Ecosystem—Landing Page | Data Science at NIH*. https://datascience.nih.gov/data-ecosystem

National Library of Finland. (n.d.). *Skosmos: Open source web-based SKOS browser and publishing tool*. Retrieved November 3, 2022, from https://skosmos.org/

National Science Foundation. (2022). *Findable Accessible Interoperable Reusable Open Science Research Coordination Networks (FAIROS RCN) (nsf22553)*. https://www.nsf.gov/pubs/2022/nsf22553/nsf22553.htm

ODISSEI. (2020). ODISSEI - Open Data Infrastructure for Social Science and Economic Innovations. https://odissei-data.nl

OECD. (2006). *Recommendation of the Council concerning Access to Research Data from Public Funding*. Organisation for Economic Co-operation and Development Paris, France.

Petersen, W. (1983). Thoughts on Writing a Dictionary of Demography. *Population and Development Review, 9*(4), 677–687. https://doi.org/10.2307/1973545

Poulain, M., Herm, A., & Depledge, R. (2013). Central Population Registers as a Source of Demographic Statistics in Europe. *Population, 68*(2), 183–212.

Pressat, R. (1980). Le vocabulaire de la démographie. *Population, 35*(4), 849–859. https://doi.org/10.2307/1532365

Publications Office of the EU. (2022). *eurovoc—EU Vocabularies*. https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/eurovoc

Ruggles, S. (1991). Integration of the Public Use Files of the US Census of Population, 1880-1980. *1991 Proceedings of the American Statistical Association*, 365–370.

Ruggles, S., Hacker, J. D., & Sobek, M. (1995). *General Design of the Integrated Public Use Microdata Series*. *28*(1), 33–39. https://doi.org/10.1080/01615440.1995.9955311

Ruggles, S., Sobek, M., & Gardner, T. (1997). Disseminating historical census data on the World Wide Web. *Iassist Quarterly, 20*(3), 4–4.

Statistics New Zealand. (2022). *Ariā*. http://aria.stats.govt.nz/aria/#Home:

Thésaurus *DemoVoc*. (2015). https://thesaurus.web.ined.fr/navigateur/en/about

Tor Vergata University of Rome. (n.d.). *VocBench: A Collaborative Management System for OWL ontologies, SKOS(/XL) thesauri, Ontolex-lemon lexicons and generic RDF datasets*. Retrieved November 9, 2022, from http://vocbench.uniroma2.it/

UNESCO. (2021). *UNESCO Recommendation on Open Science—UNESCO Digital Library*. https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en

Vardigan, M., Heus, P., & Thomas, W. (2008). Data documentation initiative: Toward a standard for the social sciences. *International Journal of Digital Curation, 3*(1).

Vincent, P. (1953). Conception d'un dictionnaire démographique. *Population, 8*(1), 103–120. https://doi.org/10.2307/1524983

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., & Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*.

CODATA

The Committee on Data of the
International Science Council

Connect with us at:

www.codata.org
simon@codata.org

International Science Council
5 rue Auguste Vacquerie
75016 Paris, France

www.twitter.com/CODATANews

www.facebook.com/codata.org

IUSSP

International Union for the
Scientific Study of Population

Connect with us at:

www.iussp.org
contact@iussp.org

IUSSP / UIESP
Campus Condorcet,
9, cours des Humanités - CS 50004
93322 Aubervilliers Cedex - France

www.twitter.com/iussp

www.facebook.com/iussp

Cover image: Tika, Pop. *Business woman drawing global structure networking and data exchanges customer connection on dark background.* Shutterstock. Retrieved 28 April 2023.