Aligning SDMX with FAIR Principles: Recommendations of the IUSSP-CODATA Working Group on FAIR Vocabularies in Population Research

Arofan Gregory (CODATA and DDI Alliance)

George Alter (University of Michigan)
Franck Cotton (Institut national de la statistique et des études économiques)
Edgardo Griesing (International Labour Organization)
Abdulla Gozalov (United Nations Statistics Division)
Steven McEachern (Australian National University, DDI Alliance, CODATA)

COMMITTEE ON DATA CODATA INTERNATIONAL SCIENCE COUNCIL

IUSSP UIESP

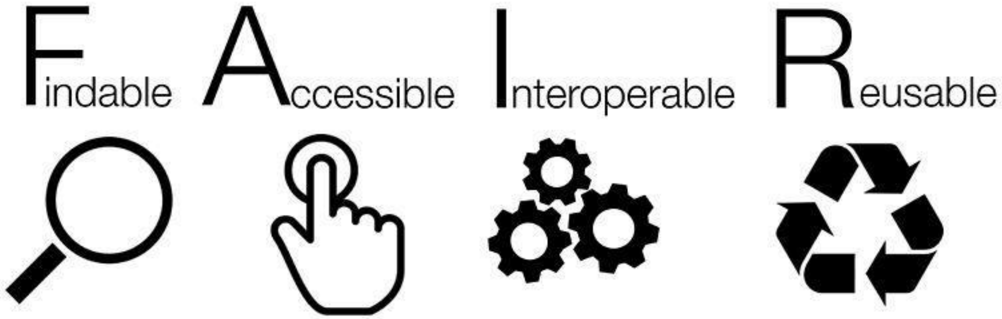
FAIR VOCABULARIES IN POPULATION RESEARCH

Report of the IUSSP-CODATA Working Group on FAIR Vocabularies

IUSSP-CODATA FAIR Vocabularies Working Group

- Goal:
  - Make demographic data more interoperable by publishing controlled vocabularies that can be found and acted upon by software.
- Jointly sponsored by
  - IUSSP – International Union for the Scientific Study of Population
  - CODATA – Committee on Data of the International Science Council
- Co-chairs
  - George Alter (University of Michigan)
  - Arofan Gregory (DDI Alliance / CODATA)
  - Steven McEachern (Australian National University)
  - **With more than 20 demographers, statisticians, and data specialists around the world**

What is FAIR ?



- **Findable**
  The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.
- **Accessible**
  Once the user finds the required data, she/he/they need to know how can they be accessed, possibly including authentication and authorization.
- **Interoperable**
  The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.
- **Reusable**
  The ultimate goal of FAIR is to optimize the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

FAIR is about making data machine-friendly

- FAIR aims to automate the process of combining data from multiple sources
    - “Data wrangling” is the most labor intensive part of empirical research
- FAIR is about metadata
    - Metadata (e.g., a codebook) describes the content and meaning of data
- FAIR data and metadata are available online with “globally unique and persistent identifiers”
    - Persistent identifiers tell machines where to look and what they will find
    - Persistent identifiers point to “landing pages” that are human- and machine-readable

Why focus on FAIR vocabularies?

- “Controlled vocabularies” create standards for terms and measurements
  - Controlled vocabularies are pervasive, e.g. industrial classifications, ICDs, geocodes (NUTS, FIPS, UN/LOCODE,…)
  - FAIR vocabularies allow machines to answer basic questions
- How do we know when two data sets are measuring the same thing?
  - If two variables have the same persistent identifier, a machine knows they are the same
- How does a machine know the difference between the Crude Birth Rate, the General Fertility Rate, the Total Fertility Rate, …?
  - Machines cannot read the PDF
  - Machines can use standards (e.g., RDF, SKOS) to navigate relationships

Use cases for Controlled Vocabularies in Population Science

- Data discovery
  - Example: Find data showing the level of infant mortality
- Data merging
  - Example: Compare infant mortality across countries
- Data harmonization
  - Combine data sets that are coded in different ways
  - Example:
    - Data set A: Infant mortality (ages 0 to one year)
    - Data set B: Neo-natal (0 to 28 days) and Post-neonatal (28 to 365 days) infant mortality

Data merging and data harmonization are very time consuming.
FAIR data can lead to automation.

The Ecosystem of Demographic Data

Demographic data is produced by two communities

  - Each has its own metadata standard: SDMX and DDI

- National and international statistical agencies (SDMX)
  - Standardized on Statistical Data and Metadata eXchange (SDMX)
  - SDMX is supported by the UN, OECD, ILO, IMF, World Bank, Eurostat, European Central Bank, Bank for International Settlements, and national statistical offices
  - International agreements on standards and content are expressed in SDMX
- Surveys and other microdata (DDI)
  - Data Documentation Initiative (DDI)
  - Data production is highly decentralized and uncoordinated
    - Use of DDI in data production is limited but growing
  - Data repositories are coordinated and standardized on DDI
    - CESSDA, DataPASS, Dataverse

FAIR Vocabularies in the SDMX community

- SDMX was created to standardize the exchange of data
  - SDMX is endorsed by international authorities, like the UN Statistical Commission, and is an ISO standard
- SDMX registries share data and metadata
  - Registries are at international organizations (UN, OECD, ILO, World Bank…)
    - Distribute SDMX concepts, codelists, and data structure definitions
    - Aggregate data from national statistical offices, national banks, etc.
  - SDMX Global Registry shares global and “cross domain” artifacts
    - E.g., UN Sustainable Development Goals (SDGs) as well as code lists for the Degree of Urbanisation, Civil Status, etc.
- SDMX is close to FAIR
  - Open source software for SDMX registries deploys common APIs
  - Most SDMX registries assign URNs but could move to resolvable URIs
  - Community is moving toward federated registries
- SDMX is still more rooted in data exchange than data sharing
  - Codelists are not always coordinated across data aggregators
  - Extracting metadata from SDMX APIs requires detailed knowledge of SDMX conventions and practices

If we don’t define demographic terms, someone else will

We searched for definitions “fertility” and “fecundity” in 19 online vocabularies

- 7 vocabularies followed the conventions in demography
- 12 vocabularies reversed the meanings of the terms

Fertility	Fecundity
Demopaedia (IUSSP’s online thesaurus) 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.	The capacity of a man, a woman or a couple to produce a live child is called fecundity**.**
Wikidata natural capability to produce offspring	actual reproductive rate of an organism or population, measured by the number of gametes (eggs), seed set, or asexual propagules.
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.
Medical Subject Headings (MESH) The capacity to conceive or to induce conception. It may refer to either the male or female.	

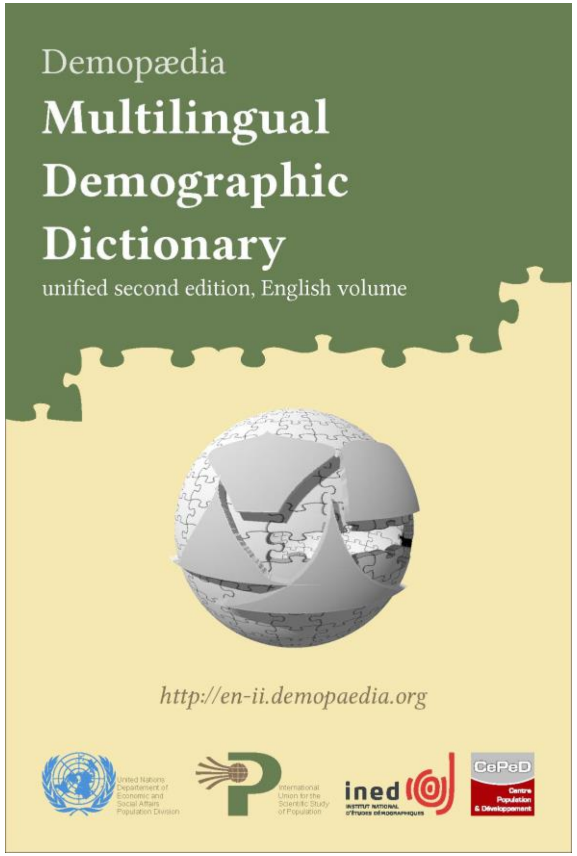
Recommendations for IUSSP

1. IUSSP should create a “FAIR Vocabulary of Demography”
    - A FAIR Vocabulary of Demography will be an authoritative resource that other services can rely on
    - IUSSP has a long history of producing multilingual dictionaries of demography
    - Partnership between IUSSP and UN Population Division
    - A FAIR vocabulary can be built on existing online resources
        - Demopaedia
        - DemoVoc

2. Outreach to SDMX and DDI
    - The FAIR Vocabulary of Demography should export content in SDMX and DDI formats
        - The aim is to lower the cost of adoption of IUSSP definitions by other communities

3. Promote data harmonization
    - Facilitate and publicize standards for concepts, measures, and classifications in demography

4. Appendix on Governance of a FAIR vocabulary



Demopædia
Multilingual Demographic Dictionary
unified second edition, English volume

http://en-ii.demopaedia.org

ined

CePeD

Recommendations for SDMX

1. SDMX registries should become fully FAIR
    - Adopt resolvable identifiers for SDMX Codelists and Concepts (and beyond?)
    - Generate from SDMX URNs?
2. Adopt semantic web standards
    - RDF, SKOS, XKOS
3. Promote best practices to simplify data dissemination
    - SDMX was developed for data exchange
    - Differences in implementation will hinder use of SDMX APIs

SDMX sponsors and users such as Eurostat and INSEE are already exploring these steps

- The technical solutions and expertise already exist
- The biggest need is agreement to take action

Benefits for the SDMX Community

- Better connection to other domains –> increased use of official data
  - Many scientific researchers use classifications from official statistics
  - Official data is very important to baseline and give context to scientific research
- SDMX can access harmonized concepts from demographic/social science
  - “FAIR Vocabulary of Demography” in SDMX formats
  - Tools for linking across domains
- Official statistics can benefit from new sources of data
  - Inputs to broad data sets (SDG Indicators, etc.)
  - Scientific data for quality checks on official data
  - Better inputs for methods like small area estimation, etc.
- The “Grand Challenges” need both official and scientific data to inform policy
  - Climate change
  - Infectious disease
  - Disaster risk reduction

Harmonizing Vocabularies to Harmonize Data

Expect a proliferation of persistent identifiers

- A centralized registry for the social sciences is very unlikely
- Both SDMX and DDI are moving toward federated registries
- Identical concepts, variables, and codelists will have different persistent identifiers in different places
- FAIR will not work if there is no way to link objects in different places

IUSSP and CODATA should encourage research and development on automated tools for linking items across vocabularies

- A large body of research on this problem already exists, especially in bioinformatics
- Agreement on FAIR practice and standards will help support a practical solution

Thanks to:

- Leadership of IUSSP and CODATA
- Co-chairs
- Consultants
- Members of the WG

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)
Franck Cotton (INSEE)
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)

Thank you!

Arofan Gregory

ilg21@yahoo.com

IUSSP-CODATA FAIR Vocabularies Working Group

https://iussp.org/en/iussp-codata-fair-vocabularies-working-group

International Union for the Scientific Study of Population (IUSSP)

IUSSP UIESP

COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL