

Report for the 2021 IUSSP General Assembly of Members

IUSSP-CODATA FAIR Vocabularies for Population Research Working Group

Co-chairs:

George Alter (University of Michigan, IUSSP)

Arofan Gregory (DDI Alliance, CODATA)

Steven McEachern (Australian National University, DDI Alliance, CODATA).

The FAIR Vocabularies for Population Research Working Group is a joint initiative with CODATA, the Committee on Data of the International Science Council. This IUSSP Scientific Panel is a 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 that this tradition can be enhanced by taking advantage of emerging standards and technologies. This Panel will focus on the development of FAIR Vocabularies for population data, which is an essential step in making data reusable and interoperable. The Working Group aims to build upon the work of the CODATA’s FAIR Vocabularies group, who recently released “Ten Simple Rules for making a vocabulary FAIR” (<https://arxiv.org/abs/2012.02325>).

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 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 found and acted upon by software.

The Working Group began meeting regularly in May 2021 (twice a month on average), and it includes more than 25 scientists with representatives from academia and international agencies. The group has discussed the FAIR principles as they apply to controlled vocabularies and related technologies used to make them accessible. We examined three controlled vocabularies of importance to population studies: Demopaedia, IUSSP’s multilingual demographic dictionary; the European Language Social Science Thesaurus (ELLST) developed by the Consortium of European Social Science Data Archives; and the SDMX Global Registry developed international statistical organizations to improve the exchange of data. Our current focus is on the development of use cases that will demonstrate the value of FAIR vocabularies and guide future development. These use cases include data collection, automated data harmonization, multilingual databases, and interoperability of resources such as Demopaedia, ELLST, and the SDMX Global Registry.