Open Classifications for Open Statistical Data

Caterina Caracciolo, Valeria Pesce, Mukesh Srivastava, Carola Fabi
Food and Agriculture Organization of the UN
{name.surname}@fao.org

Abstract

While international organizations and national governments alike are paying increasing attention to making statistical data available as Open and Linked Data (e.g., by making data available in open formats, accessible through public APIs, and endowed with open licenses), the modernization of metadata and statistical classifications is lagging behind. Despite their ubiquitous use, and the consolidated tradition of considering them as intrinsically associated to data, their treatment remains largely ad-hoc and "local" – they are often published in formats that are barely open and hardly machine-actionable, by means of ad-hoc structures, and usually with no explicit public license. Two major consequences are that much manual work is involved when reusing existing classifications within information systems, and that the maintenance of classifications (and associated information, including mappings and multilingual terminologies) is often a cumbersome an time consuming process, often implying massive duplication of effort. In this paper, we report on ongoing work carried on at the Food and Agriculture Organization (FAO) of the UN, aimed at improving the way statistical classifications for agriculture are managed, published and accessed by information systems and humans alike. This work is carried on with the support of the Bill and Melinda Gates Foundation, and all project outputs are publicly available.

In the approach we propose, statistical classifications are made available in open, standard and machine-readable formats, in order to promote their smooth reuse in third party applications. Given that most large organizations collecting statistical data (or branch in them) are at the same time users of some statistical classifications and custodians of others, a network of organizations adopting this approach will ensure that user access constantly up-to-date information. As a consequence, the duplication of work currently happening to make classifications available in information systems, it is expected to decrease dramatically.

FAO is currently experimenting with this approach in-house, and plans on extending it to a network of interested organizations. At the time of writing, a number of statistical classifications (including mapping and multilingual terminologies) have been made available as RDF resources[1], and tool to support associated functionalities are under development and testing. All project outputs are available through the platform Caliper[2].

[1] These include the sector/purpose codelist of CRS (by OECD-DAC), the UN Central Product Classification (CPC), the International Standard Industry Classification of All Economic Activities (ISIC v4), the Indicative Crop Classification for the Agricultural Census (ICC v1.0 and v1.1), the FAOSTAT Commodity List (FCL).

[2] See http://stats-class.fao.uniroma2.it/caliper/

The technology stack adopted is based on the linked open data style of data publication. It includes RDF, the de-facto standard for publishing data and metadata on the web, to be used as the “master” format from which different formats can be generated (e.g., CSV or JSON). Classification systems, and their entries, are given global identifiers, i.e., URIs (Uniform Resource Identifiers) to make statements on resources uniquely identified on the web. API and tools for editing, querying or provide human-friendly visualizations can consume the RDF and streamline each step in the data life cycle.