ESA/STAT/AC.234/22
11 May 2011

UNITED NATIONS
DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS
STATISTICS DIVISION

Meeting of the Expert Group on International Economic and Social Classifications
New York, 18-20 May 2011

Data formats for exchanging classifications

UNSD

A. Introduction

1. Classifications frequently need to be shared and exchanged with a variety of users – within a statistical organization, to other statistical agencies and to non-statistical agencies and individuals, such as in academia or business. There are many formats in which to exchange classifications that are appropriate for different uses depending on whether the information is intended for human consumption or machine reading and processing. The exchange can also take place between different platforms, which adds to the complexity of the problem since a single format has to be suitable for different users or applications.

2. In the past, classifications have mostly been distributed as printed documents, intended for a human audience. In this case, the format was designed to correspond to human needs and conventions (e.g. for making connections between linked objects) and instructions on how to interpret the printed text were easily built into the document, for instance as part of an introduction. Formats such as portable document format (PDF) and websites (HTML) are essentially a continuation of this approach, using more modern forms of storing and transmitting data, but still using the same visual concepts and instructions intended for a human audience.

3. Many audiences, on the other hand, have a need for automatic retrieval and processing of information about classifications, which is difficult to accomplish with the formats described above. Other format have therefore been chosen (or developed) to allow for an automatic processing of such information.

4. Depending on the needs of the users, different formats have emerged over time. The most notable difference in the content provided is the scope of the information included in such a "machine-readable" document.[1] Basic classification documents often include only a list of codes and the corresponding title of the category. These are often used if only a labeling of data presented according to this classification is required. More extensive documents also include information about the definition of the individual categories, such as explanatory notes, possibly grouped into inclusions and exclusions or similar categories.

5. Examples for these two formats are the ubiquitous TXT, CSV or EXCEL files provided for many classifications. Classifications on most websites are available in such formats.

6. While these formats have gained wide acceptance (possibly also because of their ease of use and link to popular software, allowing for easy manipulation), they still have

[1] While also PDF and HTML formats are strictly speaking machine-readable, they are not suitable for easy extraction of information due to the large amount of styling information embedded in such documents. The term "machine-readable" is therefore often used with the meaning of "not designed for human reading".

one strong disadvantage – they require additional knowledge about the classification to be used properly.

7. To understand this, one has to remember that a classification is not just a collection of categories with properties like "code", "title", "explanatory note", but that a classification is also defined through the relationships between the different categories, the structure of the classification. It is the latter that is typically not discernable from these files, unless the user has additional knowledge about the classification. In TXT or Excel files, the structure is sometimes implied by the order in which categories are listed in the file, but that is not a concept a machine may understand.

8. Of course, each machine-readable format needs to be understood and translated by a machine, but the format also needs to be able to convey all the information necessary for the machine to do so. The formats listed above are not sufficient for this purpose.

9. UNSD has been distributing classifications also in the form of an Access database, where an additional table describes the relationships within the classifications. Still, this requires a human making the right connections and instructing the machine to read the information appropriately.

10. Other formats have been developed that allow storing information about relationships and lending themselves to transmitting information about classifications. Many of these are variations of extensible markup language (XML). Some classifications are available in an XML format, but the content is often still based on the flat TXT format, without using the full capabilities of the XML format.

B. New options for suitable file formats

Statistical Data and Metadata eXchange (SDMX)

11. The SDMX standard has been created with the purpose of developing "more efficient processes for the exchange and sharing of statistical data and metadata" among institutions.[2] The United Nations Statistical Commission recognized SDMX as the "preferred standard for exchange and sharing of data and metadata in the global statistical community."[3] SDMX version 2 has introduced a number of new concepts that may be used as a format for classification exchange.

12. SDMX defines metadata that can exist independently of data and may be exchanged on its own. SDMX provides a complex information model of metadata concepts made up of objects that allow information about classification to be expressed. Of the many objects in the information model, *category scheme* and *category* may be particularly

[2] "SDMX User Guide, Version 2009.1". January 2009. Available from http://sdmx.org/?page_id=38
[3] United Nations. "Statistical Commission: Report on the thirty-ninth session." February 2008. Available from http://unstats.un.org/unsd/statcom/sc2008.htm

relevant for classifications. A *category scheme* is made up of a hierarchy of *categories*, and *categories* are a generic term for classification codes at any level of a hierarchy.

13. The *category scheme* and *category* objects are only one method that may be useful to code classification into SDMX. The SDMX format is rich and *category schemes* do not exhaust the possible coding formats for classifications; other innovative methods using SDMX may be created. An advantage of using SDMX for classification exchange is that it has been created with an explicit focus on the exchange of metadata concepts. However, examples of SDMX being used for the production and exchange of classifications are not readily available and it is not clear exactly how to concretely transform classifications into SDMX and how to exchange classifications in the SDMX format.[4]

Resource Description Framework (RDF)

14. RDF is an XML based format that is intended to represent metadata about a resource (in our case, classifications). RDF can be used to represent information in a hierarchical manner and is conceptually represented by a node-network graph where two nodes (a subject node and an object node) are connected by a predicate (an arc) directed from the subject node to the object node.[5]

15. Simple Knowledge Organization System (SKOS) is an RDF based format designed specifically for the exchange of organization systems such as taxonomies and classifications. SKOS represents concepts that are linked hierarchically using the properties "broader" and "narrower". The SKOS format has been used by classification custodians such as Physics and Astronomy Classification Scheme.[6] SKOS documentation and use cases are maintained by the World Wide Web Consortium (W3C).[7]

16. RDF, and in particular SKOS, can be used to code classifications and exchange classifications. An advantage of RDF is that examples of classifications in this format exist and lessons learned can be gleamed from these examples. However, examples of RDF being used by statistical organizations need to be reviewed to find best practices for classification exchange.

Web ontology language (OWL)

17. OWL[8] is another file format that can be interpreted by programs like Protégé, which is a free, open-source platform that provides a growing user community with a suite of tools to construct domain models and knowledge-based applications with ontologies.

---

[4] Fitzpatrick, Bryan. "End-to-End Management of the Statistical Process." Presentation given at the "Work Session on Statistical Metadata," UNECE March 2010. Available at http://www.unece.org/stats/documents/ece/ces/ge.40/2010/wp.10.e.ppt

[5] More information on RDF can be found in "RDF Primer" available from http://www.w3.org/TR/2004/REC-rdf-primer-20040210/

[6] See http://www.w3.org/2001/sw/Europe/reports/thes/8.5/draft01.html

[7] See http://www.w3.org/2004/02/skos/ for more information.

[8] http://www.w3.org/2001/sw/WebOnt/

Since ontologies describe the concepts and relationships that are important in a particular domain, they are useful tools for conceptualizing classifications. In addition, these tools are commonly used in the academic community and providing classifications in a suitable format may broaden the group f users (and later contributors) of the classification.

C. Questions

18. The Expert Group is invited to provide feedback on the following questions:

    a) What experiences exist with data formats for the exchange of classifications? Which data formats are being used?

    b) Have you received requests for classifications to be made available in SDMX, RDF, OWL or similar XML-based formats? If so, are there specific user groups?

    c) Can recommendations for best practices be made?