I Have A Dream…

Will linked open data stir up the way we work with statistical classifications?

Danny Delcambre; Agne Bikauskaite; Jan Planovsky (internal reviewer)

Statistical Office of the European Union (Eurostat), Luxembourg
Danny.Delcambre@ec.europa.eu; Agne.Bikauskaite@ext.ec.europa.eu;
Jan.Planovsky@ec.europa.eu

Abstract. Modernizing the statistical industry in the area of classifications means making more efficient use of existing information resources, federating scattered human expertise into collaborative networks, and implementing new tools and technologies. In this context linked open data (LOD) could make the dream of generations of classification experts come true, i.e. actively contribute to the creation of a powerful tool for classifying goods, services and activities in classifications by making available a global and multilingual alphabetical index as well as a very efficient tool for navigating across correspondence tables. The input resources for developing the application sketched in this paper already exist; these are an integrated system of international statistical classifications, a set of classification repositories and well established international expert groups. This paper also contains a side-proposal about production chains supporting the delineation of classification categories. It is also expected that linked open data will make it possible to set up a decentralized but yet fully harmonized collaborative network of classification experts. Such a new way of working would drastically improve the everyday life of classification users and be a blessing to all those who have to deal with statistical classifications. The objective of this paper is to sketch this ideal world. How this can be implemented in practice is not as such the subject of this paper; however it is expected that it will stimulate international discussion on the implementation of this vision and generate constructive feedback.

Keywords. Statistical classifications, correspondence tables, linked open data (LOD), resources description framework (RDF), interlinked assets, harmonization, global initiative, collaborative networks, semantic assets, modernization of official statistics.

1 Introduction

Statistical classifications are the foundations on which all statistical systems, be they national or international, are built. Thus, it is essential that statisticians have access to extensive, reliable and regularly updated sources of information on classification systems for use with data such as industrial, foreign trade or occupational data. This information is also of the upmost importance to data users for properly understanding the content of data sets made publicly available by statistical organizations.

Since the first international statistical classifications were established in the forties, much progress has been achieved in the convergence of classifications at global, regional and national levels. The numerous rounds of discussions conducted by international organizations culminated in the nineties with the implementation of an integrated system of international classifications where all product[1] and economic activity classifications are interlinked, either closely or loosely.

Following the implementation of this integrated system, the need for collaboration between international organizations and within their constituencies increased significantly. Various international classification expert groups are now active and trying to yet improve the convergence between classifications. Still the way national and international statistical organizations interact with each other is still very much of a stovepipe-like nature, meaning that assistance tools for users continue to be developed nationally with few efforts to coordinate with others.

In the late nineties, informatics stirred up drastically the working habits of classification experts by offering improved capacities for creating classifications and interlinking them via correspondence tables using the power of relational databases. Internet was the next revolution that offered a unique opportunity to make easily and publicly available various assets relating to classifications such as classifications structures, explanatory notes, correspondence tables, methodological material.

[1] It is important to note that the concept of "product" covers both goods and services. When a specific distinction has to be made, the concepts of "goods" and "services" will be used.

In this framework, correspondence tables play a central role as they provide bridges between classifications, thus allowing statisticians to rearrange data based on a classification system into data based on another classification system.

The importance of these correspondence tables has been recognized for a long time but due to various limitations it proved impossible to develop tools capable of accommodating in one place the vast amount of information scattered across various organizations.

In this context linked open data can be seen as the perfect catalyst to create a strong collaborative network of classification expert groups where classification assets produced by various actors and made available e.g. in RDF format will be reused by applications capable of putting in correspondence all these assets, hence adding value to the individual pieces of information.

One could also exploit the power of collaborative networks for improving the classification of objects into classifications by creating multilingual glossaries of terms that would be linked to some pivot classification(s), thus providing a gateway to all the classifications in the integrated system of classifications.

One last proposal consists in rearranging the concepts in the glossaries to describe production chains that in turn will help classification experts in delineating the categories of classifications with a view to future revisions.

After presenting the general framework in which this proposal fits (Section 2), this paper will describe the present situation with regard to classifications: presentation of the integrated system of international statistical classifications put in place over the last decades by the main international organizations (Section 3), description of the main repositories for classification resources made available by these organizations (Section 4), and introduction to the main international expert groups on classifications whose task it is to manage the integrated system of international classifications and improve its convergence whenever possible (Section 5). The remaining sections will show how linked open data could be used to feed tools that will make a much more efficient use of the existing material on classifications, and trigger the creation of a large collaborative network potentially federating the global classification community for the benefit of all those who have to deal every day with statistical classifications.

The feasibility of using RDF for representing classifications and correspondence tables has been analyzed in detail by Franck Cotton et al. [7]. This aspect will therefore not be re-considered here.

2 Potential links with other initiatives on linked open data

The approach developed in this paper fits well in the general framework of two major international initiatives in the area of linked open data and could potentially serve for the description of use cases. These two initiatives are:

- the UNECE[2] "Implementing Modernstats Standards" initiative which is the subject of a major international collaboration project under the UNECE High-Level Group for the Modernisation of Official Statistics[3]
- The DIGICOM project[4] of the European Commission

One of the deliverables of the "Implementing Modernstats Standards" project is to provide a central repository of key standards in the form of "linked open metadata", thus reducing the need for statistical organizations to develop such a resource individually (Work Package 1: Build a dissemination system for core structural metadata) [5].

The DIGICOM project's goal is to create new, innovative dissemination products, tools and services for ESS[5] statistics. The idea is that Eurostat and the National Statistical Institutes across Europe work together to develop solutions in four fields, or four "work packages". These main areas of work are: a) innovative user interaction; b) modern visualization tools; c) easy access to data, (linked) open data; d) communication and promotion [6].

Useful expertise with RDF could also be sought from some national initiatives such as the classification management system recently put in place by Statistics New Zealand and called "Aria"[6].

---

2 United Nations Economic Commission for Europe.

3 The mission of the UNECE High-Level Group for the Modernisation of Official Statistics (HLG-MOS) is to oversee the development of frameworks, tools and methods, to support modernisation in statistical organizations. The aim is to improve the efficiency of statistical production, and the ability to produce outputs that better meet user needs.

4 DIGICOM stands for Digital communication, User analytics and Innovative products.

5 European Statistical System, i.e. the 28 Member States of the European Union, EFTA countries and EU candidate countries. http://ec.europa.eu/eurostat/web/ess/

6 Statistics New Zealand - Tatauranga Aotearoa: www.stats.govt.nz/

3 Integrated system of international statistical classifications

3.1 Before the nineties

The United Nations (UN) Statistical Office (now called United Nations Statistics Division, UNSD), established in 1947, and the statistical service of the European Economic Community (now called Eurostat), established at the end of 1952, both produced various classifications, the most important ones being the economic activity classifications: International Standard Industrial Classification of All Economic Activities (ISIC) for the United Nations and General Industrial Classification of Economic Activities within the European Communities (NACE) for the European Union.

The UN International Standard Industrial Classification of All Economic Activities (ISIC), first published in 1948, and later revised in 1958 and 1968, was recommended for use by all the Member Governments either by adopting the classification as a national standard or rearranging their statistical data in accordance with this system for purposes of international comparability [2]. However, apart from a few exceptions, there was no alignment between the national classifications and the UN standard. Data compiled on the basis of the national classifications had to be heavily reconstructed to fit the international standard.

When drafting the first version of NACE, Eurostat made the decision to take as starting points the classifications already in force in the Member States and also the United Nations classification ISIC.

Classifications were made available only as hardcopies. The only classification "repositories" were the libraries in the statistical offices.

To cut a long story short, harmonization between classifications was embryonic and informatics meant basically playing with punch cards…

3.2 The integrated System of International Statistical Classifications

In the eighties various rounds of discussions were conducted at international level to explore possible ways of improving the consistency between the various classifications used at global, regional and national levels.

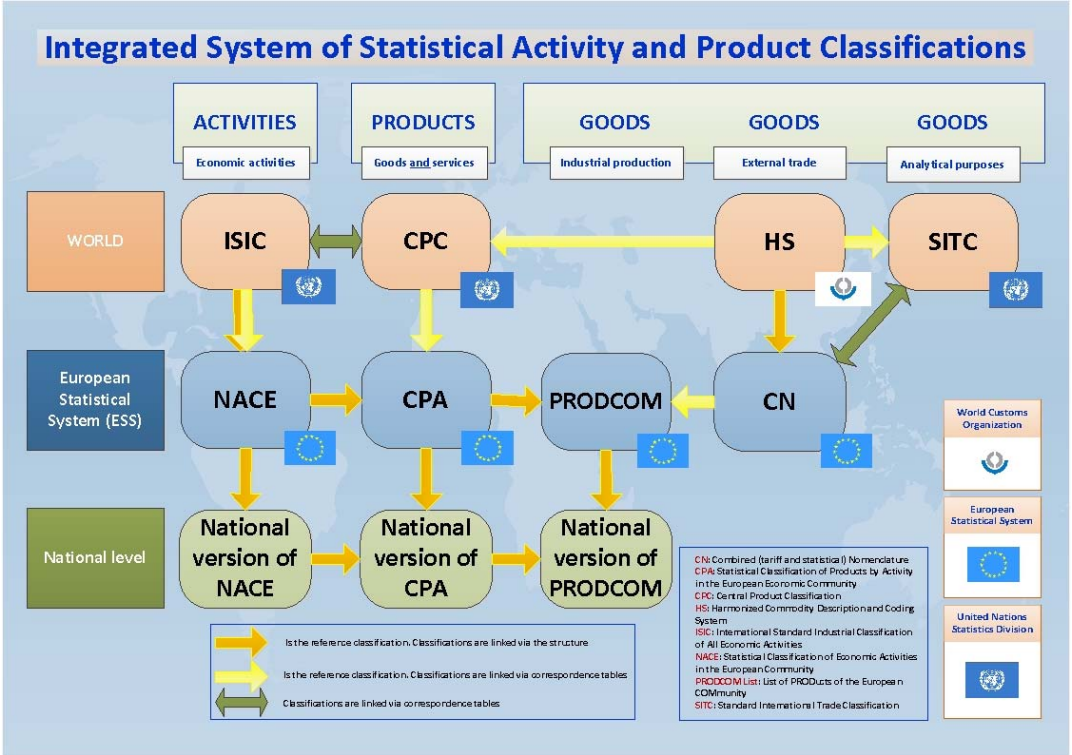

Fig. 1. Integrated System of Activity and Product Classifications[7]

[7] For a view of this schema in higher resolution: http://ec.europa.eu/eurostat/ramon/miscellaneous/Integrated_system_of_international_classifications.jpg

The outcome of this exercise was a thorough revision of the international statistical classifications, with the result that the new classifications were developed as an integrated system of statistical classifications (See Figure 1 below), whereby:

- a) the various product classifications have been harmonized and
- b) the central product classifications have been related to the classifications of economic activities by the economic origin criterion.

In addition, the European Union's classifications have been harmonized with global classifications. This also applies to the national classifications of the EU Member States [3].

The table shown below is well known to anyone involved in classifications; it was drafted in the nineties and gives a perfect representation of the close interlinking between classifications at global, regional and national levels. However it should be noted that since this synthetic table was drafted, additional classifications have been created or revised that qualify for inclusion in this table (e.g. Classification by Broad Economic Categories [BEC], Classification of Individual Consumption by Purpose [COICOP] elaborated by the United Nations; Main Industrial Groupings [MIGs], Standard Goods Classification for Transport Statistics [NST], Trade, installation, maintenance, repair and rental described by CPA categories, elaborated by Eurostat). Furthermore this representation is focused on the European Union case but the same principle also applies to other regional classification systems such as the Australian and New Zealand standard classifications which are also linked to the UN standards.

The representation above is a horizontal representation, i.e. it describes the situation at one specific point in time. However, statistical classifications are revised at regular intervals, meaning that there is also a vertical dimension in the system of integrated classifications, as shown in the figure below. It should however be noted that correspondence tables between the various temporal dimensions are developed by the classification custodians. In short, this means that navigation across classifications is possible both horizontally and vertically. This certainly adds a level of complexity to the proposed application but will in no way prove to be an insurmountable obstacle.

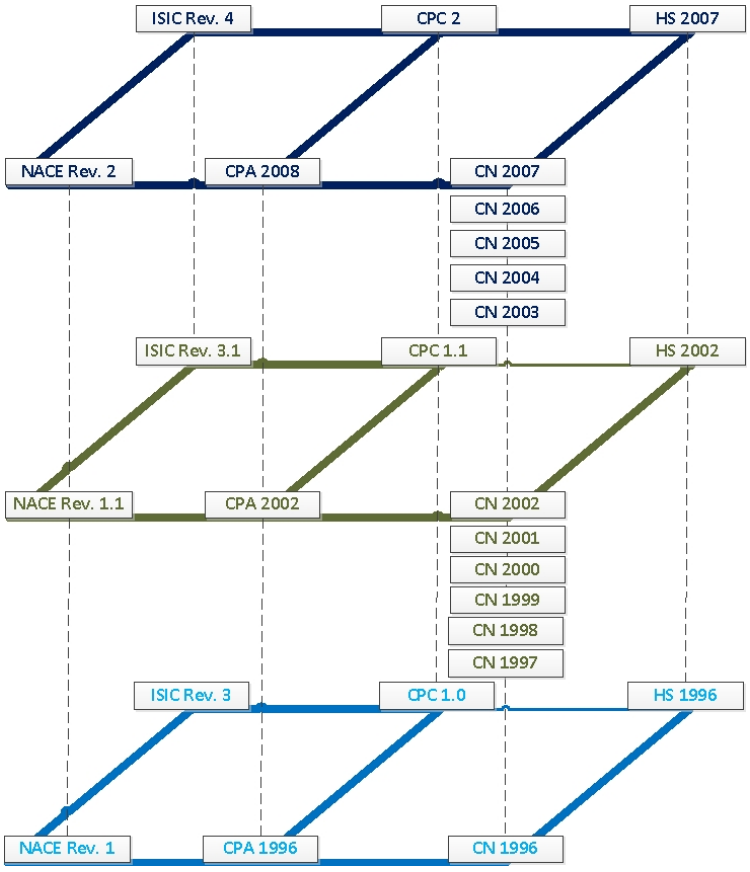

Fig. 2. Vertical dimension of the Integrated System of Statistical Classifications (partial view)

4 Repositories for classification assets

The first repositories for statistical classifications were created at the end of the nineties with the objective of making available to the general public the information available on classifications. Compared to the past where only hardcopies were made available, the creation of these repositories was a tremendous improvement.

The main such repositories are the Classifications Registry[8] maintained by the United Nations and the RAMON[9] server maintained by Eurostat. While the UN registry is limited to UN assets, the RAMON server has a much broader scope, encompassing all international statistical classifications. Both registries also make available numerous correspondence tables as well as general information or methodological information relating to classifications.

The files offered for download are available in various formats: txt, xls, mdb, html, doc, xml, csv. Until recently feedback received from users showed that they were quite satisfied with the range of formats offered. However, over the last months, the number of requests for additional formats such as RDF has increased significantly, thus showing an interest from users to get more value added from the assets made available, especially in the area of correspondence tables where users would like to take advantage of the interlinking of correspondence tables to link classifications for which there exist no direct correspondence tables. There is also a clear demand from the academic world for reference linked statistical metadata (concepts, code lists, etc.): this was strongly expressed at a recent Semantic Statistics (SemStats) workshop [5].

Attempts have been done of course in the past to create such databases of correspondence tables and this raised no major issue from a technical point of view as correspondence tables are very strictly organized set of data that can be dealt with very easily using relational databases. The problem was that such resources had to be developed individually due to a lack of collaboration culture between international organizations. Nowadays the collaboration between international organizations has improved a lot and there is a strong will from all stakeholders to actively collaborate towards a common goal of a modernized statistical industry. Still the development of common tools or websites is generally hampered by financial resource problems as well as by internal administrative problems within the various organizations.

5 International Expert Groups on Classifications

The two main international expert groups for statistical classifications are the UN Expert Group on International Statistical Classifications and the Eurostat Classifications Working Group.

The Expert Group on International Statistical Classifications was established to ensure harmonization and convergence among the classifications in the International Family of International Statistical Classifications[10]. It examines the status of the work on international classifications, makes recommendations concerning future directions to the United Nations Statistical Commission, and serves as the central coordinating body in the work on international classifications. It agrees on strategies for updating and revising classifications and reviews the underlying principles as well as practical proposals to bring about convergence of existing classifications. It is composed of members from international organizations - custodians and major users of international classifications -, as well as representatives from developed and developing countries and regional agencies [4].

If needed the Expert Group can be assisted by Technical Subgroups to work on specific issues (e.g. technical groups for ISIC and COICOP created in 2013).

The Eurostat Classifications Working Group has broadly the same tasks as its UN counterpart but for a geographic coverage limited to the ESS. It is made of one representative from each ESS country. The Working Group may also set up *ad hoc* task forces to address specific issues (e.g. on the identification of factoryless goods producers, FGPs).

When the first expert groups met in the fifties and sixties, the national interests superseded the international ones. But progressively the national experts realized that there was more value added in union than in division. And this change of attitude resulted among other things in the convergence of the main international statistical

---

[8] http://unstats.un.org/unsd/cr/registry/

[9] http://ec.europa.eu/eurostat/ramon/

[10] The international family of economic and social classifications is comprised of reference classifications that have been registered into the United Nations Inventory of Classifications, reviewed and approved as guidelines by the United Nations Statistical Commission or other competent intergovernmental board on such matters as economics, demographics, labour, health, education, social welfare, geography, environment and tourism. It also includes those classifications on similar subjects that are registered into the Inventory and are derived or related to the reference classifications and are primarily, but not solely, used for regional or national purposes.

classifications in the nineties. Today one can say that there is a strong culture of collaboration towards a common goal which is the production of high-quality statistics in the interest of the general user community and the decision-makers.

This is yet another asset on which LOD could rely.

6 What value added can linked open data bring?

In the previous sections we have seen that the main resources about classifications are the following:

- an integrated system of national and international classifications
- a network of experts ready to take part in collaborative networks
- repositories making publicly available various types of information about classifications, such as
  - structure of the classifications, consisting of codes, official labels and various derived labels (e.g. short descriptions, self-explanatory descriptions),
  - explanatory notes and similar material (e.g. classification decisions, case laws, rulings made by established international classification expert groups; various databases such as the EU BTI[11] and ECICS[12] databases),
  - alphabetical indexes and dedicated search engines,
  - methodological information (e.g. on the criteria applied to construct a given classification),
  - correspondence tables between classifications.

So, a vast amount of information already exists and extensive use is made of all these resources every day by statisticians from all over the planet. What was so far missing is a magic binder to amalgamate all these resources. Linked open data is more and more emerging as a possible option which will make it possible for IT applications to make full use of this input material.

The following sections are an attempt to describe an ideal world where the best use would be made of all the above-mentioned resources in order to supply people working with classifications with a powerful tool that could evolve into the main building block of a metadata-driven architecture.

[11] The BTI (Binding Tariff Information) database is a tool developed by the European Commission's Directorate-General "Taxation and Customs Union" to assist companies engaged in import/export activities to obtain the correct tariff classification for their goods.
`http://ec.europa.eu/taxation_customs/customs/customs_duties/tariff_aspects/classification_goods/index_en.htm`

[12] The ECICS (European Customs Inventory of Chemical Substances) database is an information tool managed by the European Commission's Directorate General "Taxation and Customs Union" which allows users to: clearly and easily identify chemicals; classify them correctly and easily in the Combined Nomenclature; name them in all EU languages for regulation purposes.
`http://ec.europa.eu/taxation_customs/dds2/ecics/chemicalsubstance_consultation.jsp?Lang=en`

7 High-level schema

The high-level architecture for this proposal is made of three sections: alphabetical index, classifications, and production chains. It can be schematized as follows:

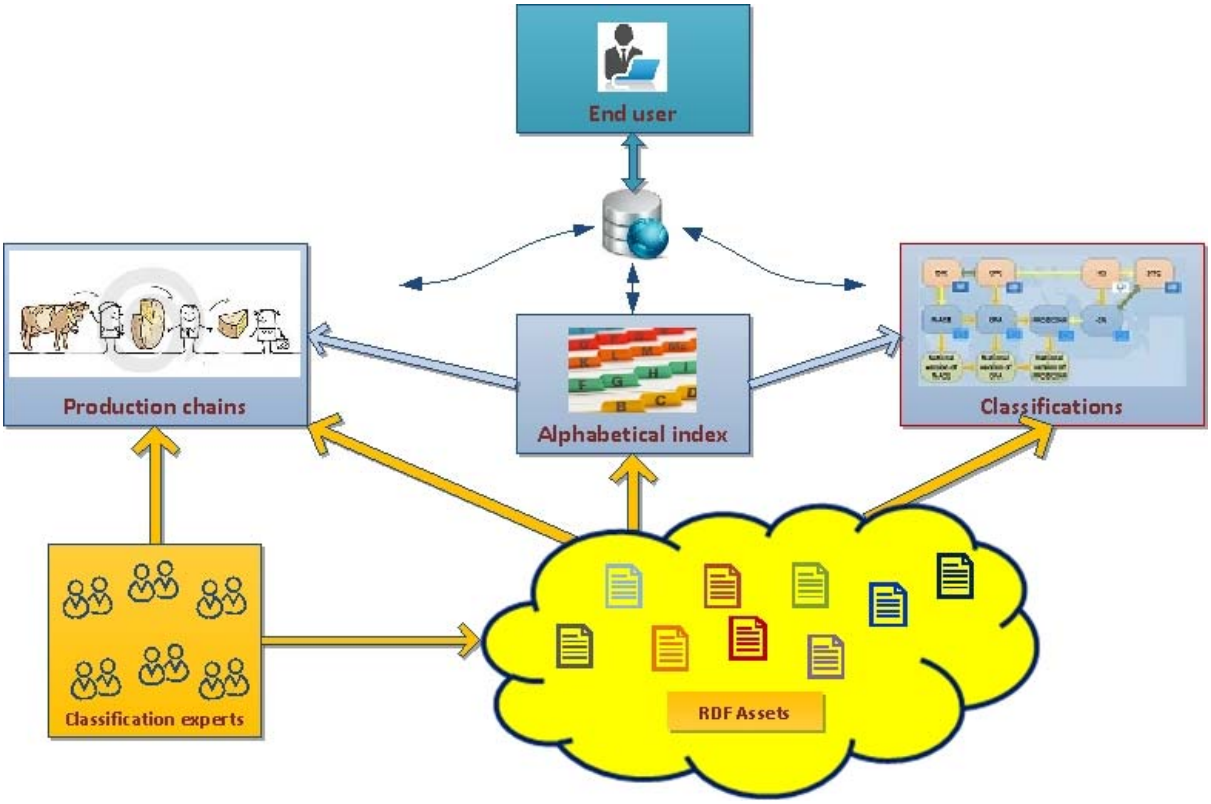

Fig. 2. High-level schema

The pivot will be the alphabetical index which will serve as the starting point for navigating the information. The information displayed under the two other sections will depend on the entry selected from the alphabetical index.

The input for the whole system will be provided by classification experts on the one hand and RDF assets on the other hand.

The following sections will provide more information on the three dimensions of the model.

8 More about the "alphabetical index" section

Basically this section will contain an alphabetical list of activities, goods and services together with one or more codes referring to the classification(s) in which they are mentioned. Reference to at least one classification is mandatory as otherwise the entry would be useless because it would not refer the user to any classification. The original allocation will preferably be in the most disaggregated classification of the integrated system because it is easier and more convenient to go from the very detailed information to the more general.

For populating this section, various information sources can be considered:

- alphabetical indexes (CPC, ISIC, SITC, HS, national indexes),
- automatic extraction of keywords from the classification descriptions using scripts, ad hoc software or existing database indexes such as the Oracle index table of RAMON,
- explanatory notes to Combined Nomenclature and Harmonized System,
- BTI and ECICS databases mentioned earlier; TARIC database[13],

[13] TARIC is a multilingual database developed by the European Commission's Directorate-General "Taxation and Customs Union", covering all measures relating to tariff, commercial and agricultural legislation. It is based on the CN which it supplements with two additional digits.

- classification decisions made by the various working groups or committees managing the classifications which are part of the integrated system of international statistical classifications.

It should be noted that all the information sources mentioned above already exist. The main task will consist in formatting this information in RDF.

This index will be constructed and maintained by a collaborative network of classification experts. The main index will be maintained in English. However the existence of a collaborative network will make it possible for any national statistical office to add translations of the index into its national language(s) in a way similar to the Wikipedia approach. Nowadays international organizations refrain from translating classifications material as much as they can due to lack of resources, so the participation from national organizations would benefit the whole statistical community.

Temporal update of the alphabetical index will be based on traditional correspondence tables between classifications, i.e. the index entries corresponding to a revised category in a classification will be reviewed by the classification experts and accordingly updated if necessary.

9 More about the "Classifications" section

Once a user has selected an entry from the alphabetical index, he/she will be presented with the existing metadata relating to this entry (e.g. short label, explanatory text) and a graph of all links with other classifications, as shown in Figure 3 below.

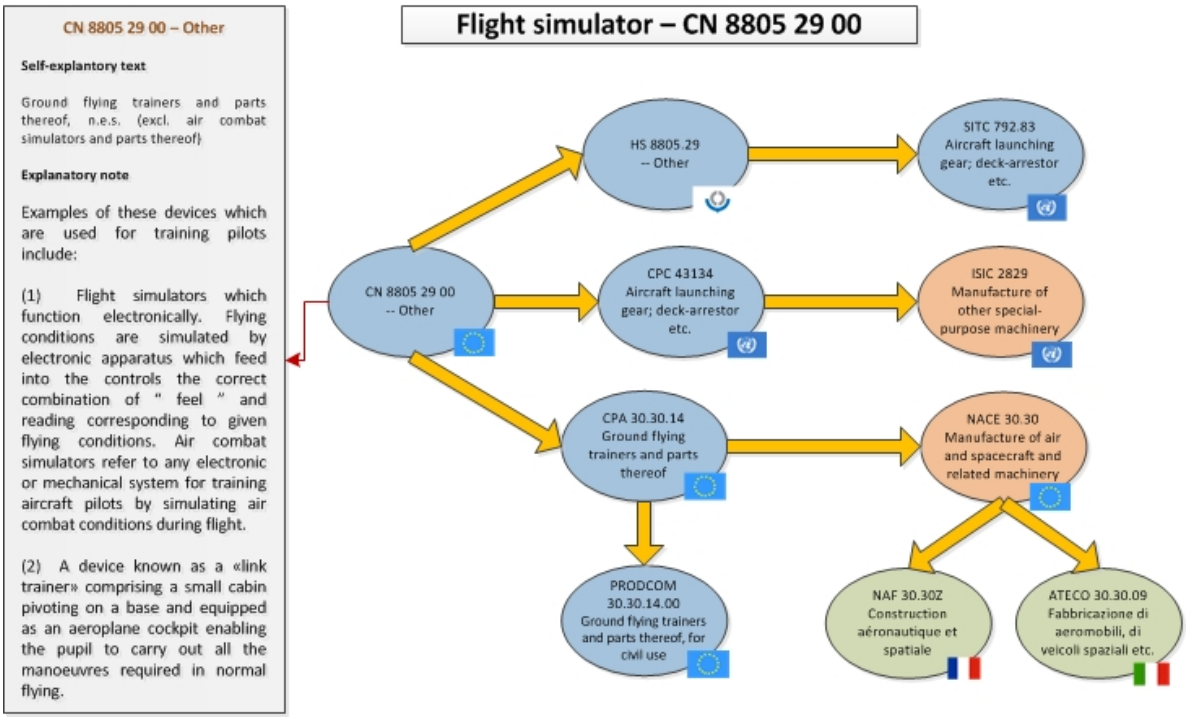

Fig. 3. Classification of flight simulators in some of the classifications which are part of the integrated system

In the example above Combined Nomenclature was chosen as input classification because it is the most detailed classification of the integrated system of international statistical classifications. However this classification has its limits because it classifies only goods and not services. So for services, another classification has to be chosen; the most disaggregated classification for services so far is the EU classification of products CPA.

A more detailed representation of the navigation possibilities offered by such an application is presented at this address in much higher resolution (the image is designed for printing in A0 format): `http://ec.europa.eu/eurostat/ramon/documents/poster.jpg`.

For displaying the corresponding codes in other classifications, the application will make use of the correspondence tables developed by classification experts and made available as RDF assets. These correspondence tables can be of various types: 1-to-1, 1-to-n, n-to-1, n-to-m. It should be noted that all the necessary correspondence tables to make this system work already exist and are made freely available by international organizations (but unfortunately not in RDF format).

The model described here is based on the central role of the alphabetical index as a tool for finding the allocation of an object in a classification. However one could imagine that the user could also click on any classification proposed in the results of the search (e.g. ISIC) and from this code navigate either in the selected classification or switch to all other classifications linked to it. As a visualization example, the figure below shows a graph of existing correspondence tables in RAMON[14].

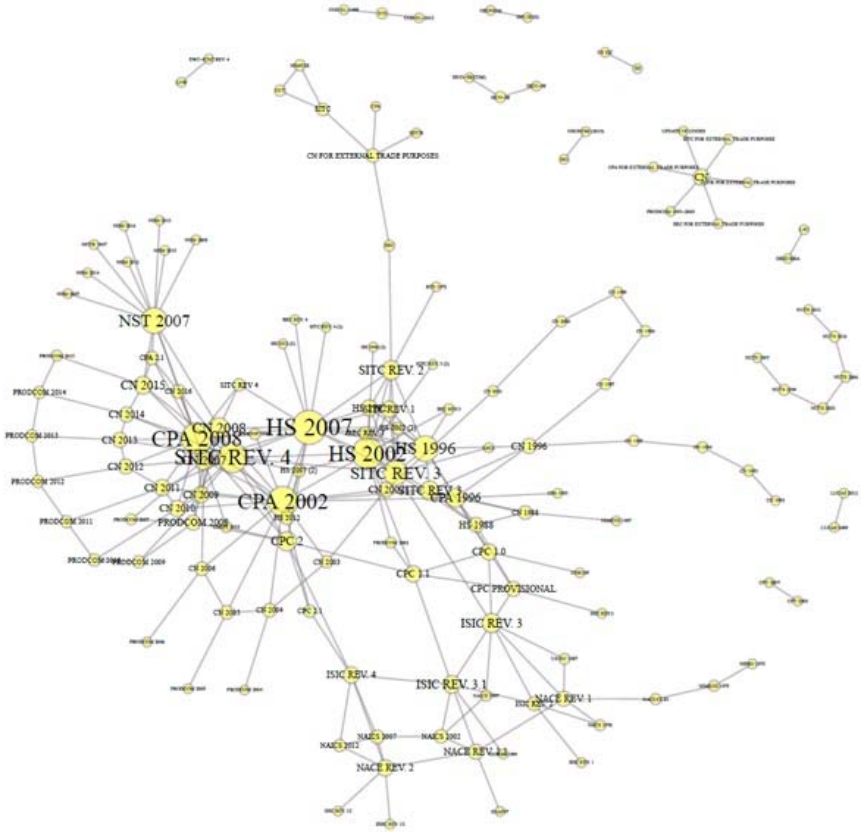

Fig. 4. Graph of existing correspondence tables in RAMON

Graph representations can also help to show relationship density to quickly assess the impact of changes to certain classifications.

But visualization and easy navigation would not be the only uses of this tool; indeed it could also be used to identify mistakes in correspondence tables as in the example below based on ISIC - NACE correspondence table..

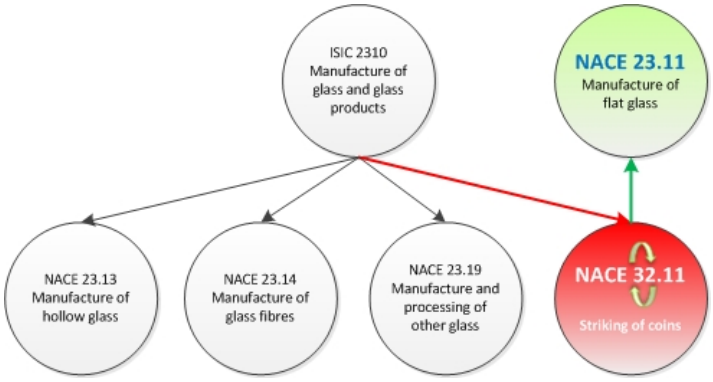

Fig. 5. Identification of a mistake in correspondence table ISIC - NACE

[14] Further examples: General graph showing links between ISIC Rev. 4 and various other classifications (`http://ec.europa.eu/eurostat/ramon/documents/graph_ISIC_general.jpg`); detailed view for one selected code from ISIC Rev. 4 (`http://ec.europa.eu/eurostat/ramon/documents/graph_ISIC_detail.jpg`).

Another very useful tool for classification users (e.g. academic people, researchers) would be to have the possibility to use the power of the interlinkages between classifications to generate ad hoc correspondence tables. Many requests received via the functional address of RAMON are for correspondence tables which are not linked by direct correspondence tables (e.g. NACE and SITC). Taking as a starting point the diagram in Figure 1 one might ask for a correspondence table between NACE and the Italian version of PRODCOM.

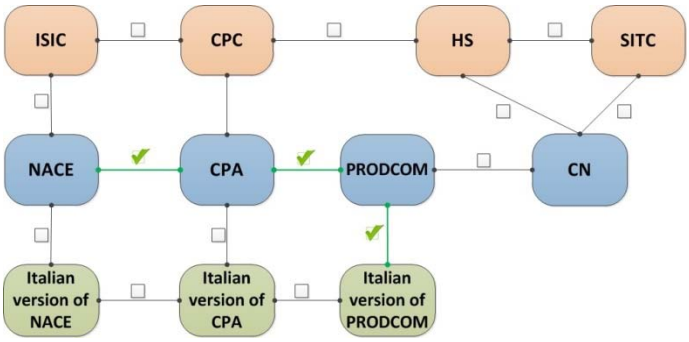

Fig. 6. Selection process of an *ad hoc* correspondence table

The user will select the path he wants to use to go from the source classification to the target classification based on existing official correspondence tables between the classifications which are part of the system. The application will then fetch the requested correspondence tables as RDF files, put them together and manipulate the information (e.g. remove duplicates) so as to provide the user with the requested information. The result of such manipulations will prove a very useful input e.g. for economic analyses of data sets based on classifications of different natures.

10 More about the "Production chains" section

Ideally a classification expert should be a specialist in all domains of economic activity, which of course is not possible. Furthermore the retirement of a classification expert always means a loss of expertise for the whole statistical community. This "Production chains" proposal is an attempt to retain as much as possible of this expertise in the public domain.

The philosophy of this proposal is to try and show, for each physical product listed in the alphabetical index, its production chain, based on the traditional `input` → `process` → `output` model.

Clear definitions of production chains will help delineating precisely the coverage of the categories of a classification by showing similarities in production. Knowing how a product is produced can be of great help to classify it. This project could also be a way of identifying missing entries in the alphabetical index.

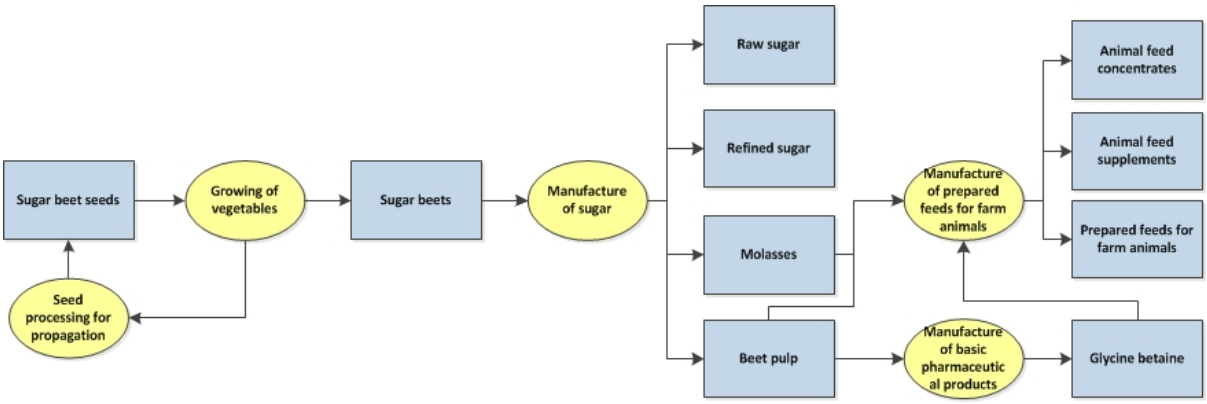

Fig. 7. Partial view of a production chain

The business case identified so far for this "Production chains" section is rather weak and would certainly not justify the huge time investment required for the construction of these chains. Therefore, unless no further and solid business cases can be suggested by the conference or after circulation among the wider statistical community, this sub-project will be abandoned.

11 Conclusion

The aim of this paper is to suggest ways of improving the way the statistical classification community is presently working with classifications and correspondence tables by making use of modern and innovative tools such as linked open data.

The proposal presented in this paper lays the path to the creation of an integrated and collaborative network of classification experts that would make optimal use of various resources such as RDF assets. Such a network would generate substantial benefits for the statistical community at large, especially when considering the central role played by classifications at all stages of the statistical process.

The proposal describes the philosophy of the project without entering into its technicalities. It is based on the long experience of the author with classifications and correspondence tables, and is intended to trigger reactions from external stakeholders as to the technical feasibility of the project and the challenges to address.

Bearing in mind the restrictions imposed on conference papers (novelty vs. limited space), we are aware that it is not possible to enter into the full detail of the proposal (e.g., the possibility offered by RDF to link the classification corpus to other external linked metadata such as thesauri, taxonomies, etc. could have been investigated). On the other hand we hope that our paper could be considered as a starting point intended for gathering constructive feedback from the conference before being submitted to the larger statistical community.

References

1. De Michelis, A., Chantraine, A.: Memoirs of Eurostat - Fifty years serving Europe. Office for Official Publications of the European Communities. Luxembourg (2003). http://ec.europa.eu/eurostat/product?code=KS-49-02-183&mode=view
2. Statistical Office of the United Nations: International Standard Industrial Classification of all Economic Activities. Statistical Papers, series M, No. 4. Lake Success, N.Y. (31 October 1949). http://unstats.un.org/unsd/cr/registry/regdntransfer.asp?f=204
3. RAINER, N.: The Revised System of International Classifications. Eurostat. Luxembourg (1995). http://ec.europa.eu/eurostat/ramon/other_documents/article_norbert_rainer/en_word.zip
4. United Nations Statistics Division. Methods and Classifications. http://unstats.un.org/unsd/methods.htm, accessed: 2016-06-12
5. UNECE. HLG Project: Implementing ModernStats Standards; Towards a unified implementation and global integration of HLG-MOS and other global models and standards, and providing a Roadmap for their implementation. Geneva (2016). http://www1.unece.org/stat/platform/download/attachments/120128748/HLG%20project%20proposal%20on%20implementing%20modernstats%20standards.pdf?version=1&modificationDate=1452783015060&api=v2
6. Eurostat. Business Case; Digital communication, User analytics and Innovative products (ESS.VIP DIGICOM). Version: 1.0.0. Luxembourg (22 October 2015). https://ec.europa.eu/eurostat/cros/system/files/DIGICOM-BC-v1.0.0.pdf
7. COTTON, F. et al.: XKOS. An SKOS extension for representing statistical classifications (28 May 2014). http://rdf-vocabulary.ddialliance.org/xkos.html, accessed: 2016-06-21