Publishing Official Classifications in Linked Open Data

Giorgia Lodi[1], Antonio Maccioni[1]
Monica Scannapieco[2], Mauro Scanu[2], Laura Tosco[2]

[1] Agenzia per l'Italia Digitale
(giorgia.lodi,antonio.maccioni)@agid.gov.it
[2] Istituto Nazionale di Statistica
(scanu,scannapi,tosco)@istat.it

Abstract. Data interoperability is well recognized as a basic step for developing integrated services supporting inter-organizations communication. The issue of ensuring data interoperability has been tackled by many different communities in order to address various problems. In particular, the (over-)national institutes of statistics deeply concern the issuing of official and shared classifications (i.e., taxonomies, schemes, code-lists) to be used in the jurisdiction of reference. On a different perspective, there has been much work from the Web data management community to publish data on the Web in an interoperable way. The efforts have converged on a series of standards and practices gathered under the Semantic Web stack. Clearly, the two mentioned scenarios are complementary as they can benefit one to another. To this concern, the Italian Institute of Statistics (Istat) and the Agency for Digital Italy (AgID) have launched an initiative aiming at producing official classifications under the form of Linked Open Data to be published in the Web of data using standard ontologies. The paper describes and motivates this initiative.

1 Introduction

In the Official Statistical (OS) domain, the issue of data interoperability has been present since decades: both National and International exchanges of data resulting from statistical processes are made possible only by adopting common metadata models and formats.

From this point of view, a specific effort has been performed in such a domain to standardize classifications, and hence to introduce the concept of "official" classification, that is one conform with internationally accepted standards. Indeed, most of the exchanged data in OS are multidimensional data represented as measures and related dimensions. These latter are coded according to specific classifications, and hence official ones have played a significant role in data exchange processes among National Statistical Institutes (NSIs).

The Linked Data initiative [7] is more and more affirming as the principal mean for data interoperability, by permitting to create and interlink arbitrary volumes of structured data across the Web. In particular, the Linked Data initiative is made possible by the widespread adoption of Web standards for publishing data according to the Resource Description Framework (RDF) model. RDF allows to uniquely identifying resources on the Web, by means of a specific URI (Uniform Resource Identifier). This feature has several advantages, including (i) the possibility of a direct access to resources via a query language and (ii) the ability to link data together in order to access them in an integrated way (with the clear positive side-effect of higher quality, more information more easily accessed, and so on).

The Linked Open Data (LOD) project is concerned with the publication of Linked Data that are "Open". There are several LOD datasets already available. The so-called LOD cloud [8] covers more than an estimated 50 billion facts from many different domains like geography, media, biology, chemistry, economy, energy, etc.

The LOD project has had also an immediate and widespread success in the e-government sector: several public administrations (PAs) and institutions are starting publishing their data as LOD. To this end, in the specific Italian context, at the end of 2012, the Agency for Digital Italy (AgID) published national guidelines [11] that paved the way to the use of LOD as the data paradigm for enabling semantic interoperability in the collaboration between PAs. Since then, AgID continued to exercise its role of national Public Sector Information (PSI) enabler by annually releasing a number of strategic documents for PAs. One of these documents is the so-called national agenda which includes principles (e.g., interoperability, usability, accessibility, data quality) and objectives to be achieved by PAs within a year in order to implement, and sustain in the long term, the PSI enhancing process. Following the G8 open data charter experience [3], the agenda introduces the concept of key datasets to be released as high quality open data. Among the key datasets, AgID identified "official" classifications as cross-domain data to be published in LOD so that to foster an effective integration between even heterogeneous data.

In view of this scenario, Istat, the Italian National Institute of Statistics, and AgID launched a joint project whose objectives can be summarised as follows:

- to model "official" classifications such as Ateco 2007 (Classification of Economic Activity) and COFOG (Classification of the Functions of Government) in LOD using standard ontologies (e.g., SKOS - Simple Knowledge Organization System, XKOS - eXtended Knowledge Organization System, ADMS - Asset Description Metadata Schema, etc);
- to certify data by provenance using the PROV framework. In particular, the framework has been used to document the overall process of the creation of the classifications by Istat, as well as of the creation of their LOD versions and of their publication on the SPARQL endpoint by AgID.

The project also helped AgID and Istat to delineate guidelines and artefacts for LOD publication that could raise the awareness on the need to certify data quality and reliability, thus enabling an effective data reuse and interoperability in the Italian PSI context.

Similar projects have been carried out by other institutions, such as the publication of official classifications by INSEE [5], or the publication of the NACE official classification by the European Community [9].

The paper describes this project and the data model that has been adopted and implemented in LOD for the specific Ateco 2007 standardized classification. The remaining of the paper is structured as follows. Sections 2 and 3 provide the background on Ateco 2007 and PROV, respectively. Section 4 introduces the data modelling in LOD of the Ateco classification and Section 5 concludes the paper.

2 Background: Official Classifications and ATECO 2007

Classifications have been one of the first metadata set that NSIs started to store, with the objectives of (i) reusing them in different production processes and (ii) promoting harmonization. The main result was the Neuchâtel Group that issued version 2.1 of the Neuchâtel Terminology Model Classification [10] database object types and their attributes in 2004. The principal purpose of the work was to arrive at a common language and perception of the structure of classifications as well as of the links between them.

More recently, the Generic Statistical Information Model (GSIM) [4] was proposed with the objective to describe the information objects and flows in the statistical business process: in the Concepts module, classifications are described in detail. Some of the elements characterizing GSIM classifications are:

- A *Classification Family* is a group of *Classification Series*. Classification Series is an ensemble of one or more *Statistical Classifications*.
- A *Statistical Classification* is a set of *Categories*. The Categories at each *Level* of the classification structure must be mutually exclusive and jointly exhaustive of all objects/units in the population of interest.
- A *Statistical Classification* has *Categories* that are represented by *Classification Items*. These *Classification Items* are organised into *Levels* determined by the hierarchy. A *Level* is a set of *Concepts* that are mutually exclusive and jointly exhaustive.
- *Statistical Classifications* can be *versions* or *variants*. A variant type of *Statistical Classification* is based on a version type of *Statistical Classification*. In a variant the *Categories* of the version may be split, aggregated or regrouped to provide additions or alternatives to the standard order and structure of the original *Statistical Classification*.

Figure 1 shows a part of the GSIM model schema representing the above mentioned concepts.

The Classificazione delle ATtività ECOnomiche (ATECO) [1] is the Italian counterpart of the NACE [9] classification. Its series is composed of the different versions: 2002, 2007, and so on every time there is a new definition. Each of these versions is organized into levels. As far as the ATECO-version 2007 is concerned, its levels are: *sezioni*, *divisioni*, *gruppi*, *classi*, *categorie* and *sottocategorie*. Up

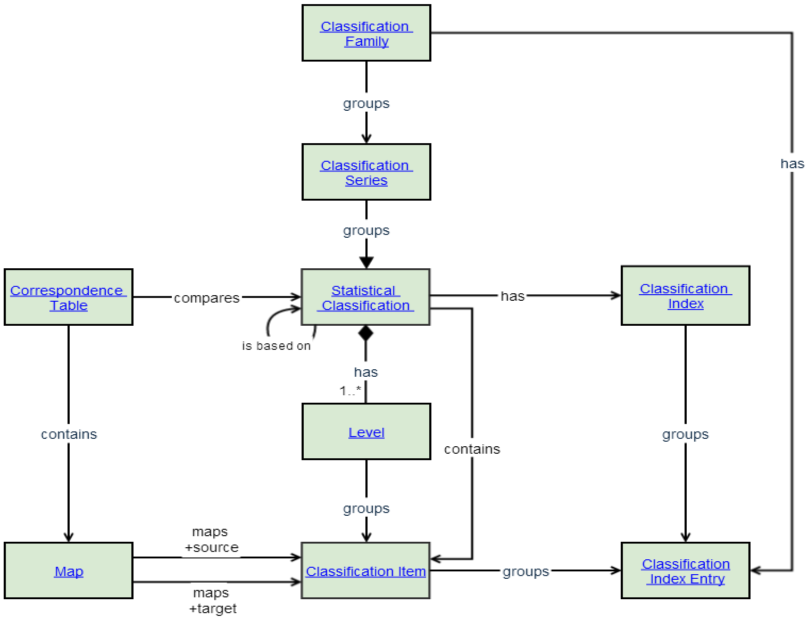

Fig. 1. GSIM Model Schema

to the level *classe*, it coincides with the NACE classification, which is a modification of the ISIC Rev. 4 [6] managed by the United Nations. Variants of these classifications managed in the Istat Sistema Unitario dei Metadati (SUM) are those used for specific purposes in Istat with the following distinctions:

- Variants organized by specific Istat systems (e.g. the system devoted to dissemination, the one devoted to data collection and so on). Each of these variants include additional categories, as the codes in the Main Industrial Groupings.
- Variants used in specific data structures (either of macro or micro data) in order to show the level of detail in which the dimension Economic Activity is given. For instance some disseminated data use all the codes of the first ATECO level, another a subgroup of them and some Main Industrial Groupings, another one focuses on one or two codes of the first ATECO level and then decomposes it up to the the last ATECO level.

3 Background: PROV and XKOS Ontologies

As mentioned, standardizing a classification is a complex process that involves different actors: the PROV ontology described in Section 3.1 helps tracking this aspect.

Moreover, statistical classifications are a specific type of classifications that needs to be modeled in an ad-hoc way. This is described by the focus on the XKOS ontology in Section 3.2.

3.1 PROV Ontology

Provenance information is relevant to certify data quality and reliability. In more detail, it is very important to publish, together with data, who is the responsible for them and which are the entities, activities and agents involved in the generation/manipulation processes; namely, the following concepts:

- *Responsible of the data*: person/institution/administration that manages/creates/manipulates the data.
- *Certified data*: i.e. data published by their responsible.
- *Provenance*: set of detailed information regarding entities and processes involved in the production and publication of data.

We tested the PROV Ontology [12], a W3C recommendation, to certify the role of Istat as official producer of ATECO 2007 classification published as linked open data.

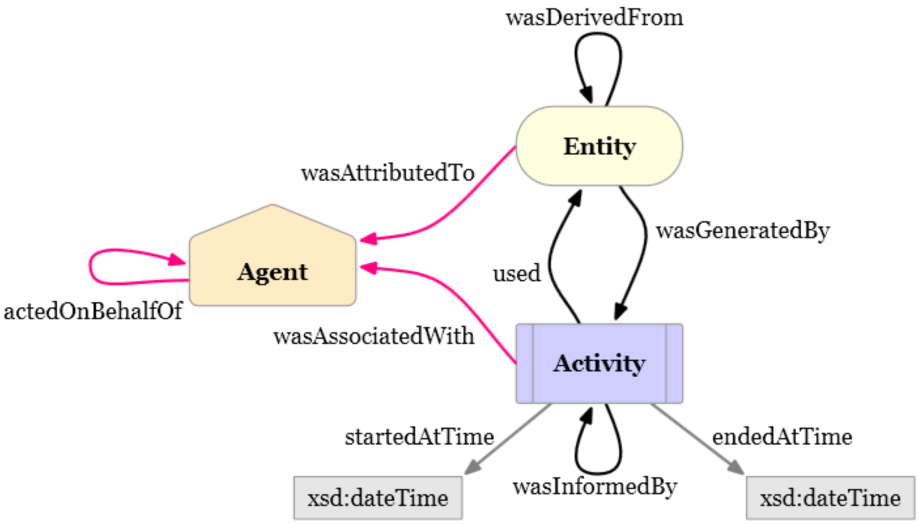

Fig. 2. Provenance Ontology Model Schema

In Figure 2, the principal elements of the PROV data model are shown. The PROV Ontology (PROV-O) is an OWL2 ontology that expresses the PROV Data Model (PROV-DM), which provides a set of classes, properties, and restrictions to represent provenance information. A PROV framework is available and is composed by a set of documents describing different aspects of the provenance issue. In detail, the framework consists of the following documents:

- *PROV-DM*: describes the data model; that is, entity, activity and agent concepts.
- *PROV-O* (PROV-Ontology): describes the data model using the OWL2 language.
- *PROV-N*: describes the data model using the human-readable notation N3.
- *PROV-XML* describes the data model using the XML notation.
- *PROV-Constraints*: describes the integrity constraint for writing correctly the provenance information using the PROV ontology.
- *PROV-Sem*: describes the PROV data model semantic.
- *PROV-Dictionary*: defines an extension of the PROV-DM for collections and dictionaries.
- *PROV-Link*s: describes an extension of the PROV-DM for the correct description of multiple data sources.
- *PROV-AQ*: describes some methods to access and query data.
- *PROV-DC*: allows to referencing provenance data expressed in the Dublin Core Ontology.

The specific usage we made of the PROV ontology is described in Section 4.

3.2 XKOS Ontology

XKOS (eXtended Knowledge Organization System)[2] is an extension of SKOS (Simple Knowledge Organization System) [13] for managing statistical classifications and concept management systems.

With respect to SKOS, XKOS enables the representation of statistical classifications with their structure and textual properties, as well as the relations between classifications. Moreover, XKOS refines SKOS semantic properties allowing the usage of more specific relations between concepts for describing statistical classifications. In more detail, SKOS concepts are defined from the point of view of a thesaurus, thus it only defines the following relationships: (i) broader, (ii) narrower, and (iii) related to. Otherwise, statistical classifications rely on hierarchical relations (generic-specific and whole-part), thus XKOS introduces the definition of these relations structuring data into levels; a level corresponds to all those concepts that are at the same distance from the top of the hierarchy. Finally, XKOS defines causal, sequential, and temporal relations. See Figure 3 showing the XKOS concepts.

4 Modeling the Classification

This section describes the process that we conducted in order to create the LOD version of the official classification Ateco 2007 we call *LinkedAteco2007*,

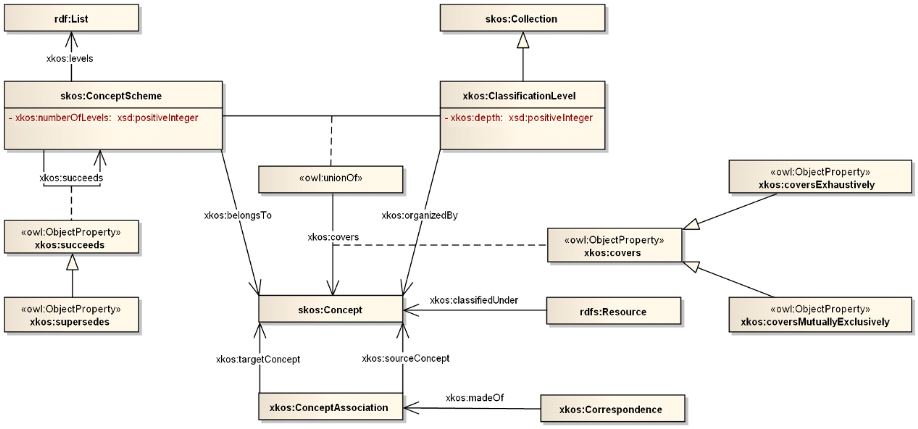

Fig. 3. XKOS Ontology Model Schema

Note that, the same process has been also applied to another classification named COFOG that has been recently published by AgID in collaboration with Istat. Both classifications are available at AgID SPARQL endpoint: http://spcdata.digitpa.gov.it/.

The activity involved the use of different ontologies and vocabularies and a customization of existing models. It is worth noticing that this activity was also described as a best practice within the Italian national guidelines for PSI valorization in order to guide PAs in (i) certifying the provenance of their data (crucial especially in the collaboration between the local and central levels of government), and (ii) using standard and common ontologies when describing their data.

The following subsections detail the activity carried out by Istat and AgID.

4.1 Provenance Modeling

In order to deal with the complexity of the process of standardizing classifications, in our modeling we gathered as much of the information related to such a process as possible, leaving in any case to other users the flexibility to extend the classification with variants and further versions. In this respect, the first phase of the modeling was to represent the provenance of the data that form the LinkedAteco2007. Figure 4 shows the resulting diagram that includes the activities, entities and actors involved in the process, all modeled through the PROV ontology. From the diagram, we can observe that the activity of publishing the *LinkedAteco2007* (i.e., *Generation Linked Ateco2007*) was carried out by two different actors: *Ufficio RST/B* and *AgID*. *Ufficio RST/B* is an organizational unit of the *DCIT* belonging to *ISTAT*. The diagram also specifies that *AgID* is the publisher of the *LinkedAteco2007*, whereas *Ufficio RST/B* is the creator. In addition, the original classification was influenced by the *Eurostat* classifica-

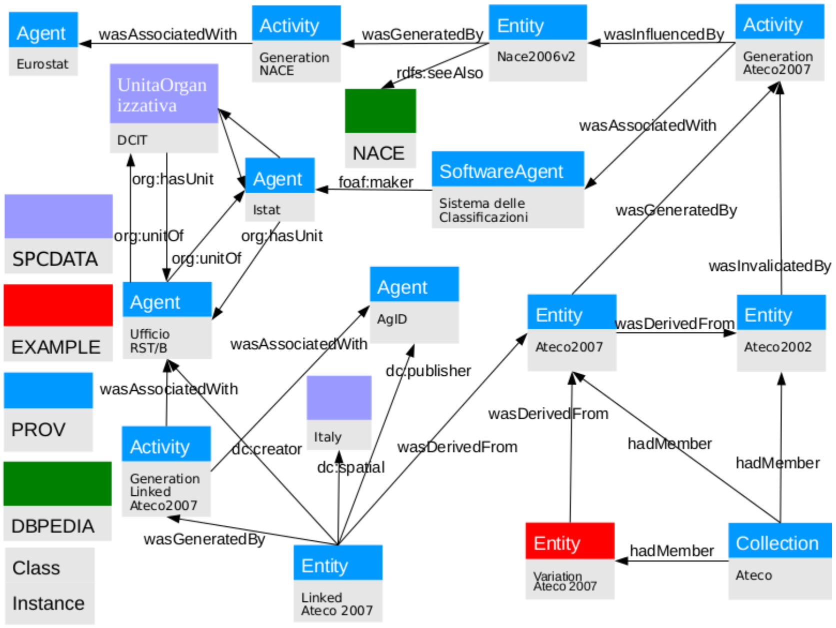

Fig. 4. Provenance Modeling

tion *NACE2006v2* and derived from the *Ateco2002*, a previous variant of the Ateco classification. All the Ateco classifications are grouped together through the *Ateco* collection.

The model is sufficiently flexible to allow an external user (e.g., a local administration) to extend the *Ateco2007* for its own purposes (e.g., add a non-existent activity): in this case, the user can define a variant from the *LinkedAteco2007* as illustrated by the *Variation Ateco 2007* box at the bottom of Figure 4.

4.2 Classification Distribution

To enrich the metadata of the classification, we exploited the use of standard and well-known ontologies. Figure 5 illustrates the metadata enrichment. In particular, we used two ontologies; namely, DCAT[3] and ADMS[4].The former allows us to insert the classification as dataset of our data catalogue and to decouple the abstract entity notion of dataset from its actual implementation. This also allows us to state the fact that the conceptual model of the dataset is expressed using the RDF framework and that we have produced such classification in different formats (e.g., RDF/N3, RDF/XML, etc.). These different productions are instances of the class *dcat:Distribution* and some of them are downloadable.

[3] http://www.w3.org/TR/vocab-dcat/
[4] http://www.w3.org/TR/vocab-adms/

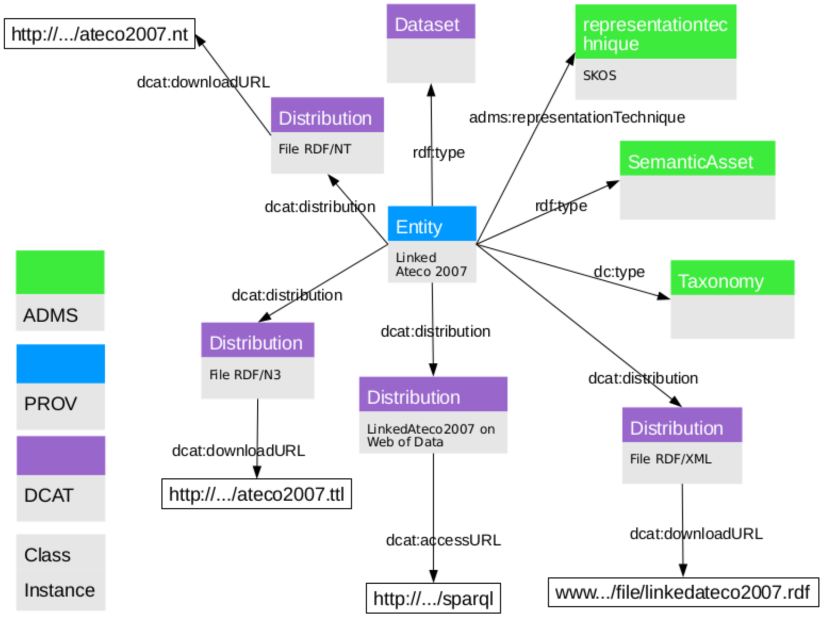

Fig. 5. Classification Distribution

The ADMS ontology allows us to specify and remark that the classification is a semantic asset, since it can be effectively used as integration element between different data, thus enabling semantic interoperability.

Finally, Figure 5 shows that the LinkedAteco2007 entity is a *Taxonomy* represented using *SKOS* (see next subsection).

4.3 Classification Modeling

To model the content of the classification we used the wide-used ontology SKOS [13] and its extension for statistical data XKOS [2].

SKOS allows us to express that *LinkedAteco2007* is a *ConceptSchema* and *XKOS* allows us to represent the full hierarchy of classification levels (i.e., the instances of the class *xkos:ClassificationLevel*). Each level of the classification is defined by its depth (i.e., *xkos:depth*) and is connected to all its member (i.e., *skos:member*).

Figure 7 completes the description of the classification by showing how each member is described. Specifically, a member has a notation (i.e., *skos:notation*), a label (i.e., *skos:prefLabel*), a textual description in the attributes note (i.e., *skos:note*) and a comment (i.e., *skos:comment*). It is worth seeing that every element is explicitly related to the upper level member; that is, it is a specification expressed through *skos:specializes*, and to the lower level member; that is, it is a generalization expressed through *skos:generalizes*.

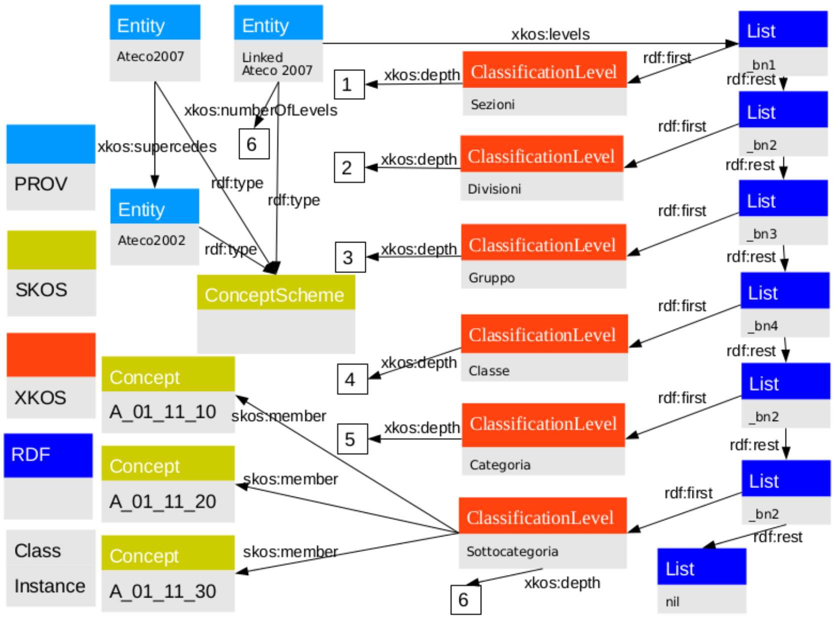

Fig. 6. Classification Modeling using SKOS and XKOS

Implementation. In general the implementation process of the previously described modelling adheres to a methodology proposed and used by AgID in other LOD projects [15]. In particular, the implementation consisted in the following four steps:

1. import the tabular data (i.e., a CSV file) from the public available data source[5] in a relational database;
2. clean and prepare the data (for instance adding new attributes with composed existing fields) for processing;
3. model the RDF data following the diagrams illustrated above and transform them accordingly using common tools. In our case, we used the D2RQ framework [14].
4. publish the resulting classification on the Web portal SPCData[6] and on the corresponding linked data cloud, and making available the LOD dataset for querying through the SPCData SPARQL endpoint.

5 Conclusions

The paper describes a real project by Istat and AgiD related to the publication of Official Classifications as LOD in the Italian Public Sector context. The technical

[5] http://sistemaclassificazioni.istat.it/class/sistemaclassificazioni/index.php

[6] http://spcdata.digitpa.gov.it/classificazioni.html

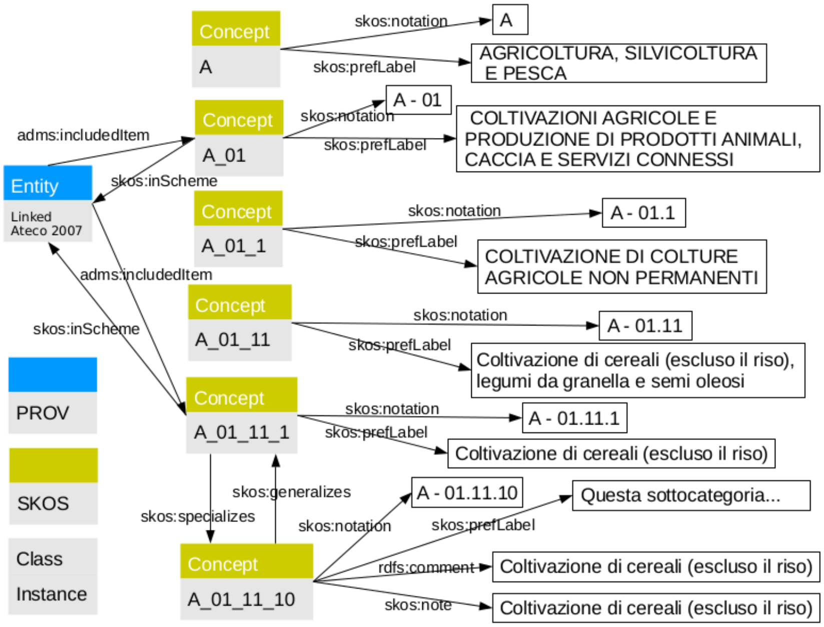

Fig. 7. Classification Levels with XKOS

contributions of the paper is focused on the modeling aspects of the classifications by making use of ontologies that are specific of the statistical domain, like XKOS, as well as of more general purpose ontologies, like PROV.

Besides technical contributions, the paper also provides a relevant methodological contribution to foster data interoperability among different institutions. Indeed, data interoperability is made possible by twofold efforts:

- shared formats and models, i.e., technological standards like RDF framework and ontologies like XKOS.
- content harmonization, i.e., common domain-specific information concepts.

We think that publishing OS classifications in LOD is a first important step towards "content" harmonization, which can consistently speed up the cooperation among different institutions based on data sharing and exchanges.

References

1. Classificazione delle attività economiche 2007 (ateco), http://www3.istat.it/strumenti/definizioni/ateco/ateco.html?versione=2007.3
2. Extended knowledge organization system (xkos), http://www.ddialliance.org/Specification/RDF/XKOS/
3. G8 open data charter and technical annex, https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex

4. Generic statistical information model (gsim) v. 1.1, http://www1.unece.org/stat/platform/display/gsim/Generic+Statistical+Information+Model
5. Insee official classification site, http://www.rdf.insee.fr/codes/index.html
6. Isic rev 4, http://unstats.un.org/unsd/cr/registry/isic-4.asp
7. Linked data initiative, http://linkeddata.org/
8. Lod-cloud, lod-cloud.net/
9. Nace official classification, http://www.ec.europa.eu/eurostat/ramon/rdfdata/nacer2.rdf
10. Neuchâtel terminology model classification, http://www3.ssb.no/DOCS/Neuchatelversion2.1.pdf
11. An overview of the italian guidelines for semantic interoperability through linked open data, http://www.agid.gov.it/sites/default/files/documentazione_trasparenza/semanticinteroperabilitylod_en_3.pdf
12. Provenance ontology (prov), http://www.w3.org/TR/prov-o/
13. Simple knowledge organization system (skos), http://www.w3.org/2004/02/skos/
14. Bizer, C.: D2r map - a database to rdf mapping language. In: WWW (Posters) (2003)
15. Lodi, G., Maccioni, A., Tortorelli, F.: Linked open data in the italian e-government interoperability framework. In: 6th International Conference on Methodologies, Technologies and Tools enabling e-Government (METTEG) (2012)