An Ontology for Representing and Extracting Knowledge Starting from Open Data of Public Administrations

Patrizia Agnello and Silvia Ansaldi
Department of Innovative Techonologies, Research Center, Inail, via Fontana Candida, 1, 00040 Monteporzio Catone, Rome, Italy

Keywords: Occupational Accidents, Open Data, Ontologies, Public Administration Knowledge.

Abstract: As proposed by European Commission through the institution of Europe's Digital Agenda, the Italian Digital Agenda has promoted the publication and the use of Open Data (OD) owned by the Public Administration (PA), providing with the appropriate guidelines for describing and managing Open Data under criteria of transparency, usability, accessibility. Inail (Italian Institute for Insurance against Accidents at Work) has recently started to publish periodically OD related to occupational accidents and diseases. The Inail OD, as they are defined, are mainly suitable for statistical studies but together with the vocabulary and the typological tables, that explicitly describe some codes in the dataset, it has been possible to develop a conceptual model which may be linked to schemas provided by other PA's for achieving the interoperability among those data through their semantics. Starting from the OD on occupational accidents, an ontology has been developed for capturing the meaning (semantics) of information, both contained into the OD available from Inail and defined in the vocabulary. The formalism of the ontology adopted for representing the knowledge, in terms of concepts and their relations, ensures to share a common language avoiding misinterpretations and misunderstanding.

1 INTRODUCTION

The challenge that we propose in this work is to structure a set of ontologies starting from the "static" vision of a phenomenon (accidents at work) that provide open data. The basic idea is to establish a robust platform, complete link for interoperability with respect to shared data with other PA deriving from open data set metadata.

In spite of other Italian PA, where ontology is mainly adopted for presenting data in a quick and trendy mode, linking them from different sites, the original idea of our proposal is that the ontology becomes a means of analyzing the "real" accidents datasets, not only a means of presenting open data.

The goal is to allow, later, to assess cause and effect, with the aim to prevent and, if possible, reduce the high number of accidents themselves.

The choice of starting from the open data stems from the fact that the information selected for the open data describing the accident happened, and in addition, some elements that determine the severity.

Are left out of the administrative, medical and economic, being its internal aspects of the scope of insurance, would contribute only to have a very complex knowledge domain. Moreover, these latter aspects would not contribute to the analysis of the cases, however, for the prevention of injury.

The experience gained in the development of ontologies comes as part of the industries and the risk of a major accident. Developments have been produced on the one hand for the management of knowledge hidden in large quantities of documents and the other as an aid in operational choices, for example the most useful devices PPE (Agnello et al., 2014) or more effective methods for the analysis of the risk in industrial areas increasingly complex (Ansaldi et al., 2012).

2 OPEN DATA

The Public Administrations, in order to perform their tasks, gather, organize and manage a huge amount of data, of different type and format. Most of that data are public (by law), and therefore they may be made open, that means available for others to use (Dati.gov, 2015). These considerations are valid

Agnello P. and Ansaldi S..
An Ontology for Representing and Extracting Knowledge Starting from Open Data of Public Administrations.
DOI: 10.5220/0005557501950200
In *Proceedings of 4th International Conference on Data Management Technologies and Applications* (DATA-2015), pages 195-200
ISBN: 978-989-758-103-8

both for national and European level (Reddick and Anthopoulos, 2015).

The Open Data can create adding value in various areas, e.g. to assure transparency of the administrative actions, to share information among the PA's, or to reduce the gap of PA with citizens (SPCData, 2015).

2.1 Inail OD

Inail has quite recently published on the owner site the OD related to the occupational accidents and occupational diseases, but only the first type has been the objective of this research.

The OD are available in the standard formats (i.e. RDF, XML, CSV), and can be downloaded free. Each Inail OD record represents a case of an accident, provided with a unique identification number, and other 24 fields corresponding to different types of information. The peculiarity of the OD is the fine granularity, which means that occupational accidents are individually represented in separate records, without applying any rules for combining them into some clusters.

The OD are delivered on monthly or half-yearly, and are split per Italian regions. At the download of the record set, a textual description provides with the explanation of each field contained in the OD, i.e. metadata name and its value type (OD Inail).

There exist two types of OD available, those related to the occupational accident data and those, which explain some codes, used in the OD.

Both of them are formally treated in a similar way; indeed, they use the same format types and textual description for the fields or metadata, but the former correspond to the real data, while the latter give the definitions, which are common to all OD. Those data correspond to reference tables, each for a specific topic, represented by a set of attributes.

Further to the textual explanations, enclosed with the OD, a vocabulary (Ciriello et al., 2013) provides detailed descriptions of all the terms used in the context of insurance of occupational safety (at least in the Inail terminology).

2.1.1 OD for Occupational Accidents

Inail decided to deliver open data in twofold aspects, as individual information and as aggregate format. As individual information, OD correspond to each occupational accident; as aggregate, they are organized into pre-defined clusters grouped by specific topics. The table 1, in the first column, shows the groups that characterize the types of attributes expressed in the OD; in the second column, some examples of the attributes are given. With "Ref. Tables", we indicate that the attributes refer to one or more reference tables which are properly of Inail "language".

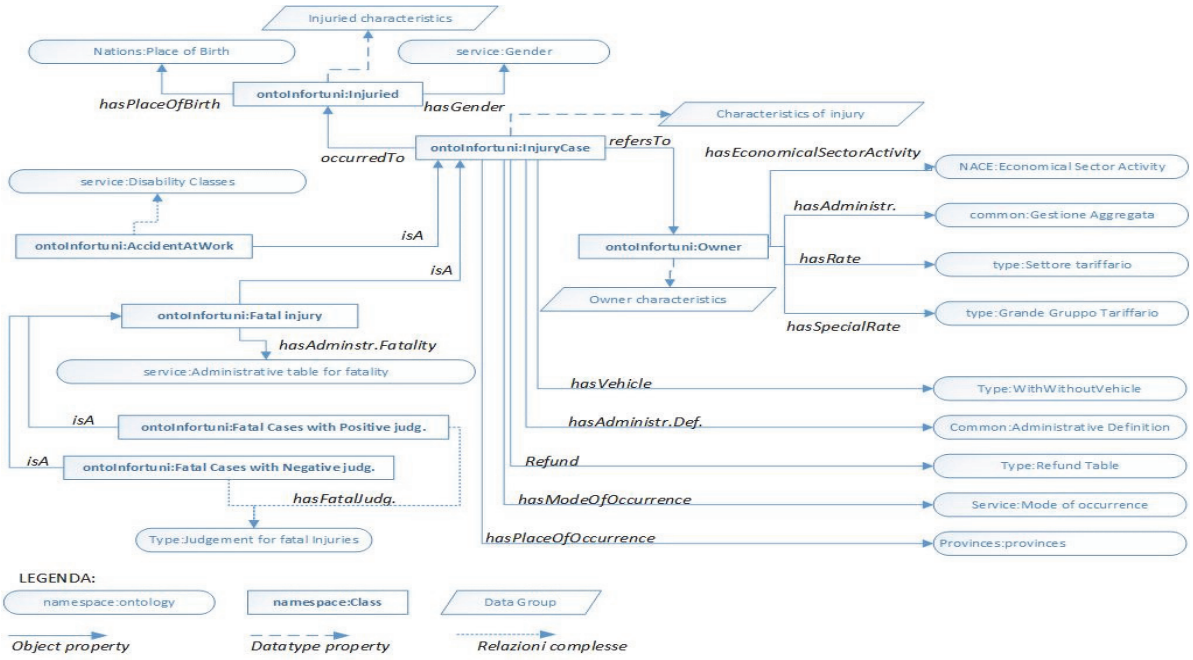

Figure 1: A diagram for the developed ontology.

Table 1: Groups of attribute types and some examples.

Group	Example
Temporal position of accident	Dates
Geographical location of accident	Province
Injured characteristics	Nationality
Modality of accident	At workplace
Administrative data of accident	Ref. Tables
Medical features of accident	Compensation
Owner characteristics	Ref. Tables

2.1.2 Reference Tables

As previously described, some reference tables are published using the same methods for representing the OD. There are three types of reference tables:

- Simple tables: which represent all the values that an attribute may assume;
- Complex tables: which contain multi-level structures;
- Tables with intervals: which are used to evaluate a parameter with respect to given ranges and determine the right classification of the injured

2.1.3 Statistical Tables

The OD, available for occupational accidents, are expressed in a fine granularity, indeed, each record corresponds to a single accident.

Furthermore, as previously discussed, some tables, that group the data on different characteristics, are also published on the site. They reflect the statistical results, which are considered of the major impact for the public interest.

For example, a set of tables groups the accidents on the basis of the modality and period of occurrence, others deal with the fatality accidents.

3 THE ONTOLOGICAL MODEL

The OD, as illustrated above, are well organized and structured data, mainly useful for developing statistical analysis, also sophisticated. Due to their fine granularity, diverse aggregate types can be studied, further to those statistical tables already available on the site.

On the other hand, the high specificity of the domain (insurance of occupational accidents) with the particular vocabulary adopted can arise some difficulties for the interoperability with other contexts. Indeed, for the current version of OD, the interoperability is represented only with the information related to NACE code, for defining the occupational sectors, the Italian Provinces, for describing where the event happened, and a classification of Nations for identifying the nationality of the injured.

What we have done is to define an ontology starting from the OD and the vocabulary, rather than model the domain of "occupational accidents" in general. The figure 1 shows the overview of all the concepts modelled into the ontology and their relations, some of them will be described in more details in the paragraphs §3.3 and §3.4.

The reason for this choice is that the domain of "occupational accidents" is characterized by sub-domains of diverse nature, such as management, legal and medical assessment, and economical evaluation. To model this domain starting from each single sub-domain in a complete form would be a very complex project, requiring the efforts of different types of knowledge and expertise. Since the data available are only those represented in the OD, it seems that the maintaining the consistency between the sets of data available and the domain of knowledge would be more effective.

Furthermore, since the granularity of the OD, the data contained can be considered the main and the sufficient information, which describes the event (accident). Probably, in order to define the administrative process for evaluating accident cases, an ontology fully devoted for covering all the administration aspects would be required.

All the fields contained into the dataset represent a result of the evaluation process of the accident, however, other information, which are considered during the evaluation phase, are actually out of scope of the OD, such as the description of cause/effect related to each case. The defined ontology can be easily extended to model those new concepts; this approach seems to be more feasible rather than trying to describe a general ontology, which covers all the aspects.

3.1 Definition Process

The design and development of an ontology may follow different methodological approaches (Noy and McGuinnes, 2001). In this research, both bottom-up and top-down approaches have been adopted, combining the two methods in an iterative process.

Indeed, starting from the data (contained into the OD and the reference tables) and the definitions enclosed into the vocabulary, through a bottom-up approach, it was pointed out the concepts and the instances to be modelled.

The top-down approach has been adopted for

getting the overview of the domain. Indeed, three elements are the key concepts to which all the other data refer. They are the case of accident, the injured, and the owner. The first two information are necessary for the domain, otherwise data are not meaningful. An identification number also uniquely represents these three elements in the OD.

3.2 Architecture

The methodology followed to define the ontology is to define many ontologies, each covers a domain, even if it is small, but it has its own consistency and conceptual autonomy of existence. This methodology seems to have advantages both in order to the extendibility of the ontology, previously mentioned, and in the concept of its reusability.

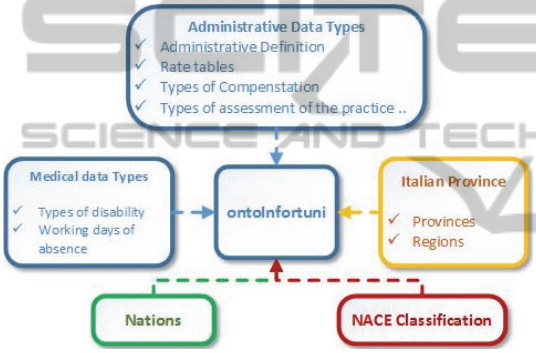

Figure 2: Architecture of *ontoInfortuni.*

In figure 2, the architecture of the ontologies developed is depicted. The central box indicates the accidents ontology (*ontoInfortuni*), and the other those representing the reference tables described in §2.1.2. The different colours correspond to different PA who are in charge of those types of data. The upper boxes refer to data of property of Inail, while, for example, NACE code is of Istat (Italian National Institute of Statistics).

The box *Administrative data Types* refers to concepts used for classifying injuries from the administrative point of view, while *Medical data Types* refers to models adopted for evaluating the severity degrees of injuries.

The following paragraphs illustrate the main ontology (*ontoInfortuni*) and the subsidiary ontologies developed. The ontology editor used for their definition is Protégé.

3.3 *ontoInfortuni*

Three concepts characterize the main ontology (*ontoInfortuni*): the case of accident, the injured, and the owner. All of them correspond to classes of the ontology.

In turn, the case of accident is subdivided into two subclasses: the occupational accident and the fatal accident. The concept "occupational accident" indicates that the injury has been formally recognized by the administration process, and can be insured by the Institute.

The "fatal accident" is itself subdivided into two disjoint subclasses, depending on the outcome of the insurance inquest, with positive or negative types. There is a reference table containing all the codes adopted for defining the outcomes of inquest of fatal events (*DecisioneIstruttoriaEsitoMortale*).

Therefore, each subclass is defined as equivalent to a class, which has relations with positive or negative outcomes, that means a partition of the codes in the reference table. The figure 3 shows a graphical representation where the orange balls correspond to the subclasses; the boxes represent the individuals of the reference tables. The diverse colours used to represent the individuals (dark for background and light for foreground, or vice versa) is only a graphical expedient for underlining that each code correspond to a specific type of outcome, positive or negative.

All the three classes have relations with the appropriate reference tables or with simple data type values (e.g. integer, string, date), which reflect the attributes present in the OD. Furthermore, the *reasoner* tools would be able to classify automatically each case of fatal accident, as positive or negative depending on their associated codes.

Nevertheless, in the OD each record corresponds to an accident report, all data are at the same level in sort of a flat representation. The identification numbers (corresponding to the injured and the owner) are unique for each individual but their values are repeated in case of multiplicity of the occurrence.

The ontology developed takes advantages of the OD representation, further to the vocabulary, however, the approach adopted is to model the concepts enclosed into the OD's, not their records. So far, each identification number, through a necessary and sufficient condition, uniquely characterizes and classifies each individual of a class (accident, injured or owner).

Each record of the OD implies that those three main concepts, and so far their classes, are related to each other. Two relations are explicitly modelled; both of them have the case of accident as a *domain*. One relation (*refersTo*) makes the event in correspondence with owner (*range*); the other

(*occurrsTo*) relates it with the injured (*range*).

This solution offers some advantages, further to avoid repeating the same individuals, but also to query the ontology.

3.4 Modelling Reference Tables

As discussed in the paragraph §2.1.2, some reference tables contain the codes, which correspond to pre-defined elements. Those tables have various meaning, diverse proprietary and different structure, as discussed in §3.2.

To represent the tables the same approach used for the main ontology is adopted that means to privilege the model of the concept rather than the record of the table. In practice, the simple tables, that contains a list of codes and their descriptions are modelled as individuals belonging to a class with two data type properties, the code and its description.

The other more complex tables, e.g. NACE classifications or Italian provinces, the approach is to model them starting from the concepts they describe; such as the hierarchical levels for NACE representation, or the provinces, regions and macro regions for the second type of table . However, those models represent only an exercise of modelling, since they are officially represented as ontologies available in the public domain.

The table, whose codes are not absolute values, but vary on a range of data, represents the last type. The class deals with the disability degree (*p*) of the injured. The table 2 shows, in the first column, the intervals assumed by *p*, in the second column, a description of type of disability.

Table 2: Disability table.

Range (*p*) (%)	Disability
[1 - 5]	Micro permanent
[6 - 15]	Minimal but superior than micro
[16 – 25]	Lower medium
[26 - 50]	Upper medium
[51 - 85]	Macro permanent
[86 - 100]	Loss of health

It is important to note that the OD contains only the value *p*, so far the classification of the events with respect to the table would be a process developed separately.

The model of this table is a class with six disjoint subclasses; each of them is equivalent (necessary and sufficient condition) to a class where the parameter *p* (the disability degree) belongs to the corresponding range. In addition one more subclass for accidents without disability.

Even in this case, any *reasoner* tool is able to classify automatically the events (accidents).

3.5 Browsing the Ontology

The system Protégé, used as ontology editor, has interesting plug-in for querying and browsing the model In our project, we used OntoGraf for graphical representations of the classes, individuals and relations; SPARQL Query for quickly interrogate the ontology; OWLDoc functionality

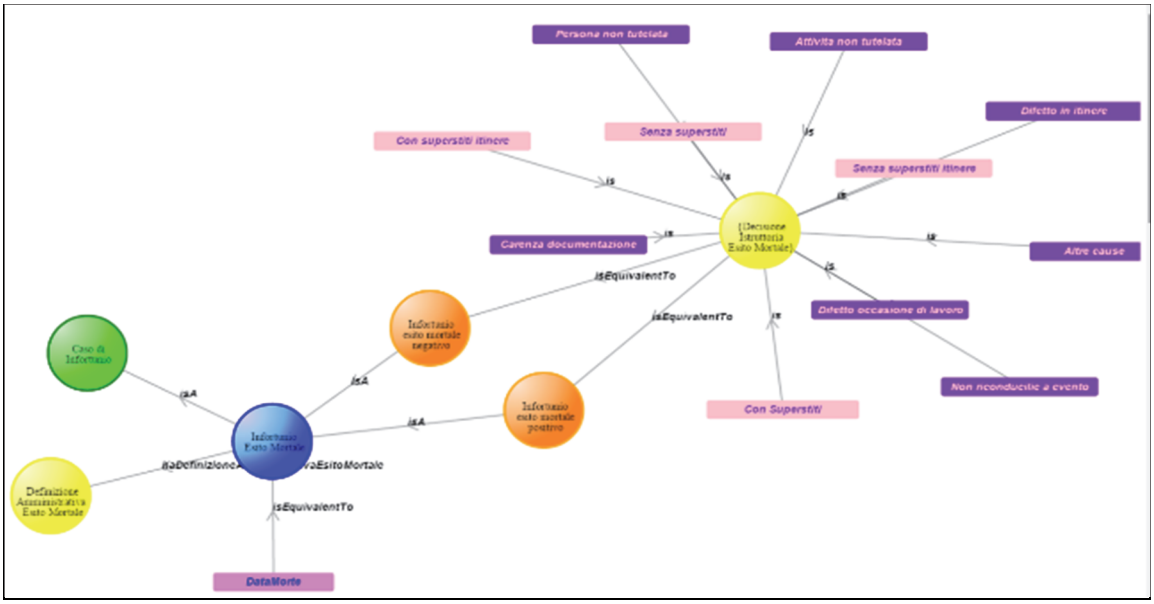

Figure 3: Graphical representation of fatal accidents and the positive or negative outcome of inquest.

which automatically generates pages in HTML for browsing into the model. However, for better represent and interrogate the ontology some applications have been implemented. One application is a graphical representation that tries to show pictures which cannot be produced by using OntoGraph plug-in, as illustrated in figure 3.

The other application is a prototype developed for querying the ontology using SPARQL language. As discussed by (Boselli et al., 2014), SPARQL is powerful but requires a deep knowledge of both the syntax of the language and the ontology model.

Both of these requirements are not common in the users who could take advantages of the Open Data. For this reason, this prototype provides the user with a list of both predefined questions in a natural language and statistical tables.

Each selection automatically generates an appropriate SPARQL query, which can be processed by the user. Furthermore, to write directly SPARQL queries is always possible.

4 CONCLUSIONS

The proposed work is a different approach to open data, at least for the Italian solutions. The ontology presented does not want to be the instrument of data presentation, as being those of fine granularity (each record in the table regional accident corresponds to a single injury) do not require display modes other than standard, and there is no pre knowledge developed by associate.

The OD Inail, related to accidents, publish a selection of information extracted from a much more complex and varied domain, covering different aspects in the procedure following the accident of the occurrence to liquidation economic.

The set of ontologies that are *ontoInfortuni* model part of the information predisposing to linked open data with other administrations, and a part of knowledge based on the characteristics of the accident itself will become the key to access to those informative dataset, from where open data were extracted for analysis of prevention. The latter is the one on which you are currently working.

ACKNOWLEDGEMENTS

Our acknowledgements are for the authors G. Ciriello, M. De Felice, R. Mosca, M. Veltroni of the Inail Research Book n.1 since this ontology is based on the vocabulary of Inail special language on occupational accidents. The authors are also grateful to Prof. De Felice and Mrs. Mosca for their support for the discussion and clarification of specific topics related to occupational accident insurance.

REFERENCES

Agnello P., Ansaldi S., Bragatto P.A., 2014. Pooling knowledge and improving safety for contracted works at a large industrial park. In *Special Issue WORK: A Journal of Prevention, Assessment, and Rehabilitation*. Karen Jacobs, Editor.

Ansaldi S., Monti M., Agnello P., Giannini F., 2012. An ontology for the identification of the most Appropriate Risk Management Methodology, *P. Herrero et al. (Eds.): OTM Workshops, Lecture Notes in Computer Science 7567, 444–453.* © Springer-Verlag Berlin Heidelberg.

Boselli, R., Cesarini, M., Mercorio, F., & Mezzanzanica, M., 2014. Are the Methodologies for Producing Linked Open Data Feasible for Public Administrations?. In Proceedings of *Special Session in Knowledge Discovery meets Information Systems (KomIS) at DATA2014*. ISBN 978-989-758-035-2

Ciriello, G., De Felice, M., Mosca, R., Veltroni, M., 2013. Infortuni sul lavoro. Un modello di lettura (della numerosità) su "open data" dell'Inail, *Quaderni di ricerca*. *Inail*, Roma.

Noy, N. F., McGuinnes, D. L., 2001. Ontology Development 101: A Guide to Creating Your First Ontology. *Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report* SMI-2001-0880.

Reddick, C. G., Anthopoulos, L., 2015. Information and Communication Technologies in Public Administration. *Taylor and Francis Group*. ISBN: 978-1-4822-3930-0

Protégé: http://protege.stanford.edu/

Agid: http://www.agid.gov.it/agenda-digitale

OD Inail: http://dati.inail.it/opendata/default/Infortuni/index.html

SPCData: http://spcdata.digitpa.gov.it/index.html

Dati.gov: http://www.dati.gov.it/