

# Metadata for Content Description In Legal Information

Aldo GANGEMI\*, Maria-Teresa SAGRI\*\*, Daniela TISCORNIA\*\*

\*Laboratory for Applied Ontology/ISTC

(Institute for Cognitive Sciences and Technology)

of the Italian National Research Council, Rome Italy

\*\* (Institute for Theory and Techniques for Legal Information)

of the Italian Research Council, Florence, Italy

\*gangemi@ip.rm.cnr.it

\*\*{sagri,tiscornia}@ittig.cnr.it

## Abstract

The paper describes the Jur-Wordnet (Jur-IWN) project, whose objective is an ontology-based extension to the legal domain of the Italian version of EuroWordNet. It aims at providing both a content description model for legal information and a resource for accessing multilingual and heterogeneous information sources. Jur-IWN involves a double interpretation of the model, and of its features. Conceived as a lexical resource, terms are linked to each other throughout lexical relation, while, as a content description model, the concepts are organised according to stronger assumptions about the ontological nature of entities that populate the legal domain, and about their relationship. The crucial problem is to save the global consistency of the model: the lexical view needs a clear definition of the boundaries between common language and technical legal terminology, while the conceptualisation of the core entities needs to be linked in a coherent way to upper level categories. The upper-level ontology chosen for the project is DOLCE, with its extensions.

## 1. Introduction

The WordNet (WN) and EuroWordnet (EWN) projects offer wide-reaching, standardised linguistic resources for searching information on the Internet. WordNet is a semantic network developed by the linguist George Miller and his colleagues at Princeton University. Developed a decade ago, it is available free of charge on Internet. ILI is the “inter-language” that, from WN, interrelates the lexical/semantic networks developed for the other European languages in the *EuroWordNet* project founded by the European Community. Currently, more than twenty languages share the same methodology and development structure and are linked to each other through the English language.

ItalWordNet (IWN) is the Italian section of EuroWordnet, developed at the Institute for Computational Linguistic of the CNR of Pisa [17].

Jur-(Ital)Wordnet (Jur-IWN) project is an extension to the legal domain of the Italian version of EuroWordnet, linked to the Interlingual Index (ILI) records; other specialised sectors dealing with technical domains have been developed (e.g. EcoWordNet<sup>1</sup> for economic/financial information).

On the other hand, some so-called *formal ontology* projects have addressed the problem of formally defining a set of concepts and relations that can be used to share heterogeneous conceptualizations from the same or close domains, and to correlate complementary conceptualizations in different domains. Contrary to the assumption made by wordnet-like projects, formal ontology does not limit its scope to “lexicalized” concepts, thus enabling a more explicit, interlingual, and powerful analysis of the legal domain.

For this reason, Jur-IWN has been based on the DOLCE foundational ontology [13] and its extensions, developed within the EC WonderWeb project<sup>2</sup>, which have been already used in several domain-oriented projects, and is being used in the OntoWordNet project [11] that aims at transforming WordNet into a set of axiomatic theories (though biased to common lexical knowledge).

Jur-IWN is presented here as joint work between the ITTIG-CNR legal information experts, and the Laboratory for Applied Ontology (LOA) at ISTC-CNR, where existing projects also address the ontological refinement of wordnets and domain taxonomies.

---

<sup>1</sup> Developed by Istituto per la Ricerca Scientifica e Tecnologica of Trento (IRST), <http://www.irst.it>

<sup>2</sup> <http://wonderweb.semanticweb.org>

## 2. The Jur-Wordnet Project

The starting point of the project was the *Norme in rete* (Law on the Net) project, launched in 1999 as part of the Italian *E-government Plan*. *Norme in rete* involves the most important Italian Institutions (the Chamber of Deputies, the Senate, the Department of Justice, the Computer Science Authority of the Public Administration); its goal is to “create a portal which, through a single and simple user interface, allows research on all the documentation of normative interest published free on Internet, particularly by institutional sites.”[1]. The portal allows free access to normative information through standard methods of editing, processing, and distributing data, which will lead to a direct, self-updating access, by reaching the sites of law makers (and distributors) themselves. In fact, the project provides codification language: standards for source types, identifiers (uniform references notation: *urn*), structure, links, and *meta-information*.

System design, by now consolidated, [2] consists of classes of XML DTDs for structuring normative<sup>3</sup> texts and of metadata, the most relevant part of which deals with the formal/structural features of each type of source, and with *urns* for the identification of the partitions of texts. Currently, content description is limited to the adoption of classification schemes (topics) as previously defined by institutional databases (Teseo). The domain of such topics is made of normative texts, taken as wholes.

The aim of Jur-IWN is providing the NormeInRete system with a knowledge base for semantic tagging at the level of articles or even dispositions; that is, recognising normative entities (the dispositions) inside the text, which is not necessarily the same as the structural entities, and assuming a double view of the text, both as a document and as a collection of dispositions.

Thus, the Jur-IWN terminology database supplies a source of metadata for semantic tagging of legislative texts (which may also be used in the legislative drafting phase as an enrichment of the specialised XMLeditor now in the development phase [3], and of other legal sources<sup>4</sup>.

Furthermore, the database can be a support tool for information retrieval systems, in order to facilitate access to heterogeneous and multilingual data, and a conceptual

source for information extraction, automatic tagging, knowledge sharing, norm comparison, etc.<sup>5</sup>

## 3. Jur-IWN as a lexical resource and a content description model

The double perspective of application of Jur-IWN involves a double interpretation of the model, and of its features: as a lexical source, links between terms are defined throughout lexical relations, while as a content description model, concepts are organised according to stronger semantic assumptions about the ontological nature of entities that populate the legal domain and about their relationship. Previous work on this has been carried out by the LOA [11]

The crucial point to solve in the project is defining the relationships between the lexical/conceptual levels, and the criteria for organising more abstract terms, where word/concept distinction is highlighted.

This tension is already perceived by ongoing work on existing wordnets. For example, WordNet was created as a lexical net: in the original project they did not believe necessary to organise hierarchies according to non-strictly linguistic or meta-level categories. Recent proposals are assessing the use of formal ontologies on WordNet upper classes to eliminate ambiguities and redundancies [e.g. 20], as well as for transforming WordNet data types into ontological metadata.

EuroWordnet made also a step towards ontology by referring to a *Top-Ontology* based upon Pustejovsky's [18] *qualia* structure, which appear to be more ‘perspectives’ than to real top-level ontological categories. The *Top-Ontology* patterns the trees of ILI, which is the interlingual Index of the English terms shared by all European languages.

As shown in Fig. 1, these research programs and models allow to envisage a multilingual, ontology-based legal ontology, at least as far as the lexical knowledge of Law is concerned.

### 3.1. Jur-IWN as a lexical resource

According to the twofold interpretation of Jur-IWN, different connotations and aspects are focuses.

As a *linguistic tool*, it will improve legal information retrieval from heterogeneous (legislation, legal cases, policies) and multilingual sources it is the task closer to the development of the Wordnet initiatives. In the legal domain, a standard lexicon, able to handle linguistic phenomena as polysemy and synonymy, can also establish

---

<sup>3</sup> A detailed description of DTDs for legislative texts is published at:

[http://www.normeinrete.it/standard/standard\\_xml.htm](http://www.normeinrete.it/standard/standard_xml.htm); similar initiatives are:

[www.lexml.de](http://www.lexml.de), <http://www.legalxml.org/>, <http://lri.jur.uva.nl/METALex/>

<sup>4</sup> The *Norma in rete* Project also includes the definition of XML DTDs for judicial decisions, local regulation, and public contracts, still in progress.

---

<sup>5</sup> The (Jur-IWN methodology will be tested in the E-Psinet Project (E.Content Program), aimed at comparing the regulatory environment of Public Sector Information in the EC. ([www.publicsectorinfo.com/](http://www.publicsectorinfo.com/)).

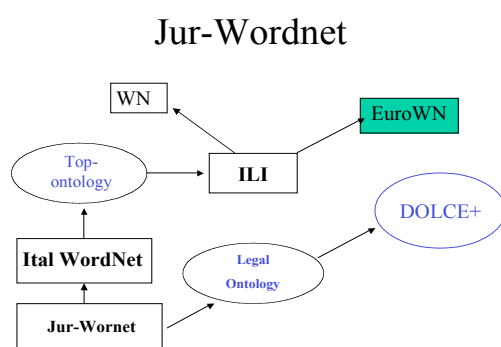
a bridge between the common language often used from the non-jurist ones -in order to place legal questions- and the technical language of the law.

From this viewpoint, the *eq-plug* and *hypo-plug* functions from Jur-WN to ItalWordNet allow a more precise definition of technical meanings of terms used in the common Italian, such as *authorisation*, *alienation*, or the specification in sub-meanings of terms acquiring specific law meaning such as *alimony*, or *delay*. Moreover, it allows the insertion of phrases, which ensures that the common semantics of the main term, e.g. *acceptance* in “*witness, evidence acceptance*”, still pertains to the legal domain. Such legally-stipulated words are linked through “hypo-plugs” to the IWN trees, and to the top-ontology shared by wordnet-like initiatives.

### 3.2. Jur-IWN as an ontology

As a source of *metadata for content description*, usable for semantic tagging, we need a standard of metadata based not upon lexical relations, but upon the definition of the content, embedded in the textual units. Such standard must be based on the ontological nature of the entities of the legal world; concepts such as ‘*license, authorisation, delegation*’ that acquire a specific meaning in the legal domain, and that roughly match the classical partitions of legal theory, are organised in a *legal core ontology* [15]. The development of the core ontology takes into account methodologies proper of the upper level ontologies [11, 12], and proposals in the field of legal ontologies [14]. The Core Ontology organises the juridical concepts in classes on the base of formal (meta) properties defined in the *foundational ontologies* [13]. Likely, the basic entities that populate the domain of law can be considered universal and clearly identifiable, and, as such, they are pointed out through a minimal generalised series of properties and relations. The lowest level of the Core Ontology is specialised according to the domain of application and widened on a lexical level through the Jur-IWN taxonomies.

Fig.1



### 4. The Legal Core-Ontology for Jur-IWN

In line with the “bottom-up” approach, the base of the ontology is the higher level of concepts obtained through the conceptualisation of the terminology: from the 1500 synsets structured so far, (we expect to reach a satisfactory coverage when about 3000 synsets will be defined) about 40 concept have been organised into the ontology.

The classes of entities are specialisations of the DOLCE<sup>6</sup> foundational ontology [11] and some recent extensions, notably the “Descriptions and Situations” [19] DOLCE and its extensions will be referred here as “DOLCE+”. The methodological choices, as well as the exploitation of properties suitable for the legal domain are based upon the approach of legal theory and of the philosophy of law. Legal world is conceived as a *representation, or a description of the reality, an ideal view of the behaviour of a social group, according to a system of rules that are commonly accepted and acknowledged.*

**Agents:** physical existence is not a necessary but a sufficient condition for being a *legal subject*: legal agent is therefore a *role*, created by (constitutive) rules and played both by human and social individuals; a *natural legal person* is a legal subject because of its only physical existence (even before birth and after death), whereas the *legal person* role needs to fit strict requirements, such as age, mental non-illness, or artificial existence. The legal subject subsumes the legal person but not the contrary.

**Roles (functional roles and, as sub-sets, legal roles):** these are descriptions of either physical or non-physical objects (the DOLCE category for objects and substances being called *endurant*). *Descriptions* are first-order entities in the DOLCE extension (Descriptions & Situations Ontology) that is being used to create a core ontology of Law. Among legal roles, some of them constitute the basic entities of the legal world, as *legal subjects* and *legally constructed persons* are: they are created by constitutive norms that justify their existence and validity, e.g. Ministry, Body, Society, Agency); a further class of legal roles (relational legal roles) are played by legal subjects and are referred to specific situations and states (defendant, representative, commission). Norms themselves are represented as non-physical and non-agentive objects. Among norms, *constitutive norms* and *regulative norms* (deontically modalized) are distinguished; *definition* and *power-conferring rules* are sub-classes of constitutive norms.

<sup>6</sup> DOLCE stands for “Descriptive Ontology for Linguistic and Cognitive Engineering.” Version 2.0 of the deliverable (Wonderweb D17) is downloadable from the Web (<http://wonderweb.semanticweb.org>.)

**Mental objects** are internal descriptions, e.g. *agreement* and *mistake*, which are results of mental processes or cognitive states. Upon cognitive states a distinction is grounded between general *legal acts*, as products of non intention-driven activity, and *legal transactions*. Among cognitive states (that are *perdurants*), intentionality is subsumed by will, which is subsumed by consciousness: the distinction between will and intentionality grounds the distinction between fault and intention fraud in crime law.

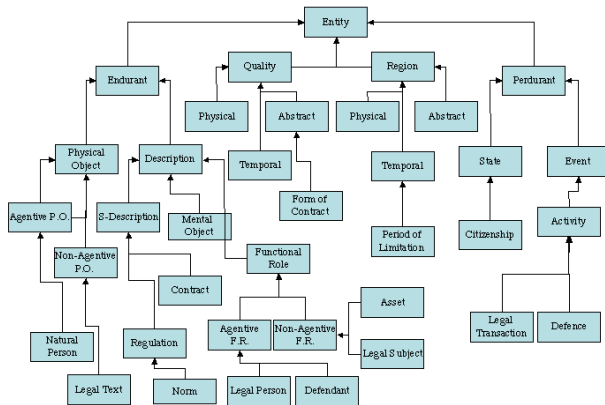


Fig.2. A preliminary linking of legal concepts to DOLCE+. Arrow semantics is *IS\_A* (subsumption).

The figure below shows the classification of the Jur-IWN concepts as entities of a *Legal Core Ontology* under DOLCE (without extensions):

<b>PHYSICAL-ENDURANT :</b>
<b>:PHYSICAL-OBJECT:</b>
<b>:: Agentive :</b>
Natural person
<b>:: Non-Agentive :</b>
Property
Legal text
<b>NON- PHYSICAL-ENDURANT:</b>
<b>:Description:</b>
<b>: :Functional-Role:</b>
<b>:::Agentive :</b>
Legal person
Agency
Authority
Government
Body
Nation
Association
<b>::: Non agentive :</b>
-----
<b>::: S-Description:</b>
<b>::: Regulation:</b>
Law
Legislation
Obligation
Right

Duty
Power
Interest
<b>::: Internal Description:</b>
<b>::: Mental Object:</b>
Mistake
Agreement
<b>QUALITY:</b>
<b>:Physical:</b>
Movable
<b>:Non-Physical:</b>
Form
Requirement
<b>ABSTRACT</b>
<b>:Temporal-Region:</b>
Time interval
Lapse of time:
Period of limitation
<b>:Physical-Region:</b>
Space-Region:
Location
Abode
<b>PERDURANT:</b>
<b>:Event:</b>
<b>::: Accomplishment:</b>
<b>::: Activity:</b>
Legal Act:
Crime
Amendment
Legal Transaction
Contract
Performance
Defence
<b>::: Phenomenon</b>
<b>:Achievement</b>
<b>:Stative:</b>
<b>: Process</b>
<b>::: Mental process</b>
<b>: State:</b>
Legal status
Capacity
Citizenship
<b>::: Cognitive- State:</b>
Mental illness
Consciousness:
Will:
Intentionality

#### 4.1 Relations between entities

The taxonomic work in Jur-IWN is now assuming a core ontology of Law that is being developed in collaboration between ITTIG-CNR and the Laboratory for Applied Ontology of ISTC-CNR. The current version of the core ontology is based on DOLCE+ [11, 19], and an exemplification of the main relations is presented here (Fig. 3):

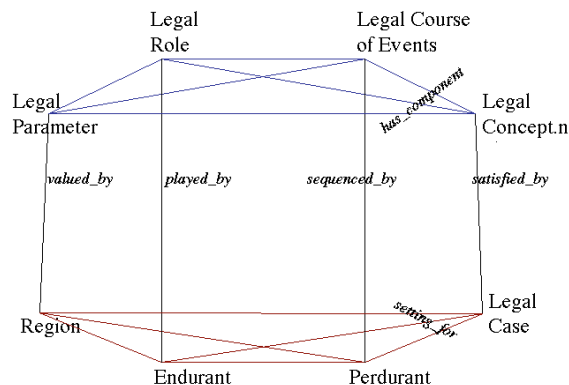


Fig.3. A conceptual template for legal descriptions (conceptualizations) and situations (cases)

The current version of the *legal core ontology* assumes the distinction between the legal and non-legal worlds, and maps it to the DOLCE+ distinction between descriptions (in this domain *legal conceptualizations*), which encompass laws, norms, regulations, crime types, etc., and situations (*legal facts* or *cases* in this domain), which encompass not only non-legal states of affairs that are relevant to the right, but also purely juridical states of affairs that occur. This enables us to use that distinction to represent meta-juridical conceptualizations (*meta-norms*) as well. From the logical viewpoint, DOLCE+ ontology for descriptions and situations operates a non-structure-preserving reification of logical theories and states of affairs [19].

A legal conceptualization is composed of legal *roles*, legal *courses of events*, and legal constraints or *parameters* on entities that are bound to the setting created by a legal case. This enables us to build a complex, *functional* representation of the Law and of its facts. For example, a certain *traffic law* will be satisfied by a fact or case in which every component of the setting respectively satisfies the components of the law: traffic events and states (*perdurants*) respect a “sequence” established in the legal traffic course of events, traffic objects (*endurants*) “play” their legal traffic roles, and the attributes (*regions*) of traffic-related events, states, and objects are “values” within the legal traffic parameters.

Here we suggest some examples of use of DOLCE+ relations (including not only relations between descriptions and situations):

#### Descriptions and situations relations:

Norms as *legal conceptualizations* are expressed by a normative text that is *physically represented* by a document.

Norms may be *satisfied* by purely juridical situations, as for norms that describe others norms: amendments, law-making norms, and validity norms.

Norms may have *parts* and *components* that are the representation of:

- Institutional functions (constitutive norms)
- Institutional powers (power-conferring norms)
- Behaviours (regulative norms)
- Incrimination acts (incriminating norms).

#### Dependency and legal constitution relations:

*Legal persons* depend on physical entities.

*Legal subjects* just depend on norms.

*Legal facts* (including *cases*) are situations *depending on* norms (only facts relevant for legal systems are legal facts). Subclasses are:

*Natural facts* (e.g. death)

*Human facts*, depending on consciousness (but not on will), which can be distinguished into:

- *Institutional facts*: legally constituted by (*satisfying*) constitutive rules
- *Legal acts* (in a strict sense) *depending on* will
- *Legal transactions*, depending on intentionality
- *Crimes*: legally constituted by (*satisfying*) incriminating norms

*Legal functional roles* are *components* of norms and then are created by (and depending on) them (are *constituted* in legal terms).

*Written form* is constitutive of (*inherent in*) some legal acts information, or put differently, some legal acts *depend on a written physical representation*.

#### Participation relations:

*Legal subjects* participate in human activities and can be *set within* legal facts (as well as within other social facts).

#### Inherence relations:

Communication form is a physical quality of documents, *inherent in* legal acts.

Jur-IWN database is still under development: we expect to reach a satisfying coverage of the basic legal contents through the definition of about 3000 synsets. The enrichment of the lexical database will probably act as a testbed for the ontological level, and will allow the refinement and completion of the work done.

## References

- [1] Report on “*Il progetto Norme in rete*”, Rome, January 2000 ([www.normeinrete.it/documenti](http://www.normeinrete.it/documenti)).
- [2] <http://www.normeinrete.it/standard/circolare-xml.htm>

- [3] [www.ittig.cnr.it/organizzazione/personale/biagioli/normeinrete](http://www.ittig.cnr.it/organizzazione/personale/biagioli/normeinrete)
- [4] Enciclopedia giuridica, 1995, Treccani, Roma, I.
- [5] Enciclopedia del diritto, 1989, Giuffrè, Varese, I.
- [6] Grande Dizionario enciclopedico del diritto, Fratelli Fabbri Editore, Milano, I.
- [7] De Mauro T., Il Grande Dizionario italiano dell'uso, UTET, Torino, I.
- [8] Il Dizionario della lingua Italiana, 2002, Garzanti, Milano, I.
- [9] Il Nuovo Zingarelli, 2002, Vocabolario della lingua italiana, Zanichelli Ed. Milano, I.
- [10] [www.ittig.cnr.it/banche/LLI/](http://www.ittig.cnr.it/banche/LLI/).
- [11] Gangemi A., Guarino N., Masolo C., Oltramari, A., Schneider L. 2002. Sweetening Ontologies with DOLCE. in *Proceedings of EKAW 2002*, Sigüenza, Spain, pp 166-178.
- [12] Gangemi A, Pisanelli DM, Steve G: An Overview of the ONIONS Project: Applying Ontologies to the Integration of Medical Terminologies. *Data and Knowledge Engineering*, 1999, 31, pp. 183-220 (1999).
- [13] Masolo C., Borgo S., Gangemi A, Guarino N, Oltramari A, Schneider L. The WonderWeb Library of Foundational Ontologies. <http://wonderweb.semanticweb.org/deliverables/D17.shtml>.
- [14] Visser P., T. Bench Capon, 1999, Ontologies in the Design of Legal Knowledge Systems, towards a Library of Legal Domain Ontologies, in *Proceedings of Jurix 99*, Leuven, Belgique.
- [15] Gangemi A., Pisanelli DM., Steve G., 2001, A formal Ontology Framework to represent Norm Dynamics. *Proceedings of Second International Workshop on Legal Ontologies*, Amsterdam, NL.
- [16] Sagri M.T., 2003, Progetto per lo sviluppo di una rete lessicale giuridica on line attraverso la specializzazione di ItalWornet, in *Informatica e Diritto*, ESI, Napoli, forthcoming.
- [17] Roventini A., Alonge A., Bertagna F., Calzolari N., Girardi C., Magnini B., Marinelli R., Speranza M., Zampolli A. (in press), *ItalWordNet: Building a Large Semantic Database for the Automatic Treatment of Italian*, in "Linguistica Computazionale", Istituti Editoriali e Poligrafici Internazionali, Pisa-Roma, ISSN.
- [18] Pustejovsky H., 1995, *The Generative Lexicon*, Mit Press, Canbrige MA
- [19] Gangemi A., Masolo C., Steve G. 2003, *Reified contexts: an ontology of Descriptions and Situations*, LOA Internal Report 2003-1, available at <http://ontology.ip.rm.cnr.it>