Bridging legal documents, external entities and heterogeneous KBs: from meta-model to implementation

Gioele Barabucci, Angelo Di Iorio, Francesco Poggi

July 15, 2012

Abstract

Every legal document contains references to external entities: people, organizations, concepts and so on. In this paper we present ALLOT, an ontology to describe non-documental entities referenced in Akoma Ntoso documents and legal documents in general. We also discuss how to develop new ontologies, XML and DB schemas that follow the current best practices and avoid the common pitfalls found in the legal domain.

1 Introduction

Any system used to manage legal documents, for instance to edit, store, sign or publish them, has to deal with so called *external entities*, i.e. entities that exist outside of a document and independent of it. For example, while marking up a document, a system may help the user selecting which of the three members of parliament named "John Doe" is the one being referenced in a certain sentence as "Mr. Doe". To help the user, the system could show them a list of three possible choices, each annotated with some disambiguation information, for example "Do you mean John Doe (born in 1980, Dem party), John Doe (born 1959, Rep party) or John Doe (born 1959, Lib party)?". In order to present such choices, all the needed pieces of information must be stored in a place accessible by the system and in an understandable format.

While formats and methodologies for legal documents are becoming widespread, shared and standardized, the data that records the information about the external entities is often stored in proprietary schemas and formats, locked in inaccessible databases. This means that many organizations are now publishing documents (acts, reports, judgments) that refer to entities that are not easily accessible, reducing the usefulness of the documents themselves [41].

This problem is mitigated by the increasing diffusion of Open Government principles [32] and Open Data principles in general [7]. More and more governments are now publishing their documents and proceedings in freely available knowledge-bases (KBs), often as Linked Data silos. This approach makes it possible to connect and integrate such legal data with other sources and external entities, by exploiting current Semantic Web technologies.

Unfortunately, a deeper analysis of such legal data shows that many of them are based on partially flawed conceptualization schemas. In particular, we notice that most

of the existing legal KBs undervalue the temporal dimension. Although some ontologies provide constructs to properly deal with time (i.e. to contextualize events and roles in a precise time point or interval), the final assertions in the KBs radically simplify such data and miss important information. It is very common, for instance, to read "John Doe is an MP", instead of "John Doe is an MP in the 54th legislation, that started in 2009 and is supposed to end in 2013". The relation between legal entities and text fragments occurring in legal documents is another aspect that is often overlooked. Strings referring to entities are mistook for the entities themselves and it is very common to find assertions, in particular references to external entities, that take over-simplified forms. For instance, sentences like "document 34 has author 'Mel Smith"' are much more used than "document 34 has author person-35, person-35 has name 'Mel Smith"'. These and other similar issues found in current databases will be discussed in more details in section §3.1 and section §6. For now it suffices to say that all these limitations make it harder to exploit such KBs at their best and make them less useful (and used) than they could be.

The main obstacle to a more effective integration of legal KBs still remains the heterogeneity of these resources. It is not only a matter of formats but also a problem of intrinsic incompatibility: in several cases radically different views of the world are employed, making it very hard or even impossible to link entities in a dataset with entities in other datasets.

With regards to formats, Akoma Ntoso is one of the most prominent formats for legal documents. Akoma Ntoso is an open XML standard for parliamentary, legislative and judiciary documents [12]. The standard is based on a clear distinction between legal texts, metadata and ontological information built on top of such metadata. The choice of which legal ontology should be used for Akoma Ntoso, or whether an ontology should be used at all, was rather difficult due to the heterogeneity of legal sources and of legal ontologies we already pointed out. The Akoma Ntoso designers decided to give implementers the maximum degree of freedom, allowing them to use any ontology and any formalism, that such an open scheme is referred to as "the Akoma Ntoso non-ontology".

This paper discusses an actual implementation of the Akoma Ntoso non-ontology and introduces ALLOT, a proof-of-concept OWL ontology based on it. ALLOT is meant to be used to expressing all the non-documental semantic data found in Akoma Ntoso documents and its surrounding systems. It is particularly useful to connect Akoma Ntoso documents to the external entities they refer to, entities that are stored in legacy knowledge-bases, newly created Linked Data silos and relational databases. In fact, one of the most interesting and challenging aspects of ALLOT (and the overall "non-ontology" approach) is the integration between heterogeneous sources of semantic data.

The goal of our work is manifold. First of all, we investigate motivations, design principles and difficulties of the "Akoma Ntoso non-ontology". Second, we report our experience in building ALLOT from such a scheme, with particular attention to the design of entities URIs and references. Third, we discuss some extensions of ALLOT, rooted again in the "non-ontology" model, that allow users to align the ontology with existing models. The final goal is to distill a set of guidelines for developing clear, unambiguous and long-lasting knowledge-bases for the legal domain.

The discussion is then structured as follows. In section §2 we give an overview of existing legal ontologies, focusing on their relation to external models; in section §3.1 we discuss some issues and possible improvements in this context; in section §3.2 we introduce the Akoma Ntoso non-ontology, while in the following section §4 we explain

how Akoma Ntoso deals with external entities; in section §5 we present ALLOT and show, in section §6, how ALLOT can be used within Akoma Ntoso in conjunction with existing legal ontologies; lessons learnt and advices for the developers of new knowledge bases are summarized in the conclusions.

2 Legal and para-legal ontologies

The construction and integration of legal ontologies is a hot research topic. Multiple ontologies have been proposed at different levels of generality, combined together with different methodologies.

It is hard to define clearly the scope of so called legal ontologies as they live on the intersection of many different domains, having to deal with documents, concepts, agents, references, outcomes, rules and many other aspects of the law. For this reason, legal ontologies often embed ontologies specific to other domains or are used in conjunction with them.

Legal ontologies are usually divided in two categories: core and domain-specific. Ontologies in the former group describe the "common conceptual denominator of the field" [23] and provide definitions of general legal concepts such as *norm*, *legal role*, *legal entity*, while ontologies in the latter group cover more context-oriented concepts such us *punishment* (useful when dealing with criminal law), *work of mind* (useful in copyright regulations) and similar abstract ideas peculiar to certain field. The overall picture is completed by top/upper ontologies that define even more abstract notions (e.g. *place*, *event*, *change*, *containment*) referred by core/domain ontologies.

We propose an orthogonal classification of legal ontologies centered around their relation to legal sources and external entities. In particular, we divide them in three groups: document-centric, content-centric and integration-centric.

2.1 Document-centric ontologies

We classify as *document-centric* those ontologies whose main goal is to describe the documental part of the legal documents. In particular, document-centric ontologies are used to model the evolution of legal sources, from their creation through their modifications and steps in the legal processes they belong to. For instance, it makes possible to relate multiple versions of a bill, to express temporal constraints among laws, to express information about the authorship of an act, to model the approval process and so on. We could say that such an ontology focus primarily on legal documents as carrier of legal knowledge, not on the legal knowledge embedded in them.

It is not a case that such ontologies usually exploit bibliographic models already established outside the legal domain. The most common approach relies on the IFLA's Functional Requirements for Bibliographic Records (FRBR) [34]. FRBR is a general model for describing the evolution of any document. It works for both physical and digital resources, and it is not tied up with a particular metadata schema. FRBR distinguishes the concepts of work (a distinguishable intellectual or artistic creation), expression (the intellectual or artistic form that a work takes each time it is realized), manifestation (a physical embodiment of an expression of a work), item (a single exemplar of a manifestation of an expression). The Legislation.gov.uk project [5] developed a light ontology based on such distinction: a legislation is defined as work, different versions of the legislation (current, valid in a specific point in time, etc.) are expressions, different publishing formats for that expression (HTML pages, PDF, etc.) are

manifestations, while specific copies of those files are items.

The refinement of the FRBR levels in the context of legislative documents was first proposed in the MetaLex Ontology [14, 1]. MetaLex is a generic and extensible framework for the XML encoding of the structure of, and meta- data about, legal sources. It also includes an ontology that classifies bibliographic entities by using the FRBR layers and defines different types of reference between such entities. Great relevance is given to events and actions in order to model the activities of creation, publishing and revision (amendment) of legislative documents. The powerful model for legislative modifications of MetaLex has been adopted by the Legislation.gov.uk initiative too.

Other legal ontologies could be classified as *domain-specific document-centric* as they are explicitly designed to model documents used in a particular environment. It is the case of the Ontology of Greek Public Administration [44]. This ontology defines the types of documents produced by each unit (office) of the Administration, the organization of each unit and the publishing/approval workflow of these documents. The model is formalized in OWL and OWL-S [35] is used to define procedures. Another such ontology is the Legal Case Ontology[46], used to enable the semi-automatic generation of paper-based reports on legal cases. Part of the ontology is designed to capture the information that can be stored in these reports and to create reliable references between reports, that can be used for sophisticated information retrieval.

2.2 Content-centric ontologies

Most of the existing legal ontologies can be classified as *content-centric*. These ontologies primarily define legal concepts carried by legal sources, these concepts are used to identify and make explicit the meaning that can be found in the legal text present inside a document. Some of the things described by these ontologies are the relations between different concepts (e.g. “abigeat” is a particular kind of “theft”) or the legal purpose of a piece of text (e.g. “the first Monday of the year 2012” is a “date of enter into operation”; “attachment C of document 36” is a “minority report”). They are particularly important as they form the basis for legal knowledge acquisition and legal reasoning.

One of the most relevant to our work is LKIF-Core Ontology [29]. The Legal Knowledge Interchange Format (LKIF) is a format enabling the translation between legal knowledge bases, that use their own models and formalisms. Within the overall initiative, a central role is played by a core ontology that acts as main reference for such translation processes. The ontology defines ‘basic concepts of law’ and is organized in three levels, each covering orthogonal aspects of the domain of interest: Top, Intensional and Legal. The upper level is borrowed from the upper level of the LRI-Core ontology, which is in turn inspired by FOLaw [15]. Multiple upper ontologies could have been reused instead, such as DOLCE [36] and SUMO [48]. Authors preferred not to include them because of the ‘intentional inclination’ of LKIF-Core: this stresses more on communication and social interaction, rather than on physical aspects as other ontologies do. The intentional level of LKIF-Core models the behavior of agents, while the legal one introduces domain-specific concepts such as rights, powers, norms, legal agents, etc. Another core ontology has been reused within LKIF-Core, for the definition of powers: the Ontology of Fundamental Legal Concepts [42]. This is formalized in OWL-DL and introduces a set of fundamental legal concepts related to deontic modalities, different kinds of rights and different kinds of powers. Notice also that it imports some concepts from DOLCE and LRI-Core.

CLO (Core Legal Ontology) [26] is another widely adopted content-centric core

ontology. It defines legal entities and relations by exploiting classes and properties from the DOLCE foundational ontology [36]. The internal organization in three levels (Top-level, Core and Domain-specific), as well as its overall goal, makes CLO similar to LKIF-Core. One the other hand, the fact that the upper level uses a foundational approach is a very important difference. CLO is strongly based on the rich axiomatization and reification provided by DOLCE, and is an extremely powerful logically-sound framework. Another characterizing aspect of CLO is the rigorous application of design patterns, called Content Ontology Design Patterns (CODeP) [25].

While LKIF, CLO, and the other ontologies mentioned so far define core legal concepts, several examples of content-centric ontologies that focus on specific domains can be cited. The ontology of Dutch criminal law, OCN.NL [9], describes concepts related to depositions and hearings and 'anchors' these concepts to more general ones defined in LRI-Core (for instance: a criminal-court is a collection of legal agents); the ALIS IP ontology [19] aims at modeling the intellectual property and copyright domain, and is rooted in the LKIF-Core ontology (for instance, the concepts of 'work of mind' is linked to the more general concept of 'expression'). The OntoPrivacy ontology [17] is particularly relevant. It aims at modeling the Italian Privacy legislation domain and takes multiple core ontologies into account, referring to LRI-Core and CLO for the definition of top level concepts such as events, roles, mental states, etc.

In other cases, domain-specific (content-centric) ontologies have been developed without a direct reuse of core ontologies. For instance CLIME [51] regards the "design, construction, maintenance, repair, operation and inspection of ships" and includes a simplified model to describe artifacts, agents and functions.

2.3 Integration-centric ontologies

The third group includes ontologies that give much relevance to the integration of legal sources and concepts to external entities, that exist outside the legal domain and are independent of it. These ontologies are widely used, for instance, to model organizations and people involved in the creation and consumption of legal information. The Parliament Ontology (PARL) [8] and Central Government Ontology (CGOV) [2] are two examples developed within the *data.gov.uk* initiative [3]: they respectively model the organization of the UK Parliament and the Central Government. The nature of these ontologies led designers to reuse straightforwardly the FOAF (Fried of a Friend) ontology [16] to model people, their activities and their relationships with other people and documents. A similar approach has been used for the formalization of "Ontologia della Camera dei Deputati" (OCD) that describes the organization of the Chamber of Deputies in the Italian Republic [6]. The ontology uses FOAF to model people and connections among them, and Dublin Core terms [40] to further characterize deputies' roles and functions. Notice that OCD could have also been classified as 'document-centric' in our taxonomy since a large part of the ontology is devoted to describe legal documents such as acts, bills, parliament inquiry, decisions, and so on.

The ontologies mentioned so far, as well as their counterparts in other nations or similar models describing the structure of the Senate, local governments, etc., are characterized as *domain-specific* as they focus on the constitutional/political organization of Countries, instead of introducing more general legal concepts.

In all these cases, designers need to align different ontologies and express relations among terms and concepts: terms are sometimes used in different contexts with slightly different meanings, in other cases apparently different terms refer to the same concept, in others a term is a specialization of another one, and so on. A very common

solution, adopted by the aforementioned projects as well, consists of reusing the SKOS vocabulary [13] to express those relations.

This last point help us introducing another branch of legal ontologies we classify as 'integration-centric': the legal lightweight ontologies (or lexical ontologies). A lightweight ontology consists of a set of terms related to each other by semantic relations – such as hyperonymy, meronymy, instance-of, etc. – and aggregated in 'synset' when they refer to the same concept. Such ontologies, even if not rigorous and powerful as a formal ones, are able to capture the most relevant and used concepts in a given domain.

In the legal domain, it is worth citing Jur-IWN (Jur-ItalWordnet) [43]. The project aims at extending EuroWordnet, the European counterpart of WordNet, with legal information. It defines a large set of legal terms that can be recognized within texts and related to each other. These concepts are directly mapped into the main entities of CLO. While Jur-IWN primarily deals with core legal concepts, other lightweight ontologies have been proposed for specific domains. For instance, Legal Taxonomy Syllabus (LTS)[10] deals with the transposition of EU Directives into national laws. It proposes a simplified model to relate terms and concepts, and exploits a two-level approach: it defines a central lexical ontology defining concepts at EU level and, for each national language, a specific one that aligns those concepts to concepts and terms in the national document base.

The integration between textual resources and core ontologies have also been successfully exploited for micro-ontologies merging [23]. In this paper, authors propose a construction method to build legal ontologies by merging small sets of entities and relations (i.e. micro-ontologies) that describe a narrow domain-specific context. The overall process consists of three steps: (1) the extraction of the micro-ontologies from texts through NLP techniques, (2) the alignment of those two ontologies to the CLO core ontology through well-known alignment techniques and (3) the creation of a single coherent ontology linked to the previous ones.

We can conclude that, although different ontologies have widely different focuses, none of the ontologies we mentioned falls exclusively in one category. For instance, an ontology for criminal law might be classified as *content-centric* as it define concepts that are relevant even outside the legal documents they are referred by (for instance *crime*, *punishment*, *court*, etc.) but also *document-centric* as it defines notions like *act*, *article*, *comma* for modeling the legal sources where a criminal law was published in. Our intent was to identify the main characterization of each ontology, and to investigate how they are formalized and how they relate to each other and to external entities. Table 1 shows how the discussed ontologies can be classified according to their main and secondary traits.

3 Akoma Ntoso documents and legal ontologies

Akoma Ntoso [11] is an open legal XML standard for parliamentary, legislative and judiciary documents, originally promoted by the Kenya Unit of the United Nations Department for Economics and Social Affairs (UN/DESA) in 2004 and now the basis for the upcoming OASIS LegalDocumentML XML standard [4]. Akoma Ntoso documents, as any other legal documents, contain many references to other documents and to external entities. Akoma Ntoso is a multi-faceted format: it aims at being an useful format for legislation drafting and storage of legal documents in general; it also aims at being useful as format usable in all the moments of the life of a legal document, from

Ontology	Top	Core	Domain	Document	Content	Integration
Ontology of Greek Public Administration [44]			X	X		x
Legal Case Ontology [46]			X	X	x	x
CLIME [51]			X		X	
OntoPrivacy [17]			X		X	
ALIS IP [19]			X		X	
OCL.NL [9]			X		X	
LTS [10]			X			X
PARL [8]			X			X
CGOV [2]			X			X
OCD [6]			X			X
Legislation.gov.uk Ontology [5]		X		X		
MetaLex [14]		X		X		
Core Legal Ontology [26]		X			X	
FOLaw [15]		X			X	
LKIF-Core [29]		X			X	
LRI-Core [15]		X			X	
Ontology of Fundamental Legal Concepts [42]		X			X	
Jur-IWN [43]		X				X
Micro-ontologies [23]		X				X
FRBR [34]	X		x	X		
DOLCE [36]	X				X	
SUMO [48]	X				X	
SKOS [13]	X					X
FOAF [16]	X					X

Table 1: Classification of existing ontologies

its drafting to its publication to its archival.

Akoma Ntoso documents are strongly organized in layers, each addressing a specific problem. The *text* layer provides a faithful representation of the original content of the legal text, the *structure* layer provides a hierarchical organization of the parts present in the text layers, the *metadata* layer enriches underlying layers with ontological information so that semantic data can be shared and semantic tools can perform automatic reasoning on them. One of the peculiar features of Akoma Ntoso metadata layer is the ability to record multiple (and even contrasting) interpretations of the the same legal text; this feature make Akoma Ntoso documents extremely rich sources of legal information, not only plain legal texts.

In order for Akoma Ntoso documents to function as a knowledge base for reasoning tools, it is important that all the resources referenced implicitly or explicitly inside the legal text are property identified, marked and classified. The classification of these resources is a non trivial task that requires at least a shared vocabulary in order to be carried out. These days the most common ways to define shared vocabularies are ontologies or DB and XML schemas; as shown in the previous section, there is no shortage of available ontologies to describe legal concepts or para-legal entities. The Akoma Ntoso specifications, instead of forcing document authors to stick to one particular ontology, specify a set of general guidelines that an ontology should conform to in order to be used in conjunction with Akoma Ntoso documents. These informal guidelines are nicknamed the Akoma Ntoso *non-ontology*. section §5 shows a proof-of-concept ontology based on such guidelines. Before discussing the Akoma Ntoso non-ontology, it is important to analyze what are the reason that led the Akoma Ntoso designers not to use one of the many existing ontology.

3.1 What legal ontology for Akoma Ntoso?

The previous discussion has shown that legal ontologies differ radically in their objectives, scope and application domain. In such a complex panorama, a question arises whenever a new format to mark-up legal documents is introduced: which is the best ontology to use with it?

Such issue has been faced by the developers of the Akoma Ntoso project. Defining a set of concepts and properties required for reasoning on Akoma Ntoso documents would have been quite easy, alternatively, one of the many legal ontologies discussed in the section 2 could have been used. On the other hand, choosing a particular ontology would have limited the expressivity (and adoption) of the format. It is very easy, in fact, for different actors (lawyers, law makers, judges, citizens) to disagree on the interpretation of a legal text. Different actors may even need to model different legal concepts taken from different ontologies (e.g. date of enter into force as modeled in the LKIF-core ontology or High Court of South Africa as modeled in an ontology about the judiciary system in Africa). What the Akoma Ntoso project was seeking was a more flexible model, able to let designers connect specific assertions to more general concepts, relationships, properties and axioms defined in a external ontology of their choice.

Making Akoma Ntoso independent from a particular ontological model might not be enough. Even the format in which ontologies are stored might become an issue, since the long-term preservation of ontological data cannot be guaranteed by the adoption of technologies such as OWL and RDF. These technologies and their tools are widespread now, can we say the same in two or ten decades? The Akoma Ntoso specifications go to great lengths to argue why it is of fundamental importance not to

let current technologies creep into documents that are meant to be preserved and used for decades or centuries (as many legal documents are expected to). In summary, it is important not to force future users and toolmakers to rely on current technologies, unless they are very basic. Current newfangled technologies, while interesting and useful today, will surely be an hindrance in the future: probably they would be regarded as too imprecise, too cumbersome or just not understandable at all should the knowledge about those technologies and techniques be lost in the meantime, something that one can reasonably expect if one takes into account that there are already files created in the '80s that cannot be read because their format specification are now lost.

A better alternative to embedding data using one of the current ontological formats is to store the legal text in plain UNICODE text and use simple XML markup to link certain parts of the text to an URI. This solution allows for the simple retrieval of the text in case knowledge of the XML markup get lost in the future (as long as future toolmakers figure out that the text between angle brackets is to be ignored). At the same time, the use of links to external URIs allows for the evolution of the knowledge bases independently of the format and technologies used to store the legal text. Currently, the preferred way to extract such kind of knowledge from XML documents is to convert implied assertions into RDF using a GRDDL transformation. GRDDL (Gleaning Resource Descriptions from Dialects of Languages) [20] is a standardized way to glean RDF assertions from XML documents: it is a W3C Recommendation that explains how to extract semantic data from XML documents using a combination of one or more XSLT stylesheets, in order to obtain a new document containing those data expressed by RDF statements. We are believe that similar extraction techniques will be available in the future, probably not based on XSLT nor on RDF but made for whatever technology will be the most useful and widespread at the moment.

There is also another issue to consider when dealing with electronic legal documents: digital signatures. One of the main requisites for a good management of the chain of custody that can preserve trust in the use of electronic legal documents is that their authenticity or their conformity to the originals should be easily testable, as discussed in great detail in [33]. Digital signatures are one of the main technological means used to make sure that the documents have been stored and preserved unchanged. Should the ontological data be encoded using current formats and technologies, it would be impossible to update these documents to use newer ontologies or data formats without modifying their content and thus breaking their digital signatures.

Another problematic detail, often not addressed in current ontologies, is the fact that most of the entities properties are not rigid and immutable, they can change as the time passes by. Some of these properties can change frequently, for example an MP political association, while others can change more rarely, but it is important to record these changes even as infrequent as they are. The most common of such properties is a person name. Person names are often regarded as "rigid" properties and employed as stable IDs for the entity that represent the person with that name. Aside obvious problems such as homonyms, multiple transliterations from other alphabets and omitted middle names, the problem with using a person name as an ID is that it is not meant to be stable. In fact, in some cultures, married women change their surname to that of their husband and, in general, many people change their name and surnames during their life. A common mistake made when dealing with name changes is to *modify* the record of the person, overwriting the old name with the new one. This is a bad practice because, first, it loses information that may be important in some applications and, second, it fails to show that a change have happened. For properties that can change in time (most of them can) it is important to contextualize such properties with time

boundaries (en example of such contextualization will later be shown in section §5). Similar considerations can be made for subclass relations: it is common to see things like `MemberOfParliament` as subclass of `Person`. While it is true that every member of parliament is a Person[1], to be a member of parliament is not a rigid property of a person, a property that does not change as time passes by[2]. Given that "to be a member of parliament" is a temporary position, it is better described as an association of a time-bound role to a person rather than as a specialization of the class.

A further problem that is rarely taken into account by ontologies and datasets is the fact that almost none of the knowledge present in a legal document is composed of universally true statement of facts. It is wrong, for example, to extract from the XML file that contains the Akoma Ntoso version of the 2011 Taxation code a statement like "the fee for late payments is 400 euros"; what should be extracted is "Ben Marks (author of the version 2012-02-16 of the XML file containing the 2011 Taxation code) affirms, in that particular version of the XML file, that the Parliament (as author of the particular version of the Taxation code recorded in the XML file) states in section 9, paragraph 38, that the fee for late payments is 400 euros". The ability to annotate each basic statement with information about its authors and the versions of the file is of paramount importance. Without it, it would be impossible for multiple versions of a document to be stored in the same dataset or processed at the same time. It would also make it hard to tell apart official interpretations from additional interpretations coming for other sources.

Another problem inherently present when dealing with electronic legal documents is the trade-off between preservation for the future and usability with current tools. It is fundamental for a collection of documents to be useful in conjunction with current technologies and the associated tools. On the other hand, nobody likes to be stuck on older formats and technologies, especially implementers. For example, hardly anyone in 2012 would like to be constrained to use a particular DAML+OIL ontology, DAML+OIL reasoning features and none of the OWL features. At the same time this would be the situation if a format for legal documents developed before 2002 forced the adoption of the DAML+OIL technologies, then latest and best at that time. It can be argued that this kind of problems can be solved by updating the formats and the documents written using these older technologies. Unfortunately this undermines many of the requirements that needs to be in place to use and maintain digital signature or other measures that maintain an healthy "chain of trust" [33].

For all these reasons, the Akoma Ntoso specifications do not require the use of a particular ontology but specify, instead, a set of very broad guidelines that must be fulfilled by any ontology that one may want to use in conjunction with Akoma Ntoso documents. This guideline is indeed called the Akoma Ntoso *non-ontology*, to highlight the fact that it not a proper ontology but only a set of guidelines on top of which other can build their own ontologies.

3.2 The Akoma Ntoso *non-ontology*

The informal ontological structure defined by Akoma Ntoso for representing metadata is grounded in a basic set of concepts called Top Level Classes (TLC). The word "informally" is used because, on purpose, there is no mandated, exhaustive and shared ontology that defines these classes and the relation among them: what exists is a guide-

[1] except, maybe, for some senators during Caligula's reign in ancient Rome

[2] although some politicians try their best to make this come true

line that allows users (especially producers) of Akoma Ntoso documents to develop their own ontology according to their particular needs or to adopt one of the already existing ontologies, as long as compatible with the principles behind the TLC.

These top level classes do not have a formal definition, they only have a broad description, useful to identify in very general terms what is their purpose and how one TLC differs from another.

All this informality is needed to allow a great degree of flexibility in what can be expressed in the metadata layer of Akoma Ntoso documents, in order to adapt any legal document to many different ontological representation of concepts. It is the duty of a third party (e.g. the document creator or the document users) to associate a clear and formal semantics to each class using a specific formalism (e.g. OWL). This semantical detachment is an important feature that allows Akoma Ntoso to maintain documents understandable and consumable independently from the passing of time: future toolmakers ("The 'future toolmaker' is 10 years old now." [38]) will have clues about the intended meaning of a marker even in the unfortunate case the formal ontology is no longer available.

Any ontology that is used to model an Akoma Ntoso knowledge base must be compatible with its non-ontology. Compatibility with the non-ontology is a straightforward concept: an ontology is compatible as long as it is possible to associate every TLC to at least one of its "class" and this does not cause any "inconsistency". In this definition the terms "class" and "inconsistency" must be interpreted in a very liberal way. In lay terms, a class is a way to group individuals and inconsistency is a state of unrecoverable error in an automated reasoner. The exact meaning of these two terms depends on the technology used to implement the ontology: in OWL "classes" refers to a class and "inconsistency" to a Description Logic inconsistency, in Topic Maps "class" would be topic types and the meaning of "inconsistency" would depend on the tool used to defined the consistency constraints.

3.2.1 TLC: Top Level Classes

The basic set of concepts required by Akoma Ntoso are the TLC: Top Level Classes. There are 14 top level classes: 10 main TLC used to describe external resources and 4 document-related TLC based on FRBR.

The 10 main top level classes allow document creators to identify individual entities present in the document:

Concept any non-tangible notion or idea: e.g. "the approval of an act", "peace", "child".

Event something that "happened", "will happen", "may happen"or "have lasted": e.g. "World War II", "the coming into force of act 27", "Sunday 26th of August 2012".

Organization a recognizable group of individuals; organizations can be formal or informal, have a strong degree of internal organization or be completely anarchic, have their own name or be anonymous, have their own legal status or be impromptu groups: e.g. "the workers' union", "France", "the Socialist party", "the proponents of bill 103/32".

Person a human being, regardless whether they are alive or deceased, named or unnamed, fictional or real: e.g. "John Doe", "the person with ID RSSMRA72-H12L116B".

Place a location that can be referred to also using geographical coordinates: e.g. “the Rio river”, “Marrakesh”, “the entrance to the Black Forrest”.

Process a series of actions or steps directed to some end: e.g. “the approval of act 317”, “the election of the 11th president of the senate”.

Reference a reference to a resource; usually the resources referenced are other documents, at one of the FRBR level.

Role a part played by a person, an organization or an agent in general, in a certain situation: e.g. “member of parliament”, “speaker”, “head of office”, “bill proposer”.

Term a word or group of words whose meaning is defined in a formal and precise manner: e.g. “opening sentence”, “rebuttal”, “impeachment”.

Object anything that can referred to but that does not fit the other top level classes.

There are also 4 additional document-related top level classes that mimic the FRBR group-1 abstraction levels: Work, Expression, Manifestation and Item.

These TLCs have been devised in such a way that choosing which TLC fits which entity is an obvious decision in all cases except the most intricate and perverse.

Another of the main points behind the use of TLCs instead of complete and more refined ontologies is that this solution allows for a gradual evolution of the tools that operate over Akoma Ntoso documents. Having to describe each reference in terms of a TLC creates a minimal set of semantic data that can be used as a starting point to reason over these documents. With this small set of assertions even the simplest tools can answer basic queries like “what are the people referenced in this document?”. With a little more effort, smarter tools can extract more information from the same references (for instance dereferencing the entities’ URIs) and answer more complex queries like “what are the people referenced in this document that do not belong the the Lib party?”. Tools that are even more smart can use external KBs to link these assertions to related assertions in other documents in other datasets, making it possible to reason over the all the ontological information made available.

The simplicity of the TLC model allows systems to start from the basics, using simple tools and simple ontologies to describe the resources they references, and later, when the need arises, to switch over to more powerful tools and complicated ontologies.

4 How Akoma Ntoso deals with references

The independence of Akoma Ntoso from a specific ontology is only part of the solution. Further flexibility, maintainability and long-term preservation are guaranteed by the way Akoma Ntoso deals with external entities. The term ‘external entities’ indicates those entities that exist outside of a document and independent of it. They play a central role in legal documents: just think about persons referenced by laws, places, events, third-party documents, and so on.

Allowing designers to refer to such resources in a clear, unambiguos and flexible way is fundamental. More important, it is crucial that changes on the systems (and technologies) that store such entities are handled correctly and do not affect the document containing the references.

The solution proposed by Akoma Ntoso is in line with the idea of *non-ontology*. Instead of adopting an immutable naming structure, Akoma Ntoso implements a flexible scheme that is better suited for documents that are meant to be useful now but also that have to be preserved unchanged for decades.

The Akoma Ntoso naming convention is actually built on top of the *non-ontology* and TLC. Although Akoma Ntoso documents' authors are not forced to follow a precise ontology, in fact, they are still required to identify external entities with one of the top level classes described in section 3.2.1. Users are required to indicate one of these classes for each external entity (as type of the reference). This fact creates a minimal basis of knowledge made available to tools that operate on these documents.

Before discussing Akoma Ntoso URIs, along with some problems they try to avoid, it is helpful to recall how references are stored within documents.[3] This mechanism is based on 'external references' and consists of three parts:

- a piece of text is marked as a *reference text*, for example as a `<person>Mr Smith</person>`;
- a *reference handle* is added to the metadata (in the `<references>` section), pairing a local name to a URI, for example `<TLCPerson id="mp-smith3" href=/ontology/person/ak.smith-1968>`;
- the reference text is annotated with a *link* pointing to the reference handle's local name, for example `<person refersTo="#mp-smith3">Mr smith</person>`.

The most important piece of information in a reference is the content of the reference handles found in the metadata. Each of these handles is composed of three parts:

local name: connects without ambiguities elements in the body section to elements in the metadata section of the same document.

type: provides a first classification of the referenced entity, assigning a top level class to the entity.

URI: defines in a precise and univoque way the entity that is being referenced. It should be noted that *univoque*, injective in mathematical terms, does not imply *unique*: univoque means that a certain URI will always reference the resource Y, independently of the context and time in which the URI is found; differently, unique means that there exists only one URI X that can be used to identify the resource Y. A national identification number is unique because it will always refer to the same person; it is not unique because there are other way to identify that person, for example its name and birth date.

4.1 Naive URI naming convention

External entities can be divided in three big groups, partially overlapping: (i) entities that already have a well known URI, (ii) entities for which there exist some kind of univoque ID inside a DB and (iii) entities that do not have any ID or that are known to have an ID but such ID is not public nor accessible. It is common for legal documents to have references to entities from all these groups. This means that there must be

[3] This mechanism is problematic because it forces a single interpretation of a given piece of text, breaking the layer independence principle. Hopefully this problem will be addressed during the upcoming LegalDocumentML standardization.

a naming convention in place to guide document writers in the choice of which URI should be used to refer to which entity.

The most basic URI naming convention one can conceive is the following:

- for entities with a well known URI, that URI is used unchanged;
- for entities with an ID we construct a new URI using the scheme *{baseURI}/{ID}*;
- for entities that do not have any ID we create new URIs based on UUIDs.

The use of a URI naming convention like that in Akoma Ntoso documents can lead to various problems.

What would happen if some URIs will change in the future, for example moving from one domain to another? The easiest solution would be to replace the old URIs with their new counterparts.

Another problem that could arise is that some URIs will, over the time, be no longer valid or dereferenceable. In this case the most sensible remedy would be to substitute the old ones with new URIs that are valid and dereferenceable now.

Yet another problem is the advent of new standards to address entities that were previously addressed via ad-hoc or proprietary URIs. For this problem the easiest solution would be to stop using the old URIs in favor of the new ones and set up a redirection mechanism from the old URIs to the new URIs. Most of the solutions to these problems require changing the content of the documents, an action that would invalidate the integrity of the documents, both their legal integrity and their digital signatures. This is something that should be avoided as much as possible because of all the technological problems it poses but also because it creates problem in the management of the chain of custody of documents [33], with the possible drawback of reducing the trust placed by citizens and institutions in the electronic storage of legal documents.

4.2 Wanted URI properties

The main purpose of the references' URIs is to identify with precision an external entity. URIs that cannot be used to identify a resource in a reliable way are of no use in the context of Akoma Ntoso documents.

Another important purpose of the URIs is to act as identifiers that can be used to access the pointed entities, for example through dereferenciation or through some resolution mechanism. Please note that an URI can be precise enough to function as a pointer to an entity but, at the same time, lack enough information to be used to interrogate a system and retrieve the referenced entity.

A third purpose of URIs is to identify entities and make them accessible worldwide, not only inside a particular system.

In order to be stored in documents that are meant to be accessed and useful in the far future, URIs must change as little as possible. At the same time, they also need to be usable in current systems and over current protocols.

4.3 URIs based on the Akoma Ntoso TLC naming convention

The Akoma Ntoso specifications provide a mechanism that should be used to reference external entities. This mechanism is composed of a naming convention for URIs and a resolution layer.

The naming convention is based on the top level classes described in 3.2.1. Under this naming convention, URIs are built using the scheme `/ontology/`*`{top level class}/{global name}`*.

Each URI starts with the fixed part `/ontology/`. This first component identifies the pointed resource as an entity that is not a document (the Akoma Ntoso naming convention requires every document URI to begin with the identifier of the issuing country). Please note that this fixed first part forces all links to the external entities to be under the same directory `/ontology/` independently of the URI of the document in which these reference are found. In addition, this first part forces all the URI to be relative (more precisely, relative URIs with absolute paths), making it easy to deploy documents over any protocol (HTTP, HTTPS, URN, any future protocol) as the protocol name is not embedded in the stored URI but added at runtime during the URI dereferenceation process. Together these two things simplify the deployment of the resolution layer that will be illustrated later.

The second part of the URI identifies which top level class the entity belongs to: it should be the same of the reference handle where this URI is used, for example `<TLCOrganization id="#un" href="/ontology/organization/un">`.

The last part is a global name used to identify the exact entity we are referring to. Global names must, as the whole URI, but univoque identifiers: an entity can be referred to by more than one global name, but each global name must always identify the same entity, in every context. The Akoma Ntoso specifications suggests how to create good global names for each top level class.

By itself, the use of these URIs in Akoma Ntoso documents fulfills only the first of the requirements set out in Section 4.2. In order to achieve also all the others, one must add another component: a resolution layer. The role of this resolution layer is to link the URIs used in the documents to concrete entities (when available). Two types of mechanisms can be used to implement this resolution layer: a redirection service or a mapping dataset. The choice between which of these two mechanisms should be employed depends on the available technologies but also on political considerations that can constraint the number of published entities and their level of detail.

A redirection service is what is currently used in the semantic web and the linked data communities to link stable and immutable URIs to their current concrete representations. At the moment, such redirection services are implemented using chains of HTTP redirects (i.e. 300, 302, 303 and 307 redirects). Internally, the redirection service maintains a set of mappings from the published URIs to the latest or most refined versions of the entities being referenced.

The other mechanism, the mapping dataset, consists in the publication of the mapping database in an open format. Instead of performing the redirection on the server for each incoming URI, clients are informed of all the mappings between the published URIs and the URIs where the entity data can be found. For example, OWL-based datasets can use the `owl:sameAs` property to state that the URI `/ontology /organization/it.legislature-15` is in fact just a nickname for the entity whose canonical URI is `http://dati.camera.it/ocd/legislatura.rdf/repubblica_15`.

With these technologies in place, Akoma Ntoso URIs can act both as simple identifiers but also as meaningful handles to concrete data about all the entities referred by the documents. At the same time, should these additional services be no longer available, these URIs would still serve as good, precise and univoque identifiers inside immutable Akoma Ntoso documents.

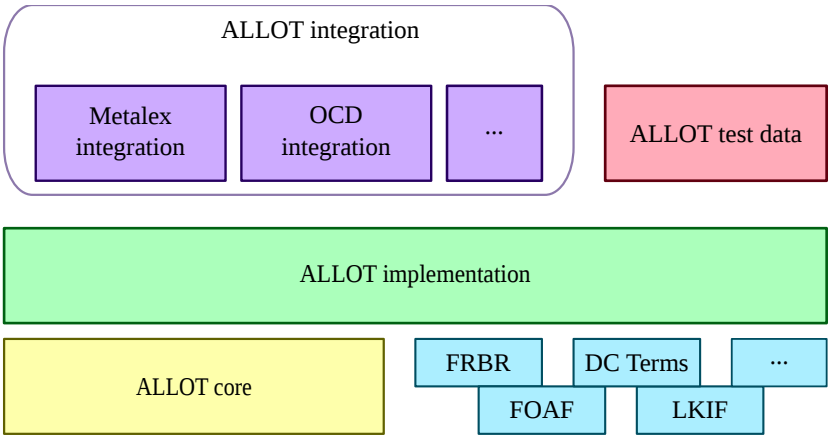

Figure 1: ALLOT layers

5 ALLOT: A Light Legal Ontology On TLC

ALLOT is a proof-of-concept ontology we developed based on the Akoma Ntoso non-ontology following the associated guidelines and best practices; it is available at http://akn.web.cs.unibo.it/allot/. The ALLOT ontology is meant to be used to describe in detail the references present in Akoma Ntoso documents, both documental and non-documental references. ALLOT can also used to bridge KBs extracted from Akoma Ntoso documents to KBs that use other ontologies such as Metalex or PARL (data.gov.uk). As expected, it uses the naming conventions discussed in the previous section and it is strongly tied up with the TLC.

ALLOT is composed of three layers, depicted in 1: *core* (where the TLC are declared and documented), *implementation* (where the TLC are implemented in terms of well known ontologies such as FOAF or SKOS) and *external integration* (small ontologies used to align the ALLOT implementation to existing datasets based on other ontologies).

In addition to these layers, there is also a *test dataset* that contains test data and examples. This dataset contains many entities described using the ALLOT ontology. We used this dataset to check the consistency of our ontology during its development. This dataset also constitutes a set of examples that can be used by implementers to see how the various pieces of the ALLOT implementation fit together and how to create similar datasets from internal databases.

5.1 The core layer

The core layer is a basic transposition of the TLCs in OWL classes. It contains one OWL class for each TLC and few other properties. It also documents what is the intended use of each TLC. The classes and properties included in this layer form the basic classification made available by any ALLOT-based dataset; query writers can be sure that every knowledge base based on ALLOT contains at least this "skeleton" of organization.

From a conceptual point of view, this layer is equivalent to the non-ontology described in section §3.2.

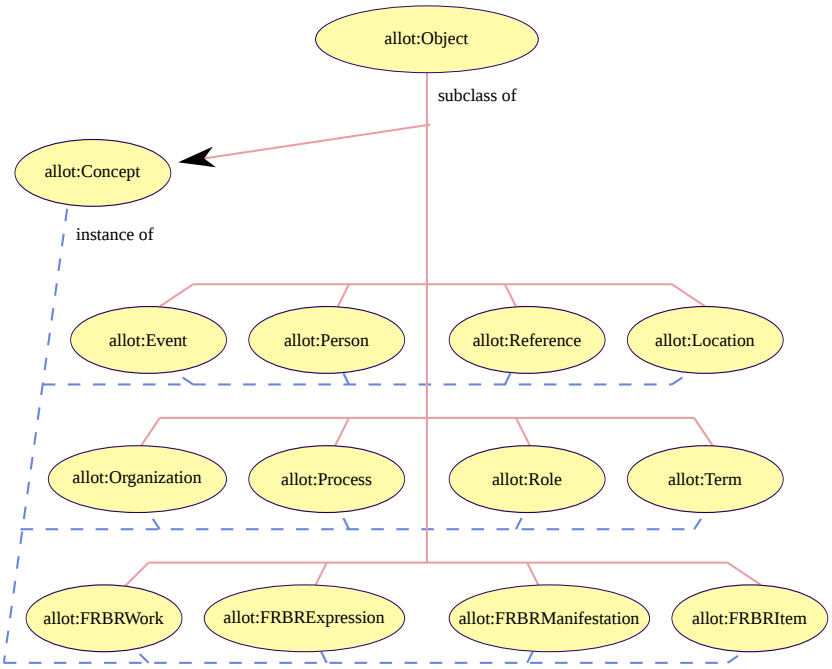

Figure 2: ALLOT core layer

From a more technical point of view, the ontology that realizes this layer exploits at least one of the advanced features of OWL2: name punning. All the top level classes are both OWL classes and individuals of the class `allot:Concept`. The rationale behind this is to allow designers to distinguish between, for example, the concept of what is a process (represented by the class allot:Process itself) from the concrete instances of these processes, i.e. the approval of act 345 (represented by individuals of type allot:Process).

5.2 The implementation layer

The implementation layer is the main component of ALLOT. Its role is to give a more precise definition to the top level classes originally defined in the core layer. This layers makes ALLOT-based datasets interoperable with current semantic web and linked data datasets. It also serves as an example of how to model an ontology for non-documental entities that does not incur in the problems described in section §3.1.

In practice, each top level class has been linked to similar classes from other ontologies, for example the TLC Person has been linked to FOAF Person and the birth of a person has been described in terms of BIO Events [22], as can be seen in figure 4.

Various kinds of "links" have been used: in some cases a TLC has been made subclass of other external classes; in other cases class or property equivalences have been used; in yet other cases external classes have been used as ranges or domains of newly defined properties. Various ontologies and vocabularies have been used: FOAF to describe people, BIO for biological events (i.e. birth and death), LKIF-core for intervals that can interoperate with other legal knowledge bases, PRO [47] for transient roles, FRBR [18] for documents, DC Terms [40] for various accessory datatypes, etc. In ad-

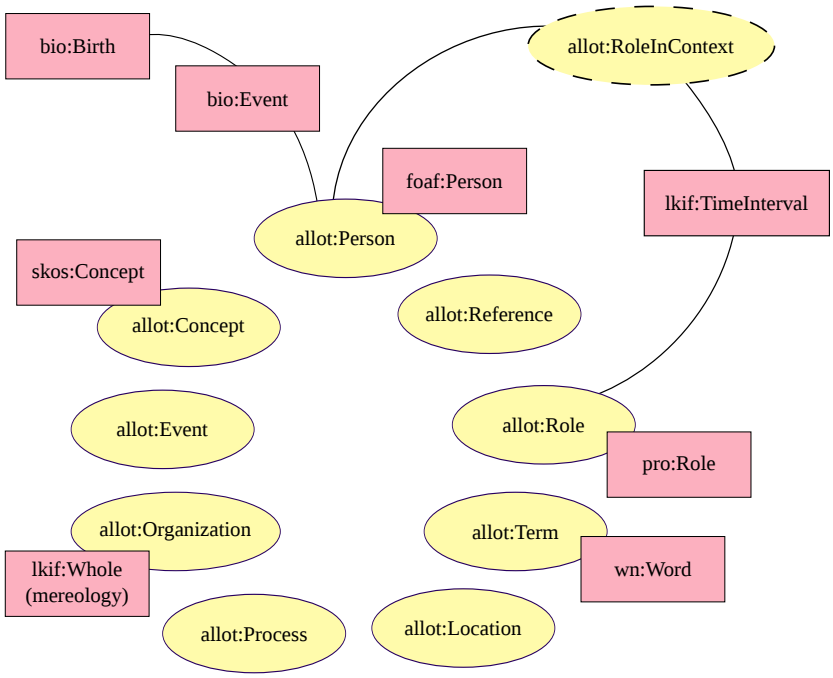

Figure 3: ALLOT implementation layer

dition to these external ontologies, some definitions has been developed from scratch as we have found no published ontology that could satisfy our criteria for fitness.

The development of this layer has been quite challenging: most of the ontologies we wanted to use to increase the interoperability of ALLOT-based datasets cannot be used together without making the datasets inconsistent, as we will show later in this section. These incompatibilities and issues made impossible the direct use of these external ontologies. We made adjustments to the way these ontologies are used to make ALLOT itself immune to the problems discussed in section §3.1. One of such adjustments is the use context objects (also called situations [25]) to keep track of properties whose value can change in the future, for instance names or political affiliations; an example of such context objects can be found in the ALLOT characterization of persons' names, as show in the example in figure 5. People in ALLOT are individuals of the FOAF:Person class. In the FOAF model each person can have more than one family name, this is accomplished attaching multiple `foaf:familyName` properties to a FOAF:Person individual. The problem with this approach is that one does not know when each of these name is or had been valid. On the other hand, in the ALLOT model each person (that is still a FOAF:Person) has one or more contextualized name objects that hold all the name information (the context of such objects can be an interval of time, a place or a social environment). This creates a problem: if we were to reuse the `foaf:familyName` property to link the contextualized names to the string with the actual surname, the reasoners could conclude that a contextualized name object, being the subject of a `foaf:familyName` assertion, is also a FOAF:Person leading to an inconsistency. In this case we did not reuse the `foaf:familyName` property and implemented our own set of name properties. We could have created a property chain that stated the equivalence between `foaf:familyName` and `allot:name`

```
Individual: rdfs:ke.susan-doe-i28i19k59382f

    Types:
        allot:Person

    Facts:
         bio:birth   test:birth-ke.susan-doe-i28i19k59382f,
         allot:name   rdfs:name-2-ke.susan-doe-i28i19k59382f,
         allot:name   rdfs:name-ke.susan-doe-i28i19k59382f,
         allot:role   rdfs:member-of-parliament-ke.susan-doe-↩
             ↪ i28i19k59382f
```

Figure 4: Example of an ALLOT Person individual

+ `allot:familyName` but OWL does not allow the use of data properties in property chains.

In some cases we were able to use existing ontologies without too many problems because their underlying conceptual model was similar to that of ALLOT. For example, roles are described using a mixture of PRO roles and a pool of ad-hoc roles created for the legal domain. Similarly, we described the various TLC FRBR classes as equivalent to their respective classes in the SPAR FRBR ontology, without the need for any additional modification.

6 Linking ALLOT to existing ontologies

ALLOT is meant to be used with datasets extracted from Akoma Ntoso databases. At the same time, there are now datasets published using other ontologies, for example Metalex or PARL. It is important to have a mechanism in place that can make all these datasets talk to each other. Interoperability is especially important when dealing with current international law: many of the entities and documents referenced in local law are often derived from international treaties or directives, or they refer to the same supranational entities. All these common entities could act as joining points for datasets, making it possible to query not only national or regional knowledge bases but also foreign datasets. However, this kind of interoperability requires that the datasets share at least part of their underlying ontologies or that their ontologies have been aligned somehow.

The integration layer of ALLOT is composed of independent modules, i.e. small ontologies that describe how to align ALLOT to other existing ontologies, connecting classes and properties of an external ontology to their equivalent in ALLOT. There are many techniques that can be used to achieve this goal; in our case we developed linking modules for the OCD (Ontologia Camera Deputati) ontology [6] and for the Metalex ontology [1] using a mixture of them. All the ontologies of the integration layer are available at http://akn.web.cs.unibo.it/allot-ext/.

The alignment with Metalex has been done using SKOS properties to describe the relation between ALLOT classes and properties to the respective classes and properties in Metalex. For the OCD ontology we developed, instead, a set of SPARQL CONSTRUCT queries to translate the available OCD individuals into equivalent individuals modeled using ALLOT.

Listing 1: "Record for Susan Doe in ALLOT"

```
Individual: test:ke.susan-doe-i28i19k59382f

    Types:
        allot:Person

    Facts:
        allot:name  test:name-ke.susan-doe-i28i19k59382f,
        allot:name  test:name-2-ke.susan-doe-i28i19k59382f

Individual: test:name-ke.susan-doe-i28i19k59382f

    Types:
        allot:NameInContext

    Facts:
        allot:interval  test:interval-name-ke.susan-doe-↩
            ↪ i28i19k59382f,
        allot:givenName  "Susan",
        allot:familyName  "Doe"

Individual: test:name-2-ke.susan-doe-i28i19k59382f

    Annotations:
        rdfs:comment "The name Susan Doe got after her first ↩
            ↪ marriage."

    Types:
        allot:NameInContext

    Facts:
        allot:interval  test:interval-name-2-ke.susan-doe-↩
            ↪ i28i19k59382f,
        allot:givenName  "Susan",
        allot:familyName  "Smith"
```

Listing 2: "Record for Susan Doe in FOAF"

```
Individual: test:ke.susan-doe-i28i19k59382f
    Types:
        foaf:Person
    Facts:
        foaf:firstName  "Susan"
        foaf:firstName  "Susan"
        foaf:surname "Smith"
        foaf:surname "Doe"
```

Figure 5: People names in ALLOT and in pure FOAF

6.1 SKOS alignment to Metalex

Metalex covers only part of the TLCs and, thus, of the ALLOT ontology: it deals only with documents as bibliographic things and events. Moreover, the conceptual model that underlies the Metalex ontology is very different from that of ALLOT: none of its classes, properties and implied relations have exact equivalents in ALLOT. The problem here is that, although, there is certain degree of overlap between the concepts expressed in Metalex and those of ALLOT, none of these similarities match the technical and semantic constraints imposed by OWL relationships such as `equivalentClass` or `sameAs`.

Anyhow, to document and preserve the similarities that exist between Metalex and ALLOT, we followed the recommendations of 'Ontology Matching' [24] and used SKOS properties to express in broad terms what is the relations between the classes of Metalex and those of ALLOT. For example we state that `metalex:Event skos:narrowMatch allot:Event`. In other words, we state that, even if it is not possible to create a mathematically precise relation between these two classes, tools should know that the definition of what is considered a `allot:Event` is similar but narrower than that of what constitutes a `metalex:Event`.

While these SKOS annotations are not directly usable by DL reasoners, there are additional tools and API (such as [21]) that can help automated analysis system in the use of datasets that contain eterogenous data, partly based on Metalex and partly based on ALLOT, allowing fuzzy matching between entities based on these two different ontologies.

6.2 SPARQL alignment to OCD

The alignment of ALLOT with OCD was not straightforward. There exist many intrinsic differences between these models that make a direct translation between them practically impossible. First of all, the ontology and the mixes the concept of person with that of role. For example, there are 8 different deputies whose name is "Massimo D'Alema", all with different titles, descriptions and duties, yet they all have the same birth date. Obviously the real meaning behind all these records is that there is a person called "Massimo D'Alema" that has been elected 8 times in the Camera. While this is easy to understand for a human being, it is very hard for a computer program to "collapse" all this slightly different records into a single one without resorting to ad-hoc fixes valid only for this dataset. Another problem is the way this ontology deal with memberships: similarly to people records, details about groups and affiliations are recorded in many different yet equivalent assertions, making the dataset unnecessarily big and impossible to navigate using standard tools and reasoners. As a final note, the 'depth' of the ontology should be noted as another problem. OCD, in fact, is overly flat compared to ALLOT: most of the classes are direct subclasses of `owl:Thing` and only a small subset of classes specialize others. Some glaring examples are: the class `ocd:incarico`, that indicates an 'assignment within a parliamentary group', the class `ocd:incaricoDiGoverno`, indicating 'governative assignments given to Deputies', and even the class `ocd:presidenteRepubblica`, representing the assignment of 'President of the Italian Republic'. These classes are all at the same level, directly linked to the class `owl:Thing`. Even relations among entities are underdeveloped within OCD: a very small set of horizontal relations exist, and each class is characterized by a few properties.

The correct management of the roles that change over the time was another tricky

issue in the alignment process. Let us discuss the case of the 'Deputy' role that a person may have multiple times in her life. This example is very helpful to explain how to link ALLOT with other ontologies that use similar approaches for this kind of modeling. OCD defines the class `ocd:person` to indicate 'a person who had some role within the Chamber of Deputies'. The biographic details about each person are expressed through FOAF [16] and Bio Events [22]constructs. The fact that a person is a Deputy is expressed by the property `ocd:rif_mandatoCamera` that connects an instance of a `ocd:person` to an instance of `ocd:mandatoCamera`, the class defining a parliamentary mandate. Thus, each person will be connected to multiple mandates if she was given multiple assignments during her political career.

The class `ocd:mandatoCamera` defines three properties to indicate: (i) the parliamentary legislature the mandate is valid for (`ocd:rif_leg` linked to the class `ocd:legislatura`), (ii) the election the mandate was given (`ocd:rif_elezione` linked to the class `ocd:elezione`) and (iii) the status of Deputy associated to that mandate (`ocd:rif_deputato` linked to the class `ocd:deputato`). The excerpts in figure 6 and figure 7 show examples of such definitions, omitting a lot of information not relevant at this stage.

In these examples, the deputy 'Giorgio La Malfa' as elected in the 15th Italian legislature (whose URI is `deputato.rdf/d3240_15`) is not linked in any way to the person 'Giorgio La Malfa' (whose URI is `persona.rdf/p3240`), elected several times as Deputy. At the same time, information about the mandate itself are kept in a separate entity, identified by the URI `mandatoCamera.rdf/mc8_3240_19790617`.

In our opinion, this ontology is an example of ontologies developed working too closely to a relational database whose schema has evolved "organically" over the years and without constant supervision. The connection with the internal organization of the backend is testified not only by the flat organization of classes but even by the names of the entities and their identifiers.

Regardless of the quality of the OCD, to work with it is mandatory is one is interested in using the data made available by the Italian lower chamber of the Parliament, as that is the ontology used to model their datasets. Given the irreconcilable differences in the way OCD and ALLOT model entities, we resorted link these two ontologies using mini-datasets generated on the fly using SPARQL CONSTRUCT. Basically, a SPARQL query is used to interrogate the OCD repository. Once the data is returned it is converted into a set of ALLOT-based assertions. These new assertions will then used to query ALLOT-based datasets and the OCD dataset at the same time using a single query.

To illustrate this technique, we will query the dataset about a member of parliament. As already pointed out, the crucial point is that the OCD dataset contains multiple instances, one for each time that person has been elected. This information is actually redundant and not completely matching the ALLOT model: the guidelines discussed in 3.1 would rather suggest to define a role 'Deputy' and to indicate the context (legislature and time period) a person was given that role. ALLOT assertions stating that 'a person was given a role (of Deputy) in a specific time frame can be generated straightforwardly. The TLC classes involved in such statements will be: `allot:Person` directly mapped into `ocd:person`, and `allot:Role` to indicate the role of Deputy, linked by the property `allot:RoleInContext` of the class `allot:Person`. The snippet in figure 8 shows the SPARQL query to achieve this goal from the OCD dataset.

We used a similar approach to align other (flat) pieces of data with the ALLOT hierar-

```
Individual: ocd:deputato.rdf/d3240_15

    Types:
        ocd:deputato

    Annotations:
        foaf:gender "male",
        foaf:firstName "GIORGIO",
        foaf:surname "LA MALFA",
        dc:title "GIORGIO LA MALFA, XV Legislatura della ↩
            ↪ Repubblica",
        rdfs:label "GIORGIO LA MALFA, XV Legislatura della ↩
            ↪ Repubblica",
        dc:description "Laurea in giurisprudenza ed economia ↩
            ↪ politica; docente universitario - professore ↩
            ↪ ordinario o di prima fascia"@it,
        ocd:rif_leg <http://dati.camera.it/ocd/legislatura.rdf/↩
            ↪ repubblica_15>,
        ocd:rif_incarico <http://dati.camera.it/ocd/incarico.rdf/↩
            ↪ i332_3240_28_20070419>,
        ocd:aderisce _:http://dati.camera.it/ocd/deputato.rdf#↩
            ↪ ader1

Individual: _:http://dati.camera.it/ocd/deputato.rdf#ader1

    Annotations:
        ocd:componente _:http://dati.camera.it/ocd/deputato.rdf#↩
            ↪ mem2,
        ocd:rif_gruppoParlamentare <http://dati.camera.it/ocd/↩
            ↪ gruppoParlamentare.rdf/gr332>,
        rdfs:label "MISTO (03.05.2006-28.04.2008)"

Individual: _:http://dati.camera.it/ocd/deputato.rdf#grp2

    Annotations:

        rdfs:label "MISTO-REPUBBLICANI, LIBERALI, RIFORMATORI ↩
            ↪ (19.03.2007-10.03.2008)",
        ocd:rif_componente <http://dati.camera.it/ocd/↩
            ↪ componenteGruppoMisto.rdf/cgm402>
```

Figure 6: OCD: a member of parliament record

```
Individual ocd:mandatoCamera.rdf/mc15_3240_19790617

    Types:
        ocd:mandatoCamera

    Annotations:
        ocd:rif_leg <http://dati.camera.it/ocd/legislatura.rdf/↩
            ↪ repubblica_15>,
        ocd:rif_deputato <http://dati.camera.it/ocd/deputato.rdf/↩
            ↪ d3240_15>,
        ocd:rif_elezione <http://dati.camera.it/ocd/elezione.rdf/↩
            ↪ e15_3240_2006>,
        ocd:motivoTermine "Fine Legislatura"
```

Figure 7: OCD: an electoral mandate record

```
PREFIX allot: <http://akn.web.cs.unibo.it/allot/>
CONSTRUCT {
        `iri(bif:concat("allot-ocd-bridge/",?title))`
                a allot:RoleInContext ;
                allot:RoleType `iri(bif:concat("allot-↩
                    ↪ ocd-bridge/",?roleName))` .

        `iri(bif:concat("allot-ocd-bridge/",?roleName))`↩
            ↪  a allot:Role .

        `iri(bif:concat("allot-ocd-bridge/",?name,"_",?↩
            ↪ surname))`
                a allot:Person;
                allot:role `iri(bif:concat("allot-ocd-↩
                    ↪ bridge/",?title))` ;
                allot:name `iri(bif:concat("allot-ocd-↩
                    ↪ bridge/",?title,"_",?name))` .

        `iri(bif:concat("allot-ocd-bridge/",?title,"_",?↩
            ↪ name))`
                a allot:NameInContext .
}
WHERE {
        SELECT DISTINCT ?name ?surname ?roleName ?title
        FROM <http://dati.camera.it/ocd/>
        WHERE {
                ?person a ocd:deputato ;
                        foaf:firstName ?name ;
                        foaf:surname ?surname ;
                        ocd:ruolo ?roleName ;
                        dc:title ?title .
        }
}
```

Figure 8: Example of SPARQL CONSTRUCT query for ALLOT-OCD

chical data model. The same conversion, for instance, was applied to model the government assignments (processing the property `ocd:rif_incaricoGoverno` connected to the class `ocd:incaricoGoverno`) or the membership of a Deputy to a Parliamentary group and/or to a Parliamentary committee. Space limits prevent us to go into details of such alignment. Some interesting issues, for instance, have also been raised by the translation of places and biographic information, as well as the normalization of missing data. Further details can be found in the online version of this ontology module.

7 Conclusions: best practices for new ontologies and schemas

The number of legal ontologies already developed and still being developed shows that this field is active both in the theoretical camp and in its practical implementations; the same can be said for the collections of electronic legal documents: their sizes grow daily. Notwithstanding all this interest, section 6 has shown how many of these ontologies are subject to flaws, even serious flaws, that can undermine the usefulness of these ontologies and document collections in the long haul.

It is important for newly-developed ontologies to avoid the mistakes made in the development of previous ontologies. To help the authors of new ontologies, we distill in this section a short list of suggestions and details that should be taken into account.

First of all, it is important for every new ontology to take into account modern modeling methodologies such as OntoClean [27, 28] and legal design patterns [25]. Although these methodologies have some difficulties in modeling certain relations, it is better to start the development using them and then make exception to their prescriptions only later in the development process, when there are solid basis to justify deviations from the methodology.

A second important requisite for current ontologies is to be "compatible" or "linkable" with other ontologies employed nowadays in the Linked Data world and existing data silos. For example it is fundamental to have a way to relate people's instances to FOAF classes, as there is widespread support for those classes in many tools. The problems found in creating links with these widespread but less expressive classes can be alleviated using ontology alignment techniques [37, 39, 45, 49, 50, 31], as we discussed in section §6.

Another detail that is frequently overlooked is the fact that most entity properties are time bound, i.e. their value can change, even more than once, as the time passes by. It is important to record these changes and to allow past and current information to coexist at the same time in the same dataset. Similarly, temporal roles such as being a member of parliament are often erroneously codified using a subclass relation instead of an association between a time-bound role record and a person record.

A last problem that developers of new ontologies should address is the fact that most of the assertions that can be extracted from legal documents are not universal truths. These assertions are assertions made by a particular actor (say, the author of the document) in a precise context (a particular version of the document) at a precise time (the date and time of the document). As discussed in section §3.1, there is no standardized way to record and transmit this information in current semantic technologies; ad-hoc descriptions must be devised and used, possibly following some of the already existing design patterns[30].

A possible justification for the fact that many ontologies do not tackle all these problems is that current semantic technologies like RDF or OWL lack the features needed to express in a simple way the kind of statements these details require. Design patterns and other technological workaround are often used to overcome the lack of such features. The problem with these workarounds is that they are harder to use compared to their simpler counterparts that cannot preserve all the information contained in the original legal text. However, the understandable designers' desire to keep the ontologies "simple" for their users, make the designers stick to the simplest features of the current technologies, leaving out of the current ontologies all the refinements that are needed to address the concerns raised in this. Our hope is that future technologies will have more expressive features to make it possible to express in simple ways all the accurate data that can be extracted from legal texts.

References

[1] CEN MetaLex: Open XML Interchange Format for Legal and Legislative Resources. 2.1, 6

[2] CGOV: Central government ontology, an ontology of UK central government. 2.3

[3] data.gov.uk: opening up government. 2.3

[4] Oasis legaldocumentml technical committee charter. 3

[5] The official home of uk legislation. 2.1, 2.3

[6] Ontologia Camera dei Deputati, an ontology for the Italian Chamber of Deputies. 2.3, 6

[7] Open definition: Defining the open in open data, open content and open services. 1

[8] PARL: Parliament ontology, an ontology of UK parliament. 2.3

[9] Joost Breuker Abdullatif, Joost Breuker, Abdullatif Elhag, Emil Petkov, and Radboud Winkels. Ontologies for legal information serving and knowledge management. In *In Legal Knowledge and Information Systems, Jurix 2002: The Fifteenth Annual Conference. IOS*, pages 73–82. IOS Press, 2002. 2.2, 2.3

[10] Gianmaria Ajani, Leonardo Lesmo, Guido Boella, Alessandro Mazzei, and Piercarlo Rossi. Terminological and ontological analysis of european directives: multilinguism in law. In *Proceedings of the 11th international conference on Artificial intelligence and law*, ICAIL '07, pages 43–48, New York, NY, USA, 2007. ACM. 2.3

[11] Cervone L. Palmirani M. Peroni S. Vitali F. Barabucci, G. Multi-layer markup and ontological structures in akoma ntoso. In *Proceeding of the International Workshop on AI approaches to the complexity of legal systems II (AICOL-II).*, Rotterdam, The Netherlands, 2009. 3

[12] Gioele Barabucci, Luca Cervone, Monica Palmirani, Silvio Peroni, and Fabio Vitali. Multi-layer markup and ontological structures in Akoma Ntoso. In *Proceedings of the 2009 international conference on AI approaches to the complexity of legal systems: complex systems, the semantic web, ontologies, argumentation, and dialogue*, AICOL-I/IVR-XXIV'09, pages 133–149, Berlin, Heidelberg, 2010. Springer-Verlag. 1

[13] Sean Bechhofer and Alistair Miles. SKOS Simple Knowledge Organization System Reference. Recommendation, W3C, August 2009. http://www.w3.org/TR/2009/REC-skos-reference-20090818/. Latest version available at http://www.w3.org/TR/skos-reference. 2.3

[14] A W F Boer, R J Hoekstra, E De Maat, F Vitali, Monica Palmirani, and B Ratai. Metalex (open xml interchange format for legal and legislative resources). *Management Center*, 2010. 2.1, 2.3

[15] Joost Breuker and Rinke Hoekstra. Epistemology and ontology in core ontologies: FOLaw and LRI-Core, two core ontologies for law, 2004. 2.2, 2.3

[16] Dan Brickley and Libby Miller. FOAF Vocabulary Specification 0.98. Namespace document, August 2010. 2.3, 6.2

[17] Amedeo Cappelli, Valentina Bartalesi Lenzi, Rachele Sprugnoli, and Carlo Biagioli. Modelization of domain concepts extracted from the italian privacy legislation. In *Proceedings of Seventh International Workshop on Computational Semantics*, IWCS-7, pages 305 – 308, January 2007. 2.2, 2.3

[18] Paolo Ciccarese and Silvio Peroni. Essential FRBR in OWL2 DL, 2011. 5.2

[19] Migle Laukyte Regis Riveret Claudia Cevenini, Giuseppe Contissa and Rossella Rubino. Development of the ALIS IP Ontology: Merging Legal and Technical Perspectives. In *IFIP 20th World Computer Congress, Proceedings of the Second Topical Session on Computer-Aided Innovation*, volume 277 of *Computer-Aided Innovation (CAI)*, pages 169–180, September 2008. 2.2, 2.3

[20] Dan Connolly. Gleaning Resource Descriptions from Dialects of Languages (GRDDL). Recommendation, W3C, September 2007. http://www.w3.org/TR/2007/REC-grddl-20070911/. Latest version available at http://www.w3.org/TR/grddl/. 3.1

[21] Jérôme David, Jérôme Euzenat, François Scharffe, and Cássia Trojahn dos Santos. The alignment api 4.0. *Semantic Web – Interoperability, Usability, Applicability*, 2(1):3–10, 2011. 6.1

[22] Ian Davis and David Galbraith. BIO: A vocabulary for biographical information. 5.2, 6.2

[23] Sylvie Despress and Sylvie Szulman. Merging of legal micro-ontologies from europen directives. *Artif. Intell. Law*, 15(2):187–200, June 2007. 2, 2.3

[24] Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2007. 6.1

[25] Aldo Gangemi. Introducing pattern-based design for legal ontologies. In barabucci, editor, *Proceedings of the 2009 conference on Law, Ontologies and the Semantic Web: Channelling the Legal Information Flood*, pages 53–71, Amsterdam, The Netherlands, The Netherlands, 2009. IOS Press. 2.2, 5.2, 7

[26] Aldo Gangemi, Maria-Teresa Sagri, and Daniela Tiscornia. A constructive framework for legal ontologies. In *Law and the Semantic Web*, pages 97–124, 2003. 2.2, 2.3

[27] Nicola Guarino and Christopher Welty. Ontological analysis of taxonomic relationships. In *Proceedings of the 19th international conference on Conceptual modeling*, ER'00, pages 210–224, Berlin, Heidelberg, 2000. Springer-Verlag. 7

[28] Nicola Guarino and Christopher A. Welty. A formal ontology of properties. In *Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management*, EKAW '00, pages 97–112, London, UK, UK, 2000. Springer-Verlag. 7

[29] Rinke Hoekstra, Joost Breuker, Marcello Di Bello, and Alexander Boer. LKIF Core: Principled Ontology Development for the Legal Domain. In *Proceedings of the 2009 conference on Law, Ontologies and the Semantic Web: Channelling the Legal Information Flood*, pages 21–52, Amsterdam, The Netherlands, The Netherlands, 2009. IOS Press. 2.2, 2.3

[30] C. Huitfeldt, S. Peroni, and F. Vitali. Documents as timed abstract objects. *to appear in the Proceedings of Balisage Series on Markup Technologies, The Markup Conference 2012*, 2012. 7

[31] Jaehong Kim, Minsu Jang, Young-Guk Ha, Joo-Chan Sohn, and Sang Jo Lee. Moa: Owl ontology merging and alignment tool for the semantic web. In *Proceedings of the 18th international conference on Innovations in Applied Artificial Intelligence*, IEA/AIE'2005, pages 722–731, London, UK, UK, 2005. Springer-Verlag. 7

[32] Daniel Lathrop and Laurel Ruma. *Open Government: Collaboration, Transparency, and Participation in Practice*. O'Reilly Media, Inc., 1st edition, 2010. 1

[33] Heather MacNeil. Providing grounds for trust ii: The findings of the authenticity task force of interpares. *Archivaria*, 54:24–58, 2002. 3.1, 4.1

[34] Olivia M.A. Madison. The IFLA Functional Requirements for Bibliographic Records: International standards for universal bibliographic control. *Library Resources & Technical Services*, 44(3):153–159, July 2000. 2.1, 2.3

[35] David Martin. OWL-S: Semantic Markup for Web Services. Member submission, November 2004. http://www.w3.org/Submission/2004/SUBM-OWL-S-20041122/. Latest version available at http://www.w3.org/Submission/OWL-S/. 2.1

[36] Claudio Masolo, Laure Vieu, Emanuele Bottazzi, Carola Catenacci, Roberta Ferrario, Aldo Gangemi, and Nicola Guarino. Social roles and their descriptions. pages 267–277. AAAI Press, 2004. 2.2, 2.3

[37] Natalya Fridman Noy and Mark A. Musen. An algorithm for merging and aligning ontologies: Automation and tool support. In *In Proceedings of the Workshop on Ontology Management at the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*. AAAI Press, 1999. 7

[38] Monica Palmirani and Fabio Vitali. Akoma Ntoso Users. 3.2

[39] Rahul Parundekar, Craig A. Knoblock, and José Luis Ambite. Linking and building ontologies of linked data. In *Proceedings of the 9th international semantic web conference on The semantic web - Volume Part I*, ISWC'10, pages 598–614, Berlin, Heidelberg, 2010. Springer-Verlag. 7

[40] A. Powell, M. Nilsson, A. Naeve, and P. Johnston. Dublin core metadata initiative: Abstract model, 2005. White Paper. 2.3, 5.2

[41] Dorian Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann, 1999. 1

[42] Rossella Rubino, Antonino Rotolo, and Giovanni Sartor. An owl ontology of fundamental legal concepts. In *Proceedings of the 2006 conference on Legal Knowledge and Information Systems: JURIX 2006: The Nineteenth Annual Conference*, pages 101–110, Amsterdam, The Netherlands, The Netherlands, 2006. IOS Press. 2.2, 2.3

[43] Maria-Teresa Sagri and Daniela Tiscornia. Metadata for content description in legal information. In *Proceedings of the 14th International Workshop on Database and Expert Systems Applications*, DEXA '03, pages 745–, Washington, DC, USA, 2003. IEEE Computer Society. 2.3

[44] Ioannis Savvas and Nick Bassiliades. A process-oriented ontology-based knowledge management system for facilitating operational procedures in public administration. *Expert Syst. Appl.*, 36(3):4467–4478, April 2009. 2.1, 2.3

[45] François Scharffe, Jérôme Euzenat, and Dieter Fensel. Towards design patterns for ontology alignment. In *Proceedings of the 2008 ACM symposium on Applied computing*, SAC '08, pages 2321–2325, New York, NY, USA, 2008. ACM. 7

[46] Yueting Shen, Robert Steele, and John Murphy. Building a Semantically Rich Legal Case Repository in OWL. In *Proceedings of AusWeb08, the Fourteenth Australasian World Wide Web Conference*, AusWeb08, April 2008. 2.1, 2.3

[47] David Shotton and Silvio Peroni. PRO, the publishing roles ontology, 2012. 5.2

[48] John F. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks/Cole, August 2000. 2.2, 2.3

[49] Bach Thanh Le and Rose Dieng-Kuntz. A graph-based algorithm for alignment of owl ontologies. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, WI '07, pages 466–469, Washington, DC, USA, 2007. IEEE Computer Society. 7

[50] Rubén Tous and Jaime Delgado. A vector space model for semantic similarity calculation and owl ontology alignment. In *Proceedings of the 17th international conference on Database and Expert Systems Applications*, DEXA'06, pages 307–316, Berlin, Heidelberg, 2006. Springer-Verlag. 7

[51] Radboud Winkels, Alexander Boer, and Rinke Hoekstra. CLIME: Lessons Learned in Legal Information Serving. In Frank van Harmelen, editor, *ECAI*, pages 230–234. IOS Press, 2002. 2.2, 2.3