Enhancing Publication Description with Resources Metadata

Christian Cote,
ELICO, University Lyon III,
6, cours Albert-Thomas
69008 Lyon

Richard Dapoigny,
LISTIC, Polytech'Annecy-Chambéry
University of Savoie, Po. Box 80439,
74944 Annecy-le-vieux cedex,

Caroline Wintergerst
MODEME, University Lyon III
6, cours Albert-Thomas, Po. Box 8242
69355 Lyon cedex 08

Abstract

In this paper, we suggest to increase the quality and the precision of a document description using publication's context description. Today, a lot of linguistic resources are both available on line and described by specific metadata. We first integrate them into an ontology which describes how linguists consider their primary data and tools. Then, we add to this ontology an inference system based on the information flow theory in order to establish causal relations between heterogeneous data. The result of the inference is characterized by a small set of properties which are embedded into three sequences of metadata enhancing the usual metadata describing publications.

1 Introduction

Primary data (e.g., corpora, maps, protocols) and tools (e.g., annotation tools, syntactic analyzers in linguistics) are an essential part of scientific activity and constrain the issues in information retrieval, e.g., for on-line digital libraries. While these two notions add interpretative value to existing digital libraries tools, their descriptions are not exploited to increase the quality of the content description of their related primary data. To address this challenging problem, we suggest to enhance the quality of content description for publications by using their related tools and primary data descriptions.

Despite the increasing role of ontologies in Information Retrieval, these ontologies rarely refer to metadata excepted in (Chang et al. 2007) in which a core ontology, constructed from the metadata of learning resources aims at representing some general semantics of the domain that goes beyond a domain ontology. In order to take advantage of resources descriptions, a domain ontology and an inference system based on the Information Flow (IF) theory are proposed, both for enhancing the description of the publications and for proving the influence of these resources on information retrieval. The domain ontology characterizes how linguists represent their primary data and tools while the IF-based mechanism is supplied with the ontology. We demonstrate that some features of the publications can be described by the conveyance of related features of the resource description in the domain of publication descriptions. This property will convert the role of metadata into an information retrieval process, i.e., metadata allow conveyance of descriptive contents among related documents (e.g., corpora and publications) in a way to increase the relevance of the classified issues. We overcome the actual limits of publication content description (cf. Dublin Core or "open archives" formats) by the integration of publication context descriptions.

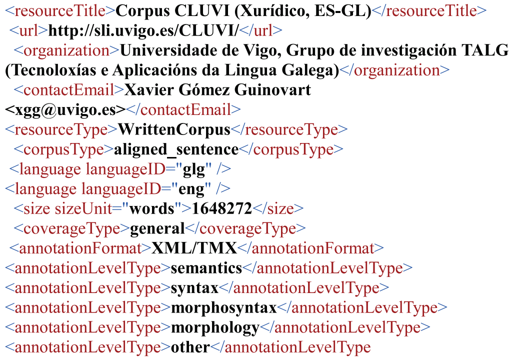
2 Notion of Linguistic Resource

The present work is restricted to *linguistic resources* defined in a web context as primary data which are specifically built for a linguistic analysis (e.g., a collection of newspapers of the nineteenth century with semantic annotation). Raw data (e.g., a language record without description elements) and primary data (e.g., data with description elements like annotation) compose corpora. Since annotated corpora are included in the collection, then we must also include tools of annotation, clustering, etc. In summary, a *linguistic resource* characterizes any structured data and any software functionally structured for a linguistic use (e.g., a tool for lexicon/terminology extraction and terminology management in English, or an east European computational lexicon with morphologic information and morpho-syntactic descriptions to manage lexical resources).

This definition of *linguistic resources* requires both the acceptation of the diversity of lexical entries and a common description. To solve this duality, we focus on their descriptive metadata which are consensual in the frame of linguistic because they have been elaborated in a way to represent the specificities of digital entities w.r.t. Dublin Core and bibliographic metadata. Linguistic communities have their proper specification of terms and acceptance but metadata have a descriptive goal which overlays the communities' specific definitions and terms. These specificities are described at a lexicon level. Let us consider the following publication:

```
@INPROCEEDINGS{Girju07improvingthe,
AUTHOR = {Girju, Roxana},
TITLE = {Improving the interpretation of noun phrases
with crosslinguistic information},
BOOKTITLE = {Proceedings of the 45th Annual Meeting of
the association on Computational Linguistics},
YEAR = {2007},
pages = {568-575}
}
```

This running example will illustrate how the proposed approach yields the characterization of a publication. The name of the corpus is identified into the publication (here, ”CLUVI-Lexicon”) and is further associated to the corpus description into the ontology.



```
<resourceTitle>Corpus CLUVI (Xurídico, ES-GL)</resourceTitle>
 <url>http://sli.uvigo.es/CLUVI/</url>
  <organization>Universidade de Vigo, Grupo de investigación TALG
(Tecnoloxías e Aplicacións da Lingua Galega)</organization>
  <contactEmail>Xavier Gómez Guinovart
<xgg@uvigo.es></contactEmail>
<resourceType>WrittenCorpus</resourceType>
  <corpusType>aligned_sentence</corpusType>
 <language languageID="glg" />
<language languageID="eng" />
  <size sizeUnit="words">1648272</size>
  <coverageType>general</coverageType>
 <annotationFormat>XML/TMX</annotationFormat>
<annotationLevelType>semantics</annotationLevelType>
<annotationLevelType>syntax</annotationLevelType>
<annotationLevelType>morphosyntax</annotationLevelType>
<annotationLevelType>morphology</annotationLevelType>
<annotationLevelType>other</annotationLevelType
```

Figure 1: The CLUVI Bio-Lexicon description.

3 Basis of Information Flow

Prior to explain how objects can be processed with IF in the linguistic framework, some basic knowledge is required. The foundational principle of the IF theory (Barwise and Seligman 1997) states that information flow results from regularities connecting parts of a system and make the existence of this flow possible. Regularities and connections are respectively modeled in a system with the help of two concepts, classification and infomorphism.

Classifications result from a categorization of entities within a system in objects (also called instances or tokens) and types (also called properties or classes). Major works stem from the idea that a classification of information must exist in each of the components of a distributed system (Barwise and Seligman 1997; Kent 2000). Above this assumption, the authors give a domain-neutral definition of a classification:

Definition 1. *A classification A is a triple $\langle tok(A), typ(A), \models_A\rangle$, which consists of:*

1. *a set $tok(A)$ of things to be classified known as the instances of A,*
2. *a set $typ(A)$ of things used to classify the instances, the types of A,*
3. *a binary classification relation $\models_A$ between $tok(A)$ and $typ(A)$.*

The notation $a \models_A \alpha$ must be understood as ”instance a is of type α in A”. Classifications are related through infomorphisms.

Definition 2. *Let A and B be IF classifications. An infomorphism $f = \langle f^\wedge, f^\vee\rangle : A \rightleftarrows B$ is a contravariant pair of functions $f^\wedge : typ(A) \rightarrow typ(B)$ and $f^\vee : tok(B) \rightarrow tok(A)$ which satisfies the fundamental property:*

$$f^\vee(b) \models_A \alpha \text{ iff } b \models_B f^\wedge(\alpha) \qquad (1)$$

for each $\alpha \in typ(A)$ and $b \in tok(b)$

Infomorphisms can be combined to form another infomorphism and give rise a particularly suitable model for data integration problems. Connecting different classifications relies on the assumption that for each token, we are speaking of the same object through infomorphisms. This presupposition can be modeled with the notion of information channel.

Definition 3. *An IF channel consists of two classifications A_1 and A_2 connected through a core classification C by means of two infomorphisms f_1 and f_2.*

Given two classifications A_1 and A_2, then these classifications can be combined into a single classification $A_1 + A_2$. The tokens of $A_1 + A_2$ consist of pairs $\langle a_1, a_2\rangle$ of objects from each, whereas its types consist of the types of both (if there are common types, then they are distinguished). Each property of a corpus (or metadata instantiated) has no systematic relation to publication description. How we decide that some properties are relevant and some other not? The relevance of some metadata is defined by the fact it conveys information from a type of metadata to another. Information flow allows the representation of inferences between heterogeneous descriptive classes at the medium level of the ontology. The issues of these inferences are conveyed to the publication descriptions in a way to propose a new set of metadata. These inferences are controlled by the relations of the high level of the ontology. According to the IF theory, inferences are made of infomorphisms while the high level control is achieved with a channel.

4 Ontology of Linguistic Data and Tools

We elaborate an ontology founded on three levels of description of these primary data: contained linguistic facts, bibliographic description (or metadata) and tool. Metadata integrate a middle level of abstraction of the ontology. We are concerned only by a family of tools used into the linguistic discipline for data analysis. This empirical foundation entails a bottom-up approach together with a realistic foundation. However, realism fails to represent in that case the high level of abstraction: some knowledge of a high abstraction level is attached to any process of tool elaboration. The domain we represent is a collection of material objects (essentially digital) used by linguists to achieve their search. They are defined like tools following an anthropological point of view (Keller and Keller 1996) in which a tool is an externalization of a cognitive competence and is integrated into a productive act, considered as an accomplishment (anything material that contributes to a research task is defined as a tool: it designates both a corpus and a software tool, but a corpus contains data or material fact for the use of a tool and can be described by the linguistic facts it contains).

The ontology is not founded on the identification and the description of the domain entities and their taxonomy, but on technical and functional properties common for a family of tools used in a particular activity. The relations that are represented by the ontology affect material and functional parts of an object and not different objects. The relations represent how heterogeneous constituents build a homogeneous object considering its functionalities. The ontology is systematically in relation to linguist professional activity: requests are formulated using the lexicon of a specific linguistic community.

A three levels ontology represents distinctively the lexicon of the discipline (characterizing the possible content of the resource), the descriptive elements (or components) of the resource and the functional characterization of the resource (why and how the resource has been built). These levels are justified by the logic of use of the ontology: requests (about publications) are formulated using the lexicons of the discipline. Lexicon reflects the diversity of terminology that characterizes the linguists'community. The retrieved publications are analyzed by a resource name extractor. Extracted resource names are conveyed to the resource database connected to the ontology. The model verifies the relevance of the identified resource to the lexicon. An inference (or information flow) allows the conveyance of resource properties to the description of the retrieved publications in a way to increase their description, classification and at last the user choice. The figure 2 represents how the model works.

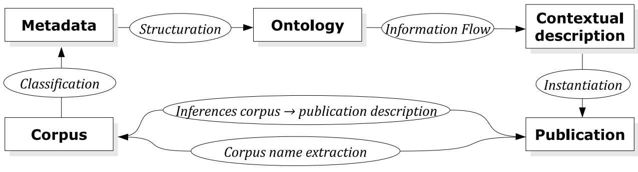


Figure 2: The model overview.

The model is characterized by three levels: lexical, descriptive (or predicative) and conceptual.

- Low level of abstraction: lexical categories that correspond to all the terms used to represent a linguistic fact (anaphora, tense verbs, preposition, etc.). If a linguistic fact is contained into a corpus, then it can be connected to the corpus description, or medium level of the ontology. These categories are publications keywords and represent instances of linguistic facts observed into primary data.
- Medium level: semantics of the primary data description. It contains established metadata that describe how primary data are presented and represented. Metadata describe resources considering different types of data properties (language, annotation, discourse, genre, etc). They are less precise than linguistic facts because primary data contain an unidentified set of facts and then it is impossible to identify all these. Medium level of the ontology is founded on heterogeneous properties because these properties are relevant for different uses and tasks (different linguistic research, data management, comparisons, etc). Then they are partial too.
- High level of abstraction: conceptual foundations of a process description. It characterizes how and why we have this sort of data presentation. This is a conceptual level that unifies the heterogeneous descriptive classes presented at the medium level. It explains the coherence between these classes by their relations in the process of resource building. This level presents a process, and more precisely a process of conception. It can be used for the organization of other domains.

4.1 Low level: lexicons and categories of corpus components

Domain ontologies are generally founded on taxonomy, lexicon or terminology. They characterize the low level of abstraction and the ontology is a progressive abstraction from this level. The low level contains denominations of linguistic facts described by publications and contained into primary data. These lexical entries denominate diverse scientific objects that are represented distinctively among the different corpora and tools (because primary data and tools depend on theoretical, methodological and technical choices).

4.2 Middle level: metadata and properties of categories

Metadata are the new external structured data integrated at the middle level of the ontology. Assuming that the characterization of the content of a resource results from its use, metadata describe how linguistic facts (i) are collected, (ii) processed and (iii) ascribed to a type/instance (or property/value) classification. During the past decades, corpus and NLP research communities have striven for developing annotation schemes and guidelines for metadata. Some of these annotation schemes have been proposed as quasi standards in these communities, such as the OLAC[1] and IMDI[2] sets of metadata, and the propositions of BAMDES initiative[3] integrating many catalogs. We use a synthesis of these annotation schemes to build the domain ontology. In this ontology, classes and values are issued from available metadata sets or repertories specialized on corpora and tools descriptions. Metadata are founded on hybrid lexicons and collect information about distinctive scientific domains (e.g., linguistics).

The ontology is structured considering that terms describe any corpus material features. High-level classes consider resources as the different steps within a process of linguistic resource elaboration. The obvious consequence is that the ontology (i) will support the representation of causal relations and (ii) is able to characterize the fundamental steps of this elaboration. Furthermore, it should be able to represent both standard ontological relations (i.e., partonomic and subsumptive relations) and causal relations. Based on the analysis of the above sets of metadata and on our process-based view, we isolate five basic classes for the ontology. The relationships between these five classes are illustrated on figure 3. In each rectangle, the class (strong caps) is composed of sub-classes, themselves defined by properties. Relations between classes are constraints. For the linguistic tools, we have a direct relation from the `DEFINITION` to the `APPLICATION`, and for corpora, the relations from `SEMANTIC CONTENT` or `DESCRIPTION ELEMENT` to `APPLICATION` are unnecessary. The distinction between classes and properties reflects

[1] http://www.language archives.org/

[2] http://www.mpi.nl/IMDI/

[3] www.lrec conf.org/proceedings/lrec2010/workshops/W20.pdf

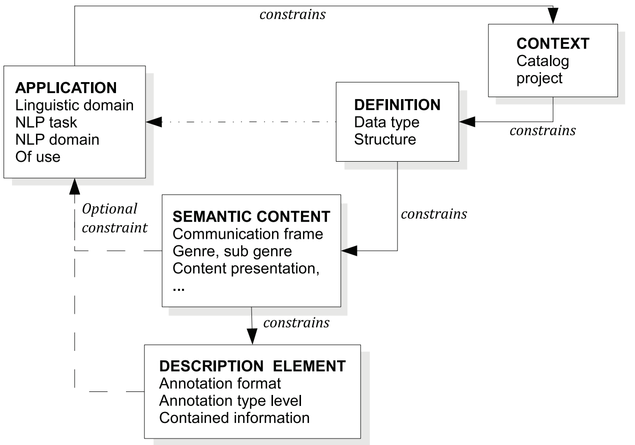


Figure 3: Constraints and causal relations.

the distinction between a descriptive term and a step in the process which integrates a specific property for this term. Here, properties are seen as ontological endurants[4]. If we consider for instance the class SEMANTIC CONTENT, then the subclass added content is such that added content [isa] SEMANTIC CONTENT and adding the properties commentary and translation we obtain the respective relations: commentary [isa] added content and translation [isa] added content. For the current example reported on figure 4, the ontology is populated from the description of the lexicon (cf. figure 1).

CONTEXT
Project
Language: **eng**
Language: **glg**

DEFINITION
DATATYPE
Size: entries: **1648272**
resource ID: **http://sli.uvigo.es/CLUVI**
Resource link: organization: **Universidade de Vigo, Grupo de investigación TALG (Tecnoloxías e Aplicacións da Lingua Galega)**
Written: **WrittenCorpus**

SEMANTIC CONTENT
Content presentation: **aligned_sentence**
status : coverageType: **general**

DESCRIPTION ELEMENTS
annotation format : **XML/TMX**
annotation type level : discourse based annotation : **morphology**
annotation type level : discourse based annotation : **semantic**
annotation type level : discourse based annotation : **morpho-syntax**
annotation type level : discourse based annotation : **syntax**

Figure 4: Ontological description of the BioLexicon corpus.

Causality is defined by the adjunct of information to a belief while the result of this operation is knowledge. Causality reflects necessary inference towards the further step at a lower level than constraints. For example, if a particular DATATYPE causes a specific STRUCTURE (of data), then this belief becomes knowledge because lexicon (with lexicon [isa] DATATYPE) conveys this information to lexicon resource type with lexicon resource type [isa] STRUCTURE. Then, a single choice is allowed into the lexicon resource type class, i.e., the subclass computational lexicon, concordance, terminology, wordlist, glossary or dictionary). It follows that the lexicon resource type class is the cause of a limited choice in SEMANTIC CONTENT. Considering the information that we have a lexicon resource, the content presentation subclass will permit the values parallel or comparable. Conversely communication frame and genre will be irrelevant. This causal relation is independent to the chosen instances into the lexicon resource type class, that is any instance of the lexicon resource type class is relevant for the subclass content presentation of SEMANTIC CONTENT.

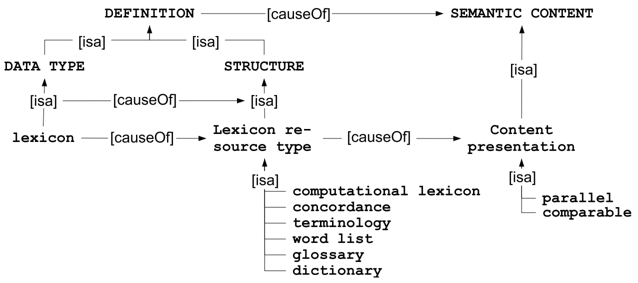


Figure 5: Causality.

4.3 High level: conceptual structure, activity and causality

As advocated above, steps in the process of linguistic resource elaboration are defined at the high level of abstraction of the ontology. This level proposes relations between the different descriptive classes that compose the medium level. As advocated in (Lando et al. 2007), the conceptual structure characterizes the intention of the project and how this intention is realized by a succession of constrained choices. In a project of resource building a single goal requires different steps, each of them performing a particular component of the resource. Each component specifies one particular dimension then heterogeneous descriptive classes compose an homogeneous object. These components are both ordered from the global to the more specific and from the context of production to the user.

The high level of the ontology, i.e., constraints during the emergence process explains causal relations between heterogeneous classes at the middle level. Constraints are defined (Devlin 1991) as an inference mechanism from a situation to another using information. For example, the APPLICATION constrains the CONTEXT because it conveys information about the goal of the resource to the framework where this resource takes place. But these constraints must

[4] An endurant is an entity which exists wholly in every instant at which it exists at all.

also operate at a lower abstraction level for identifying exactly what content elements are conveyed and their consequences on the constrained situation.

5 Using Infomorphisms for Metadata Inference

The IF theory assumes that information flow results from regularities in a distributed system. The distributed system refers here to the ontology whose classes are types (e.g., SEMANTIC CONTENT or DEFINITION) assuming that regularities rely on the concepts of ”type” and ”token”. The causal dimension of IF shows that the constraints which appear into an ontology can be represented by regular relations between sets of types and tokens (Jayez and Mari 2005; Collier 2010). If we consider any instance of a corpus and a publication using this corpus, we classify this corpus by its metadata into an ontology in order to infer relevant properties for a contextual description of the publication. A publication description accepts only a summary of the corpus description (for economic reasons) and must be structured into sequences considering that they must be used into different configurations of metadata. For that purpose, we isolate three sequences:

FOUNDATION which characterizes the resource context involving relations between sources and the scientific context of this resource.

MATERIAL which presents how the linguistic theme is observed. The sequence details the material content of the resource and how this material is structured.

METHOD which states the intellectual objective of the resource building and the author’s point of view on the resources use.

Then a vocabulary is chosen to refer explicitly to the resource in an heterogeneous frame. As detailed on table 1, resource descriptive metadata like ”value” for the metadata components lead to new metadata attributes and refinements (or extensions of attributes). Assuming a closed vocabulary of values for each attribute, any conclusion of an inference characterizes a corresponding value for the attribute. We use the DCMI resource model (i.e., Dublin Core) as reference for the conceptual representation. These new metadata attributes and refinements which are defined by $Property$ and $SubPropertyOf$ into RDF Schemas will populate the C classification. High level types of A_i are not relevant for the description of the publication contents since they characterize resource structural components. Any issue of an inference is classified as an attribute or a value in C. Every refined concept of the ontology middle level is considered as a value in the metadata sequence, and every higher concept, as an attribute. $Property$ and $SubPropertyOf$ are defined as complex elements and accept a sequence of typed simple elements conveyed from the ontology. Each sequence is structured and contains types, properties and values. Each descriptor results from an inference which both constructs the resource description and witnesses its influence on the publication content. The information flow assigns constraints on the quantities of admitted properties for a particular resource. Functions between types characterize constraints at the higher level while subtypes are related with causal relations at the middle level (subtypes represent material features of resources).

The IF mechanism operates from an ontology to a publication description. The representation inherits the ontology structure but only makes IF explicit. The classification channel C characterizes relevant components of the resource for the publication description. Functions f^{sup} are characterized only in ontology while functions F^{sup} represent the infomorphisms which create the channel. Functions f^{inf} relate the publication to corpora at the token level (they represent citations about the corpus in the publication). Classifications $\models_{A_1}$, $\models_{A_2}$ and $\models_{A_3}$ are relations from the corpus to the ontology whereas $\models_C$ concerns the relation publication - metadata.

We isolate three functions at the token level: context of use of the corpus (degree of implication of the resource in the publication, defined by the locations of citation in the publication), modalities of use (how the resource is used in the research, considering sequences of words around the corpus name), and specific interest (contents of the resource with a particular interest, attested by examples or description of functionalities). These functions will be obtained by a tool of information extraction. These functions characterize how the resource is used in the publication and the relevance of the relation between the two resources.

6 Discussion

The benefits of the present approach relies on the introduction of a bibliographic ontology enhanced with an inference mechanism which operates between descriptions of documents and documents, instead of considering relationships between structured data. Bibliographic ontology like BIBO and DBLP add an ontology level on the pairs of attribute/values that characterize metadata. It increases relations between resources and formal description of these resources, but like any metadata set, they are not able to propose a more precise content description than the related concepts of Dublin Core (*description*, *keyword*, *subject*). This precise content characterization becomes possible because of the inference that is conveyed into the ontology to another document description. In such a way, the relation between documents is duplicated by a content conveyance.

The high level of the ontology proposes reusable model of a tool elaboration and use. Functional concepts are ordered by a constraint relation. The different steps of the resource elaboration are characterized by their conditions. The ontology level is sufficient to support a structure that has some relation with the business ontology integrating the tool description and then the material dimension of the described activity. A crucial difference is that our high level ontology has an immediate sub-level in which the material dimension characterizes how the present choices are causally explained by previous ones.

Property	SubpropertyOf	Value (issued from Ontology Classes)
FOUNDATION	ResourceType ResourceStatus	T(A0) CONTEXT
		T(A1) DEFINITION
		T(A2) SEMANTIC CONTENT
MATERIAL	ResourceAnnotation ResourceAdditionalInformation	T(A3) DESCRIPTION ELEMENT
		SubT(A3.1) Annotation Format
		SubT(A3.2) Annotation Type Level
		SubT(A3.3) Contained Information
METHOD	ResourceDomainOfUse ResourceRelatedTasks	T(A4) APPLICATION
		SubT(A4.1) Linguistic Domain
		SubT(A4.2) NLP DomainOfUse
		SubT(A4.3) NLPTask

Table 1: Determinating values.

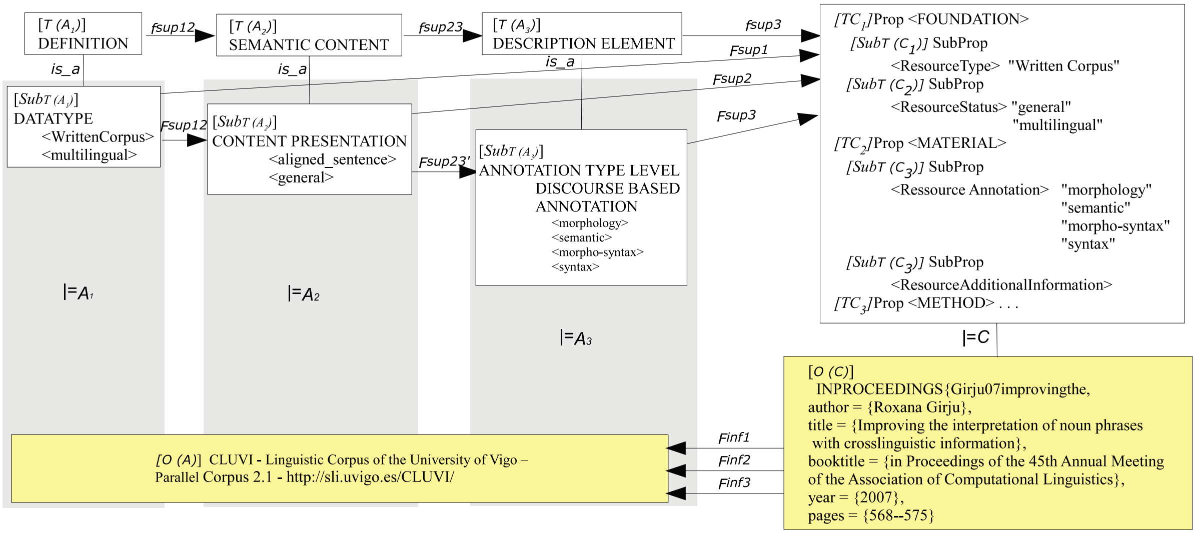

[T (A1)] DEFINITION
fsup12
[T (A2)] SEMANTIC CONTENT
fsup23
[T (A3)] DESCRIPTION ELEMENT
fsup3
Fsup1
Fsup2
Fsup3
is_a
is_a
is_a
[SubT (A1)] DATATYPE <WrittenCorpus> <multilingual>
Fsup12
[SubT (A2)] CONTENT PRESENTATION <aligned_sentence> <general>
Fsup23'
[SubT (A3)] ANNOTATION TYPE LEVEL DISCOURSE BASED ANNOTATION <morphology> <semantic> <morpho-syntax> <syntax>
[TC1]Prop <FOUNDATION>
[SubT (C1)] SubProp
<ResourceType> "Written Corpus"
[SubT (C2)] SubProp
<ResourceStatus> "general" "multilingual"
[TC2]Prop <MATERIAL>
[SubT (C3)] SubProp
<Ressource Annotation> "morphology" "semantic" "morpho-syntax" "syntax"
[SubT (C3)] SubProp
<ResourceAdditionalInformation>
[TC3]Prop <METHOD> . . .
|=A1
|=A2
|=A3
|=C
[O (C)]
INPROCEEDINGS{Girju07improvingthe,
author = {Roxana Girju},
title = {Improving the interpretation of noun phrases with crosslinguistic information},
booktitle = {in Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics},
year = {2007},
pages = {568--575}
Finf1
Finf2
Finf3
[O (A)] CLUVI - Linguistic Corpus of the University of Vigo – Parallel Corpus 2.1 - http://sli.uvigo.es/CLUVI/

Figure 6: IF-based enhancing of publication descriptions.

7 Conclusion

The present proposition integrates information conveyance into ontology. The question of the inheritance from relation at the high level of the ontology to the medium is clarified by the information flow. The knowledge-based relations at the high level are converted into information at the medium level. Information represents how and why heterogeneous structured sets of data have relations: they both contribute to a process and are ordered following some principles of information theory (choices, causal chains, etc.). Finally, information flow allows a concrete characterization (i.e., integrating material dimension) of the reasoning: how constraints on plan and process are translated into objects.

References

Barwise, J., and Seligman, J. 1997. *Information Flow*, volume 44 of *Cambridge tracts in Theoretical Computer Science*. Cambridge University Press.

Chang, B.; Ham, D.; Moon, D.; Choi, Y. S.; and Cha, J. 2007. Using ontologies to search learning resources. In Gervasi, O., and Gavrilova, M., eds., *ICCSA'07*, volume 1 of *LNCS*, 1146 1159. Springer.

Collier, J. 2010. *Information, causation and Computation*. World Scientific. chapter Information and Computation: Essays on Scientific and Philosophical Understanding of Foundations of Information and Computation.

Devlin, K. 1991. *Logic and Information*. Cambridge University Press.

Jayez, J., and Mari, A. 2005. Togetherness. In E. Maier, C. B., and Huitink, J., eds., *Proceedings of SuB9 (Sinn und Bedeutung 9)*, 155 169.

Keller, C., and Keller, J. 1996. *Cognition and tool use, the blacksmith at work*. Cambridge University Press.

Kent, R. 2000. The information flow foundation for conceptual knowledge organization. In Verlag, E., ed., *Procs. of the 6th int. conf. of the int. society for knowledge organization*.

Lando, P.; Furst, F.; Kassel, G.; and Lapujade, A. 2007. Premiers pas vers une ontologie générale des programmes informatiques. In *in 18es Journées Francophones d'Ingénierie des Connaissances*.