Bootstrapping semantics on the Web: meaning elicitation from schemas

Paolo Bouquet[1]
Joint work with: Luciano Serafini[2] and Stefano Zanobini[1]

[1]University of Trento, Italy
[2]ITC-Irst, Trento, Italy

WWW2006
Edinburgh (Scotland), 26 May 2006

Objective

Deeper Semantics

- A wide variety of schemas (such as classifications, directory trees, web directories, relational schemas ...) are exposed on the Web.
- They convey a clear meaning to humans (e.g. help in the navigation of large collections of documents).
- However, they convey only a small fraction of their meaning to machines, as meaning is not formally/explicitly represented.

Objective

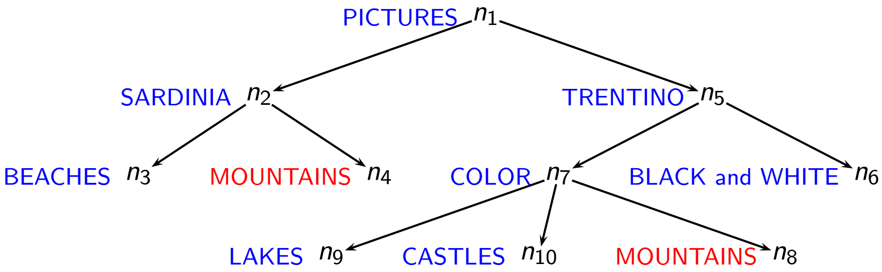
Deeper Semantics

- A wide variety of schemas (such as classifications, directory trees, web directories, relational schemas . . .) are exposed on the Web.
- They convey a clear meaning to humans (e.g. help in the navigation of large collections of documents).
- However, they convey only a small fraction of their meaning to machines, as meaning is not formally/explicitly represented.

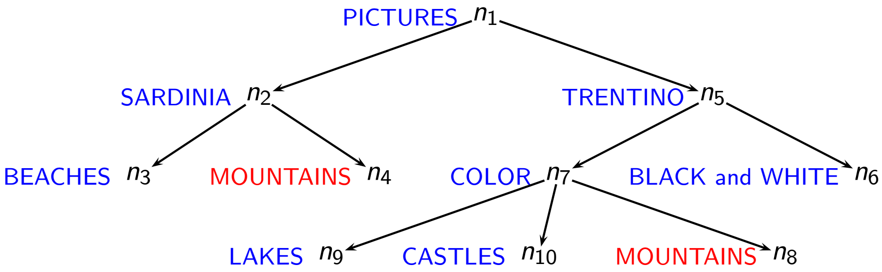
Our goal

Design a general methodology for automatically eliciting and representing the intended meaning of schema elements and making it available to machines.

Directory Structure



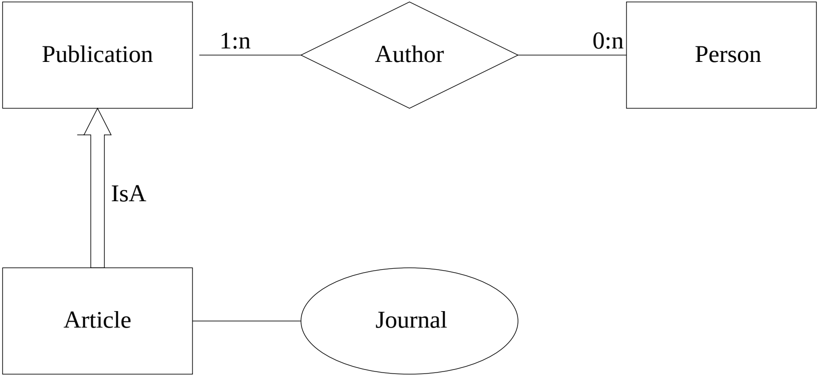
Directory Structure



Intended meaning

Pictures			[depicting]	mountains	[located in]	Sardinia
Pictures	[in]	color	[depicting]	mountains	[located in]	Trentino

ER schema



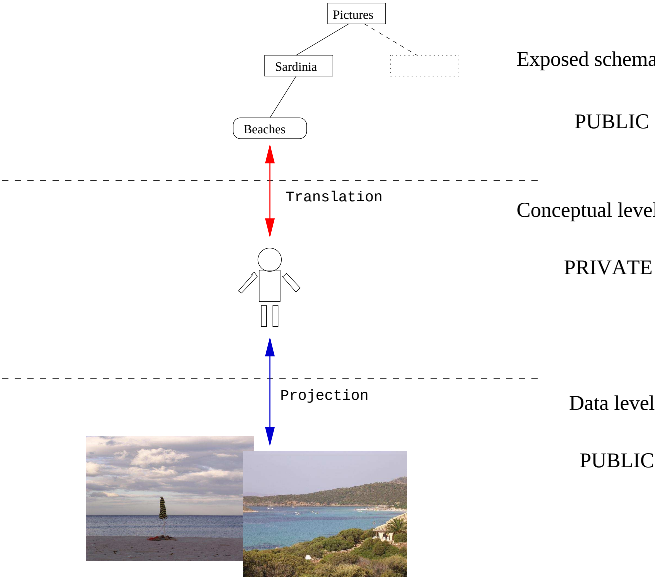
Problems

- Eliciting the meaning of an exposed schema requires that we formally/explicitly represent the intended meaning of each of its elements
- Part of element meaning (the *structural meaning*) is exposed with the schema (and for some types of schemas, like ER schemas or RDFS, even formally codified)
- However:
  - typically, part of the structural meaning is not exposed (e.g. the relation between pictures and Sardinia)
  - the conceptual content is “hidden” in the choice of (natural language) labels

Our proposal (version 0.9)

- Construct all meaning skeletons which are compatible with the structure of a schema
- Construct the conceptual content of labels from their linguistic formulation
- Use any available domain knowledge to filter out meaning skeletons which are not compatible
- Use the combination of structural meaning and conceptual content to produce a formal and explicit representation of each schema element's deep semantics.

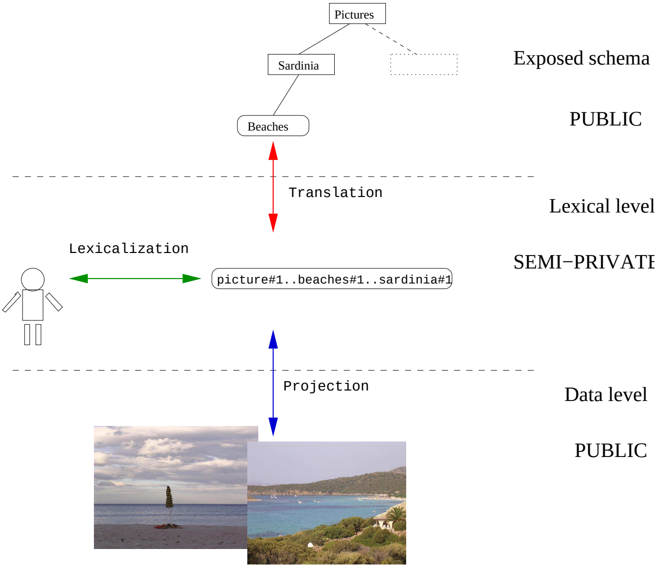
A problem with this idea



Dictionaries as semantic coordination tools

- Concepts are not directly accessible (they're mental constructs) nor comparable
- The only access we have to other people's concepts is through their use of (natural) language
- Luckily, for natural languages, we have a very powerful tool for semantic coordination: dictionaries (lists of words + list of acceptable senses for each word)
- We propose to systematically use dictionary senses as surrogates of concepts

The intuitive model

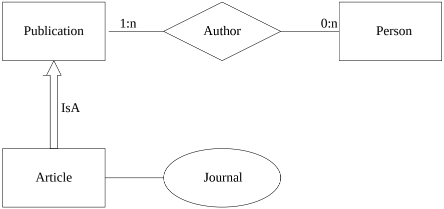


Our proposal (version 1.0): WDL

Meanings are represented in a formal language (called WDL, for WordNet Description Logic), which is the result of combining two main ingredients:

- a logical language, with a precise (formal) semantics and a sound a complete decision procedure (Description Logics)
- WordNet senses as the vocabulary of the descriptive language

WDL example - ER

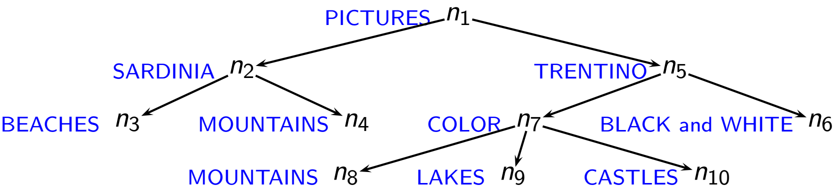


The meaning of the node labeled with "Publication" in this ER schema is

$$\text{Publication\#1} \sqcap \exists \text{Author\#1}^{-}.\text{Person\#1}$$

and the intuitive semantics is "a copy of a printed work offered for distribution" that "a human being", "writes ... professionally ..."

WDL example - Directories



The meaning of the node n_3 of the hierarchical classification is

$$\text{image\#2} \sqcap \exists \text{subject\#4}.(\text{beaches\#1} \sqcap \exists \text{Located\#1}.\{\text{Sardinia\#1}\})$$

The intuitive meaning is "a visual representation produced on a surface" [image#2] whose "subject" [subject#4] is "an area of sand sloping down to the water of a sea or lake" [beach#1] "situated in" [Located#1] "an island in the Mediterranean west of Italy" [Sardinia#1]

Meaning Elicitation

The problem of meaning elicitation can be restated as the problem of finding a WDL expression $\mu(n)$ for each element n of a schema, so that the intuitive semantics of $\mu(n)$ is a good enough representation of the intended meaning of the element.

Semantic Elicitation in Practice

Three main steps

- Meaning Skeletons: encode the structural information contained in a schema, namely the information carried by a schema with meaningless labels. This information comes from the (in)formal semantic of the schema.

Semantic Elicitation in Practice

Three main steps

- Meaning Skeletons: encode the structural information contained in a schema, namely the information carried by a schema with meaningless labels. This information comes from the (in)formal semantic of the schema.
- Local meaning: encodes the meaning of the label associated to an element when taken in isolation. Information on local meanings can be derived from a lexicon (e.g. WordNet).

Semantic Elicitation in Practice

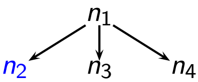
Three main steps

- Meaning Skeletons: encode the structural information contained in a schema, namely the information carried by a schema with meaningless labels. This information comes from the (in)formal semantic of the schema.
- Local meaning: encodes the meaning of the label associated to an element when taken in isolation. Information on local meanings can be derived from a lexicon (e.g. WordNet).
- Relations between local meanings (R_{mn}): relations that may hold between local meanings (e.g. the relation Located#1 between beach#1 and Sardinia#1). Relations between local meaning can be extracted from the domain knowledge (ontologies).

Meaning Skeletons

- Meaning skeletons are associated to each node *n* of a schema,
- A Meaning skeleton is a DL concept whose basic components are the nodes of the graph, and the possible relations between them.
- The meaning skeleton associated to a node *n* represents the structural information carried by this node (independent from its label).

Meaning Skeletons (cont'd)



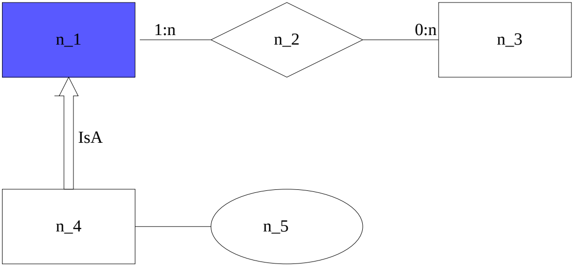
Example

In directories, the meaning skeleton of the node n_2 is:

$$n_1 \sqcap \exists R_{n_1,n_2}.n_2$$

n_2 acts as a "modifier" of n_1, and R_{n_1,n_2} is role connecting the two nodes.

Meaning Skeletons



Example

The meaning skeleton of the blue node (identified by n_1), according to the formal semantics of ER schema described by Alex Borgida et. al. is the following:

$$n_1 \sqcap \forall n_1.n_4 \sqcap \exists n_2.n_3$$

Local Meanings

- The local meaning of a node *n* in a schema, denoted with $\lambda(n)$, is a DL description representing all possible meanings of the label associated to a node.
- $\lambda(n)$ is computed by exploiting a linguistic resources
- A *linguistic resource* as a function which, given a word, returns a set of *senses*, each representing an acceptable meaning of that word.
- WordNet is probably the best electronic lexical available to date.

Local Meanings - Examples

Example

$$\text{WORDNET}(\text{"picture"}) = \text{picture\#1}, \text{picture\#2}, \ldots, \text{picture\#9}$$
$$\text{WORDNET}(\text{"Sardinia"}) = \text{Sardinia\#1}, \text{Sardinia\#2}$$

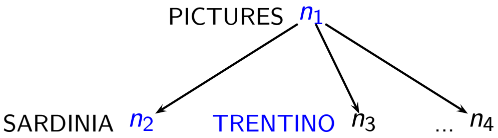
If the label of m is "picture" and the label of n is "Sardinia" then

$$\lambda(m) = \text{Picture\#1} \sqcup \text{Picture\#2} \sqcup \cdots \sqcup \text{Picture\#9}$$
$$\lambda(n) = \text{Sardinia\#1} \sqcup \text{Sardinia\#2}$$

Relations between local meanings

- Domain knowledge is used to discover semantic relations holding between local meanings.
- Intuitively, given two primitive concepts C and D, we search for a role R, denoted with $\rho(C, D)$ that possibly connect a C-object with a D-object.
- As an example, the relation that connects the concept picture#2 and the concept Sardinia#1 can be subject#4.

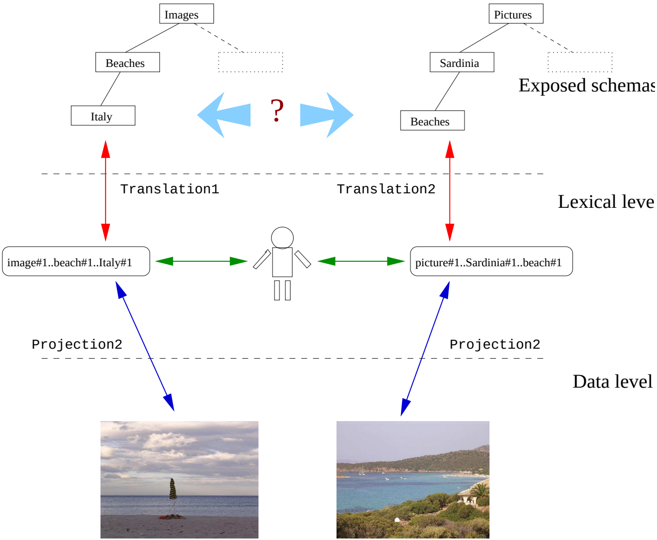
Putting things together



1. Meaning skeleton $n_1 \sqcap \exists R_{n_1,n_2}, n_2$
2. Instanciate the skeleton with all possible combinations of local meanings (e.g. picture#1 $\sqcap \exists R_{n_1,n_2}$.Sardinia#1, ..., picture#5 $\sqcap \exists R_{n_1,n_2}$.Sardinia#2, ...)
3. fill the meaning skeleton with the semantic relations between the local meanings and discard all the local senses which do not have semantic relations:

picture#1 $\sqcap \exists$subject#4.Sardinia#1

An application: schema matching



Schema matching (continued)

- Once the meaning of two schemas is elicited and represented in WDL, discovering semantic relations across them is a matter of logical reasoning
- We can use any standard DL reasoner to discover equivalence or subsumption between any pairs of nodes of different schemas
- The relations computed by this method are meaningful (have a clearly defined semantics) and can be used for distributed DL reasoning

Schema matching (continued)

Concept Γ from the first schema:

$$\text{image\#2} \sqcap \exists \text{subject\#4}.(\text{beaches\#1} \sqcap \exists \text{Located\#1}.\{\text{Italy\#1}\})$$

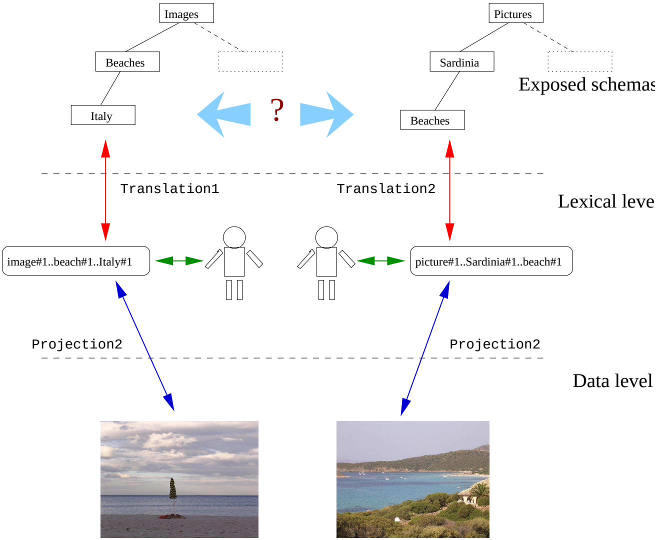
Concept Δ from the second schema:

$$\text{picture\#1} \sqcap \exists \text{subject\#4}.(\text{beaches\#1} \sqcap \exists \text{Located\#1}.\{\text{Sardinia\#1}\})$$

Using lexical + domain knowledge, we can easily infer that:

$$\text{image\#2} \equiv \text{picture\#1},\ \text{Sardinia\#1} \sqsubseteq \text{Italy\#1} \models \Delta \sqsubseteq \Gamma$$

Peer-to-peer schema matching



Implementations

- A first implementation called CtxMatch1.0, which uses WPL (propositional logic) encoding
- Our current implementation CtxMatch2.0, which uses a WDL encoding (WordNet + "lexicalized" OWL ontologies)
- GUI for CtxMatch2.0 which allows creating, editing and matching schemas

Projects

- Matching classifications in Distributed Knowledge Management (Project: EDAMOK – Provincia di Trento)
- Extracting knowledge from information and content sources (Project: VIKEF – EU funded integrated project)
- Ontology alignment via elicitation in e-learning environments (Project: APOSDLE – EU funded)
- Intelligent queries across heterogeneous web sites (Project: WISDOM – Italian Ministry of Research and University)
- Database integration through DB schema elicitation and matching (Project: RISICOM)
- Ontology extraction from texts using elicitation (Project: ONTOTEXT – Provincia di Trento)

Conclusion

- The method presented here can be used on many schemas which are already available on the web (e.g. in most portals or e-business web sites)
- The main message is: ontologies MUST be complemented with lexical information
- We need a principled way for “lexicalizing” ontologies (and store the results in OWL) to close the gap between structural and intended meaning