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
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.
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.
Intended meaning

<table>
<thead>
<tr>
<th>Pictures</th>
<th>color</th>
<th>[depicting]</th>
<th>mountains</th>
<th>[located in]</th>
<th>Sardinia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pictures</td>
<td>[in]</td>
<td>3</td>
<td>mountains</td>
<td>[located in]</td>
<td>Trentino</td>
</tr>
</tbody>
</table>
ER schema

Publication \( \rightarrow \) Author \( 1:n \) \( \rightarrow \) Person \( 0:n \)

IsA

Article \( \rightarrow \) Journal

Paolo Bouquet  Meaning elicitation from schemas
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

Exposed schema

PUBLIC

Conceptual level

PRIVATE

Data level

PUBLIC

Exposed schema

PUBLIC

Translation

PRIVATE

Projection

PUBLIC

Paolo Bouquet

Meaning elicitation from schemas
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

Exposed schema
PUBLIC

Lexical level
SEMI-PRIVATE

Data level
PUBLIC

Lexicalization

Projection

Translation
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
The meaning of the node labeled with “Publication” in this ER schema is

\[
\text{Publication} \cap \exists \text{Author}. \text{Person}
\]

and the intuitive semantics is “a copy of a printed work offered for distribution” that “a human being”, “writes ... professionally ...”
The meaning of the node $n_3$ of the hierarchical classification is

\[ \text{image} \sqcap \exists \text{subject} \sqcap \supseteq \text{beaches} \sqcap \exists \text{Located} \sqcap \text{Sardinia} \]

The intuitive meaning is “a visual representation produced on a surface” [image] whose “subject” [subject] is “an area of sand sloping down to the water of a sea or lake” [beach] “situated in” [Located] “an island in the Mediterranean west of Italy” [Sardinia]
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.
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.
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).
Example

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

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

$n_2$ acts as a “modifier” of $n_1$, and $R_{n_1,n_2}$ is role connecting the two nodes.
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:

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

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.
Example

\[ \text{WordNet}(\text{“picture”}) = \text{picture#1, picture#2, \ldots, picture#9} \]
\[ \text{WordNet}(\text{“Sardinia”}) = \text{Sardinia#1, Sardinia#2} \]

If the label of \( m \) is “picture” and the label of \( n \) is “Sardinia” then

\[ \lambda(m) = \text{Picture#1 \Box Picture#2 \Box \cdots \Box Picture#9} \]
\[ \lambda(n) = \text{Sardinia#1 \Box Sardinia#2} \]
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, \text{Sardinia}\#1}, \ldots, picture\#5 \sqcap \exists R_{n_1, n_2, \text{Sardinia}\#2}, \ldots)
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:

\[
\text{picture}\#1 \sqcap \exists \text{subject}\#4. \text{Sardinia}\#1
\]
An application: schema matching

Exposed schemas

Lexical level

Data level

Projection1

Projection2

Translation1

Translation2

Beaches

Images

Italy

Beaches

Sardinia

Pictures

picture#1..Sardinia#1..beach#1

image#1..beach#1..Italy#1

Paolo Bouquet

Meaning elicitation from schemas
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.
Concept $\Gamma$ from the first schema:

\[ \text{image} \sqcap \exists \text{subject}. (\text{beaches} \sqcap \exists \text{Located}. \{\text{Italy}\}) \]

Concept $\Delta$ from the second schema:

\[ \text{picture} \sqcap \exists \text{subject}. (\text{beaches} \sqcap \exists \text{Located}. \{\text{Sardinia}\}) \]

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

\[ \text{image} \equiv \text{picture}, \quad \text{Sardinia} \sqsubseteq \text{Italy} \mid\mid \Delta \sqsubseteq \Gamma \]
Peer-to-peer schema matching

Exposed schemas

Lexical level

Data level

Projection2

Translation2

Translation1

Observation

Matching can be asymmetric (directional), depending on the two agents' domain knowledge.
Implementations

- A first implementation called \texttt{CtxMatch1.0}, which uses WPL (propositional logic) encoding
- Our current implementation \texttt{CtxMatch2.0}, which uses a WDL encoding (\texttt{WordNet} + “lexicalized” OWL ontologies)
- GUI for \texttt{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: APOSDELE – 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)
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.