# Types and Roles of Ontologies in Web Information Extraction

Martin Labský, Vojtěch Svátek and Ondřej Šváb

Department of Information and Knowledge Engineering,
University of Economics, Prague, W. Churchill Sq. 4, 130 67 Praha 3, Czech Republic
{labsky,svatek,xsvao06}@vse.cz

**Abstract.** We discuss the diverse types and roles of ontologies in web information extraction and illustrate them on a small study from the product offer domain. Attention is mainly paid to the impact of domain ontologies, presentation ontologies and terminological taxonomies.

## 1 Introduction

Web information extraction (WIE) is a sub-discipline of web mining that applies *pre-existent* patterns on web data with the aim of populating structured models, typically databases or ontologies, with records or class/relation instances, respectively. The research in WIE and in applied ontology are closely related, since WIE transforms the content of 'legacy' web to machine-understandable form and ontologies are the conceptual backbone of semantic web, the web 'for machines'. However, since the notion of ontology is very ambiguous and the term is interpreted differently in various communities, the nature as well as role of ontologies in WIE may significantly vary from one project to another.

In a general overview of ontology types, van Heijst [13] distinguishes among terminological, information and knowledge ontologies. *Terminological ontologies* are centered around human-language terms, without direct reference to real world. Their main constructs are synonym sets and (hyponymy/meronymy) hierarchies. *Information ontologies* and *knowledge ontologies* both deal with classes directly mapped to sets of entities (instances) in some universe of discourse. Knowledge ontologies however differ from information ontologies by presence of formal axioms, most particularly, by the possibility to define the extent of a class via a logical expression over its properties (relations to other classes).

The range of models possibly appearing in different phases of WIE (as specific type of application) seems to be somewhat analogous to the general categorisation. Stevenson & Ciravegna [10] already raised the issue of ontologies 'for customer service' that do not satisfy the needs of information extraction components, namely, they point out the contrast between *domain ontologies* suitable for reasoning over real-world objects (in the 'customer' application) and *linguistic ontologies* applicable on (presumably, continuous) text. This contrast however becomes less sharp when considering semi-structured web content in the form of lists, tables or forms. Ontologies directly usable for analysis of web structures

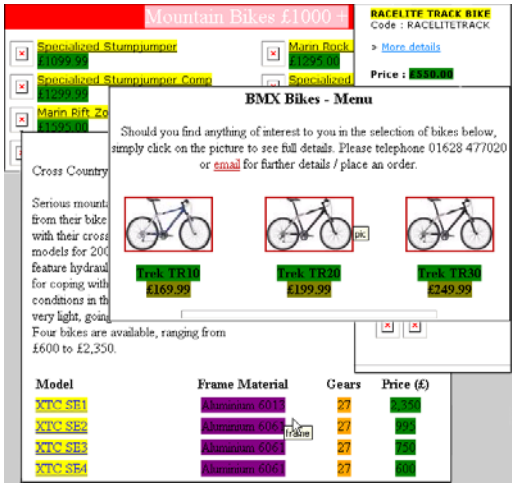


**Fig. 1.** Samples of annotated training data

are likely to borrow a lot from 'customer-service' ontologies, since the fragments of HTML code will often directly map on ontology classes, attributes/relations and instances. We will call them *presentation ontologies*, since their universe of discourse is that of web objects as *presented* on the web (e.g. bicycle offers encoded in HTML) rather than of real-world objects (real bicycles). Finally, at the level of plain text strings, *terminological ontologies* may come into play.

In the rest of the paper, we illustrate this simple typology of WIE (uses of) ontologies on our experiments in the bicycle domain and on related projects.

## 2 Experience from the Bicycle Sale Domain

*Web product catalogues* contain names, prices, pictures and other characteristics of products. When performing information extraction, text fragments corresponding to these items have to be discovered and composed into instances of 'product offer' (or similar). Complete instances, stored in a database or ontology, are then subject to retrieval or inference.

In our ongoing experiment, we processed 100 catalogues from 40 British *bike shop* websites containing more than 900 instances of 'bike offer'. Examples of catalogue pages (with different data items marked with different colours) are on Fig. 1. Let us now discuss different 'bicycle' ontologies related to the extraction process, in the inverse order of their appearance in this process.

### 2.1 Populating the Domain Ontology

Our 'custom service' is end-user search over bicycle sale data, represented in a format suitable for the semantic web: we opted for RDF[1] and the *Sesame*

[1] `http://www.w3.org/RDF`

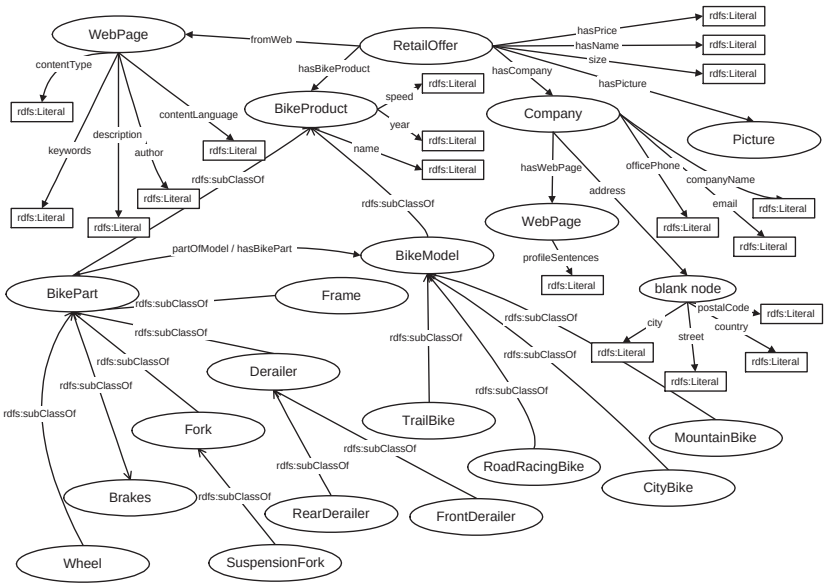


**Fig. 2.** Part of RDF schema for the bicycle domain

repository [1]. On the top of the repository, an HTML query interface was built[2]. The ontology to be populated was expressed in RDF Schema. Fig. 2 shows most of the ontology: it covers information on the product offer itself (as presented on the web), characteristics of the product, as well as those of the selling company. Consistently with the observation made in [10], this ontology links together pieces of information occurring nearby each other at a product catalogue page, as well as those located quite separately or even not directly present on the website and thus unlikely to be picked up by means of a single WIE procedure. Indeed, different parts of the ontology are assumed to be populated by different web analysis methods (some already operational and some under design) within a distributed architecture named *Rainbow* [11].

### 2.2 Template-Filling with Presentation Ontology

We assume that presentation ontologies will most likely be restricted to a smaller portion of the original domain, cut up according to web presentation factors. Our simple ontology shown on Fig 3 is specific to product catalogues[3], and only contains one 'true class', that of Bike Offer; the remaining concepts are shrunk

[2] It is available at `http://rainbow.vse.cz:8000/sesame`. Details on the RDF query technology used can be found in [12]

[3] Similar presentation ontologies could be designed e.g. for company profile (pages) or contact info (pages); the former would presumably be more linguistic-oriented as the profile information is typically expressed by free text, cf. [6].

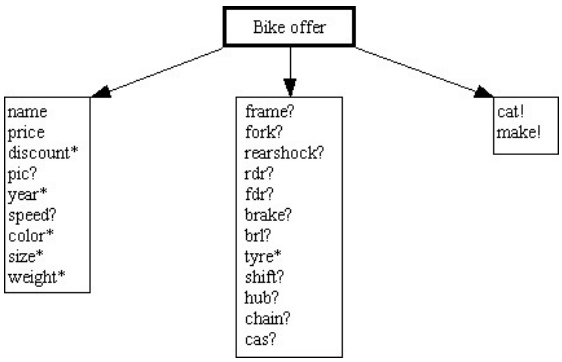


**Fig. 3.** Bicycle offer presentation ontology

to its properties. Note (in contrast to the domain ontology above) the direct link between product offer information and information about bike parts. Although not 'deeply' ontologically related, they fit together in terms of presentation: the company hopes to sell the offered bike thanks to pointing out its equipment.

While the domain ontology was destinated for direct retrieval of structured information, our presentation ontology is tuned for 'template filling' by means of a simple sequential algorithm (assigning properties to the 'current' object as long as constraints are satisfied). The expressive power of the ad hoc 'ontology language' used is thus kept limited. The central features are the *uniqueness*, *multiplicity* and *optionality* of properties, the latter two indicated with the `*` and `?` symbols, respectively. In addition, *'sticky' properties* are distinguished: as soon as the value of sticky property is discovered on a page, it is filled to all objects extracted afterwards, until a new value is discovered for this property.

The domain of product offers is simple enough (in terms of logical structure of presentation) to allow to keep only one class and to dissolve the remaining ones into properties. This assumption would certainly not hold for all domains where WIE might be applied; the presentation ontology then would have multiple 'class vertices', and the template-filling algorithm would be more sophisticated.

### 2.3 Lexical Taxonomy for Primary Annotation

In our project we have not used a lexical taxonomy in the primary annotation of bike names, prices, component names and the like. The annotation was carried out by means of statistical (Hidden Markov) models; see [12] for details. However, a collection of more than 60 bicycle *categories* (in various sense) arose as side product of annotation, and was later arranged into a hierarchy (see part of it at Fig. 4). We could easily imagine adoption of a similar taxonomy, e.g. a domain-specific part of product taxonomy such as bicycle-specific part of UNSPSC[4], for automated annotation with possibility of conceptual abstraction upto an arbitrary level of taxonomy.

[4] `http://www.unspsc.org`

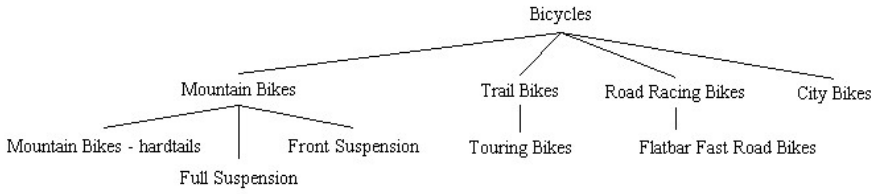


**Fig. 4.** Fragment of empirical taxonomy for bicycle 'categories'

## 3 Ontologies in Other WIE Projects

Embley [3, 4] uses the notion of 'extraction ontology' for conceptual schema with data frames hand-crafted by domain expert (i.e. presentation ontology). While [4] focuses on *HTML table* analysis (for offers of products, namely, cars), [3] deals with *free text*[5] (obituaries); the nature of 'extraction ontology' however remains the same. In the Armadillo [2] project, an (inductively learnt) presentation ontology allows to reuse a surface-logical structure from one resource to another, e.g. accross multiple bibliography resources from the same domain, containing data about overlapping sets of publications. A sort of presentation ontology is also used in the OntoBuilder project [5] aiming at 'deep web' information extraction. It defines layout rules for HTML forms used as input to online databases. On the other hand, the Crossmarc project [9] is limited to terminological level (term sets mapped on semantic classes) in its usage of ontologies[6]. In the AeroDAML approach [7], a terminological ontology (WordNet) is used for annotation and a knowledge ontology (expressed in DAML) is populated by extraction results. Since the extraction method is named-entity recognition rather than structural IE, consistency-constraints are only applied at the level of target domain ontology rather than within a dedicated presentation ontology. Similarly, Maedche et al. [8] used an ontology engine (OntoBroker) to verify 'conceptual bridges' between terms extracted via shallow syntactic analysis.

## 4 Conclusions and Future Work

We discussed the types and roles of ontologies in web information extraction. By our own experience as well as by literature study, it seems worthwhile distinguishing, at least, between genuine *domain ontologies* used for the target application, *presentation ontologies* used in heuristic template filling (or linguistic discourse analysis), and *terminological ontologies* used in text annotation.

While terminological ontologies are ubiquitous and sharable domain ontologies are also likely to proliferate, sharable presentation ontologies suitable for WIE are rare, since their aspects are typically hidden inside IE tools in proprietary languages. An interesting direction for future work thus is to partially

[5] But applies surface term-distance heuristics rather than sentence parsing.
[6] Admittedly, its main focus is multi-linguality rather than HTML-centred WIE.

automate[7] the *transformation of domain ontologies to presentation ontologies*, which could significantly improved the portability of WIE tools. For example, in our setting, in order to port the application to a different retail-offer domain, we would 'only' retrain the low-level annotator on new labelled data (and/or make it reuse a new terminological ontology), and rebuild the presentation ontology so as to reflect a new domain ontology. In long term, we believe that a *shared format* for WIE ontologies (different from e.g. OWL but with the possibility of mutual mapping) should arise, so as to alleviate the application portability problem and pave the way to semantic web bootstrapping.

*The research is partially supported by grant no.201/03/1318 of the Grant Agency of the Czech Republic.*

## References

1. Broekstra, J., Kampman, A., van Harmelen, F.: Sesame: An Architecture for Storing and Querying RDF and RDF Schema. In: ISWC'02, Sardinia, LNCS 2002.
2. Ciravegna, F., Chapman, S., Dingli, A., Wilks, Y.: Learning to Harvest Information for the Semantic Web. In: ESWS-04, Heraklion, Springer LNCS 2004.
3. Embley, D.W., Campbell, D.M., Jiang, Y.S., Liddle, S.W., Lonsdale, D.W., Ng, Y.K., Smith, R.D.: Conceptual-model-based data extraction from multiple-record Web pages. *Data and Knowledge Engineering*, Volume 31, Issue 3 (November 1999).
4. Embley, D.W., Tao, C., Liddle, S.W.: Automatically extracting ontologically specified data from HTML tables with unknown structure. In: ER2002, Tampere 2002.
5. Gal, A., Modica, G.A., Jamil, H.M.: OntoBuilder: Fully Automatic Extraction and Consolidation of Ontologies from Web Sources. In: Proc. ICDE 2004.
6. Kavalec, M., Svátek, V.: Information Extraction and Ontology Learning Guided by Web Directory. In: ECAI Workshop on NLP and ML for ontology engineering, Lyon 2002.
7. Kogut, P., Holmes, W.: AeroDAML: Applying Information Extraction to Generate DAML Annotations from Web Pages. In: K-CAP 2001 Workshop Knowledge Markup & Semantic Annotation, 2001.
8. Maedche, A., Neumann, G., Staab, S.: Bootstrapping an Ontology-Based Information Extraction System. In: Intelligent Exploration of the Web, Springer 2002.
9. Pazienza, M.T., Stellato, A., Vindigni, M.: Combining ontological knowledge and wrapper induction techniques into an e-retail system. In: Workshop on Adaptive Text Extraction and Mining (ATEM03) held with ECML/PKDD 2003, Cavtat 2003.
10. Stevenson, M., Ciravegna, F.: Information extraction as a Semantic Web technology: Requirements and promises. In: Workshop on Adaptive Text Extraction and Mining (ATEM03) held with ECML/PKDD 2003, Cavtat 2003.
11. Svátek, V., Kosek, J., Labský, M., Bráza, J., Kavalec, M., Vacura, M., Vávra, V., Snášel, V.: Rainbow - Multiway Semantic Analysis of Websites. In: 2nd International DEXA Workshop on Web Semantics (WebS03), Prague, IEEE 2003.
12. Šváb, O., Labský, M., Svátek, V.: RDF-Based Retrieval of Information Extracted from Web Product Catalogues. In: SIGIR'04 Semantic Web Workshop, Sheffield.
13. van Heijst, G., Schreiber, G., Wielinga, B.: Using Explicit Ontologies in KBS development, *Int. J. Human-Computer Studies*, Volume 46, 1997, 183-292.

[7] Some known heuristics on information presentation principles could presumably be adopted for this purpose.