1 Introduction *Graham Wilcock*

This chapter is based on the panel session “The Roles of Natural Language and XML in the Semantic Web” at the 2nd Workshop on NLP and XML (NLPXML-2002), held at COLING-2002 in Taipei. The workshop\(^1\) covered a wide range of topics in which both NLP and XML are central: XML-based NLP tools, corpus annotation standards, XML in document generation, XML in spoken dialogue systems. Whereas most of the workshop papers presented tools and systems that are already implemented, the main aim of the panel session was to look ahead to the future development of the Semantic Web. The panel members, nevertheless, were researchers who already have practical experience of using Semantic Web technologies.

The chapter contains six sections, each written independently. I chaired the panel session and wrote this introduction. The next four sections were contributed by the four panel members: Paul Buitelaar presents an overall vision of the Semantic Web and its implementation technologies; Antonio Pareja-Lora describes experiences with XML and RDF; Barrett Bryant describes experiences with DAML+OIL; Jimmy Lin argues for a wider role for natural language. The final section by Nancy Ide forms both a conclusion and a link to the 3rd workshop\(^2\) in the NLPXML series.

In this introduction I outline why we had a panel on “The Roles of Natural Language and XML in the Semantic Web”. The Semantic Web has become an important topic, but what do we mean by “the role of XML” and “the role of natural language”? What kinds of roles are required in the Semantic Web? The sections by the panel members will give more detailed descriptions of some of these roles.

One of the questions raised at the panel session, leading to extensive discussion, concerned the possibility or impossibility of having a single all-embracing ontology as the basis for the Semantic Web. I will summarize some of the points from this discussion.

\(^1\) http://www.ling.helsinki.fi/~gwilcock/NLPXML

\(^2\) http://www.cs.vassar.edu/~ide/events/NLPXML3
1.1 The Roles of XML

XML\(^3\) (eXtensible Markup Language) can be used in two distinct ways, which are both important for the Semantic Web. First, XML is a language which can be used, directly and by itself, to represent information. Second, XML can be used to define more specialized languages. In fact, XML has been used as the basis of many different specialized languages. Here are just a few examples:

- MathML\(^4\) (Mathematics Markup Language) for mathematical formulae
- WML\(^5\) (WAP Markup Language) for WAP mobile phones
- JSML\(^6\) (Java Speech Markup Language) for speech synthesizers
- SVG\(^7\) (Scalable Vector Graphics), a language for XML-based graphics
- XHTML\(^8\), a form of HTML that conforms to XML syntax.

In addition, XML Schema\(^9\) is a language for defining the permitted structures and data types of an XML document type. The schema definition, itself an XML document, is used for validating the contents of other XML documents. An XML Schema can therefore be used to define a specialized language like XHTML. Unlike the earlier DTD form of document type definition, an XML Schema itself conforms to XML syntax. A language that conforms to XML syntax (like XHTML but unlike ordinary HTML) can be efficiently parsed, validated and transformed by standard XML processors.

Among this multitude of XML-based languages there are only a few languages (XML itself, XML Schema, RDF and RDF Schema, DAML+OIL, OWL) that we are concerned with here because they are used particularly in the Semantic Web. In fact, when we talk about “the role of XML in the Semantic Web” we are really using “XML” to refer to this small group of XML-based languages which play a particular role in the Semantic Web. This group of Semantic Web languages and their relationships are described by Paul Buitelaar in Section 2.

---

\(^3\) http://www.w3.org/XML/
\(^4\) http://www.w3.org/Math/
\(^5\) http://www.oasis-open.org/cover/wap-wml.html
\(^6\) http://java.sun.com/products/java-media/speech/forDevelopers/JSML/
\(^7\) http://www.w3.org/TR/SVG/
\(^8\) http://www.w3.org/TR/xhtml1/
\(^9\) http://www.w3.org/XML/Schema
1.2 The Roles of Natural Language

Natural language is, of course, ordinary language like English or Chinese. We use the term “natural language” in order to explicitly exclude artificial languages like Java or XML. When we talk about “the role of Natural Language in the Semantic Web” we are referring to the use of natural languages to play some particular role in the Semantic Web, as opposed to the use of the group of XML-based languages (XML, RDF, DAML+OIL) mentioned above.

Of course, the existing World Wide Web already contains enormous amounts of natural language in the texts of many millions of web pages. The problem is that it is difficult to find relevant information and extract it from this huge mass of texts. Most of the texts are marked up in HTML, but the markup mainly specifies the presentation format of the text, not its contents. By contrast, the vision of the Semantic Web is to mark up the semantic content of the information on the Web. The information whose semantic content needs to be marked up may be in many different forms. In addition to natural language texts, the information may be in table format, or in graphical images, audio, video or other forms.

This leads to an interesting question. What form should the markup itself take? Should the markup language for the Semantic Web be XML? Or should it be one of the XML-based languages such as RDF or DAML+OIL, or some combination of these? Or would it be better to use natural language as the markup language? When we talk about “the Roles of XML and Natural Language in the Semantic Web” we are referring to this question about what form the markup language should take, not merely to the existence of natural language texts in the Web.

1.3 Layers in the Semantic Web

Sometimes the Semantic Web is described in terms of a layer model. There are different versions of this model, such as the one by Tim Berners-Lee in Figure 1.
The layers we are most interested in are the XML layer, the RDF layer, and the ontology layer. We are not concerned here with the underlying Unicode/URI layer, and we do not have much to say about the upper layers dealing with logic, proof and trust.

To some extent, the different sections in this chapter focus on different layers. Paul Buitelaar introduces the overall architecture and describes the relationship between the different representation languages. Antonio Pareja-Lora deals mainly with the XML and RDF layers, while Barrett Bryant discusses DAML+OIL used in the ontology layer. As an alternative to these XML-based languages, Jimmy Lin puts forward an alternative proposal for a greater use of natural language.

1.4 Ontologies

The question at the panel session which generated the most discussion, for which we are indebted to Eduard Hovy, concerned the possibility or impossibility of having a single all-embracing ontology as the basis for the Semantic Web. Where does this

10 http://www.w3c.org/2000/Talks/1206-xml2k-tbl/slide10-0.html
question come from? I believe it arises partly from the overall vision of the Semantic Web, and partly from the arguments put forward for the benefits which the Semantic Web will bring. The Semantic Web is based on the World-Wide Web, which is a single all-embracing web, so the vision of the Semantic Web is that it will be a single all-embracing Semantic Web. Among the arguments put forward for the expected benefits of the Semantic Web is the vision that, once all the information in it is semantically annotated clearly and unambiguously in some agreed way, the Semantic Web will make smart searching and inferencing possible, and will eliminate all the confusion and contradictions in the present chaotic World-Wide Web. This annotation will be based on ontological classification, and it will only be fully effective once the Semantic Web reaches a certain “critical mass”.

This argument can clearly be challenged from a practical point of view, as the problems involved in building such a semantically annotated web on a global scale are enormous. However, the question raised at the panel session was not whether a single all-embracing ontology is difficult in practice, but whether it is even theoretically feasible. As an example, Eduard Hovy mentioned the question: how many colours are there? It is well-known that different languages and cultures divide the spectrum in various different ways, and there is no possible way to arrive at one single globally agreed list of colours. Similarly, different languages and cultures divide everything else in the world in different ways, and there is no possibility of a global, fundamental agreement about classification of time and space, or entities and events – in short, it seems that there is no possibility of a single all-embracing ontology.

One way to approach this issue is to remember the origins of the World-Wide Web and the Internet. The World-Wide Web is a single all-embracing web, based on the Internet which is a single all-embracing network, but the Internet grew from many small local networks based on local communications protocols, which were gradually joined together into regional and national networks and in some cases into industry-specific or sector-specific networks by adopting agreed protocol standards. Eventually these large networks joined together globally by internetworking based on the Internet Protocol. Similarly, the Semantic Web is growing from many applications based initially on small ontologies stored in local databases. These local ontologies will be combined with others to produce national or industry-specific or sector-specific ontologies, by adopting ontology standards and making the agreed ontologies more widely available. These wider ontologies will become part of the Semantic Web, which will have a sufficient critical mass to produce the benefits
predicted. This growth process will be accelerated because the vital importance of standards has been understood, and local applications based on local ontologies are already using W3C standard ontology languages from the outset. The ideas and experiences of some who have pioneered the use of these technologies are presented in the following sections.