Position Paper: Integrating Ontologies for Semantic Web Applications

ENABLER/ELSNET Workshop on International Roadmap for Language Resources
Paris 28-29/9/2003

Antonio Sanfilippo, Battelle/Pacific Northwest Division, USA
Ansgar Bernardi, Deutsches Forschungszentrum für Künstliche Intelligenz, DE
Ludger van Elst, Deutsches Forschungszentrum für Künstliche Intelligenz, DE
Luis Sánchez Fernández, Universidad Carlos III de Madrid, ES
Michael Sintek, Deutsches Forschungszentrum für Künstliche Intelligenz, DE

Introduction

Ontologies represent a pivoting resource for Semantic Web applications as they make available a knowledge representation language and a dictionary of classes and relations that Web services can use to describe content and reason about it. As the application scope for ontologies grows, it becomes increasingly clear that a general ontology cannot be made to satisfy the needs of different content domains and applications. Ultimately, ontologies must be easily composable if they are to be viable in terms of scalability and sustainability. The scope of this position paper is to exemplify the issue of composability for ontologies with specific reference to the needs of the news industry.

Ontologies and Semantic Web Standards

RDF (Resource Description Framework) provides an ideal encoding standard to make available ontologies to Semantic Web applications. However, RDF offers a limited set of semantic primitives and cannot therefore meet the requirements of a markup language for the Semantic Web (Heflin & Hendler, 2000).

A number of RDF-based semantic markup languages are now available for publishing and sharing ontologies on the World Wide Web. RDF-based markup languages such as the OWL (Web Ontology Language) offer appealing inference capabilities, and are likely to exert significant leverage on the development of ontology standards for the Semantic Web.

As the Semantic Web gains momentum, extensions to existing RDF-based semantic markup languages are developed which address the requirements of specific industry sectors. In the publishing industry, the PRISM initiative was launched about 2 years ago in the US to provide a set of metadata vocabularies that assist in the automation of publishing production processes and content exchange as an extension of RDF.

PRISM constitutes an important springboard for Semantic Web application in the news industry and has attracted a wide group of organizations including content providers, syndicators and vendors of content management, digital asset management and search & categorization software. However, PRISM does not cover news content encoding standards --- such the IPTC News Mark-up Language (NewsML), Subject Reference System (SRS) and News Industry Text Format (NITF) --- which are well established in the news industry. Moreover, PRISM does not offer the rich logical expressiveness of frameworks such as OWL, which is necessary to fuel advanced content intelligence technologies for Semantic Web applications. Finally, PRISM does not address the issue of multilinguality, which is a central concern for the news industry in the multi-cultural European market.

In keeping with the needs of the news industry and the requirements of Semantic Web applications, a likely solution would be to define RDF-based ontologies for the news domain using OWL or related frameworks as the ontological representational language, and the IPTC and PRISM news standards as the domain specific content ontologies. Such a solution would have to be extended to include generic content ontologies (e.g. SUMO, WordNet). In addition, multilingual issues would have to be addressed with reference to language-specific semantic taxonomies and thesauri in use by participating news agencies. For example, we could exploit the universality of the IPTC and PRISM standards to build a language independent ontology layer that is related via inter-lingual links to language specific taxonomies and thesauri, capitalizing on the experience of multilingual ontologies such as EuroWordNet.

NEWS Ontology

NEWS[1] is a new EC-funded project in FP6, due to start in a few months. The goal of the project is to develop News Intelligence Technology for the Semantic Web that will enable users to access, select and personalize delivery of multimedia and multilingual news content using advanced semantic-based annotation, query and inference machinery. Such technology is based on the integration of domain specific knowledge models with existing upper-level ontologies and Semantic Web standards to develop a standard vocabulary for semantic annotation of news information objects.

The NEWS Ontology will serve as the knowledge representation and exchange language for all content intelligence components in the NEWS system. Here is a summary of issues to be addressed.

- **Selection of an ontology representation language**: Existing standards for the representation of ontologies in the semantic web like RDF/S and OWL (Lite) will be reviewed and evaluated with respect to their adequacy as a basis for the NEWS ontology. Requirements for the expressive power of the representation language come from a study of user requirements and the information management strategy adopted. While the requirements analysis indicates which level of expressiveness is demanded by the news industry and has a chance to be accepted by the end user, the information management strategy dictates how rich a representation language can be exploited for high-precision content intelligence services. The final decision has to form a balance between these potentially conflicting demands.
- **Acquisition and formalization of the NEWS core ontology**: As shown in Figure 1, the NEWS ontology consists of an inner core and a set of peripheral modules. The core ontology comprises a *representational ontology, basic informational and content ontologies, as well as informational* and *content ontologies* for the news domain. The representational ontology defines the class of admissible logical relations that obtain among ontological objects. The information ontologies describe which *types* of objects exist, what *metadata attributes* they have, the range of values for these attributes, and the class of *content relations* between ontological objects. The content ontologies capture explicit knowledge about world objects, which is employed to describe the content of news articles in a suitable manner for machine resoning.

  The IPTC and PRISM news standards will be the mainstay of the information and content ontologies for the news domain. Some elements of upper ontologies developed by the Semantic Web community (e.g. SUMO, CYC etc.), general description schemas such as

[1] News Engine Web Services.

the Dublin Core, and WordNet may also be considered for inclusion. The multilingual thesauri and taxonomies selected by the user partners (ANSA and EFE) will form the multilingual component of the NEWS content ontology.

The representation of *time and temporal relations* will form an important focus point in the NEWS Ontology, due to the nature of the news domain and the specific requirements of the content intelligence services envisioned (e.g. trend discovery). Time and temporal relations will be considered with reference to the need to identify events and their temporal anchoring in newswires, and will touch both the informational and content ontology modules. For example, general time encoding types and relations will be treated in the informational ontology, while instances of such types and relations which are specific to the news domain will be accounted for in the content ontology. In developing a time ontology for NEWS, the project will capitalize on the experience of the ARDA-sponsored project TERQAS and the TIDES Temporal Annotation Guidelines (Ferro et al., 2001).

- **Integration with legacy systems and multilingual support**: The integration of legacy annotation systems into the News ontology will be staged in two main phases, as indicated in Figure 1. First, the IPTC and PRISM annotation standards and the multilingual thesauri from ANSA and EFE will be formalized as modules of the core NEWS Ontology. This will be done by reformulating the IPTC, PRISM, ANSA and EFE descriptions selected for inclusion in the NEWS Ontology according to the representation language chosen (e.g. OWL or RDF/S). *Interface Links* will be established between the NEWS reformulations and the original IPTC, PRISM, ANSA and EFE encoding to enhance clarity and ease integration for new users who are already using IPTC standards and wish to use NEWS services. Second, the NEWS content ontologies created from ANSA'a and EFE's multilingual thesauri will be linked to equivalent or subsuming objects in the NEWS core ontology via *Interlingual Links*. These *Interlingual Links* will be crucial in implementing cross-lingual search and navigation capabilities.

  Software tools will be developed to automate the reformulation of news industry standards, thesauri and taxonomies into the representation language of the NEWS ontology and to facilitate the linking of multilingual ontologies with (language independent) core ontology modules.

- **Embedding of ontology services in the NEWS information landscape**: While ontologies are typically seen as passive entities (e.g., stored at a specific location referred to by a URI) the dynamic and rather open NEWS scenario demands a more active view on ontologies. Ontologies have various "clients" (annotation services, content intelligence services), but also several contributors (information providers and consumers, trade discovery services) for their maintenance. The NEWS project will therefore wrap the NEWS core ontology as an ontology web service, which will be able to handle ontology utilization requests as well as maintenance operations (e.g., update proposals).
- **Maintenance Methodology**: While an information ontology typically is relatively stable, the content ontology in the news domain will lean towards frequent extensions and changes in order to keep its level of utility. New concepts gain interest and have to be described in more detail (e.g., stock market topics in standard news media at the beginning of the 90s in Germany) while others might become less relevant or associated with newly coined terms. In the NEWS methodology all users of the content ontology –

news providers and consumers – will be seen as source for potential updates. For example, an "often asked but seldom answered" query might indicate an update need as well as novel trends found with the trend discovery tool. The NEWS ontology maintenance methodology will provide procedures for a controlled handling of ontology updates. Revision concepts like the one included in OWL will be integrated and extended.

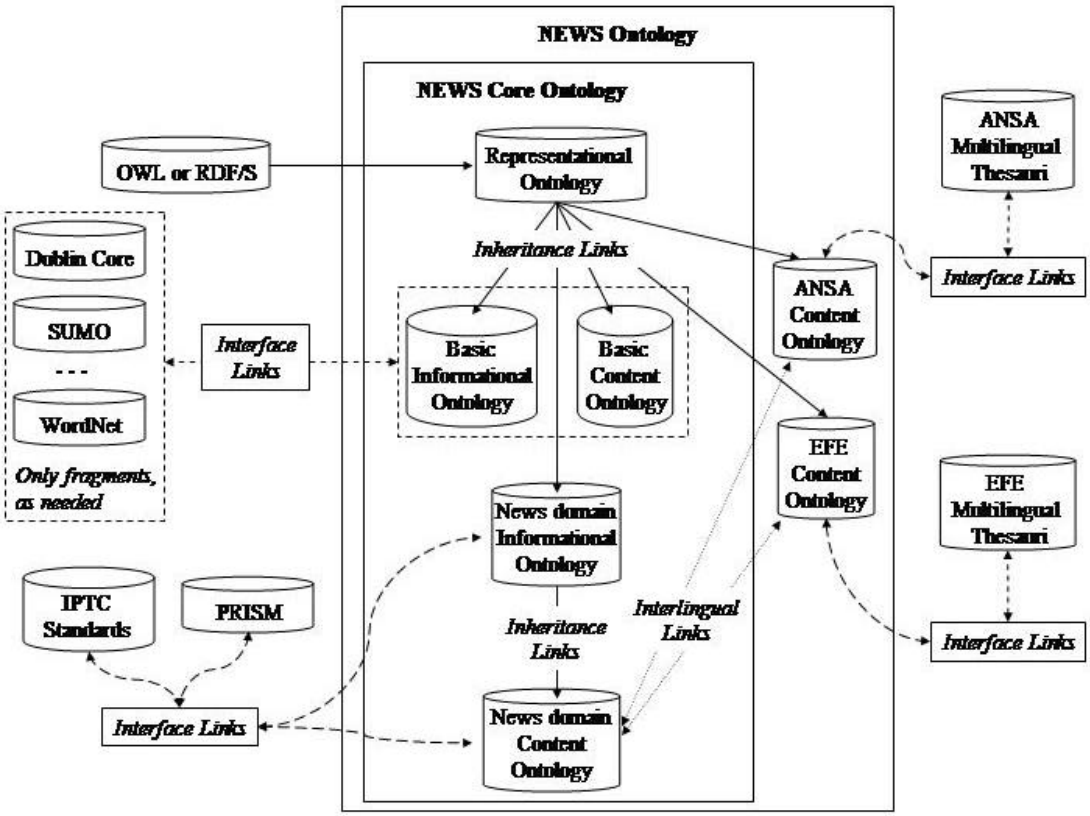

Figure 1: NEWS Ontology Structure.

References (TBA)

CYC: www.cyc.com.
DAML-S: http://www.daml.org/services/
EuroWordNet: http://www.illc.uva.nl/EuroWordNet.
Ferro, L., I. Mani, B. Sundheim, and G. Wilson (2001) *TIDES Temporal Annotation Guidelines.* MITRE Technical Report MTR 01W0000041. Available from http://www.cs.brandeis.edu/~jamesp/arda/time/readings.html.
Heflin, J, and J. Hendler (2000) Semantic interoperability on the Web. In *Proceedings of Extreme Markup Languages 2000.* Graphic Communication Association, Alexandria, VA.
IPTC NewsML, NITF, SRS: www.iptc.org.
PRISM: www.prismstandard.org.
RDF: http://www.w3.org/RDF.
SUMO: http://ontology.teknowledge.com.

TERQAS: http://www.cs.brandeis.edu/~jamesp/arda/time.
W3C: http://www.w3.org.
WordNet: http://www.cogsci.princeton.edu/~wn.