

Demo Abstract: A Web-based Document Harmonization and Annotation Chain: from PDF to RDF

Thierry Jacquin, Olivier Fambon, Boris Chidlovskii
Xerox Research Centre Europe
6, chemin de Maupertuis, F-38240 Meylan, France
{firstname.lastname}@xrce.xerox.com

ABSTRACT

We propose a demonstration of a Web-based document harmonization and annotation chain developed within the VIKEF integrated project. The chain integrates a combination of Web Services in order to access, harmonize and semantically annotate remote document collections. Annotations are then mapped onto RDF descriptions that serve as a basis for building semantic-enabled services to support community processes.

Categories and Subject Descriptors

H.3.5 [Information storage and retrieval]: Online Information Services—*Web-based services*; I.2.4 [Computing Methodologies]: Knowledge Representation Formalisms and Methods—*Semantic networks*; I.7.5 [Document and Text Processing]: Document Capture—*Document analysis*

General Terms

Design, Documentation, Management, Performance

Keywords

PDF, RDF, document annotation, Web Services

1. INTRODUCTION

We propose a demonstration of a Web-based document harmonization and annotation chain developed within the VIKEF integrated project. The chain integrates a combination of Web Services in order to access, harmonize and semantically annotate remote document collections. Annotations are then mapped onto RDF descriptions that serve as a basis for building semantic-enabled services to support community processes.

VIKEF (*Virtual Information and Knowledge Environment Framework*, <http://www.vikef.net>) is an Integrated Project within the IST Sixth Framework Programme of the European Community, which started in April 2004. Its main aim is to bridge the gap between the (partly) implicit knowledge and information conveyed in scientific and business information, content and knowledge resources and the explicit representation of knowledge required for an effective access, dissemination, sharing and reuse of such resources by scientific and business communities.

2. HARMONIZATION AND ANNOTATION CHAIN

The harmonization and annotation chain (see Figure 1) provides a number of integrated functions including the virtual document repository, reference management, two-level document harmonization, semantic annotation of textual and multimedia content, and ontology-based mapping of annotations onto RDF descriptions.

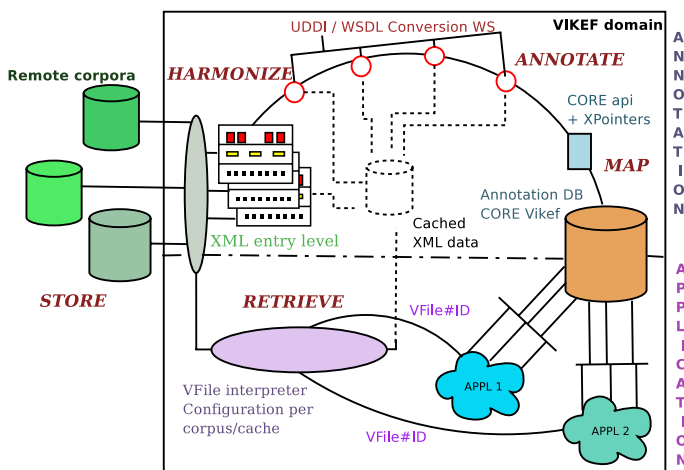


Figure 1: The harmonization and annotation chain (upper part) enables semantic services (lower part) in the VIKEF framework.

Virtual repository. The first component provides a unified and virtual storage for collections of documents, an access to remote or cached collections via virtual repositories, and the retrieval of entire documents or their fragments. Document sources (typically PDF) reside on remote hosts or in the cache, but are uniformly accessible via the URL mechanism. The URL Manager ensures an access to the different harmonized forms of the documents. The retrieval of document fragments is guided by RDF annotations which are exploited by a specific target application or service.

Harmonization. PDF documents are first converted into a rendering-equivalent XML format, by deploying a registered Web Service which makes the PDFtoXML rewriting and cleans it up with a document analysis component. Then the two-level harmonization schema is implemented to match a number of storage, access and

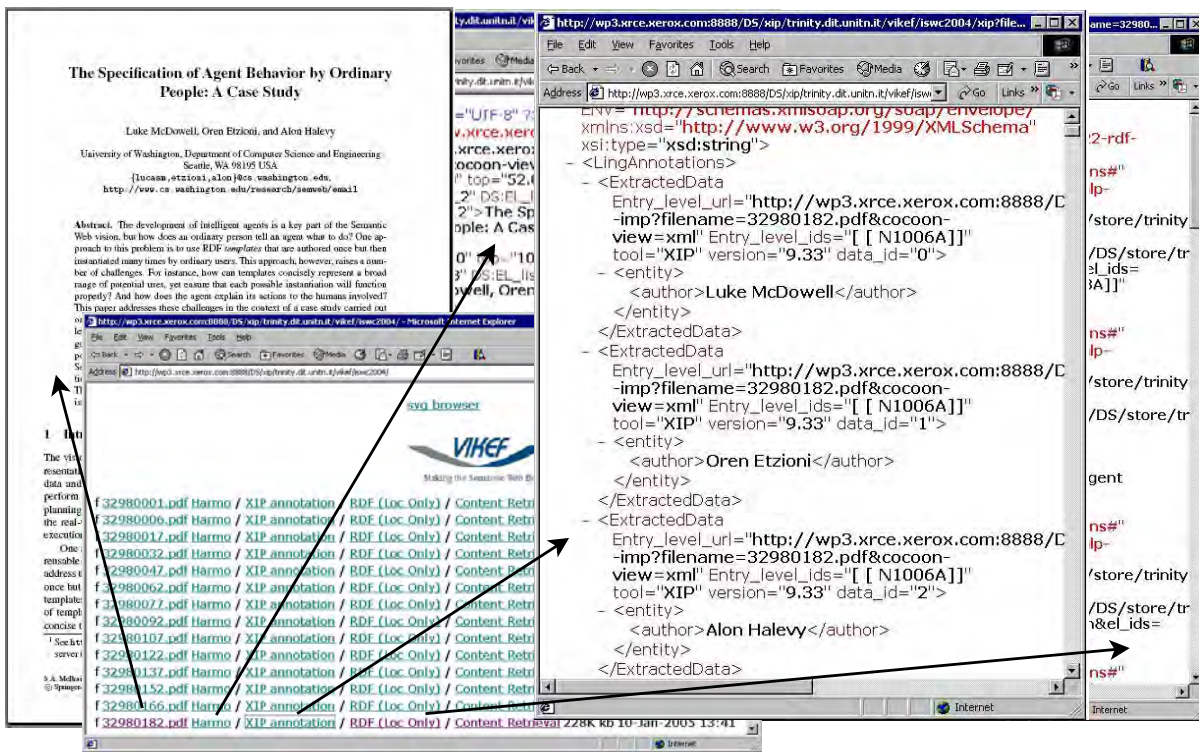


Figure 2: From a PDF document to RDF description, through harmonization and semantic annotation.

retrieval constraints. The *entry-level* content harmonization targets a clear separation between the annotation creation and the annotation reuse by applications; it guarantees that the document retrieval based on RDF annotations excludes reapplying the full harmonization and annotation chain on the remote documents. The entry-level harmonization is a generic and universal document indexing schema, according to which each XML node is assigned with a unique ID; this ID is preserved through the annotation and exploitation phase. Each annotation refers to a set of XML nodes through the mechanism of virtual ID XPointers.

The *second-level* conversion deploys additional conversion steps that may be needed for a specific annotation service. The second-level conversion preserves node IDs and stores annotations as ID sets. It allows also a word-based document indexing for the finer document annotations and supported by the corresponding schema. Additional conversion services are to be published and discovered through the VIKEF UDDI registry.

Semantic annotation and mapping. Semantic annotation refers to the categorization of document fragments according to a pre-defined ontological model and attaching the semantic tags to the classified fragments. The annotations are then used to associate different fragments and to discover semantic relationships between the fragments in the same or different resources.

The semantic annotation of the document content relies on natural language processors working on harmonized XML documents. Linguistic annotations are produced as stand-alone objects, they conform to the VIKEF XML annotation schema. On the next step of the chain, the linguistic annotations are mapped onto RDF annotation schema (see Figure 2), derived from classes in the referred ontology. The mapping operates through XLST rules that are either edited manually or generated through a dedicated mapping editor. Linguistic annotations are produced on the top of Xerox Incre-

mental Parser (XIP, see <http://wsportal.xrce.xerox.com:8080>); they cover entities, relations between entities and more advanced semantic relations (e.g. co-reference). Alternative annotation services can be integrated in the chain, in particular for processing documents in different languages or annotating the image content.

Retrieval. The fragments of source documents that correspond to selected RDF annotations get retrieved through URLs from the remote stores or from the local XML cache. The fragment identification by the URL includes the path to the source document, the corresponding entry-level IDs and, optionally, an application-specific instruction for further processing of the retrieved content. Only the entry-level harmonization step of the overall chain should be re-run at the retrieval time to ensure an access to the requested document fragments.

Web-based interface and Web Service composition. The Web interface (shown in Figure 2) allows to trace all steps of the harmonization and annotation chain, from the original PDF documents to the final RDF descriptions, through the harmonization and annotation steps. All components in the chain are declared as Web Services in the VIKEF UDDI registry. Any extension and recomposition of the chain is achieved by declaring new components and their discovery in the UDDI registry.

3. ACKNOWLEDGMENTS

We thank our XRCE colleagues Jean-Pierre Chanod, Hervé Dejean, Jean-Luc Meunier, Claude Roux, Jean-Yves Vion-Dury for their support, collaboration and contribution to various demonstration components.

This work is partially supported by VIKEF Integrated Project co-funded under the IST Sixth EU Framework Programme of the European Community.