Generating Adaptive Hypertext Content from the Semantic Web

David Millard, Harith Alani, Sanghee Kim, Mark Weal
Paul Lewis, Wendy Hall, David De Roure, Nigel Shadbolt

July 18, 2003

Abstract

Accessing and extracting knowledge from online documents is crucial for the realisation of the Semantic Web and the provision of advanced knowledge services. The Artequakt project is an ongoing investigation tackling these issues to facilitate the creation of tailored biographies from information harvested from the web.

In this paper we will present the methods we currently use to model, consolidate and store knowledge extracted from the web so that it can be re-purposed as adaptive content. We look at how Semantic Web technology could be used within this process and also how such techniques might be used to provide content to be published via the Semantic Web.

1 Introduction

The growth of the World Wide Web and the corpus of documents that it covers has increased the demand for content to be annotated to facilitate systematic search, discovery of knowledge and intelligent information processing.

Accessing and extracting knowledge from online documents is crucial for the realisation of the Semantic Web and the provision of advanced knowledge services. The collation of ontologically structured information from distributed web sites would provide the needed infrastructure for a variety of new services including the reconstruction of the original source material in new ways.

The Artequakt project seeks to create dynamic biographies by harvesting biographical information from the web, using the information to automatically populate ontologies, and then reconstruct these annotated fragments based on user preferences using story schema [10].

Annotating existing Web documents forms one of the basic barriers towards realising the Semantic Web [9, 15]. Manual annotation is impractical and unscalable, while automatic annotation tools are still in their infancy. Hence advanced knowledge services may require tools able to search and extract the required knowledge from the Web, guided by a domain conceptualisation (ontology) that specifies what type of knowledge to harvest.

We believe that the tools we are developing to generate our internal knowledge structures could also be used to automatically annotate existing pages for the Semantic Web.

The expertise and experience of three separate projects are drawn together under the umbrella of the Artequakt project. These are:

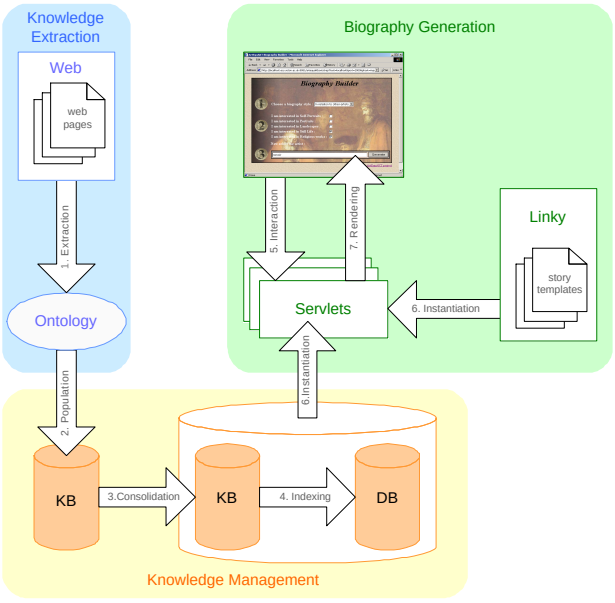

Figure 1: The Artequakt architecture

Sculpteur (formally Artiste) - A European project working on a distributed database of art images in collaboration with partners that include the Louvre, the Uffizzi Gallery, the National Gallery and the Victoria and Albert Museum.

Equator - An EPSRC funded Interdisciplinary Research Centre, exploring the relationship between the physical and the digital.

AKT - An EPSRC funded Interdisciplinary Research Centre looking at all aspects of the knowledge lifecycle.

2 System Overview

Figure 1 shows the key components of the Artequakt architecture. These can be broadly classified into three groups.

Knowledge extraction tools: The knowledge extraction tools are used to extract factual information items together with sentences and paragraphs from web documents. This information is coded up in rdf.

Knowledge management and storage: The RDF information is stored by the ontology server and consolidated into a structured knowledge base. Database indexes are used for speed of access to the original web-based content.

Biography generation: Story templates are used to structure queries into the knowledge and data bases. On user request the templates are instantiated and adaptive web pages produced.

Figure 2 illustrates a typical user interaction with the Artequakt system. In the initial screen the user enters the name of the artist and selects a type of biography to

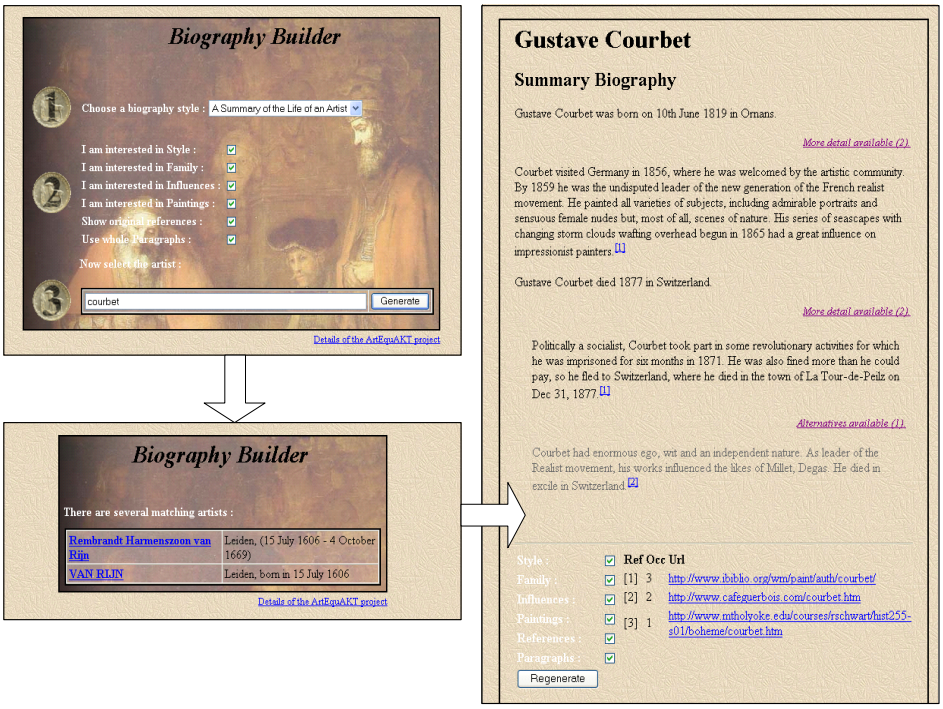

Figure 2: Screenshots of the web interface

generate, they also enter any preferences (for instance stating that they are not interested in the artist's personal life). If there is more than one artist in the knowledge base that matches the search criteria, the user is presented with a choice, shown in the second screen shot. Finally, once the user has selected an artist the generated biography is displayed for them.

3 Knowledge Extraction and Management

Ontologies play a significant role on the Semantic Web with respect to semantic enrichment and interpretation [2]. For Artequakt the ontology represents the domain of artists and artefacts used to structure the extracted information. The main part of this ontology was constructed from selected sections of the Conceptual Reference Model (CRM) ontology [5]. Artequakt uses Protégé, a graphical ontology editing tool [7]. Protégé provides an option to define ontology definitions in RDF Schema (RDFS) and has basic ontology reasoning capabilities. The extension of RDFS as a way of increasing ontological reasoning services for the full-fledged Semantic Web is exemplified in [3].

3.1 OWL

Part of the W3C's Semantic Web activity has been to define the Web Ontology Language (OWL). Defined using the Resource Definition Framework (RDF), OWL allows the formal definition of an ontology and provides the syntax for declaring instances of

Original sentences.

> Pierre-Auguste Renoir was born in Limoges on February 5, 1841. His father was a tailor and his mother a dressmaker.

```
<kb:Person rdf:about="&kb;Person_1"
         kb:name="Pierre-Auguste Renoir"
         rdfs:label="Person_1">
        <kb:date_of_birth rdf:resource="&kb;Date_1"/>
        <kb:place_of_birth rdf:resource="&kb;Place_1"/>
        <kb:has_father rdf:resource="&kb;Person_2"/>
        <kb:has_information_text rdf:resource="&kb;Paragraph_1"/>
</kb:Person>
<kb:Date rdf:about="&kb;Date_1"
         kb:day="25"
         kb:month="2"
         kb:year="1841"
         rdfs:label="Date_1">
</kb:Date>
<kb:E53.Place rdf:about="&kb;Place_1"
         kb:name="Limoges"
         rdfs:label="Place_1"/>
<kb:Person rdf:about="&kb;Person_2"
         rdfs:label="Person_2">
        <kb:has_work_information rdf:resource="&kb;Work_information_1"/>
</kb:Person>
<kb:Work_information rdf:about="&kb;Work_information_1"
         kb:job_title="tailor"
         rdfs:label="Work_information_1">
</kb:Work_information>
```

Figure 3: RDF representation of a paragraph

defined classes.

Eventually it is hoped that most information on the web will have an OWL representation (or similar structured metadata) and that the reliance on current extraction algorithms will be reduced. However, even if Semantic Web technology becomes pervasive, extraction tools will need to be employed to help people create OWL conformant documents and also to allow information extraction software to deal with the inevitable legacy information that is not marked up.

3.2 Information Extraction

Information Extraction (IE) is one of a number of promising methods for enriching Web-based documents with semantics for the purpose of future semantic interpretation. However, the time and effort needed to manually annotate large numbers of pages and the prerequisite of templates that stipulate which types of information are extractable are major challenges of exploiting such extraction techniques [15].

It is well-known that information on Web pages use effectively limitless vocabu-

laries, structures and composition styles for defining approximately the same content. This makes it hard for any IE technique to cover all variations of possible writing patterns. More importantly, traditional IE systems lack the domain knowledge required to pick out relationships between the extracted entities, which is essential for adding expressivity to the Semantic Web.

These observations lead us to the use of an ontology coupled with a general-purpose lexical database, WordNet [13] and an entity-recogniser, GATE (General Architecture for Text Engineering [6]) as guidance tools for identifying knowledge fragments consisting not just of entities, but also the relationships between them. Automatic term expansion based on WordNET is used to increase the scope of text analysis to cover syntactic patterns that imprecisely match our definitions.

When a user searches for an artist, if the given artist is new to the KB, the Information Extraction process is run. A script submits the query to search engines (currently we use 'Google', 'Altavista' and 'Yahoo'). In order to select only art-related Web pages (as opposed to pages which may match the search criteria but are concerned with other topics) we use keywords extracted from trusted sites as a basis for measuring similarity between the query and the search results.

In order to construct semantically rich information, it is necessary to extract binary relationships between any identified pair of entities to gather structured collections of information [2]. Therefore, knowledge about the domain specific semantics is required, and can be inferred from the ontology. Artequakt submits a query to the ontology server to obtain such knowledge. In addition, three lexical chains (synonyms, hypernyms, and hyponyms) from WordNet are used to reduce the problem of linguistic variation between syntactically different entities. For example, the concept of 'depict' can be matched with 'portray' (synonym) and 'represent' (hypernym).

By providing the IE process with direct access to the concepts and relations in the ontology, our approach is applicable across more than one domain.

The output RDF representation (Figure 3) is submitted to the ontology server to be inserted into the KB. It would be possible to use this RDF to annotate the existing pages for the Semantic Web. Figure 4 shows how the previously extracted paragraph of text might be represented in OWL.

3.3 Automatic Ontology Population

Some semi-automatic approaches have investigated creating document annotations and storing the results as assertions in an ontology. For example, in Vargas-Vera [14], relationships were added automatically between instances only if these instances already exist in the KB, otherwise user intervention is needed. Handschuh et al [8] describe a framework for user-driven ontology-based annotations, enforced with an IE learning tool; Amilcare [4]. However, the framework lacks the capability of identifying relationships reliably.

In Artequakt we investigate the possibility of moving towards a fully automatic approach of feeding the ontology with knowledge extracted from the Web. Information is extracted in Artequakt with respect to a given ontology (e.g. the artist ontology described earlier), and provided as RDF files, one per document, using tags mapped directly from names of classes and relationships in that ontology (Figure 3).

One of the difficulties that arises when extracting similar or overlapping information from different sources is the consolidation of duplicate information. Tackling this problem is important to maintain the referential integrity and quality of results of any

"Pierre-Auguste Renoir was born in Limoges on February 5, 1841. His father was a tailor and his mother a dressmaker."

```
<rdf:RDF
    xmlns           ="http://www.example.org/renoir#"
    xmlns:artequakt ="http://www.artequakt.ecs.soton.ac.uk/2002/artequakt#"
    xmlns:owl       ="http://www.w3.org/2002/07/owl#"
    xmlns:rdf       ="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs      ="http://www.w3.org/2000/01/rdf-schema#">
   <owl:Ontology rdf:about="http://www.example.org/renoir">
      <owl:imports rdf:resource="http://www.artequakt.ecs.soton.ac.uk/2002/artequakt.owl"/>
      <owl:Class rdf:ID="TailorJob">
         <rdfs:subClassOf rdf:resource="&artequakt;Job"/>
      </owl:Class>
      <artequakt:Region rdf:ID="LimogesFrance">
         <owl:locatedIn rdf:resource="&artequakt;France"/>
      </artequakt:Region>
   </owl:Ontology>
   <artequakt:Text rdf:ID="Text1">
      <content rdf:datatype="string">Pierre-Auguste Renoir was born in Limoges
      on February 5, 1841. His father was a tailor and his mother a dressmaker.</content>
   </artequakt:Text>
   <artequakt:Person rdf:ID="Person1" >
      <name          rdf:datatype="string">Pierre-Auguste Renoir</name>
      <birthLocation rdf:resource="#LimogesFrance"/>
      <father        rdf:resource="#Person2"/>
      <birthYear     rdf:datatype="year">1841</birthYear>
      <textsrc       rdf:resource="#Text1/>
   </artequakt:Person>
   <artequakt:Person rdf:ID="Person2" >
      <vocation      rdf:resource="#Tailor"/>
      <parentof      rdf:resource="#Person1"/>
      <textsrc       rdf:resource="#Text1/>
   </artequakt:Person>
</rdf:RDF>
```

Figure 4: Artist information represented in OWL

ontology-based knowledge service [1]. A description of Artequakts approach to this problem can be found in [10].

4 Biography Generation

While the Semantic Web promises to ease the problems of machine interaction, many of its applications will be attempting to sort, arrange and present information to people. Ontologies are appropriate vocabularies for machines, but human beings need a more natural interface.

Story telling provides this kind of intuitive mechanism. We can consider the structured information on the Semantic Web (consolidated by a system such as Artequakt

into a knowledge base) as the underlying story, waiting to be told. The fragments of text in the knowledge base can be re-ordered and combined with generated sentences to produce an eventual discourse, personalised to a particular reader and drawing on the resources of many Semantic Web sites.

The Artequakt system uses biography templates to arrange the information in the knowledge base into a narrative. It then renders that into a DHTML page so that the personalised biography can be displayed in a web browser.

4.1 Biography Templates

The structures we use to arrange the story are human authored biography templates that contain queries into the KB. The templates are written in the Fundamental Open Hypermedia Model (FOHM) [12] and stored as XML in the Auld Linky contextual structure server [11]. As the templates are stored in a structure server they can be retrieved in different contexts and thus may vary according to the user's preferences and experience.

Each biography is a tree of sub-structures and queries. The most common structure is a *sequence*, this represents a list of queries that should be instantiated and rendered in order. Each query uses the vocabulary of the Artequakt ontology to discover fragments of text concerned with a particular aspect of the artist's life. Other structures allow for more complex behaviour. *Concept* structures are used to group alternative queries together, any of which may be successfully used at that particular point in the biography. A *Level of Detail* (LoD) structure is similar, but orders the queries so that the most preferable is given the highest index.

The fact that fragments of text are associated with facts in the knowledge base is useful as it allows real text to be used in the final biography in preference to generated text. These fragments have been extracted from existing larger texts and so contain elements of discourse (focalisation, tense information etc.). We are currently looking at how we might detect these attributes to ensure that the generated biographies are consistent (e.g. to ensure that a biography in the third person does not include a paragraph in the first person).

As the attributes of existing text might preclude it from being used the Artequakt system also allows the knowledge base to be queried directly and basic natural language generation to be used to render them into the biography. This might also be useful for facts in the knowledge base that have been inferred (and for which there is no corresponding text).

The story renderer uses the information in the knowledge base to keep track of which facts are being placed into the text during generation. In this way it can minimise repetition. It uses the structures of the template to chose which content to display by default but also uses Adaptive Hypermedia techniques to optionally reveal other content which may be relevant (using stretchtext and dimming secondary fragments).

5 Conclusions

The system discussed here integrates a variety of tools in order to automate an ontology-based knowledge acquisition process and maintain a knowledge base with which to generate customised biographies.

We believe that the automatic biography generation approach presented here should be applicable to other domains with few changes. For example, the current artist on-

tology could be replaced with a researcher ontology, where the extraction is expected to focus on information about research activities. Since the relation extraction between entities in our Artists ontology is mostly determined by the main type of verb used in the source text, we can expect similar extraction performance from a research ontology if the research activities are also identifiable from such a main verb. With respect to entity recognition tools, domain specific entities (e.g. publication styles) need specialised extraction rules that have to be modified when the domain changes.

Research on the Semantic Web involves Web engineering, knowledge management, agent systems and logic programming all aiming at enriching given Web-pages with semantic expressibility. When the fully-fledged Semantic Web is realised, it might seem that IE tools are no longer required, however, we believe that the extraction results can act as additional information sources to enrich the Semantic Web and may help build it.

In addition we see Artequakt's use of ontologically described content to construct adaptive web pages as an example of how hypertext systems may take advantage of Semantic Web technology. We are currently looking at adding more navigational structures to reflect the position of the individual biographies within the wider network. As a simple first step we might add Generic Link support based on names which would resolve to other generated pages. In the longer term, and as our ontology and domain are extended, the choice of template would itself become adaptive and the choice of which terms to treat as link sources would be automated.

The Artequakt project provides a framework for investigating issues that will be important to future Semantic Web applications such as automatic ontology population, consolidating duplicated, or even contradictory information, and developing Information Extraction techniques to help annotate the Web and form the basis for adaptive hypertexts.

6 Acknowledgments

This research is funded in part by EU Framework 5 project "SCULPTEUR" , EPSRC IRC project "Equator" GR/N15986/01 and EPSRC IRC project "AKT" GR/N15764/01.

References

[1] H. Alani, S. Dasmahapatra, N. Gibbins, H. Glaser, S. Harris, Y. Kalfoglou, K. OHara, and N. Shadbolt. Managing reference: Ensuring referential integrity of ontologies for the semantic web. In *In Proceedings of the 13th International Conference on Knowledge Engineerind and Knowledge Management*, Spain, 2002.

[2] T. Berners-Lee, J. Hendler, and O. Lassila. Semantic web. *Scientific American*, 2001.

[3] J. Broekstra, M. Klein, S. Decker, D. Fensel, S. Staab, F. van Harmelen, and I. Horrocks. Enabling knowledge representation on the web by extending rdf schema. In *In Proceedings of the Tenth International Conference on World Wide Web*, pages 467–478, Hong Kong, 2001.

[4] F. Ciravegna, A. Dingli, Y. Wilks, and D. Petrelli. Timely and non-intrusive active document annotation via adaptive information extraction. In *Workshop*

on Semantic Authoring, Annotation & Knowledge Markup (SAAKM'02), 15th European Conference on Artificial Intelligence (ECAI'02), pages 7–13, France, 2002.

[5] N. Crofts, D. Dionissiadou, and M. Stiff. Definition of the cidoc object-oriented conceptual reference model. Technical report, International Organization for Standardization, 2000.

[6] H. Cunningham, K. Bontcheva, V. Tablan, C. Ursu, and M. Dimitrov. Developing language processing components with gate (user's guide). Technical report, University of Sheffield, U.K., 2002. available in http://www.gate.ac.uk/.

[7] H. Eriksson, R. Fergeson, Y. Shahr, and M. Musen. Automatic generation of ontology editors. In *12th Workshop on Knowledge Acquisition, Modelling, and Management (KAW99)*, Canada, 1999.

[8] S. Handschuh, S. Staab, and A. Maedche. Cream - creating relational metadata with a component-based, ontology-driven annotation framework. In *In Proceedings of the First International Conference on Knowledge Capture*, pages 76–83, Canada, 2001.

[9] J. Kahan and M.-R. Koivunen. Annotea: an open RDF infrastructure for shared web annotations. In *World Wide Web*, pages 623–632, 2001. http://citeseer.nj.nec.com/kahan01annotea.html.

[10] S. Kim, H. Alani, W. Hall, P. H. Lewis, D. E. Millard, N. Shadbolt, and M. J. Weal. Artequakt:generating tailored biographies with automatically annotated fragments from the web. In *In Proceedings of the Workshop on Semantic Authoring, Annotation & Knowledge Markup*, pages 1–6, France, 2002.

[11] D. Michaelides, D. Millard, M. Weal, and D. DeRoure. Auld leaky: A contextual open hypermedia link server. In *Hypermedia:Openness, Structural Awareness, and Adaptivity (Proceedings of OHS-7, SC-3 and AH-3), Published in Lecture Notes in Computer Science, (LNCS 2266), Springer Verlag, Heidelberg (ISSN 0302-9743)*, pages 59–70, 2001.

[12] D. Millard, L. Moreau, H. Davis, and S. Reich. FOHM: A Fundamental Open Hypertext Model for Investigating Interoperability Between Hypertext Domains. In *Proceedings of the Eleventh ACM Conference on Hypertext and Hypermedia, San Antonio, Texas, USA*, pages 93–102, 2000.

[13] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Introduction to wordnet: An on-line lexical database. Technical report, University of Princeton, U.S.A., 1993.

[14] M. Vargas-Vera, E. Motta, and J. Domingue. Knowledge extraction by using an ontology-based annotation tool. In *In Proceedings of the Workshop on Knowledge Markup and Semantic Annotation*, K-CAP'01,Canada, 2001.

[15] R. Yangarber and R. Grishman. Machine learning of extraction patterns from unannotated corpora: Position statement. In *In Proceedings of Workshop on Machine Learning for Information Extraction*, pages 76–83, ECAI,Berlin, 2001.