Applying the Semantic Web: The VICODI Experience in Creating Visual Contextualization for History

Gábor Nagypál
FZI Research Center for Information Technologies, Germany

Richard Deswarte
University of East Anglia, UK

Jan Oosthoek
University of Newcastle upon Tyne, UK

Abstract

Semantic Web applications in the humanities that visualize knowledge are still few and far between. The Visual Contextualization of Digital Content (VICODI) project brought together Semantic Web technologies with the concepts of contextualization and visualization of knowledge, an approach which we term visual contextualization. The goal was to enhance users' understanding of digital content in the domain of history. It succeeded in doing this by creating an ontology-based web portal of European history where extra historical knowledge or 'context' is added to resources and visualized through textual hyperlinks and interactive Scalable Vector Graphics historical maps. VICODI also created a history-specific ontology. In this article the novel approach of visual contextualization is introduced in conjunction with a detailed explanation of the core elements of the VICODI portal. The article also addresses several of the problems encountered in developing a Semantic Web application for a humanities domain.

Correspondence:
Richard Deswarte,
School of History,
University of East Anglia,
University Plain,
Norwich, NR4 7TJ, UK.
E-mail:
r.deswarte@uea.ac.uk

1 Introduction

Over the past few years the World Wide Web has become a main source of information for millions of people around the world. This increased usage has resulted in the creation of various technologies to

Literary and Linguistic Computing, Vol. 20, No. 3, 2005.
doi:10.1093/llc/fqi037 Advance Access Published on 22 July 2005

help people organize the huge amounts of information and make it more comprehensible. The development of advanced search engines, such as Google and Yahoo, has made information search and retrieval much more productive. However, simple or general queries presented to such advanced full-text search engines, still return too many irrelevant resources. Indeed the required pieces of information remain hidden among the mass of retrieved links presented in a multi-page list with little or no visual interface. For example one study on web site usability concludes that users can find the information they are searching for only 42% of the time (Spool *et al.*, 1999). This is no surprise, as the precise information requested by users usually cannot be expressed in general search terms alone.

On the other hand, advanced search interfaces that allow specific fine-grained and complex queries overwhelm most users. Instead of precise queries expressed via a powerful search interface, users usually prefer to start with relatively simple queries, refining and modifying them in favour of alternatives in subsequent iterations. In many cases, a search result is only a starting point for exploring an archive or the Internet.

The goal of the Visual Contextualization of Digital Content (VICODI) project was to enhance people's comprehension of digital content on the Internet by providing a solution to the above problem as well as addressing the deeper theoretical and technological issue of turning information into knowledge. As already discussed, present-day Internet search engines provide people with long lists of unrelated links. Further these links are in no way connected or logically structured. In effect the links represent vaguely organized information, a far cry from knowledge. Contextualization provides a solution to this lack of knowledge. It is a way of organizing these lists of information into coherent related groups of documents that are numerically limited by the nature of their content. Through this process information is raised to a higher knowledge level.

A further challenge is how to present this contextualized information in a way that is easy to comprehend and interpret by users. This can be done by visualizing it in intuitive and interactive formats, such as images, maps or colour codes to mention only a few. In addition, an interface search tool is required that effectively converts raw information into a visually intuitive format that brings, in the case of history, key issues, events, objects etc. to the immediate attention of the user. Therefore visualization aids the contextualization process of enhancing basic information.

Thus the main aim of the VICODI project was to create a semantic web application that would bring together contextualization and visualization in a way that assists in raising information to the knowledge level. This goal was achieved by specifying and implementing this novel approach, which we termed *visual contextualization*. During the VICODI project (Deswarte *et al.*, 2004) we used European history as the showcase domain and developed a web portal, which

demonstrates the idea of visual contextualisation based on historical resources (mainly textual articles) stored in our digital archive.

In this article we provide a general overview of the VICODI system, introduce the main concepts, discuss our solutions in general terms and report on the lessons learned[1]. The structure of the article is as follows. In Section 2 we discuss the main theses behind the VICODI system and in Section 3 we provide a high-level overview of the VICODI workflow from the user's perspective. Section 4 enumerates the main VICODI features and describes the components of the technical architecture. In Sections 5–7 we discuss the different interesting aspects of the system in more detail and analyze related work. Section 8 concludes the article with some critical comments and suggestions for further research.

2 The Main Theses of VICODI

The main objective of VICODI is the development of a methodology and tools for the comprehensive structuring and graphic visualization of context. This context will then facilitate the creation of new knowledge.

Underpinning many so-called context-based systems is the thesis that considering a user's context can improve the results produced by an information system. This thesis is based on the insight that while context plays an important role in human-to-human communication, context is ignored in most of the present computer systems (Dey, 2001).

In VICODI there is no explicit or implicit representation of a user's interest or context. Instead, users express their interest through sequential navigation and multiple subsequent searches while they explore the content of the VICODI information system. Thus, one of the most important insights is that by reading a document the user accepts the meaning of the document—its document context—as a part of his or her user context. Therefore the context of the actual document can be used to filter and rank the results of subsequent queries. For instance, if the user has just read an article about World War I and requests more information about Serbia, then we can assume that he is not interested in the 1998 Kosovo conflict, but in documents about Serbia from the beginning of the 20th century.

The document context is extracted semi-automatically from digital resources, i.e. semantic metadata describing their context is created. The purpose of semantic metadata is twofold. First, the resource context may assist the users in retrieving other related resources with a similar context as the original search result. Second, the context of a resource is visualized through maps and colour-coded links, which assist the user to better re-construct the context of the information, and this helps it to raise the resource to the knowledge level. In doing so the VICODI approach refers to the well-known

1 More information about the project is available from http://www.vicodi.org

metaphor in knowledge management, which states 'knowledge is information in context' (Rumizen, 2002).

Combining visualization and context-based enhancement of information system output, results in a unique approach which we have termed visual *contextualisation*. It is important to note that this approach is not limited to textual documents, but it is applicable generally to digital resources including pictures or videos, although in the project we used almost exclusively textual documents.

During the VICODI project our theses were validated by developing a web portal for European history resources. The remainder of this article will focus on the specific features of the system and on our development experience. Although history is the chosen knowledge domain it must be remembered that the approach of visual contextualization itself is a general approach and the results are transferable to other application domains as well.

3 Typical User Interaction With the VICODI Portal

There are several ways to retrieve and visually contextualize documents using the VICODI prototype portal (Eurohistory). The first method and probably the most obvious means to retrieve documents and set a historical or geographical context is through the historical maps of Europe. The user can specify the initial context visually by selecting a time period and clicking on a particular country in combination with the category buttons—politics, culture, economics and social (see Fig. 1). Alternatively users may choose to execute a simple text search for ontology labels, which initiates the context assembly process. Through this option users can build a context search query by adding instances from the ontology. The ontology elements can be found by using an ordinary full-text search of the ontology to find the requested instance. When the user is satisfied with the list of ontology instances setting the context, e.g. Elizabeth I, Spanish Armada, and Sir Francis Drake to find documents related to this famous sea battle, the system then conducts a full search of documents for a combination of the instances set as the context. Both map-based and assembled context-based searches result in a Google-style list of related resources, i.e. resources with a similar resource context. These resources can in turn be used to access additional relevant documents by means of links in the document or the maps.

When retrieved by the user, the resources, in this case historical documents, will be contextualized based on certain parameters (keywords, years, names, etc.) that the system uses to generate the appropriate context for the whole article. In the text, colour-shaded hyperlinks represent further possible search queries to the system. Contextual data in the form of interactive graphical maps displaying the most relevant geographical locations is also provided. Thereby VICODI presents a visually enriched interface to navigate both through

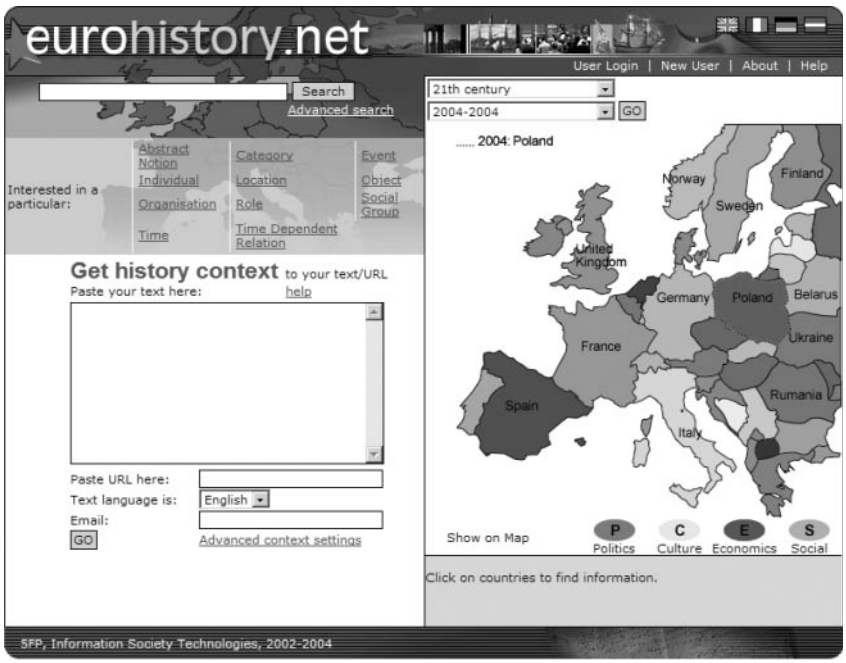

Fig. 1 The VICODI Graphical User Interface. A colour version of this figure can be found in the Supplementary Data section on the journal's website at www.llc.oupjournals.org.

space and time. For example, if the user specifies an article about Napoleon, the system will automatically generate a 19th century map from predefined Scalable Vector Graphics (SVGs) drawings and will put links below the map as well as indicating France as the most important location (see Fig. 2).

The user can further navigate through the repository of historical texts to find additional resources by clicking either on the contextualized (hyperlinked) terms in the document or on the map to the right of the text. This generates a new list of documents, which the system considers the most relevant to the context of the map or hyperlinked term on which the user last clicked. For example clicking on the link 'Russia' in a document about Napoleon will provide documents related to Russia in the context of Napoleon, e.g. documents about the occupation of Moscow by Napoleon, and not about Russia in general.

Users may also upload their own historical documents for contextualization and then use the hyperlinks, added by the system, to find additional and related texts. However the initial contextualization process demands considerable memory and processing power so the results are not instantaneous but posted via email usually within several minutes.

In summary, the user interacts with the system by setting an initial context in three ways by either selecting a map, by assembling a context description through conducting text searches on ontology labels or

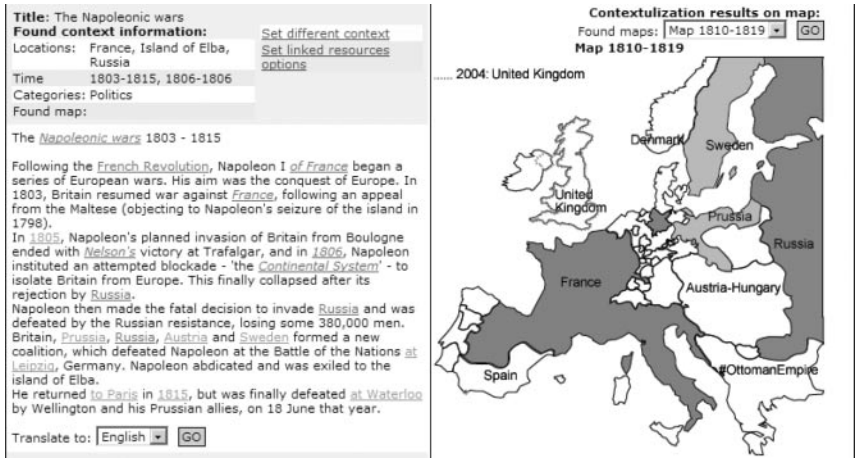

Fig. 2 A Contextualized Resource. A colour version of this figure can be found in the Supplementary Data section on the journal's website at www.llc.oupjournals.org.

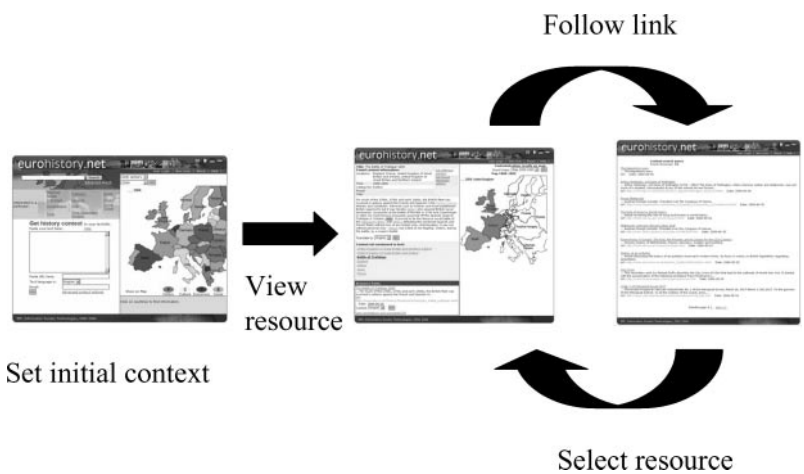

Fig. 3 VICODI portal interaction workflow. A colour version of this figure can be found in the Supplementary Data section on the journal's website at www.llc.oupjournals.org.

by uploading a document for contextualization. Based on the initial context the system will produce a list of resources, which can help to refine the search result. The links in the documents can in turn be used to find further resources when the process described above will be repeated. This defines a simple 'click on link and pick next document' workflow (see Fig. 3).

4 VICODI Features and Components

To implement the described workflow the following major features and components were needed to make the portal function properly (the key individual components will be described in subsequent sections).

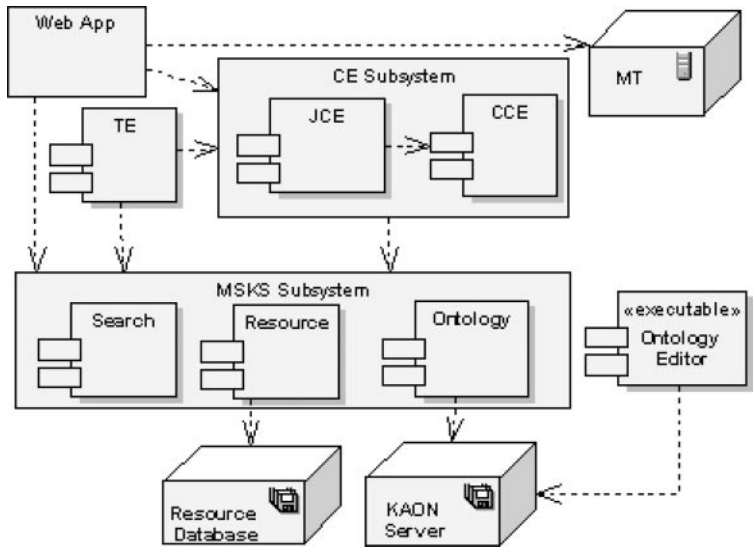

Fig. 4 The VICODI system architecture. A colour version of this figure can be found in the Supplementary Data section on the journal's website at www.llc.oupjournals.org.

4.1 Features

4.1.1 *An ontology of European history*

The domain-specific background knowledge in VICODI is encoded in an ontology of European history. While there are many definitions of an ontology (Gómez-Pérez *et al*., 2004), in essence, it is the formal hierarchical specifications of concepts in a knowledge domain, such as history, where relationships between the specific instances of the various concepts are formally defined. In VICODI the ontology specifies the vocabulary for context definitions and is used by various heuristics during context generation and context-based search.

4.1.2 *Automatic context generation*

When new documents are added to the system a context estimation has to be automatically generated to make retrieval and visual representation possible.

4.1.3 *Context visualization*

The document context is visualized to help user comprehension. Various visual cues, including hyperlinks and SVG historical maps, are automatically inserted or linked to the original document content.

4.1.4 *Context-based search*

A stored resource context is not only useful for visualization, but also makes it possible to provide a so-called 'context-based' search in addition to classical full-text search. This is executed when the user clicks on a link or one of the individual elements in the SVG maps as well as by building a query through the search tool.

4.1.5 *Multilinguality*

The VICODI system presently supports four languages: English, German, French and Latvian. The system is multilingual on all levels, providing machine translation of document content, a multilingual ontology, a multilingual user interface, a multilingual full-text search and language-independent context-based search. The translation of document content and ontology is done with the XML-based, modular, new generation machine translation infrastructure of Systran (Senellart *et al.*, 2003).

4.2 Components

The VICODI system prototype has been developed as a typical Java web application based on the Model-View-Controller (MVC) pattern. The implementation is based on the Expresso Application and Architecture Framework (Expresso Project).
The components which form the VICODI architecture (Fig. 4) are:

(1) MSKS—Management System of Knowledge Space
This component consists of three subcomponents: Ontology, Resource and Search. The MSKS ontology component provides an Application Program Interface (API) to the underlying open-source KAON framework (Motik *et al.*, 2002), which manages the VICODI ontology stored in a PostgreSQL database (other types of relational databases are also supported). The MSKS Resource module provides access to the resources stored in the repository, which is also a PostgreSQL database. This ensures access to the resource metadata including the context information. The search module allows for two types of searching—context-based search and ordinary full-text search. To execute a context-based search the MSKS component relies heavily on the Contextualization Engine component.

(2) CE—Contextualization Engine
This component includes the client CE (Java CE—JCE) and the remote Computational CE server (CCE). One function of the CE module is to automatically generate context data of resources newly submitted to the system. Another function of this module is to support the context search by providing various methods to calculate pair context similarity.

(3) TE—Transformation Engine
The TE implements the core of the text transformation visualizations and supports the XML-based SVG historical maps by producing location contexts of a resource for use when rendering the SVG historical map. This component builds on the document contexts that were produced by the CE component.

(4) MT—Machine Translation Server
The MT Server of Systran (Systran—White Papers) is accessed remotely in the system. The HTML/XML fragment translation is available for all of the supported languages.

(5) Web application
The web application is the portal user interface of the VICODI system, which is found under the URL www.eurohistory.net. The portal builds on the functionality of the already described modules (MSKS, CE, TE and MT). It also manages several off-line jobs for computing heavy and longer processes, such as contextualization, resource and ontology translation and resource as well as ontology indexing jobs.

(6) Ontology Editor
The Ontology Editor is a stand-alone JAVA Graphical User Interface (GUI) application based on the OIModeler component of the KAON framework (KAON Semantic Web). The interface allows graphical editing of the ontology structure and ontology instances. In addition to visual editing, the VICODI system also supports the mass upload of ontology instances from standard Excel spreadsheets. Using this feature large numbers of ontology instances were added easily and quickly to the VICODI history ontology.

5 The VICODI Ontology

As was discussed earlier, the ontology models a knowledge application domain—in our case European history. It plays a central role in VICODI since it defines the vocabulary of context definitions and provides background knowledge needed for sophisticated heuristics during context generation and context-based searching. In this section the methodology used in the ontology development and the lessons learned during that process are discussed (see also: Nagypál, 2004).

Developing an ontology from scratch is a time- and resource-intensive enterprise. Therefore a number of related ontologies and thesauri were examined in the hope that it would be possible to reuse them with only minor modifications. Unfortunately no suitable ones were found. They were either too general in scope and thus overly complicated, for example Cyc (Cycorp), IEEE SUMO ontology (SUMO Ontology) or CIDOC CRM (The CIDOC Conceptual Reference Model) or on the contrary, they were too specific for example the Getty Thesaurus of Geographic Names (GTGN Online), the CULTOS ontology (Cultural Units of Learning—Tools and Services), the ABC ontology (Lagoze and Hunter, 2001) or an ontology on Italian opera (ONTOPIA). Some of the relevant thesauri contained an ad hoc structure, which rendered them unusable for our purpose of storing machine-processable domain-specific background knowledge. Examples of this category are the Hassett Thesaurus (Humanities and Social Science Electronic Thesaurus) and the UNESCO Thesaurus (UNESCO Thesaurus). Moreover, most of these ontologies and thesauri did not contain any instances, which is absolutely crucial in an application ontology.

For these reasons a new ontology was needed. Of course the VICODI ontology could have integrated some special-purpose ontologies, like the Getty Thesaurus for Geographic Names. However this integration did not happen in the project due to the fact that the possible candidates—Getty Thesaurus of Geographic Names and other Getty Thesauri (Getty Vocabulary Databases)—are not freely available. In addition, for the purposes of the VICODI project and in order to demonstrate the viability of visual contextualization creating a sample set of instances was more feasible. The justification was that a narrower coverage for our demonstration purposes was needed and it had to be on a higher knowledge level. This higher level included specifying existence times for the specific ontology instances, validity times of instance relations as well as geographical locations. None of the above were available in any of the aforementioned thesauri. It was agreed that evaluating, filtering, converting, extending and integrating the above thesauri would be more time-consuming and expensive than creating our own corpus of related instances. The developed VICODI ontology is available for interested parties on the VICODI website under the GNU Free Documentation License (GFDL).

5.1 Problems of modelling history in an ontology

The goal of an ontology is to store all the knowledge needed for successful automatic context generation and context search. In VICODI this meant that the ontology had to be an *application-level ontology*, which could be directly exploited by intelligent algorithms (Guarino, 1998). To successfully automate context generation a great body of historical facts was needed.

Although we did limit the ontology scope somewhat, focussing only on European history from 500 CE to the present, it is still an immense and very complex knowledge domain. Therefore it was extremely difficult to find a good balance between an overly complex detailed ontology, which is infeasible to build and impossible to manage and exploit in algorithms, and an overly general upper-level style ontology of history, which is useless in a real-world application. In addition the historians compiling the ontology focused on a limited number of subjects that had a strong impact on European history, such as the Enlightenment, the Scientific Revolution, the World Wars and European integration amongst others.

On the whole we had to constrain ourselves to building an ontology with a stable and simple upper-level structure, which could be explored and understood by computing scientists who develop intelligent algorithms for context generation and who are usually not historians. Yet, at the same time the knowledge encoded in the ontology had to be 'proper' from a historical point of view.

In addition to the inherent conflict between full-coverage of a complex domain and a simple, understandable ontology structure,

an issue that exists for most complex application domains, history also caused many other problems as a domain including the following:

(1) Time dependence
In history practically every instance is time dependent. For example while Strasbourg is part of France today, there were periods in the past when it was a part of Germany.

(2) Uncertainty
Since history deals with issues that are based on missing or contradictory historical documents uncertainty is inherent in this domain. In particular, the time dimension is problematic because many of the temporal specifications are uncertain. A good example is Stalin's birth date. Officially in the Soviet Union it was 21 December 1879 but according to church records his birth was registered as 6 December 1878. Historians remain in disagreement over which date is the correct one. Another example is the inability to date precisely the paintings of the famous Dutch painter Vermeer.

(3) Subjectivity
Most complex historical notions are vaguely defined or open to multiple interpretations, and thus can be interpreted subjectively making them difficult to model conceptually. For example concepts like 'Enlightenment', 'Middle Ages' or 'Industrial Revolution' have no agreed start and end dates. It is also possible that the boundaries of the transition and core periods of complex events are uncertain. As a net effect, it can be uncertain whether a specific termporal instance is part of a complex event or not. Furthermore it can also be a matter of subjective opinion on the part of the historical expert. For example there is no precise consensus on the period of the Viking raids in Europe. The first raids on particular localities can be dated more or less precisely (the first in the British Isles occurred in 796), but the intense periods of raiding differed from locality to locality and were punctuated by periods of peace and changes in the character of raiding. There is also little historical consensus on the end of the period since raiding tended to die away slowly and varied widely across Europe.

(4) Why questions
Most knowledge representation formalisms are good at representing precise facts, such as axioms and rules. In this regard ontologies are no exception. This is easily done for formal taxonomies, such as the Linnaean system used in botany, but historical knowledge is difficult to model in that way due to the questions asked by historians. In history 'where, when, who, what' type questions are not the most interesting ones. Instead historians and students studying history are drawn to 'why' type questions. A typical historical question would be: 'How and why did the cultural image of the Jews change in medieval

Europe?' or 'Why did Chamberlain's appeasement policy fail?' In other words, historians are not interested in simple facts (information) rather they want to see facts in context (knowledge).

In terms of solutions, the existence time of instances were represented in a straightforward manner by connecting them to instances of a TIME INTERVAL concept. Time-dependent relations can be represented using the standard technique of relation reification, i.e. by representing relations between instances as new ontology instances themselves. However a somewhat more advanced solution was required for 'time-dependent instantiates' relationships. For example it would be straightforward to have subconcepts of persons like KING or PRESIDENT and say that Henry VIII was a king and Bill Clinton a president. It is not clear, however, how to represent the time dependency of those relations, as it is obvious that a person is not a king or a president for their whole lives. This problem was solved by introducing the role concept, and modelling the particular roles of individuals as instances of that concept. The roles are in fact used to model changes in the meaning and functions of instances over time. For example, Winston Churchill had different roles, such as Prime Minister, member of Parliament, journalist and author.

To address the issue of uncertain and subjective temporal information a novel fuzzy temporal model was explored (Nagypál and Motik, 2003). Although this model is not yet integrated into the VICODI prototype, further development and implementation is on its way.

The 'why questions' problem is not solved in the ontology itself, but by the whole VICODI system. Users searching an answer to a 'why' question must reconstruct the context of the complex question as precisely as possible (see Section 3) and from that they can get documents, which hopefully also suggest answers or more precisely interpretations to those 'why' questions.

5.2 Building the ontology

While many different ontology-building methodologies are mentioned in the literature none of them can be viewed as the standard methodology so far. Perhaps the two most mature methodologies, which have already been used in several projects, are METHONTOLOGY (Gómez-Pérez, 1997) and the On-To-Knowledge methodology (Sure *et al.*, 2003). Most of the methodologies, including the two aforementioned ones, describe a *top-down* or *middle-out*, highly iterative process for ontology building (Fernández-López and Gómez-Pérez, 2002), a strategy that was also followed for the VICODI ontology. This was not a hard decision, as the alternative *bottom-up* development strategy was not possible for us because at the beginning of the project we did not have a suitable corpus of historical documents which could have been used for semi-automatic ontology generation

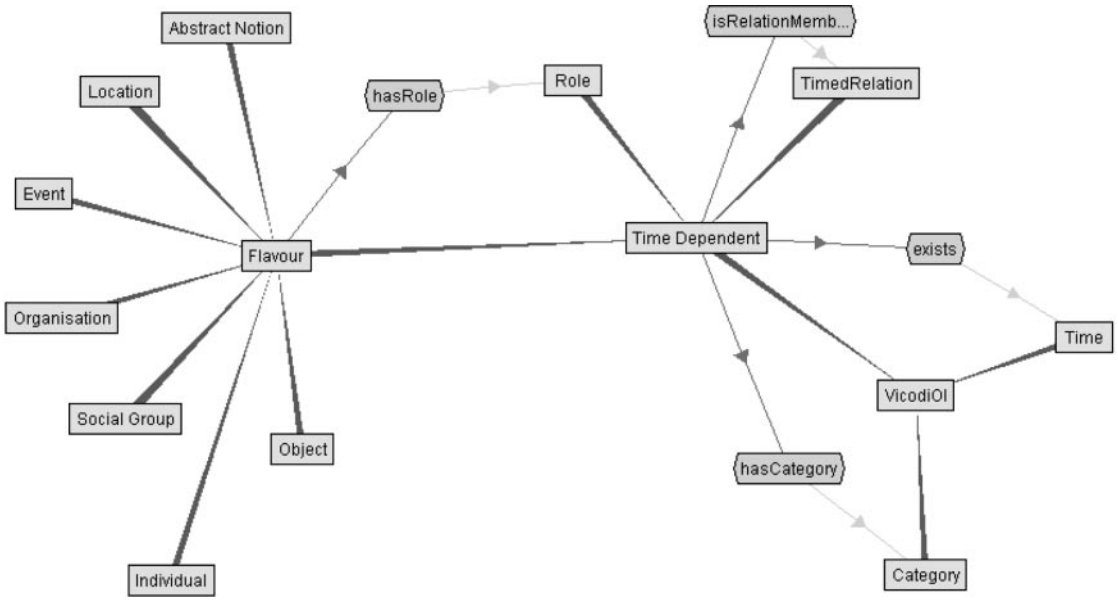

Fig. 5 High-level concept structure of the VICODI ontology. A colour version of this figure can be found in the Supplementary Data section on the journal's website at www.llc.oupjournals.org.

using tools like Text-to-Onto (TextToOnto Summary) or systems built on the GATE (General Architecture for Text Engineering) framework.

Following the middle-out strategy we introduced a shallow concept hierarchy starting from only seven basic concepts (called flavours), which are meaningful for domain/history experts: individual, event, abstract notion, organisation, object, social group and location (see Fig. 5). The hierarchy below these concepts is shallow (2–3 levels) and stops at an abstraction level, which while still meaningful for historians is still general enough to make the placement of new instances in the ontology easy thus speeding up the population of the ontology with new historical knowledge. The complexity of history is further represented by connecting instances of these flavours with a limited number of property relations (a maximum of fifteen) and a number of roles, such as King, Pope, Prime Minister etc.

This resulted in a structure that is intuitive, simple and allows for the uploading of instances and relations (representing historical facts) into the ontology in huge numbers. This approach turned out to be adequate for the purposes of VICODI and is probably adequate for most applications employing machine algorithms to handle history-specific information.

The ontology was developed using a customized version of the OIModeler component of the open-source KAON framework (Motik *et al.*, 2002), which follows the usual visual ontology editing paradigm of Protégé (Protégé Ontology Editor) or OilEd (OilEd Ontology Editor). The KAON framework is an extension of the W3C

RDFS standard (Brickley and Guha, 2004), which makes it possible to link with other ontologies using W3C standards (W3C Resource Description Framework).

It turned out that the visual ontology editing approach illustrated above was suitable only for editing the concept level of the ontology. When the historical experts started uploading large numbers of instances, relations and roles, they preferred a database or spreadsheet-style interface instead of the visual GUI. Features of existing spreadsheet tools, like grouping and sorting of thousands of records based on different criteria, or the copying and pasting of thousands of elements between sheets are aspects that proved invaluable in the preparation of instance upload operations. Thanks to KAON's capability to provide programmatic access to the ontology it was possible to add a huge number of instances and concepts to the ontology by processing databases and particularly Excel spreadsheets. This—together with the simple and intuitive concept hierarchy—significantly sped up the ontology populating process, as the historians could use their preferred software tools, such as Excel, when codifying their knowledge. The manual functions of the ontology editor were therefore only needed to carry out advanced operations, like relocating existing concepts and instances, adding new connections and visualizing the existing ontology structure.

Interestingly, Excel sheets can be viewed as an intermediary conceptual level representation above the implementation level of the ontology. The success of this approach in our project, where user acceptance was enormous and the usage of Excel sheets sped up the upload process significantly, validates the statements of approaches like METHONTOLOGY (Lopez *et al.*, 1999) or GALEN (Rector *et al.*, 2001), which advocate the usage of such intermediate models. Nevertheless, the process of developing and populating the ontology was an extremely time-consuming process taking over 24 man-months of efforts.

6 Contextualization

The concept of context and its application in the form of contextualization are key aspects of the VICODI project. Contextualization in VICODI includes three major tasks: generation of context, executing context-based search and visualizing documents based on their context. This section will begin with an initial discussion and definition of the concept of contextualization, and then the first two of the above tasks will be considered. The third task of visualization will be examined in the next section.

6.1 Context and Contextualization in VICODI

By definition, data taken out of context are just pieces of *information*, which are useful in games, such as trivial pursuit, but not to understand the complexity of the world surrounding us. Information structured

in context and connected with other *words* and *ideas* becomes *knowledge*. This knowledge provides an understanding of processes and events. This progression from information to knowledge was at the heart of the VICODI project. Indeed one of the wider objectives of the project was the development of a methodology and tools for the comprehensive structuring of context in an attempt to facilitate the creation of new 'knowledge'.

Recently, the notions of context and contextualization have become very fashionable in web-based information systems. While humans intuitively understand what 'context' means, it is very hard to give a clear, operational definition of it. Typically, the term 'context' is used to describe computational models that express the whereabouts of users (such as mobile and ubiquitous computing), the users' goals and interests (including adaptive hypermedia and personalized information systems) and semantic indices for documents and knowledge spaces (semantic web technologies). Thus in general we can roughly summarize the status quo of the term context as 'user models, user profiles and semantic indices'.

Work in computational semantics in the early 1990s has led to systems that allow for the specification of featured contexts (Tin and Akman, 1995). For example Wurman's LATCH (Location, Alphabetic order, Time, Category and Hierarchy) properties were used in identifying structural mechanisms for storage, retrieval, presentation and navigation (Wurman *et al.*, 2000). The VICODI project used a modified form of LATCH for organizing, retrieving and graphically presenting knowledge.

For context representation we chose a pragmatic approach, which follows the semantic indices view of context and based our context definition on the context model described by Jurišica (Jurišica, 1994). Context in VICODI is a weighted set of elements from a suitable ontology.[2] For example a possible (partial) context of a document describing the causes and consequences of the Russian Revolution could be the following:

{Lenin:1.0, 1919-1924:1.0, Russia:0.8, Russian Revolution:1.0}

where 'Lenin', 'Russia' and 'Russian Revolution' are historical notions from the ontology, and the numbers represent relevancy weights. Thus the ontology defines the vocabulary for the context descriptions. More formally, our VICODI document context (VDC) consists of two sets: defining the *conceptual part* (*CSet*) and the *temporal part* (*TSet*). The *CSet* is a set of weighted ontology elements (WOI—weighted ontology instance), while the *TSet* is a set of weighted time intervals (WTI). Floating weights between 0.0 and 1.0 are used and time is represented as an interval in years. For visualization purposes (see Section 6) the part of the context that specifies the L, T and C parts of the LATCH approach mentioned above has proven particularly relevant in this project and is referred to as LATCH context.

2 For technical reasons and because of the importance of the time dimension we represent time in our context not as ontology instances, but as time intervals, which can be manipulated freely by various context operations.

In this article, the term *contextualization* refers to the process of tailoring content or content visualization to specific (potentially dynamic) requirements or demands by the user. These demands are expressed as resource context in the VICODI system, and as part of the contextualization process this context is identified automatically. When the resource is displayed, temporal and location contexts are visualized via a historical map. The content context is presented (as was explained in Section 3) through a list of other historically related documents.

A typical example of how the contextualization process successfully creates knowledge in response to user and historical demands and requirements are the name changes of the city of St Petersburg. The VICODI system is able to identify and distinguish the various names of that city throughout its history from St Petersburg to Petrograd to Leningrad and then back to St Petersburg. The location function tailors the content results depending on which time period or name the user provides, yet at the same time it recognizes that this is one and the same city throughout history.

6.2 Dynamic Contextualization

In VICODI the goal of *dynamic contextualization* was pursued. The contextualized content is generated dynamically based on the actual context of a document. That means it changes when new insights about the meaning of the document change its context, such as when the context generation algorithm is improved or an expert manually changes the context. Further, the set of relevant documents to the actual one is calculated dynamically (based on the context of other documents via the context-based search feature). That means if new relevant documents are added to the repository they will be automatically found when the user clicks on the contextualized links of a document.

It must be realized that the generation of context description is far from trivial. To begin with text-to-OI mapping (OI—ontology instance is an abbreviation used in the KAON system to denote elements of an ontology) is ambiguous most of the time. For example 'the King' can refer to Richard the Lionheart in a document about the Crusades and to Elvis Presley in a document about Rock'n'Roll. The OI labels have to be disambiguated. Second, even OIs that are not explicitly mentioned in the text can belong to the context. For example a document about World War II will normally not contain that term but will report on various subevents, persons and locations related to that complex event.

Context-based search is also a difficult task because simple syntactical measures of context similarity usually do not help much. As an example consider the following two contexts:

{Lenin: 1.0, 1917-1924: 1.0, Russia: 0.8}

{Trotsky: 0.7; Bolshevism: 0.9, Red Army: 0.8, Russian Revolution: 1.0}

While they clearly describe the same context of the Russian Revolution, and therefore should be considered similar, syntactically they do not have any common elements.

As was already described in Section 4, both context generation and context similarity calculations are the tasks of the CE component. Generally speaking the internal functionality of the CE is based on a novel cross-correlation approach similar to latent semantic indexing (LSI) (Letsche and Berry, 1997), but utilizing the OIs as basic elements instead of document terms. It uses the training data (that means expert-specified OI relevances to documents) and the bag-of-words approach (Mladenic, 1999) to provide context estimates and context similarity between pairs of contexts.

The estimate of a document's context is based on a weighted (fuzzy) 1–*N* classification. This classification exploits a bag-of-words document representation and the cross-correlation features of the classifier allow the CE to identify in addition abstract notions (elements of the ontology), which are only implicitly mentioned in the documents.

In VICODI the historical experts always have the possibility to review and, if needed, correct the context estimations, which are generated automatically by the CE. Indeed the experts usually had to make such corrections. These expert-reviewed document contexts are then used to further train the CE to generate even better estimations. Given time and training data, the fully automatic context estimation generation in VICODI did start to yield adequate results for casual (non-expert) users.

The similar results provided by the CE can be used by intelligent algorithms to provide the context search functionality. In VICODI intelligent ontology-based heuristics were used to pre-filter the set of context candidates stored in the MSKS repository, which are then ranked with the help of the CE pair-wise context similarity function (Surányi *et al.*, 2004).

There are already several systems available that utilize similar strategies to that of the VICODI system. Sites following the Wiki idea (Wikimedia Foundation) also generate hyperlinks in new documents automatically, allowing easier navigation in the document space. However, these links are generated on a purely syntactic base and only a one-to-one connection is possible. Other systems, like HyperNietzsche (HyperNietzsche Project) or the SRFG LATCH Browser (LATCH Browser) also have a notion of context, which must, however, be manually specified in those systems. Further, information extraction applications—most of them are based on the open-source GATE framework—do try to achieve what we do in the context generation step. They mostly work only on the syntactical level, although the first ontology-aware algorithms are being integrated to the GATE framework, which shows that the importance of ontologies is already recognised in this area. Such novel information extraction algorithms can be integrated into our system in the future to improve

the quality of the automatically generated context estimation. Finally, in the area of information retrieval the importance of ontologies to achieve better query results has been recognized. Indeed, the first research prototypes building on this idea have already appeared, for example the Textpresso system (Textpresso). This insight was also the basis for the context-based search facility in the VICODI system.

7 SVG-based Geographical Visualization

Visual information retrieval systems play an increasingly significant role in how we access, analyse, and understand information. However, most such systems do not provide a combination of visual representation with contextualization of the retrieved information.

For a better visual enhancement, but also as a means of easier access to resources and to improve the transfer of knowledge, the VICODI historical portal provides historical maps in SVG format (see Fig. 2). These maps represent the geographical information found in contextualized historical documents thus adding a spatial contextual layer to the resource. For the VICODI portal a total of over a hundred historical SVG maps at ten-year intervals from 1000 CE to the present day were created.

The choice of SVG maps was a practical one since existing mainstream online mapping methods were not suitable for use in a server-intensive environment like that of the VICODI system. Online mapping techniques can be classified into two categories: raster-imaged based and vector-based. In a typical raster-imaged based approach, servers generate maps as pictures in one of the standard raster graphic formats supported by graphical Web browsers. Interaction is accomplished by submitting a request to the server for a new map image. Even simple user actions, such as turning on or off display attributes often require such a 'round-trip' and complete screen refreshes. This architecture typically requires considerable server capacity and places severe restrictions on map interactivity and interface design flexibility. It also imposes additional limitations, such as slow map updates, increased network and server loads, and often, time poor scalability in terms of the number of simultaneous users (Andrienko and Andrienko, 1999).

The use of vector-based SVG maps to visualise the context of Location, Time and Hierarchy solved these potential problems in VICODI (see Fig. 1). The aim of the VICODI SVG-based visualization research was to enable Internet users to visualize the retrieved information and to provide them with innovative tools to perform dynamic queries. Simplified input and rapid visual display of results should enable more users to benefit from World Wide Web information—regardless of the actual application domain.

Most Internet graphics today are in bitmap formats, such as GIF, PNG or JPEG. Bitmap visualization must contain information on every pixel needed to display an image, which makes them large in size

and slow to download. Up to now, all attempts to deliver high-quality vector-based visualization of dynamic data to the Internet were more or less unsuccessful, usually due to technical restrictions, such as slow connections and less powerful hardware (Romero, 2002). The introduction of XML-based SVG opens ways for GUI designers to concentrate on content delivery and interactions. The two main benefits that SVG images offer over conventional bitmaps are small file sizes and independent scalability, which is compelling within the Internet environment where download times are crucial and viewing platforms differ. It is expected that most browsers will support SVG in the near future since the W3C organization has recommended it as an Internet standard (Scalable Vector Graphics (SVG) 1.1 Specification). When browser versions with embedded SVG support will be released, the market for SVG visualization will increase rapidly, as SVG opens new and simpler ways for designers to create attractive and interactive web interfaces.

The SVG-based visualization technique implemented in VICODI strives to work towards universal usability. The aim was to support both occasional users with modem speed connections and low-end machines, as well as users with high-speed connections and fast machines. While long downloading time is a typical problem for most interactive geo-referenced visualization solutions that deliver vector-format maps to the browser for rendering, SVG-based interfaces make the visualisation task simpler and downloading time shorter. One of the drawbacks of the VICODI SVG maps was the fact that creating them proved extremely time-consuming.

For the above-mentioned context geo-referenced visualisation, a modified concept of the much-used choropleth maps was implemented for displaying information on the historical maps. A choropleth map shows regions or areas that contain the same characteristics. Each area in the map, such as a country, is shaded differently representing classes of quantitative data, such as birth and death rates. For example, the higher the death rate in a specific geographical area, the darker its colour is (Mersey, 1990). In the VICODI historical maps, the choropleth system was used in a qualitative way, rather than a quantitative, in order to rank the importance of a geographical entity in the context of a historical document. Darker colours indicate that a country is more important than lighter coloured geographical areas. In this way the VICODI SVG-based historical maps enable Web users to visualise the retrieved geographical information and to provide them with a tool to perform additional queries. Indeed users commented that the SVG historical maps were probably the most interesting and helpful part of the interface. However they did want more complex features, such as zoomable layered maps of individual countries or regions instead of just a single map of the whole of Europe for each historical period. Further and more critically users indicated that the shade-coded maps

were ambiguous and that colour-coded weightings would have been more accessible.

8 Conclusions

The present VICODI prototype combined several innovative technologies to create an experimental history knowledge space for the semantic web. The outcome of the project was certainly successful although in many respects the VICODI project has only initiated potential solutions. In terms of proof of principle the goal of creating a functioning semantic humanities portal was most definitely achieved. However, in terms of creating a usable tool for historical research and learning, the project was less successful even though this was only ever thought of as a useful by-product of the technological research.

The project also raised a number of new research questions, not least those related to the construction of a history knowledge system: the ontology. The VICODI ontology development process has shown that a complex humanities domain can be represented through a shallow ontology structure and a limited number of concepts and properties. One of the questions that still remains is whether constructing an ontology for any humanities domain is too labour intensive and too costly. VICODI has shown that by applying mass upload techniques, for example through Excel spreadsheets, it is possible to construct an ontology of over 15,000 instances in a relatively short time span. Furthermore, the ontology is extensible and could be expanded or integrated with other ontologies in the future. In addition, the context model used for VICODI is a very generic and extensible mechanism. The direct integration of context sensitive queries into contextualized document content eliminated the need to specify a query over an ontology browser—a solution that typically suffers reduced usability due to the large graph structure of many ontologies, which are too complicated for most users.

The project also provided interesting technological results in the areas of contextualisation and context visualisation. VICODI is one of the very few existing systems that exploit a fully fledged ontology for the purposes of (context) information extraction and intelligent information retrieval. The novel SVG-based context visualisation approach using a modified version of choropleth maps turned out to be superior to traditional raster-imaged based approaches in many aspects.

In addition to the technological advances described in this article, perhaps the most significant achievement of the project was the combination of the various elements—ontology, contextualisation and visualisation—which illustrated how a semantic system for a humanities subject, such as history, could be built and operated.

Acknowledgements

We would like to thank Bob Mulrenin and Tobias Berka from Salzburg Research who provided invaluable help in describing the CE component. In addition thanks must also be offered to Edvins Snore and Juris Zubkans of RIDEMO, as well as Gareth Prosser, for their valuable information and comments. The VICODI project was funded by the EU under the contract EU-IST-2001-37534.

References

Andrienko, G. and Andrienko, N. (1999). Interactive maps for visual data exploration. *International Journal of Geographical Information Science*, **13**: 356–74.

Brickley, D. and Guha, R. (2004). Rdf Vocabulary Description Language 1.0. *Rdf Schema. Recommendation*, World Wide Web Consortium.

Cultural Units of Learning—Tools and Services. http://www.cultos.org/ (accessed 12 April 2005).

Cycorp. http://www.cyc.com/cyc (accessed 12 April 2005).

Deswarte, R., Mulrenin, B., Nagypal, G., Oosthoek, J., and Snore, E. (2004). Visual contextualization of digital content. In *Proceedings of IADIS e-Society 2004 (ES2004)*, IADIS: Spain.

Dey, A. K. (2001). Understanding and using context. *Personal and Ubiquitous Computing Journal,* **5**: 4–7.

Espresso Project. http://www.jcorporate.com (accessed 12 April 2005).

Eurohistory. http://www.eurohistory.net (accessed 31 January 2005).

Fernández-López, M. and Gómez-Pérez, A. (2002). OntoWeb Deliverable 1.4: A Survey on Methodologies for Developing, Maintaining, Evaluating and Reengineering Ontologies. *Technical report*, EU IST Project IST-2000-29243, Ontoweb Consortium. http://ontoweb.org/About/Deliverables/D1.4-v1.0.pdf (accessed 12 April 2005).

Fernández-López, M., Gómez-Pérez, A., Sierra, J. P., and Sierra, A. P. (1999). Building a chemical ontology using methontology and the ontology design environment. *IEEE Intelligent Systems*, **14**: 37–46.

General Architecture for Text Engineering. http://gate.ac.uk/ (accessed 12 April 2005).

Getty Vocabulary Databases. http://www.getty.edu/research/conducting_research/vocabularies/ (accessed 12 April 2005).

GFDL. http://www.gnu.org/copyleft/fdl.html (accessed 12 April 2005).

Gómez-Pérez, A. (1997). Knowledge sharing and reuse. In *Handbook of Applied Expert Systems*. Philadelphia: CRC Press.

Gómez-Pérez, A., Fernández-López, M., and Corcho, O. (2004). *Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*, New York: Springer Verlag.

GTGN Online. http://www.getty.edu/research/conducting_research/vocabularies/tgn/ (accessed 12 April 2005).

Guarino, N. (1998). Formal ontology and information systems. In *Proceedings of Formal Ontology in Information Systems (FOIS'98)*, Trento, Italy, pp. 3–15. Italy: IOS Press.

Humanities and Social Science Electronic Thesaurus. http://www.data-archive.ac.uk/search/hassetSearch.asp (accessed 12 April 2005).

HyperNietzsche Project. http://www.hypernietzsche.org/doc/presentation/en/ (accessed 12 April 2005).

Jurišica, I. (1994). How to retrieve relevant information? In Greiner, R. (ed.), *Proceedings of the AAAI Fall Symposium Series on Relevance*. New Orleans, Louisiana, AAAI Press, pp. 101–4.

KAON Semantic Web. http://kaon.semanticweb.org (accessed 12 April 2005).

Lagoze, C. and Hunter, J. (2001). The ABC Ontology and Model. *Journal of Digital Information*, **2**(2), http://jodi.tamu.edu/Articles/v02/i02/Lagoze/ (accessed 12 April 2005).

LATCH Browser. http://suntrec.salzburgresearch.at/projects/LATCHBrowser/ (accessed 12 April 2005).

Letsche, T. A. and Berry, M. W. (1997). Large-scale information retrieval with latent semantic indexing. *Information Sciences*, **100**: 105–37.

Mersey, J. E. (1990). Colour and thematic map design: the role of colour scheme and map complexity in choropleth map communication. *Cartographica*, **27**, Monograph 41.

Mladenic, D. (1999). Text-learning and related intelligent agents. *IEEE Intelligent Systems*, **14**(4): 44–54.

Motik, B., Maedche, A., and Volz, R. (2002). A conceptual modeling approach for semantics-driven enterprise applications. In *Proceedings of On the Move to Meaningful Internet Systems Confederated International Conferences DOA, CoopIS and ODBASE 2002*, October 29–31, University of California, Irvine, London, UK: Springer, pp. 1082–1099.

Nagypál, G. (2004). Creating an application-level ontology for the complex domain of history: mission impossible? In *Proceedings of Lernen—Wissensentdeckung—Adaptivität (LWA 2004)*, FGWM 2004 Workshop, Berlin, Germany, pp. 287–94. http://lwa.informatik.hu-berlin.de/proceedings/LWA04_FGWM.pdf (accessed 12 April 2005).

Nagypál, G., and Motik, B. (2003). A fuzzy model for representing uncertain, subjective, and vague temporal knowledge in ontologies. In Meersman, R., Tari, Z., and Schmidt, D. C. (eds), *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*. Heidelberg: Springer-Verlag, pp. 906–23.

OilEd Ontology Editor. http://oiled.man.ac.uk/ (accessed 12 April 2005).

ONTOPIA. http://www.ontopia.net/index.html (accessed 12 April 2005).

Protégé Ontology Editor. http://protege.stanford.edu/ (accessed 12 April 2005).

Rector, A. L., Wroe, C., Rogers, J., and Roberts, A. (2001). Untangling taxonomies and relationships: personal and practical problems in loosely coupled development of large ontologies. In *Proceedings of the International Conference on Knowledge Capture* (K-CAP 2001), October 21–23, Victoria, BC, Canada. New York: ACM Press, pp. 139–46.

Romero, S. (2002). Price is limiting demand for broadband, *The New York Times*, 5 December, p. 5.

Rumizen, M. C. (2002). *The Complete Idiot's Guide to Knowledge Management*. Madison, WI: CWL Publishing.

Scalable Vector Graphics (SVG) 1.1 Specification. http://www.w3.org/TR/SVG/ (accessed 12 April 2005).

Senellart, J., Boitet, C., and Romary, L. (2003). SYSTRAN new generation: The XML translation workflow. In *Proceedings of the Ninth Machine Translation Summit*, September 23–27, New Orleans. Stroudsburg, PA: AMTA. http://www.amtaweb.org/summit/MTSummit/FinalPapers/102-Sennelart-final.pdf (accessed 12 April 2005)

Spool, J. M., Scanlon, S., Tara, S., Will, C., and DeAngelo, T. (1999). *Web Site Usability: A Designer's Guide*. San Francisco: Morgan Kaufmann Publishers.

SUMO Ontology. http://ontology.teknowledge.com/ (accessed 12 April 2005).

Surányi, G. M., Nagypál, G., and Schmidt, (2004). A. Intelligent retrieval of resources by exploiting their semantic context. In *Proceedings of On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE, OTM Confederated International Conferences, Part I*, October 25–29, Agia Napa, Cyprus, Springer, pp. 705–723.

Sure, Y., Akkermans, H., Broekstra, J., *et al.* (2003). On-to-knowledge: Semantic Web Enabled Knowledge Management. In Zhong, N., Liu, J., and Yao, Y. (eds), *Web Intelligence*. Heidelberg: Springer-Verlag, pp. 277–300.

Systran—White Papers. http://www.systransoft.com/Technology/WhitePapers.html (accessed 12 April 2005).

Textpresso. http://www.textpresso.org/ (accessed 12 April 2005).

TextToOnto Summary. http://sourceforge.net/projects/texttoonto/ (accessed 12 April 2005).

The CIDOC Conceptual Reference Model. http://cidoc.ics.forth.gr/ (accessed 12 April 2005).

Tin, E. and Akman, V. (1995). Situations and computation: An overview of recent research. In Griffith, J., Hinrichs, E. W., and Nakazawa, T. (eds), *Proceedings of Topics in Constraint Grammar Formalism for Computational Linguistics: Papers Presented at the Workshop on Grammar Formalisms for Natural Language Processing held at ESSLLI-94*, 25 September–2 October, Copenhagen, pp. 77–106.

UNESCO Thesaurus. http://www.ulcc.ac.uk/unesco/ (accessed 12 April 2005).

W3C Resource Description Framework. http://www.w3.org/RDF/ (accessed 12 April 2005).

Wikimedia Foundation. http://www.wikimedia.org/ (accessed 12 April 2005).

Wurman, R. S., Sume, D., and Leifer, L. (2000). *Information Anxiety 2*. Indianapolis: Que.