

UPGRADE is the European Journal for the Informatics Professional, published bimonthly at <http://www.upgrade-cepis.org/>

UPGRADE is the anchor point for UPENET (UPGRADE European Network), the network of CEPIs member societies' publications, that currently includes the following ones:

- **Mondo Digitale**, digital journal from the Italian CEPIs society AICA
- **Novática**, journal from the Spanish CEPIs society ATI
- **OCG Journal**, journal from the Austrian CEPIs society OCG
- **Pliroforiki**, journal from the Cyprus CEPIs society CCS
- **Pro Dialog**, journal from the Polish CEPIs society PTI-PIPS

Publisher

UPGRADE is published on behalf of CEPIs (Council of European Professional Informatics Societies, <http://www.cepis.org/>) by **Novática** (<http://www.ati.es/novatica/>), journal of the Spanish CEPIs society ATI (*Asociación de Técnicos de Informática*, <http://www.ati.es/>)

UPGRADE monographs are also published in Spanish (full version printed; summary, abstracts and some articles online) by **Novática**, and in Italian (summary, abstracts and some articles online) by the Italian CEPIs society ALSI (*Associazione nazionale Laureati in Scienze dell'Informazione e Informatica*, <http://www.alsi.it/>) and the Italian IT portal *Tecnoteca* (<http://www.tecnoteca.it/>)

UPGRADE was created in October 2000 by CEPIs and was first published by **Novática** and **INFORMATIK/INFORMATIQUE**, bimonthly journal of SVI/FSI (Swiss Federation of Professional Informatics Societies, <http://www.svifsi.ch/>)

Editorial Team

Chief Editor: Rafael Fernández Calvo, Spain, rfcvalvo@ati.es
 Associate Editors:
 François Louis Nicolet, Switzerland, nicolet@acm.org
 Roberto Carniel, Italy, carniel@dgt.uniud.it
 Zakaria Maamar, Arab Emirates, Zakaria.Maamar@zu.ac.ae
 Soraya Kouadri Mostéfaoui, Switzerland, soraya.kouadrimostefaoui@unifr.ch

Editorial Board

Prof. Wolfried Stucky, Former President of CEPIs
 Prof. Nello Scarabottolo, CEPIs Vice President
 Fernando Piera Gómez and
 Rafael Fernández Calvo, ATI (Spain)
 François Louis Nicolet, SI (Switzerland)
 Roberto Carniel, ALSI – Tecnoteca (Italy)

UPENET Advisory Board

Franco Filippazzi (Mondo Digitale, Italy)
 Rafael Fernández Calvo (Novática, Spain)
 Veith Risak (OCG Journal, Austria)
 Panicos Masouras (Pliroforiki, Cyprus)
 Andrzej Marciniak (Pro Dialog, Poland)

English Editors: Mike Andersson, Richard Butchart, David Cash, Arthur Cook, Tracey Darch, Laura Davies, Nick Dunn, Rodney Fennemore, Hilary Green, Roger Harris, Michael Hird, Jim Holder, Alasdair MacLeod, Pat Moody, Adam David Moss, Phil Parkin, Brian Robson

Cover page designed by Antonio Crespo Foix, © ATI 2005

Layout Design: François Louis Nicolet

Composition: Jorge Llácer-Gil de Ramales

Editorial correspondence: Rafael Fernández Calvo rfcvalvo@ati.es

Advertising correspondence: novatica@ati.es

UPGRADE Newsletter available at

<http://www.upgrade-cepis.org/pages/editinfo.html#newsletter>

Copyright

© Novática 2005 (for the monograph and the cover page)

© CEPIs 2005 (for the sections MOSAIC and UPENET)

All rights reserved. Abstracting is permitted with credit to the source. For copying, reprint, or republication permission, contact the Editorial Team

The opinions expressed by the authors are their exclusive responsibility

ISSN 1684-5285

Monograph of next issue (February 2006):

Key Success Factors in Software Engineering
 (The full schedule of UPGRADE is available at our website)

Monograph: The Semantic Web (published jointly with Novática*)

Guest Editors: *Luis Sánchez-Fernández, Michael Sintek, and Stefan Decker*

- 2 Presentation. The Semantic Web or The Next Web Wave – *Luis Sánchez-Fernández, Michael Sintek, and Stefan Decker*
- 5 The Semantic Web: Fundamentals and A Brief State-of-the-Art – *Luis Sánchez-Fernández and Norberto Fernández-García*
- 12 Leveraging Metadata Creation by Annotation for The Semantic Web – *Siegfried Handschuh*
- 19 The Quest for Information Retrieval on The Semantic Web – *David Vallet-Weadon, Miriam Fernández-Sánchez, and Pablo Castells-Azpilicuet*
- 24 Functional RuleML: From Horn Logic with Equality to Lambda Calculus – *Harold Boley*
- 30 Towards Semantic Desktop Wikis – *Malte Kiesel and Leo Sauerermann*
- 35 Towards Semantically-Interlinked Online Communities – *Uldis Bojars, John G. Breslin, Andreas Harth, and Stefan Decker*
- 41 A Semantic Search Engine for the International Relation Sector – *Luis Rodrigo-Aguado, V. Richard Benjamins, Jesús Contreras-Cino, Diego-Javier Patón-Villahermosa, David Navarro-Arno, Robert Salla-Figuerol, Mercedes Blázquez-Cívico, Pilar Tena-García, and Isabel Martos-Laborde*
- 48 Semantic Search in Digital Image Archives: A Case Study – *Julio Villena-Román, José-Carlos González-Cristóbal, Cristina Moreno-García, and José- Luis Martínez-Fernández.*
- 55 Configuring e-Government Services Using Ontologies – *Dimitris Apostolou, Ljiljana Stojanovic, Tomás Pariente-Lobo, Joan Battle-Montserrat, and Andreas E. Papadakis*

UPENET (UPGRADE European Network)

- 63 From **Novática** (ATI, Spain)
 ICT for Education
 An Initiative for Educational Modernization: The Ponte dos Brozos Project – *Simón Neira-Dueñas and Felipe Gómez-Pallete Rivas*
- 71 From **Pro Dialog** (PIPS, Poland)
 ICT for Education
 On The Superiority of Internet-Based Mass Enrolment to High Schools over Traditional – *Andrzej P. Urbanski*

* This monograph will be also published in Spanish (full version printed; summary, abstracts, and some articles online) by **Novática**, journal of the Spanish CEPIs society ATI (*Asociación de Técnicos de Informática*) at <http://www.ati.es/novatica/>, and in Italian (online edition only, containing summary, abstracts, and some articles) by the Italian CEPIs society ALSI (*Associazione nazionale Laureati in Scienze dell'Informazione e Informatica*) and the Italian IT portal *Tecnoteca* at <http://www.tecnoteca.it/>.

A Semantic Search Engine for the International Relation Sector

Luis Rodrigo-Aguado, V. Richard Benjamins, Jesús Contreras-Cino, Diego-Javier Patón-Villahermosa, David Navarro-Arnao, Robert Salla-Figuerol, Mercedes Blázquez-Cívico, Pilar Tena-García, and Isabel Martos-Laborde

The Royal Institute Elcano (Real Instituto Elcano, RIE) is a prestigious independent political Spanish institute whose mission is to comment on the geo-political situation in the world focusing on its relation to Spain. As part of its dissemination strategy it operates a public website. In this paper we present and evaluate the application of a semantic search engine to improve access to the Institute's informational content: instead of retrieving documents based on user queries of keywords, the system accepts queries in natural language and returns answers rather than links to documents. Topics that will be discussed include ontology construction, automatic ontology population, semantic access through and a natural language interface.

Keywords: Knowledge Acquisition, Ontology, Question Answering, Semantic Search.

1 Introduction

Worldwide there are several prestigious institutes that comment on the geo-political situation in the world, such as the UK's Royal Institute for International Affairs, <<http://www.riia.org>>, or the Dutch Institute for International Relations, <<http://www.clingendael.nl>>. In Spain, the *Real Instituto Elcano* (Royal Institute Elcano, RIE, <<http://www.realinstitutoelcano.org>>) fulfils this role. The institute provides several types of written reports which discuss the political situation in the world, with a focus on events relevant for Spain. The reports are organized into different categories, such as Economy, Defense, Society, Middle East, etc. In a special periodic report - the "Barometer of the Royal Institute Elcano" - the Institute comments on how the rest of the world views Spain in the political arena. Access to the content is provided by categorical navigation and a traditional full text search engine. While full text search engines are helpful instruments for information retrieval (<<http://www.google.com>> is the champion), in domains where relations are important, those techniques fall short. For instance, a keyword-based search engine will have a hard time finding the answer to a question such as: "*Governments of which countries have a favorable attitude toward the US-led armed intervention in Iraq?*" since the crux of answering this question resides in 'understanding' the relation "*has-favourable-attitude-toward*".

In this paper we present a semantic search engine that accepts natural language questions to access content produced by the Institute. *Semantic* in this context means related to the domain of International Relations (politics).

2 An Ontology of International Affairs

When searching for a particular datum, looking for a concrete answer to a precise question, a standard search engine that retrieves documents based on matching keywords falls short. First of all, it does not satisfy the primary need of the user, which is finding a well-defined data, and provides a collection of documents that the user must ex-

Luis Rodrigo-Aguado graduated in Computer Science from the *Universidad Politécnica de Madrid* (UPM), Spain, and is currently studying towards his doctoral thesis on the subject of the Semantic Web and Natural Language. He divides his time between the Smart Systems Lab of UPM's Dept. of Artificial Intelligence and the company Intelligent Software Components (iSOCO, S.A.), where he is currently working as a project manager, coordinating work related to Natural Language. He has authored a number of articles and presentations in national and international conferences. <lrodrigo@isoco.com>

V. Richard Benjamins is Director of Research & Development and board member at Intelligent Software Components (iSOCO, S.A.), in Madrid, Spain. He co-founded iSOCO in June 1999, and contributed to its start-up (now 70 persons) and international positioning as a Semantic Web Solutions company. He is also part time professor at the *Universidad Politécnica de Madrid*. He has acquired and managed over 5 million euro in R&D projects in advanced Information Technologies related to the Internet. Before working at iSOCO, he had a permanent position at the University of Amsterdam, Nederland, in the area of Knowledge Systems Technology (1998-2000). Between 1993 and 1998, he worked at the University of Sao Paulo, Brazil, the University of Paris-South, France, and the Spanish Artificial Intelligence Research Institute in Barcelona, Spain. He has published over 80 scientific articles in books, journals and proceedings, in areas such as Knowledge Technologies, Artificial Intelligence, Knowledge Management, Semantic Web and Ontologies. He has been guest editor of several journal special-issues, serves on many international program committees, and has been co-chair of numerous international workshops and conferences. He is member of the editorial board of IEEE Intelligent Systems and of Web Semantics (Elsevier). He received his Master's Degree (1988) and Ph.D. (1993) in Cognitive Science from the University of Amsterdam. <rbenjamins@isoco.com>

Jesús Contreras-Cino obtained a PhD in Artificial Intelligence (2004) at the *Universidad Politécnica de Madrid*, Spain. Since 1996 he was an Assistant Researcher in the Intelligence Systems Research Group, <<http://www.isys.dia.fi.upm.es>>, where he participated in projects oriented towards the development of Knowledge Based Systems and Advanced Artificial Intelligence Applications. In 1998 he joined Software A.G.'s e-business competence center, <<http://www.softwareag.com>>, where he

amine, looking for the desired information. Also, not all of the retrieved documents might contain the appropriate answer, and some of the documents that do contain it may not be included in the collection. These drawbacks suggest the need for a change in the search paradigm, evolving from the extraction of whole documents to the information contained in those documents. This approach, however, is not feasible in all conditions. It is not cost justifiable to build such a search engine for general usage, but can be justified for limited, well-defined domains. Such is the semantic search engine developed for the Real Instituto Elcano, which focuses on the topics covered by the reports written by the institute analysts i.e. international politics.

In order to be able to analyse the documents, and reach sufficient 'understanding' of them to be able to answer the users' questions, the system relies on a representation of the main concepts, their properties and the relations among them in the form of an ontology. This ontology provides the system with the necessary knowledge to understand the questions of the users, provide the answers, and associate with them a set of documents that mention the concept of the answer. Based on the ontology, each document gets its relevant concepts annotated and linked to the representing concept or instance in the ontology, allowing a user to browse from a document to the information of a concept he is interested in, and backwards, from the ontology to any of the reports that mention that concept.

2.1 Ontology Design

An ontology is a shared and common understanding of some domain that can be communicated across people and computers [6][7][3][8]. Ontologies can therefore be shared and reused among different applications [5]. An ontology can be defined as a formal, explicit specification of a shared conceptualization [6][3]. 'Conceptualization' refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. 'Explicit' means that the type of concepts used, and the constraints on their use are explicitly defined. 'Formal' refers to the fact that the ontology should be machine-readable. 'Shared' reflects the notion that an ontology captures consensual knowledge, that is, it is not private to some individual, but accepted by a group. An ontology describes the subject matter using the notions of concepts, instances, relations, functions, and axioms. Concepts in the ontology are organized in taxonomies through which inheritance mechanisms can be applied. It is our experience that building a commonly agreed ontology is not easy, especially the social part [2].

Based on interviews with experts of the Elcano Institute, we used the CIA world factbook, <<http://www.cia.gov/cia/publications/factbook/>>, as the basis for the design of the ontology of International Affairs. The CIA fact book is a large online repository with actual information on most countries of the world, along with relevant information in the fields of geography, politics, society, economics, etc.

We have used the competency questions approach [10]

(cont. from previous page)

was enrolled in various European projects as software engineer and main researcher. In November 2000, he joined the Innovation Dept. of Intelligent Software Components (iSOCO, S.A.). During his career he has published various articles about Natural Language Processing in Human-Computer Interaction. <contreras@isoco.com>

Diego-Javier Patón-Villahermosa graduated (HND) in Software Engineering. He is a Knowledge Engineer in Intelligent Software Components (iSOCO, S.A.), specialising in portals and social networks based on semantic search engines, <<http://elcanoisoco.net>>. He has participated in Knowledge Parser projects: framework software that enables data to be extracted automatically from online sources and then locally stored in structured warehousing. <dpaton@isoco.com>

David Navarro-Arno is a developer and researcher in the Innovation Department working on projects such as the *Buscador Semántico Residencia de Estudiantes*, AMASS (Associative Memory Arrays for Semantic Search), *Buscador Semántico Real Instituto Elcano*, Esperanto Services (link between the current Web and the Semantic Web), Knowledge Parser (automatic data extraction), NETCASE (a smart system based on Semantic Web technologies for application in legal environments), and the open source library KPONTOLOGY for working with ontologies, used and maintained by a number of projects (HOPS, Semantic Search Engine, SEKT, Esperanto Services, Onto-H, Iuriservice). <dnavarro@isoco.com>

Robert Salla-Figuerol Graduated as Technical Engineer (Computer Science), in the *Universitat de Lleida* (UDL), Spain, with the thesis "Selfsimilar processes applied to Internet traffic". He has a Master of Computer Science from the Computer Science Faculty of the *Universitat Politècnica de Catalunya* (UPC), Barcelona, Spain. He has a long experience in information retrieval software development for Semantic Web purposes, and has authored several papers for international congresses and journals. <rsalla@isoco.com>

Mercedes Blázquez-Cívico has been working at Intelligent Software Components (iSOCO, S.A.) as a researcher since September 2000. She graduated in Computer Science from the *Universidad Politécnica de Madrid* (UPM), Spain, in 1997 and studied for a masters degree in knowledge engineering and software engineering at the Computer Science faculty of the same university, finishing in November of 2000. She is currently studying towards her doctorate in Computer Science and Artificial Intelligence at the UPM, in knowledge management and its applications. Her research activities include the application of the Semantic Web in knowledge management, and she is currently participating as a work team leader at iSOCO in the under the SEKT 6th framework IP (IST-2003-506826). She has also participated in various PROFIT R&D projects related to the application of Semantic Web technologies and ontologies. <mblazquez@isoco.com>

Pilar Tena-García graduated in Law and Information Science (branch of Journalism) in the *Universidad Complutense de Madrid* (UCM), Spain (1977). She is Deputy Director of Institutional Relations of the Real Instituto Elcano since 2002. <pilar.tena@r-i-elcano.org>

Isabel Martos-Laborde graduated in Information Science in the *Universidad Complutense de Madrid* (UCM), Spain (1992). She is the webmaster of the Real Instituto Elcano site since its inception in 2002. <isabelmartos@r-i-elcano.org>

to determine the scope and granularity of the domain ontology. The ontology consists of several top level classes, some of which are:

- Place: concept representing geographical places such as countries, cities, buildings, etc.
- Agent: concept taken from WordNet [11] representing entities that can execute actions modifying the domain (e.g.: Persons, Organizations, etc.).
- Events: time expressions and events.
- Relations: common class for any kind of relations between concepts.

Without instances information, the ontology contains about 85 concepts and 335 attributes (slots, properties). The ontology has been constructed using Protégé [9]. Figure 1 shows a fragment of the ontology in Protégé.

3 Automatic Annotation

One of the challenges for the success of the Semantic Web is the availability of a critical mass of semantic content [17]. Semantic annotation tools play a crucial role in upgrading the actual web content into semantic content that can be exploited by semantic applications. In this context we developed the Knowledge Parser © (KP), a system able to extract data from online sources populating specific domain ontologies, adding new or modifying existing knowledge facts or instances. The Semantic Web community often calls this process semantic annotation (or just annotation).

The Knowledge Parser offers a software platform that combines different technologies for information extraction, driven by extraction strategies that allow the optimal technology combination application to each source type based on the domain ontology definition.

Ontology population from unstructured sources can be considered as the problem of extracting information from the source, its assignment to the appropriate location in the ontology, and finally, its coherent insertion in the ontology. The first part deals with the information extraction and document interpretation issues. The second part deals with the information annotation, in the sense of adding semantics to the extracted information, according to domain information and pre-existing strategies. The last part is in charge of populating, i.e., inserting and consolidating the extracted knowledge into the domain ontology. The three phases can be seen in the architecture of the system, illustrated in Figure 2.

3.1 Information Extraction

The KP system at present handles HTML (HyperText Markup Language) pages, and there are plans to extend it to handle also PDF (Portable Document Format), RTF (Rich Text Format), and some other popular formats.

To be able to capture as much information as possible from the source document, KP analyzes it using four different processors, each one focusing on a different aspect: the plain text processor, the layout processor, the HTML source processor and the natural language processor.

The plain text source interpretation supports the usage of regular expressions matching techniques. The usage of these kinds of expressions constitutes an easy way of retrieving data in the case of stable, well known pages. If the page suffers frequent changes the regular expression becomes useless.

It is very common that even when documents of the same domain have very similar visual aspect they have a completely different internal code structure. Most of the online

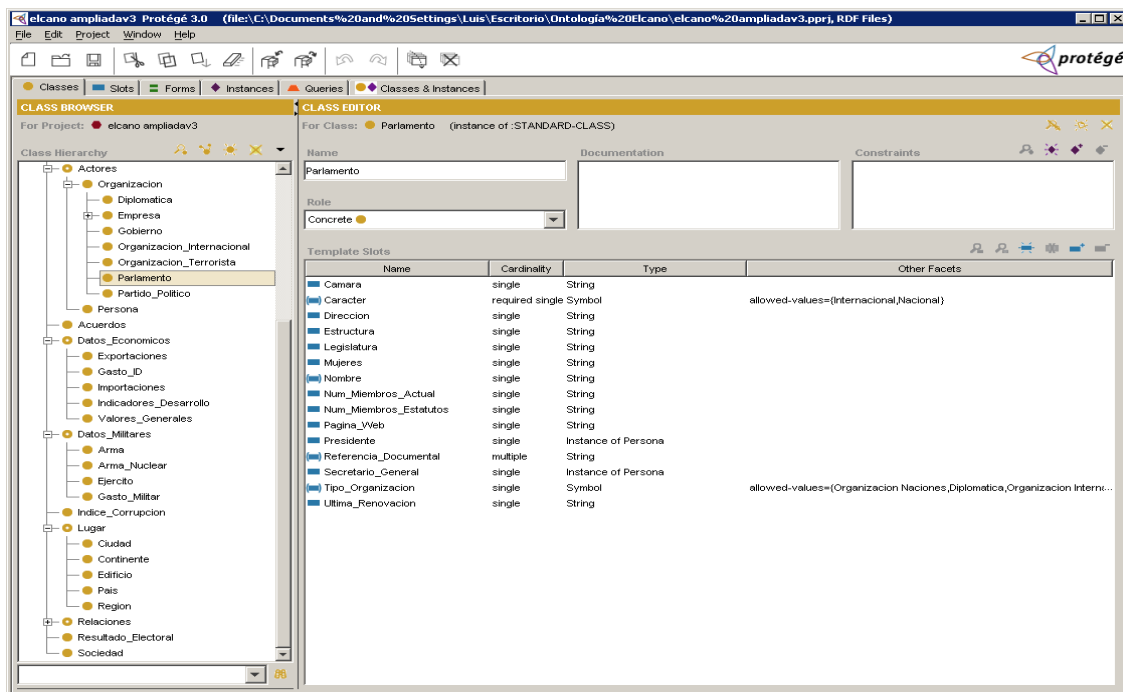


Figure 1: Ontology for International Affairs.

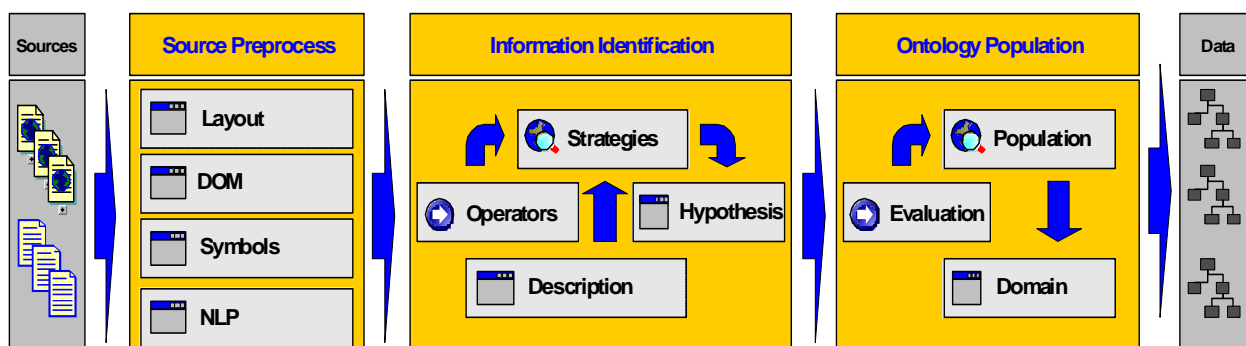


Figure 2: Architecture of The System.

banks offer a position page where all the personal accounts and their balance are shown. These pages have very similar visual aspect, but their source code is completely different. The KP system includes layout interpretation of HTML sources, which allows to determine if certain pieces of information are visually ABOVE, UNDER, RIGHT, LEFT, IN_ROW, IN_COLUMN, IN_ROW RIGHT;... of another piece of information.

In addition to HTML rendering of the source code in a visual model, the KP system needs to process the HTML elements in order to browse through the sources. The source description may include a statement that some information is a valid HTML link (e.g., a country name in a geopolitical portal), and when activated takes one to another document (a country description).

Finally, the fourth model tries to retrieve information from the texts present in the HTML pages. To do that, the user describes the pieces he is interested in in terms of linguistic properties and the relations among them (verbal or nominal phrases, coordinations, conjunctions, appositions, etc.).

3.2 Information Annotation

Once the document is parsed using different and complementary paradigms, the next challenge is to assign the extracted information piece to the correct place in the domain ontology. This task is called annotation, since it is equivalent to wrapping up the information piece with the corresponding tag from the ontology schema.

In most cases the annotation of information is not direct. For instance, a numeric data extracted from the description of a country can be catalogued as the country population, its land area, or its number of unemployed. It is necessary to have some extra information that facilitates reducing this ambiguity. This information, formulated in another model, enlarges the domain ontology with background knowledge, the same way a human uses for its understanding. The extraction system needs to know, for example, that in online banking the account balance usually appears in the same visual row as the account number, or that it is usually preceded by a currency symbol. This kind of information describing the pieces of information expected in the source and the relations among them is formalized in a, so

called, *wrapping ontology*. This ontology supports the annotation process holding information describing the following elements: document types, information pieces and relations among the pieces (any kind of relation detectable by the text, layout, HTML or NLP - Natural Language Processing - models).

According to the domain ontology and the background information added, the system should construct possible assignments from the information extracted to the ontology schema. The result of this process is a set of hypotheses about data included in the source and their correspondence with the concepts, properties and relations in the domain ontology. During the construction process the system can evaluate how much the extracted information fits the information description.

The different ways in which hypotheses can be generated and evaluated are called strategies. Strategies are pluggable modules that, according to the source description, invoke operators. In the current version of the system there are two possible strategies available. For system usages where the response time is critical we use the greedy strategy. This strategy produces only one hypothesis per processed document using heuristics to solve possible ambiguities in data identification. On the other hand, when quality of annotation is a priority and requirements on response time are less important, we use a backtracking strategy. This strategy produces a whole set of hypothesis to be evaluated and populated into the domain ontology.

3.3 International Affairs Ontology Population

Using the Knowledge Parser system, we populated the ontology of international affairs, designed as described in Section 2.1. The domain experts selected four sources where they could find most of the information that they used on their daily basis. These four sources are:

- CIA World Factbook, <<http://www.cia.gov/cia/publications/factbook/>>.
- Nationmaster, <<http://www.nationmaster.com>>.
- Cidob, <<http://www.cidob.org/bios/castellano/indices/indices.htm>>.
- International Policy Institute for Counter-Terrorism, <<http://www.ict.org.il>>.

The set of sources is, of course, not exhaustive, but it

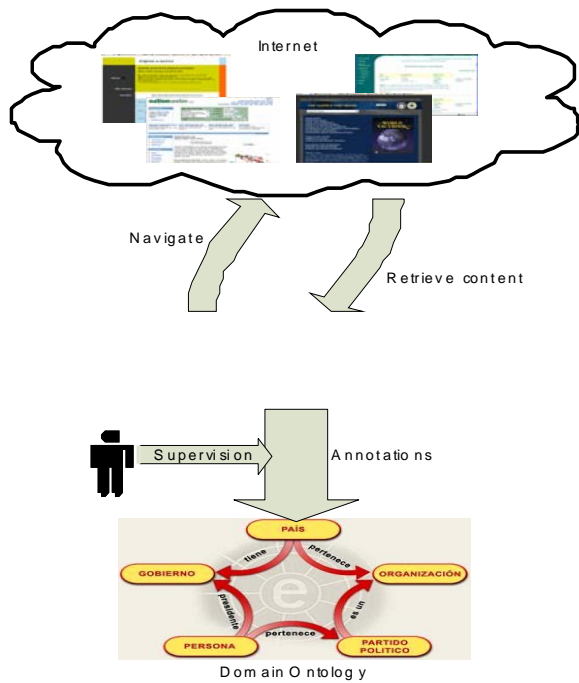


Figure 3: Domain Ontology Population Process.

tries to follow the 80-20 rule, where a few sites cover most of the knowledge needed by the users of the system.

For each of the sites, a wrapping ontology was developed, describing the data contained in it, the way to detect it and the relations among them. The development of these kinds of descriptive ontologies is at present done by experienced knowledge engineers, but future advances will be designed to develop some kind of tools that will allow the domain experts to describe a new source and populate the ontology with its contents themselves.

As a result of this process, we evolved from an empty ontology to an ontology with more than 60,000 facts, occupying more than 20 MB of RDF files.

4 The International Relations Portal

Modeling the domain in the form of an ontology is one of the most difficult and time consuming tasks in developing a semantic application, but an ontology itself is just a way of representing information, it provides no added value for the user. What becomes really interesting for the user is the kind of applications (or features inside an application) that an ontology allows.

In the following, we will describe how we have exploited the semantic domain description, in the form of enhanced browsing of the already existing reports, and a semantic search engine integrated in the international relations portal, interconnected between them.

4.1 Establishing Links between Ontology Instances and Elcano Documents

The portal holds two different representations of the same knowledge, the written reports from the institute analysts and the domain ontology, which are mutually independent. However, one representation can enrich the other, and vice versa. For example, an analyst looking for the Gross

Domestic Product (GDP) of a certain country may also be interested in reading some of the reports where this figure is mentioned, and, in the same way, someone who is reading an analysis about the situation in Latin America may want to find out the political parties present in the countries of the region.

Trying to satisfy these interests, we inserted links between the instances in the ontology and the documents of the Institute. The links are established in both directions. Each concept in the ontology has links to the documents that mention it, and each document has links that connect the concepts mentioned in the article with the corresponding concepts in the ontology. This way, the user can make a question (for example, "Who is the USA president?") and gets the information of the instance in the ontology corresponding to George Bush. From this screen, he can follow the links to any of the instances appearing in the text, George Bush being one of them. This process can be seen in Figure 4, where the information about George Bush in the ontology contains a set of links, and the document seen can be reached following one of them.

To generate these links a batch process is launched that generates, at the same time, both the links in the ontology and the links in the articles.

At present, the process of adding links is a batch process that opens a document, and looks for appearances of the name of any of the instances of the ontology in that text. For any matching, it adds a link in the text to the instance in the ontology and link in the ontology with a pointer to the text. To evaluate the matching, not only the exact name of the instance is used, but also the possible synonyms, contained in an external thesaurus, which can be easily extended by any user, i.e., the domain expert.

Future plans include the automation of this task, so that any new document in the system (the institute constantly produces new reports) is processed automatically by the link generator tool and the new links are transparently included in the system.

4.2 The Semantic Search Engine

With the objective of making available the knowledge contained in the ontology in a comfortable, easy to use fashion, we also designed a semantic search engine. Using this engine, users can ask in natural language (Spanish, in this case) for a concrete data, and the system retrieves the data from the ontology and presents the results to the user.

5 Related Work

Our Knowledge Parser is related to several other initiatives in the area of automatic annotation for the Semantic Web, including KIM [12], which is based on GATE [13], Annotea [14] of W3C., Amilcare [15] of the UK Open University (also based on GATE), and AeroDAML [16]. For an overview of those approaches and others, see [4]. All approaches use NLP as an important factor to extract semantic information. Our approach is innovative in the sense that it combines four different techniques for Information Ex-

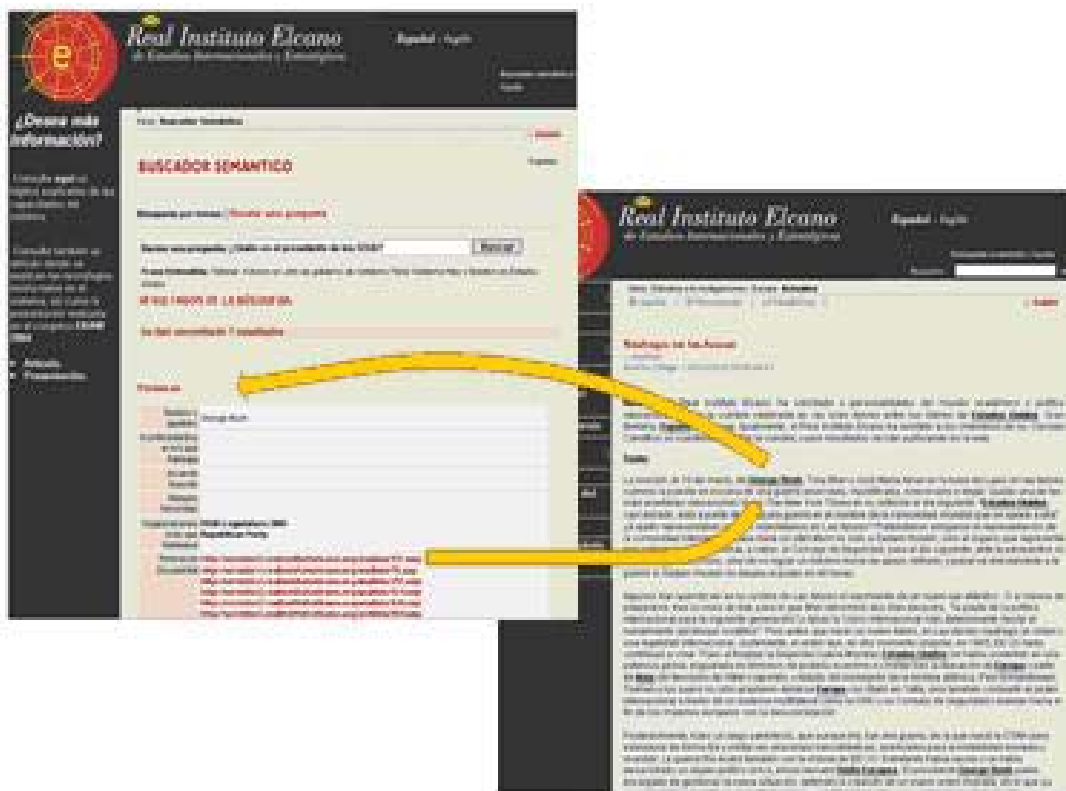


Figure 4: Links between The Instances and The Documents.

traction in a generic, scalable and open architecture. The state of the art of most of these approaches is still not mature enough (few commercial deployments) to provide concrete comparison in terms of performance and memory requirements.

6 Conclusions

A semantic search engine for a closed domain was presented. The figures of the evaluation are promising, as more than 60% of the spontaneous questions are understood and correctly answered when they belong to the application domain. However, some things need to improve, such as the automatic link generation, a more flexible mechanism for building queries, an automated process to generate complete synonym files from linguistic resources, just to mention a few.. It would also be of a high interest to completely decouple the search engine from the domain information, which are currently lightly connected, in order to be able to apply the semantic search engine to a new domain just by replacing the domain ontology and the synonyms files.

The semantic search engine is, at the same time, a proof of the utility and applicability of the Knowledge Parser © which will also be further developed in future projects.

Acknowledgements

Part of this work has been funded by the European Commission in the context of the project Esperanto Services IST-2001-34373, SWWS IST-2001-37134, SEKT IST-2003-506826 and by the Spanish government in the scope of the project: Buscador Semántico, Real Instituto Elcano

(PROFIT 2003, TIC). The natural language software used in this application is licensed from Bitext, <<http://www.bitext.com>>. For ontology management we use JENA libraries from HP Labs, <<http://www.hpl.hp.com/semweb>>.

References

- [1] A. Gómez-Pérez et al. *Ontological Engineering*. Springer-Verlag. London, UK, 2003.
- [2] V. R. Benjamins et al. (KA)2: Building ontologies for the internet: a mid term report. *International Journal of Human-Computer Studies*, 51(3):687–712, 1999.
- [3] W. N. Borst. *Construction of Engineering Ontologies*. PhD thesis, University of Twente, 1997.
- [4] Contreras et al. D31: Annotation Tools and Services, Esperanto Project, <<http://www.esperanto.net>>.
- [5] A. Farquhar et al. The ontolingua server: a tool for collaborative ontology construction. *International Journal of Human-Computer Studies*, 46(6):707–728, June 1997.
- [6] T. R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5:199–220, 1993.
- [7] N. Guarino. Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human-Computer Studies*, 43(5/6):625–640, 1995. Special issue on The Role of Formal Ontology in the Information Technology.
- [8] G. van Heijst et al. Using explicit ontologies in KBS development. *International Journal of Human-Computer Studies*, 46(2/3):183–292, 1997.

- [9] Protege 2000 tool, <<http://protege.stanford.edu>>.
- [10] M. Uschold and M. Gruninger. Ontologies: principles, methods, and applications. Knowledge Engineering Review, 11(2):93–155, 1996.
- [11] WordNet, <<http://www.cogsci.princeton.edu/~wn/>>.
- [12] Atanas Kiryakov et al. Semantic Annotation, Indexing, and Retrieval 2nd International Semantic Web Conference (ISWC2003), 20-23 October 2003, Florida, USA. LNAI Vol. 2870, pp. 484-499, Springer-Verlag Berlin Heidelberg 2003
- [13] H. Cunningham et al. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002
- [14] José Kahan et al. Annotea: An Open RDF Infrastructure for Shared Web Annotations, in Proc. of the WWW10 International Conference, Hong Kong, May 2001.
- [15] Fabio Ciravegna. "(LP)2, an Adaptive Algorithm for Information Extraction from Web-related Texts", in Proceedings of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining, held in conjunction with the 17th International Conference on Artificial Intelligence (IJCAI-01), Seattle, August, 2001
- [16] P. Kogut and W. Holmes. "AeroDAML: Applying Information Extraction to Generate DAML Annotations from Web Pages", in Proceedings of the First International Conference on Knowledge Capture (K-CAP 2001).
- [17] V.R. Benjamins et al. Six Challenges for the Semantic Web. White Paper, April 2002.