International Affairs Portal: A Semantic Web Application

Contreras J.[1], Benjamins V.R[1], Blázquez M[1], Losada S[1], Salla R[1], Sevilla, J[1], Navarro D.[1], Casillas J.[1], Mompó A.[1], Patón D.[1], Rodrigo L.[1], Tena P.[2], Martos I[2]

[1]Intelligent Software Components, S.A., www.isoco.com
{rbenjamins, jcontreras}@isoco.com
[2]Real Instituto Elcano, www.realinstitutoelcano.org
pilar.tena@r-i-elcano.org

Abstract This paper describes a semantic portal on the domain of International Affairs. This application is an integration of several technologies in the field of the Semantic Web in a complex project. We describe an approach, tools and techniques that allow building a semantic portal, where access is based on the meaning of concepts and relations of the International Affairs domain. The approach comprises an automatic ontology-based annotator, a semantic search engine with a natural language interface, a web publication tool allowing semantic navigation, and a 3D visualization component. The portal is being deployed in the Royal Institute Elcano[†] (Real Instituto Elcano) in Spain, which is a prestigious independent political institute whose mission is to comment on the political situation in the world focusing on its relation to Spain. As part of its dissemination strategy it operates a public website. The online content can be accessed by navigating through categories or by a keyword-based, full text search engine. The work described in this paper aims at improving access to the content. The semantic portal is currently being tested by the Institute.

1. Introduction

Worldwide there are several prestigious institutes that comment on the political situation in the world, such as the UK's Royal Institute for International Affairs (www.riia.org), the Dutch Institute for International Relations (www.clingendael.nl). In Spain, the Real Instituto Elcano (Royal Institute Elcano, www.realinstitutoelcano.org) is fulfilling this role. The institute provides several types of reports where they discuss the political situation in the world, with a focus on events relevant for Spain. The reports are organized in different categories, such as Economy, Defense, Society, Middle East, etc. In a special 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

[†] Juan Sebastian Elcano was a famous Spanish sailor, the first seaman who ever made the complete circuit of the globe

helpful instruments for information retrieval (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 to find 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 describe a project whose aim was to provide semantic access to content available in the portal of the Elcano Institute. With semantics, we mean here meaning related to the domain of International Affairs. In other words, we aim to construct an island of the Semantic Web for the International Affairs sector.

In order to construct this Semantic Web Island, we use an approach, tools and techniques that are being developed in the context of several European and Spanish R&D projects. Components include:

1. A domain ontology (in this case an ontology of International Affairs)
2. An automatic annotator (metadata generator), called Knowledge Parser®
3. A semantic search engine with a natural language interface, as well as a forms-based interface
4. A publication tool for publishing semantic content on the web –Duontology®-, enabling semantic navigation including a 3D visualization tool

In Section 2, we describe the ontology of the International Relations domain. In the section 3, the idea of semantic portal is described, based on the Duontology® approach. Section 4 details how we populate the ontology with instances, and how we establish relations between the current content of the Elcano Institute and the (instances of the) ontology. Section 5 concludes the paper.

2. An Ontology of International Affairs

2.1. Ontology

An ontology is a shared and common understanding of some domain that can be communicated across people and computers [6, 7, 3, and 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 especially the social part for building a commonly agreed ontology is not easy [2].

2.2 An Ontology of International Affairs

Based on interviews with experts of the Elcano Institute, we used the CIA word factbook (www.cia.gov/cia/publications/factbook/) as the basis for 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] to determine the scope and granularity of the domain ontology. Some examples of competency questions that we considered include:

```
What countries are participating on Iraq campaign?

Who is the head of the state of France?

What government type has Georgia?

How big is the population of Iceland?

Which are all European Union member countries?

Which are all agreements between Spain and Brazil subscribed
during Da Silva's govern?
```

An important design decision we took (based on [13]) was that relationships between concepts are modeled as first class objects. This decision was taken because often the relationships themselves have attributes that cannot be modeled by its involving concepts. Take for example, the relation "in_favour_of" between an agent (person, nation, government) and an event (war, boycott, treaty). This relation is qualified by a start and end date, which is not meaningful to agent nor event.

The ontology consists of several top level classes, some of which are:

1. Place: Concept representing geographical places such as countries, cities, buildings, etc.
2. Agent: Concept taken form WordNet [11] representing entities that can execute actions modifying the domain (e.g.: Persons, Organizations, etc.)
3. Events: Time expressions and events
4. Relations: Common class for any kind of relations between concepts.

The ontology has been constructed in Protégé 2000 [9]. Fig 1 shows a fragment of the ontology in Protégé 2000.

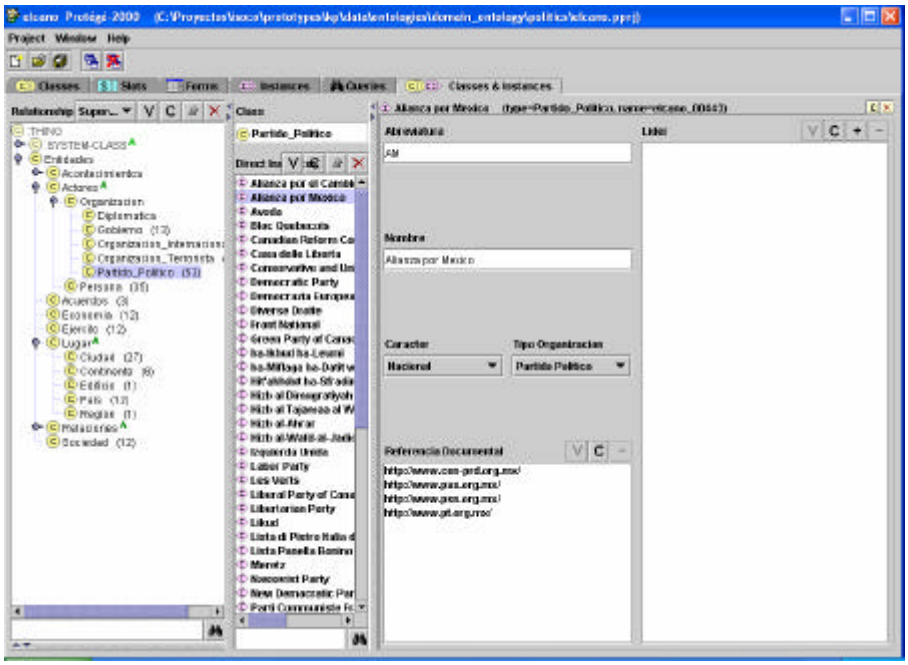

Fig. 1. Ontology for International Affairs

3. A Semantic Portal for International Relations

A semantic portal provides semantic access (as opposed to keyword-based access) to content. In our case semantic access is provided through:

1. Semantic publishing and navigation
2. Semantic search engine
3. 3D Visualization

3.1 Semantic Publishing and Navigation

Experience has shown [14] that the knowledge base as modeled by domain experts and knowledge engineers is not always a good candidate to visualize it as is.

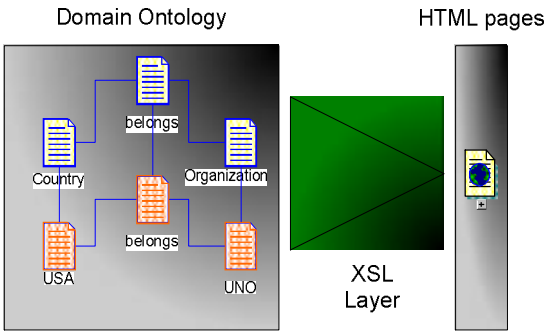

Fig. 2. Explicit visualization using direct translator

The main purpose for building ontologies is to provide semantic content for intelligent systems. The knowledge models are designed to offer the appropriate information to be exploited by the software. No visualization criteria are used to build an ontology and often the information is not suitable to be published as it is:

1. Concepts may have too many attributes
2. When relations are represented as independent concepts (first class objects) the navigation becomes tedious
3. Concepts to be shown do not always correspond to modeled ones.

Therefore we felt a need for explicit visualization rules that allow the creation of views on the International Relations ontology, in order to visualize only the relevant information in a user friendly way. We introduced the concept of "visualization ontology", which makes explicit all visualization rules and allows an easy interface management. This ontology will contain concepts and instances (publication entities) as seen on the interface by the end user, and it will retrieve the attribute values from the International Relations ontology using a query. It does not duplicate the content of the original ontology, but links the content to publication entities using an ontology query language. This way one ontology that represents a particular domain can be visualized through different views.

The visualization ontology has two predefined concepts:

1. Publication entity: Concept that encapsulates objects as they will be published in the portal. Any concept defined in the visualization ontology will inherit from it and should define these attributes
   -XSL style-sheet associated to the concept that translates its instances to final format (HTML, WAP, VoiceXML, etc.)
   -Query that retrieves all attribute values from the original ontology.
2. Publication Slot: Each attribute that is going to appear on the web should inherit from this concept. Different facets describe how the attribute will appear on the page.
   - Web label: The label that will appear with the value
   - RDQL: reference to the query used to retrieve the attribute value

- Link: When the published value should perform some action on mouse click (link, email, button, etc...), the action is described here.

Portal elements are described as children of the Publication Entity and their instances are defined according to the languages the entity will be published in (labels in English, Spanish, etc.), or the channel (whether the transformation style-sheet is going to translate into HTML, WAP, or just XML

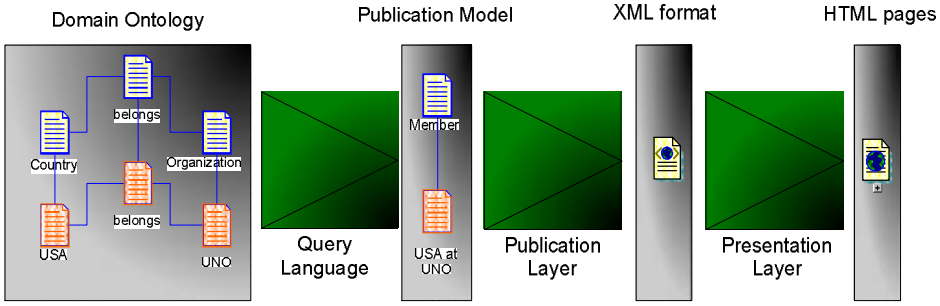

Fig. 3. Publication using specific visualization ontology

Back-office management is divided into two tasks:
1. Content management on domain ontology: adding new instances or modifying the overall schema.
2. Visualization management on publication ontology: modifying how information is shown (look and feel, layout, etc.)

Both tasks are performed using the Protégé 2000 editor, since both domain and publication models are defined in the RDF language.

3.2. Semantic Search Engine

We have developed a Semantic Search Engine for improving content access. Semantic search engines return instances that constitute answers to queries rather than documents containing searched strings as traditional keyword based engines would do. Semantic engines work with the meaning of the query terms. The meaning of each term is defined using domain ontology.

The user can ask for a list of instances of a selected concept putting general constraints on attribute values. For instance, he or she can ask for all events that happened during 1991. Traditional engines would return all documents containing that number, including birth dates, names, etc. A Semantic engine returns instances of the concept “Event” whose duration includes the year 1991. For each instance there exists a link (called reference) to documents where that event is mentioned. The user is also able to make a compound query nesting concepts through their attributes, for

example: all countries that have common border with Lithuania. Other way of searching is looking for relations. Since this ontology has been designed to model relations as first class concepts, the user can make a query about any relation between two states (e.g.: “ All relations between Spain and France”).

There are two kinds of interface for the search engine. The first is based on forms representing domain concepts and existing relations. The user chooses some concept and constructs a complex query putting values for attributes and/or nesting more concepts through relations. This kind of search is based on the conceptual model of the domain ontology.

The second type of interface accepts input written in natural language. The user can, if he or she prefers so, formulate a simple natural language query. It is the task of the system to understand the query. For this, the system parses the input sentence using NLP software and identifies those terms that are related to the ontology, such as concepts, instances, attributes or values (i.e those terms carry domain-specific semantics). Based on those terms and on the NLP analysis of the sentence, the system generates a domain-specific semantic representation of the sentence as a path between concepts, with some constraints introduced by values. This representation is then transformed into an RDQL query, which is then submitted. The current version of the NL interface allows simple sentences such as: "All countries in war with Iraq", "President of France", "Population of Mozambique". It also allows more complex structures as: "Which is the political party of the president of the French Government?”

3.3 3D Visualization

We have also developed a 3D generation module that allows navigating through the search result or ontology content. For that purposes we have implemented software that translates any given domain ontology, applying visualization rules, into the X3D [12] standard. The resulting scene shows instances in a three-dimensional net represented as geometrical bodies with an ad-hoc defined texture. The scene is highly interactive allowing users to move the focus position and interact with the object by clicking on them.

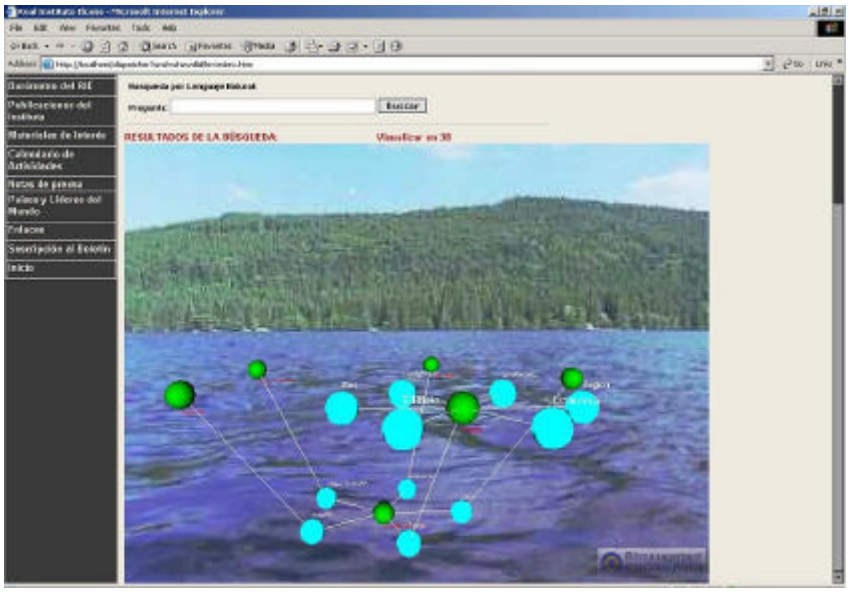

Fig. 5. 3D navigation on domain ontology

4. Automatic Metadata Generation

The annotation task for the Semantic Web takes as input existing content, either structured, semi-structured or unstructured, and provides as output the same content along with a semantic annotation based on ontologies. The semantics as such are defined in ontologies. The annotations provide pointers to these ontologies. Some fragment of text needs to be associated with ontological metadata. The result is that an instance of a concept is created, or that a new occurrence of an existing instance is recorded.

Annotation can be performed in several manners, ranging from completely manual to tool-assisted to fully automatic. As a result of the analysis performed in [4], it turns out that the type of annotation approach to be chosen depends on the rate of structure the content exhibits. More structured sources allow for more automation, while maintaining the quality of the annotations. As has been the experience of several researchers and practitioners, the annotation effort is a serious barrier to the Semantic Web [2].

Our approach for automatic metadata generation is based on a combination of several technologies from the information extraction research area: Natural Language Processing (NLP), Text Engineering, Document Structure Processing and Layout Processing. Such a combination of techniques allows processing each source with the most suitable and effective approach depending on its structure and content. For

instance, for a highly structured table-based text the most effective approach would consists on structure processing with some help of layout and NLP techniques. For large descriptions where whole sentences are usual, the most appropriate approach would be mainly NLP processing.

We named our system Knowledge Parser® since it is able to parse content and extract knowledge from it. Following we explain in briefly how our Knowledge Parser works. Figure 4 illustrates the process executed in three main steps: Source Preprocessing, Information Identification and Ontology Population.

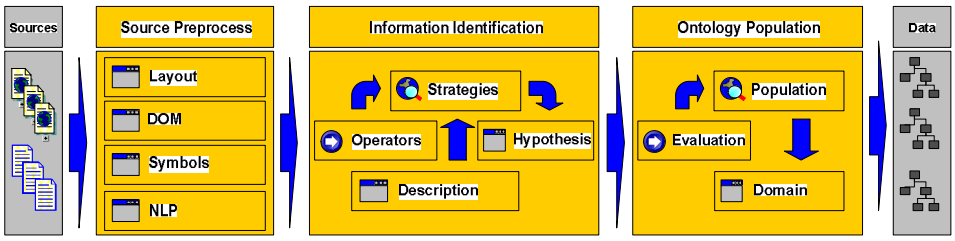

Fig. 4. Overview of the extraction and population process

4.1 Source Preprocessing

The Source Preprocess module (on the left-hand side) provides four different visions on the same source using four document access models:

1. Document Object Model (DOM): It is used for HTML understanding of the source in term of allowing the system for navigation and tag processing
2. Text Model: This model treats the source as a simple character string and allows using regular expression techniques.
3. Layout Model: Provides a special model for the source assigning two dimension coordinates to each element.
4. NLP Model: Provides access NLP information such as a list of proper names, verbal phrases, synonyms, etc.

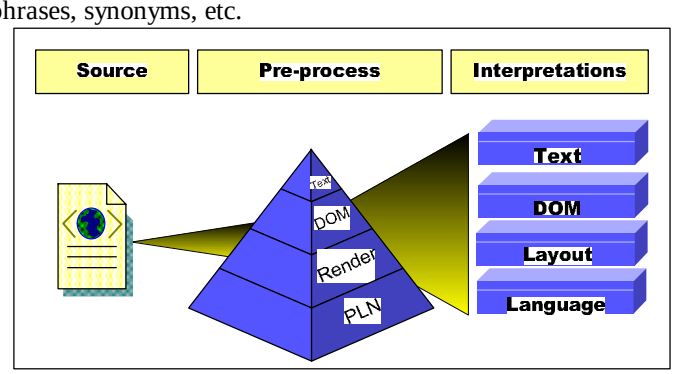

Fig. 5. Source pre-preprocessing according to different technologies

4.2 Information Identification

The goal of the central Information Identification Module is the main control of the extraction activity. The result of this step is a hypothesis set about how the extracted information pieces fits into the domain ontology. In contains three key components:

1. **Operators**: perform extraction actions on document access models provided in the first step.
2. **Strategies**: build operator sequences according to user time and quality requirements
3. **Source Description**: formal description of typical characteristics of sources (see below)

In the following, we explain each of them in more detail.

Operators are software functionalities associated to each source access model that can perform some useful action for extraction purposes. We classify operators into three categories depending on their behavior:

1. Retrieval: perform text chunk extraction on some source access model (e.g.: all proper names extraction form the NLP source access model)
2. Check: checks some constraint on the extracted data (e.g.: whether two elements are in the same visual row in the Layout access model)
3. Execute: allows executing source elements (e.g.: in the DOM access model, executing a web form or navigating through a link)

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.

Source Description: In order to perform any strategy the system needs to understand what kind of information pieces are expected in the source. The **source description** is formalized in, a so called, wrapping ontology. The wrapping ontology contains the following elements:

1. Document Types: Description of document types in the source. For instance the CIA World Factbook home page is described with its URL and some relation to country name pieces.
2. Pieces: Are the basic elements of the retrieval process. They contain the searched information with which the ontology will be populated. Each piece has defined its possible data types (number, NLP phrase, string, etc.).
3. Relations: Relation between pieces and documents are modeled here. For instance, there are two possible relations between two pieces:
4. Layout relation: (e.g.: IN ROW: Two pieces may lay in the same visual row in the documents, see Fig. 3)

5. Semantic relation: (e.g.: verbal predicate: Two pieces are related using a verbal phrase with the main verb: 'agree' or synonyms)

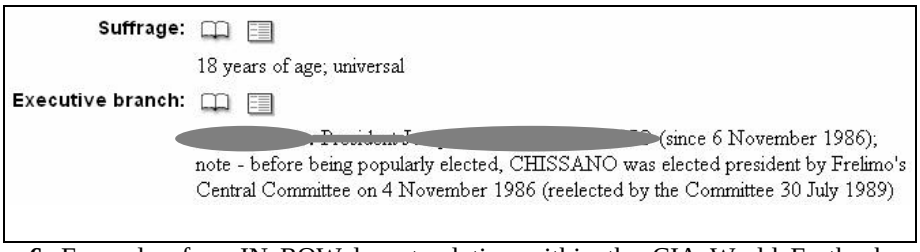

Fig. 6. Example of an IN ROW layout relation within the CIA World Factbook: Mozambique page.

4.3 Ontology Population

The final stage of the overall process is to decide which hypothesis represents the extracted information to insert into the ontology. The Ontology Population Module (on the right-hand side at Fig. 2) is in charge of evaluating and sorting hypothesis for their insertion in the domain ontology. For evaluation purposes the module simulates insertions and calculates the cost according to the number of new instance creations, instance modifications or inconsistencies found. According to the Occam's razor principle only the lowest cost hypothesis is used for population. A low cost corresponds to an insertion with minimal modifications with respect to already existing instances. The result of this step, as well as the result of the whole system, is the augmented domain ontology including new instances and values.

Next, we explain in detail how is the population process performed. The behavior of the population module heavily relies on the smartness of the running wrapping strategy. Communication is performed through hypothesis objects. For each document in the source there is constructed a set hypothesis with different possibilities. One the best, lowest cost, hypothesis for each document will be populated. The hypothesis contains information about possible assignments between expected information pieces (defined in the wrapping ontology) and source chunks (retrieved from the text).

In order to illustrate the overall population mechanism, the following example is considered (Note that some example's details are deliberately omitted for the sake of simplicity). A concept in the domain ontology is *Country* which is a subclass of the concept *Place*. This concept represents information such as: country's borders, society, military manpower, etc. Also, this concept has relations with other concepts; an example of this kind of relation is a country that takes part in an event (war, conference, etc) in a specific date.

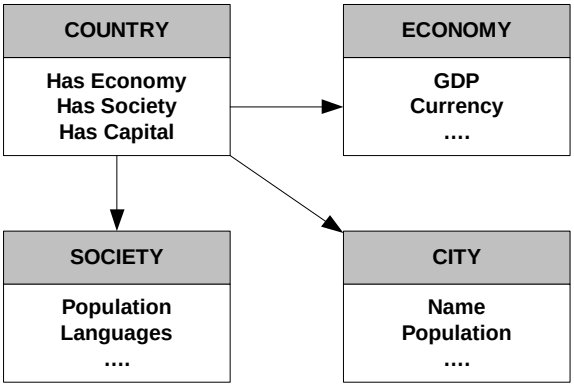

Fig. 7. Graph with the relations between the *Country* class and the connected classes *Economy*, *Society* and *City*.

Extracted information for each one of the slots is acquired by the wrapping strategy in separate decision cycles, which implies that different hypothesis, will be sent to the population module. Some difficult issues arise from this fact:

- How to relate slot values that come from different hypotheses but belong to the same class instance.
- How to decide whether a new instance of a class must be created or the population value substituted.

The population strategy adopted to solve these conflictive cases is two-fold:

- Formulate hypotheses in the wrapping strategy as tightly coupled as possible, e.g. any slot of a class related to a country will be found in a hypothesis together with the country key population value. A *recursive* slot of the population descriptor is used to link concepts that are related in the ontology, e.g. a populated city is recursively the capital of the country.
- When consistent hypothesis cannot be formulated, the *Occam's razor* principle is used, and the population values are assigned to the nearest applicable context, which can lead to misconceptions, but has shown to be useful in compromise cases.

The following diagram represents an example hypothesis, result of the wrapping for the *Capital Name Piece* in the *Country page* document, coming from a relation with the *Country Name Piece*:

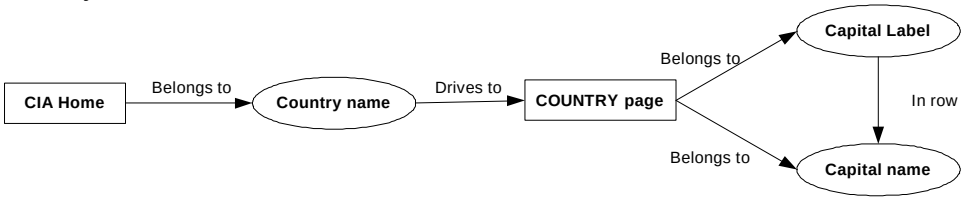

Fig. 8. Example of *CandidateConceptHypothesys*.

4.4 Wrapping International Affairs Domain

In this particular application for the Elcano Institute, our Knowledge Parser fulfils two roles. The first one is the wrapping of the CIA World Factbook in order to populate the ontology with instances. From the CIA Factbook we extract information regarding countries such as their government composition, geographical data, political and commercial agreements, etc. We basically use a combination of the Layout Model and the NLP model. Once the ontology is populated with instances, we apply the Knowledge Parser to the documents provided by the Elcano institute. At this point, we are interested in identifying occurrences of instances. For example, if we found in the Factbook "Bush" as an instance of the "Head of State" concept of the ontology, then in the Elcano documents we want to find occurrences of these instances. That is, we want to recognize that when the parser finds "Bush", "George Bush" or "The president of the US", these are occurrences of the instance “Bush”.

Summarizing, in the Elcano application, we use the Knowlege Parser in a two-step approach. Of course, we could have populated the ontology directly by wrapping only the documents of the Elcano institute. However, it would be a pity not to use the CIA Factbook, which is a high-quality, up to date and free repository with relevant information. The state of the art is still that structured information is easier to wrap than unstructured information.

5. Related Work

Our Knowledge Parser is related to several other initiatives in the area of automatic annotation for the Semantic Web, including KIM [15], which is based on GATE [16], Annotea [17] of W3C., Amilcare [18] of the Open University (also based on GATE), and AeroSWARM[1]. For an overview of those approaches and others, see [5]. 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 Extraction 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

In this paper, we presented an application of Semantic Web Technology for the International Affairs Sector. The application will be launched by the Royal Institute Elcano in the Fall of 2004 at www.realinstitutoelcano.org. Currently we are in the final testing phase (pre-deployment). The application allows visitors of the web site to

[1] http://ubot.lockheedmartin.com/ubot/hotdaml/aeroswarm.html

access the Elcano's documents in a more intelligent manner through a semantic search engine, semantic navigation and 3D graphical navigation and interaction.

The ontology of International Relations is inspired by the CIA World Fact book, which is a large online up to date source with relevant information. Documents of the Elcano Institute are automatically associated to this ontology, thereby disclosing them semantically. Semantic access is made possible through a software we call Knowledge Parser®, which is capable of "understanding" digital text.

The Knowledge Parser® is a generic architecture currently integrating four different technologies relevant for information extraction from text: NLP, Text Engineering, Document Structure and Layout. Its input is digital content in more or less structured form, and the output is an ontological classification of the content (ontological annotation).

For this particular application, we have applied the Knowledge Parser® in a two-step bootstrapping approach. First, to automatically populate the International Relations Ontology with instances from the CIA World Fact Book, and secondly, to automatically find occurrences of the ontology instances in the Elcano documents.

In future work we plan to include content of other institutes of the same area, providing a semantic one-stop shop for access to information about International Affairs.

7. Acknowledgements

Part of this work has been funded by the European Commission in the context of the project Esperonto 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 (www.bitext.com). For ontology management we use JENA libraries from HP Labs (http://www.hpl.hp.com/semweb) and Sesame (http://www.openrdf.org).

References

[1] V. R. Benjamins and D. Fensel. Editorial: Problem-solving methods. International Journal of Human-Computer Studies, 49(4):305–313, October 1998. Special issue on Problem-Solving Methods.

[2] V. R. Benjamins, D. Fensel, S. Decker, and A. Gomez-Perez. (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, Enschede, 1997.

[4] Contreras et al. D31: Annotation Tools and Services, Esperonto Project: www.esperonto.net
[5] A. Farquhar, R. Fikes, and J. Rice. 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, A. T. Schreiber, and B. J. Wielinga. Using explicit ontologies in KBS development. International Journal of Human-Computer Studies, 46(2/3):183–292, 1997.
[9] Protete 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] X3D http://www.web3d.org/x3d.html
[13] Benjamins, Contreras, et al, Cultural Heritage and the Semantic Web. In proceedings of First European Semantic Web Symposium, May 2004, Crete,
[14] Rubén Lara, Sung-Kook Han, Holger Lausen, Michael Stollberg, Ying Ding, Dieter Fensel: An Evaluation of Semantic Web Portals, submitted to IADIS Applied Computing 2004 Conference
[15] Atanas Kiryakov, Borislav Popov, Damyan Ognyanoff, Dimitar Manov, Angel Kirilov, Miroslav Goranov 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
[16] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. 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
[17] José Kahan, Marja-Riitta Koivunen, Eric Prud'Hommeaux, and Ralph R. Swick, Annotea: An Open RDF Infrastructure for Shared Web Annotations, in Proc. of the WWW10 International Conference, Hong Kong, May 2001.
[18] 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