Ontology Based Queries – Investigating a Natural Language Interface

Ielka van der Sluis
Computer Science
Trinity College Dublin
vdsluis@cs.tcd.ie

Feikje Hielkema
Computing Science
University of Aberdeen
f.hielkema@abdn.ac.uk

Chris Mellish
Computing Science
University of Aberdeen
c.mellish@abdn.ac.uk

Gavin Doherty
Computer Science
Trinity College Dublin
gavin.doherty@cs.tcd.ie

ABSTRACT

In this paper we look at what may be learned from a comparative study examining non-technical users with a background in social science browsing and querying metadata. Four query tasks were carried out with a natural language interface and with an interface that uses a web paradigm with hyperlinks. While it can be difficult to attribute differences in performance to specific design features, a qualitative analysis of the user behavior provides some insight into the task and problematic aspects of existing interfaces. In general it was found that casual subjects have difficulties recognizing typical ontology based concepts like objects, attributes and values.

Author Keywords

Querying and browsing, metadata, evaluation, natural-language interfaces, web-based interfaces.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

The advent of Semantic Web technologies [2] has generated a number of challenges relating to the use of technology by domain experts and researchers in areas such as social science [3]. Among the questions to be addressed are the extent to which these researchers are comfortable with the Web as a framework for research practice and collaboration; whether ontologies are appropriate (and acceptable) to this community as a way of representing concepts to facilitate their research activities; the utility (or otherwise) of existing metadata frameworks in use by the social sciences; and how best to integrate e-science tools and methods into existing working practices.

A key aspect is concerned with support for creation of metadata and access to resources annotated by semantic metadata. This semantic metadata is captured with RDF (Resource Description Framework; www.w3.org/RDF/), statements of the type Property (subject, object) whose semantics are defined by OWL ontologies (www.w3.org/TR/owl-features/). These ontologies consist of classes (e.g. City, State) and properties (hasCapital, Name). The RDF statements describe instances of these classes (e.g. 'The State of New York, whose capital is New York'). RDF is a subset of XML and potentially difficult to understand for most non-technical users. This paper focuses on browsing RDF and the task of constructing complex queries.

Support for these activities for casual, non-technical users is an important challenge for the entire Semantic Web research community. As most members of the social science community are unfamiliar with complex formalisms such as RDF, this makes them a representative group of non technical users of the Semantic Web. Non-technical users may benefit from what the Semantic Web offers, but may be deterred by its complexity and the need to learn to use graphical representations or controlled languages. While well-designed graphical tools can provide advantages, tools that use graphical representations (e.g. CREAM [6] or SHAKEN [13]) may be difficult to interpret for users unused to complex graphical presentations or ontologies. For instance, Petre [9] argues that graphical readership is an acquired skill, and describes experiments into reading comprehension of graphical and textual representations. These showed that for some tasks people process graphical representations significantly slower than text, with novices in particular suffering from mis-readings and confusion. Kaufmann and Bernstein [7] demonstrated via an experiment that compared four different query interfaces for the Semantic Web, that naive users preferred the interface that used full natural language sentences (as opposed to keywords, partial sentences and a graphical interface). Hence, it is worth considering whether a natural language representation of metadata could serve as a good solution for novices to the Semantic Web (such as many social scientists). In order to investigate this possibility a tool named LIBER was developed, which uses natural language to provide access to metadata. This paper presents a comparative study that was set up to assess and explore the querying and browsing interface of LIBER.

INTERFACES FOR QUERY CONSTRUCTION

LIBER (Language Interface for Browsing and Editing RDF) was developed for providing access to descriptions of social science resources (e.g. papers, statistical datasets, interview transcripts) held in a data repository. The interface (driven by a number of ontologies) enables users

----
Workshop on Visual Interfaces to the Social and Semantic Web (VISSW2010), IUI2010, Feb 7, 2010, Hong Kong, China.

to find resources in the repository through querying and browsing of metadata, and to deposit new resources with a metadata description. Each component of the LIBER interface uses natural language generation to present information to the user through the WYSIWYM (What You See Is What You Meant) approach [13]. WYSIWYM has been used by a number of other projects, such as MILE [10] and CLEF [5]. The positive results from these projects [4, 11], suggest that WYSIWYM could be a suitable approach to use for constructing and accessing metadata.

With WYSIWYM a system generates a feedback text for the user that is based on a semantic representation. The representation includes generic phrases, or 'anchors', which correspond to objects in the description. Each object has a pop-up menu which lists the properties it can have; to add information, the user selects a property and provides an appropriate value. In LIBER, properties of objects are used in queries, which may also include boolean operators ('and', 'or', 'not'), and queries may also include optional elements. Results are presented as the query is constructed.

As many other querying tools have been developed in the Semantic Web community, we could compare LIBER's querying and browsing modules to existing systems. The question of which approach (natural language, graphics, faceted browsing) produces more usable interfaces is far from settled. We were therefore interested in comparing the natural language interface of LIBER to one that uses a different approach. Kaufmann & Bernstein [7] describe an evaluation study in which they compared four querying interfaces: a graphical interface, a controlled language interface, a natural language interface that uses confirmation dialogues for disambiguation (Querix), and a natural language interface that identifies relevant key phrases in the search term. The study showed that all natural language interfaces outperformed the graphical interface and that subjects preferred Querix and achieved the best results with it. We decided to use a similar set-up and materials for our evaluation, so we could adopt a simple ontology and have a reference point for the evaluation results.

We compare the LIBER interface with Longwell [8], a web-based RDF-powered faceted browser developed by the SIMILE project at MIT. Longwell takes an RDF dataset as input, and creates a website in which the data can be browsed and filtered using classes, properties and keywords. The user browses through the dataset by clicking hyperlinks (which correspond to classes, properties and values) and keyword searching; each click and keyword search adds (or removes) a filter. Longwell thus uses the web paradigm to present information rather than natural language, and we were interested to see which would prove more effective and/or popular.

Following Kaufmann & Bernstein's study, it might be expected that users would be more accurate and complete tasks more quickly with the natural language tool LIBER than with the faceted browser Longwell. Realistically, we knew this inference might not apply as that study compared the natural language based interface to a graphical interface, while Longwell is a faceted browser; moreover, Longwell was developed by a company and has a user community, while Kaufmann & Bernstein produced their own graphical interface, so we cannot be sure that its deficiencies reflect those of such interfaces in general.

EXPERIMENTAL STUDY

Before describing the experiment, we note that there can be problems with interpreting comparison studies. Importantly, it can be difficult to attribute differences in performance to specific design features, such as the use of a natural language interface, as such choices necessitate many other differences in the design. For example, a badly executed natural language based design might be outperformed by another interface, whereas a well-executed natural language design might perform better.

Methodology

Twenty students and researchers with backgrounds in various social science related disciplines participated, one of which did not finish the experiment and was excluded (N=19). None had previous experience with LIBER or Longwell, and only two had used an ontology before. Subjects were asked to supply some background information, then were handed a one-page description of one of the tools and were asked to follow the instructions to become acquainted with its operation. They then received four questions to answer, and were asked to find the answer using the tool without relying on their own general knowledge about the world. When finished, subjects were asked to fill out a SUS questionnaire [1], a standardized usability test containing ten standardized questions (e.g. 'I felt very confident using the system') which are rated on a 5-point Likert scale. This procedure was repeated for the other tool. Afterwards, subjects were asked to complete a questionnaire in which the tools were compared directly. On average subjects needed about 45 min to finish the task.

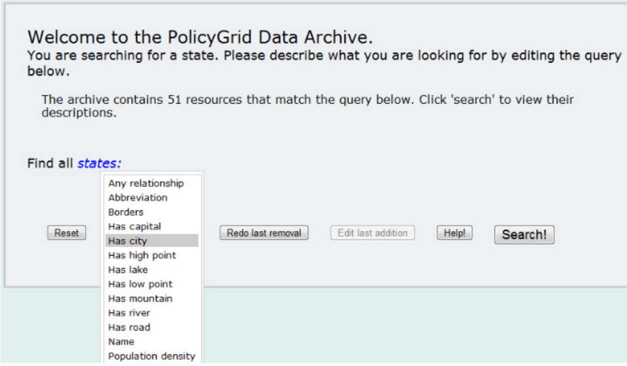
Both the order of the tools and the order of the questions were varied per subject. For both tools we recorded the answers the subjects provided and the time it took to answer a question, and made video captures of the screen for qualitative analysis. To drive both tools, we used a simple ontology that models the geography of the USA, which was developed for Kaufmann & Bernstein's study and is available online[1]. It is not faithful to the real world situation (Alaska appears to have the smallest state area, for example), but this made it easier to prevent subjects from relying on their own knowledge and thus bias the results. We used two sets of questions, which were based on those used by Kaufmann & Bernstein in their study. One of the two sets is exemplified below:

1. What is the area of Alaska?
2. How many lakes are there in Florida?
3. Which states contain a city called Springfield?

[1] http://www.ifi.uzh.ch/ddis/research/semweb/talking-to-the-semantic-web/owltest-data/

4. Which rivers run through the state that contains the largest city in the US?

'Figures 1, 2 an 3 show screenshots of LIBER and Figures 4,5 and 6 show screenshots of Longwell, where the user is searching for the answer to the question 'Which states contain a city called Springfield?'. Both interfaces support multiple strategies for finding this answer; the screenshots portray merely one of them. In LIBER this user has created a search term that provides the answer without further browsing, by searching for all states which have the property 'hasCity' with as value a city by name of 'Springfield'; the answer appears when the user presses 'search'.

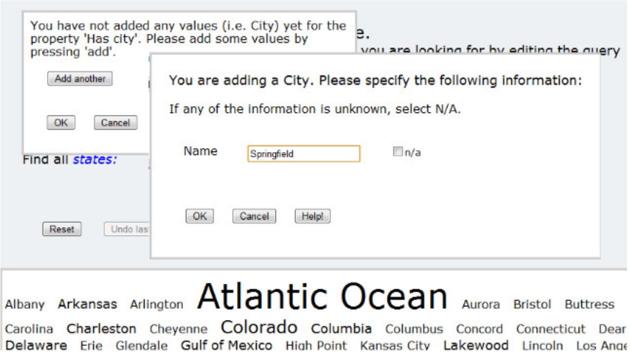


Welcome to the PolicyGrid Data Archive.
You are searching for a state. Please describe what you are looking for by editing the query below.

The archive contains 51 resources that match the query below. Click 'search' to view their descriptions.

Find all *states:*

Figure 1. LIBER: The user chooses the property 'Has city'.



You have not added any values (i.e. City) yet for the property 'Has city'. Please add some values by pressing 'add'.

You are adding a City. Please specify the following information:

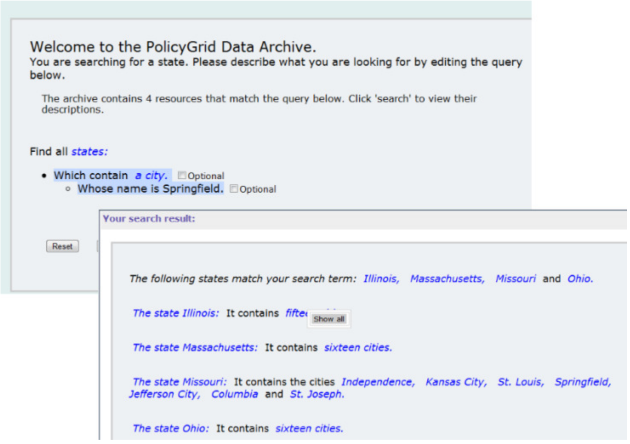
If any of the information is unknown, select N/A.

Name: Springfield □ n/a

Find all *states:*

Add another
OK
Cancel
OK
Cancel
Help!
Reset
Undo las
Albany Arkansas Arlington Atlantic Ocean Aurora Bristol Buttress
Carolina Charleston Cheyenne Colorado Columbia Columbus Concord Connecticut Dear
Delaware Erie Glendale Gulf of Mexico High Point Kansas City Lakewood Lincoln Los Ange

Figure 2. LIBER: The user specifies the name of the city.



Welcome to the PolicyGrid Data Archive.
You are searching for a state. Please describe what you are looking for by editing the query below.

The archive contains 4 resources that match the query below. Click 'search' to view their descriptions.

Find all *states:*

- Which contain *a city.* □ Optional
  - Whose name is Springfield. □ Optional

Your search result:

The following states match your search term: Illinois, Massachusetts, Missouri and *Ohio.*

The state Illinois: It contains *fifte*

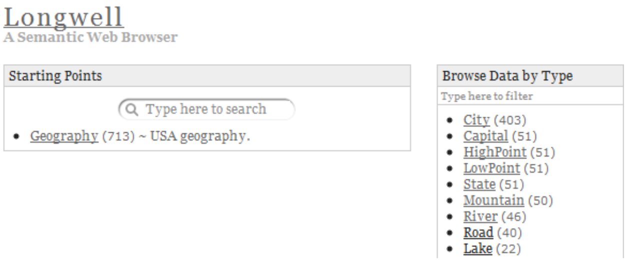
The state Massachusetts: It contains *sixteen cities.*

The state Missouri: It contains the cities *Independence, Kansas City, St. Louis, Springfield, Jefferson City, Columbia* and *St. Joseph.*

The state Ohio: It contains *sixteen cities.*

Figure 3. LIBER: Search results for question 3.

In Longwell, the user has first added a filter 'city' to select all cities, then another filter on the name (Springfield), and finally opened the facet 'cityOf' on the right-hand side to view the four states.'



Longwell
A Semantic Web Browser

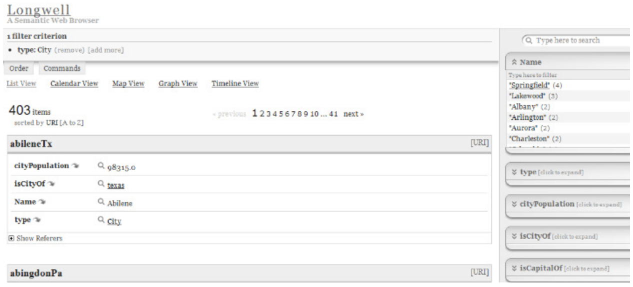
Starting Points

- Geography (713) ~ USA geography.

Browse Data by Type

- City (403)
- Capital (51)
- HighPoint (51)
- LowPoint (51)
- State (51)
- Mountain (50)
- River (46)
- Road (40)
- Lake (22)

Figure 4. Longwell: The user clicks 'city'.



Longwell
A Semantic Web Browser

1 filter criterion

- type: City (remove) [add more]

403 items
sorted by URI [A to Z]

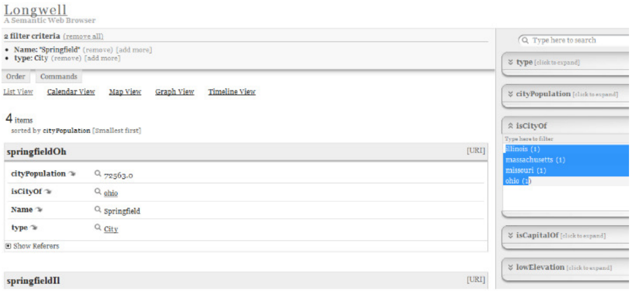
abileneTx [URI]

cityPopulation	98315.0
isCityOf	texas
Name	Abilene
type	City

abingdonPa [URI]

Name
"Springfield" (4)
"Lakewood" (3)
"Albany" (2)
"Arlington" (2)
"Aurora" (2)
"Charleston" (2)

Figure 5. Longwell: The user clicks 'Springfield'in the 'Name' filter.



Longwell
A Semantic Web Browser

2 filter criteria (remove all)

- Name: "Springfield" (remove) [add more]
- type: City (remove) [add more]

4 items
sorted by cityPopulation [Smallest first]

springfieldOh [URI]

cityPopulation	72563.0
isCityOf	ohio
Name	Springfield
type	City

springfieldIl [URI]

isCityOf
illinois (1)
massachusetts (1)
missouri (1)
ohio (1)

Figure 6. Longwell: The user opens the facet 'cityOf' to view the results

Results: Comparative Analysis

Two-tailed paired t-tests show that the Longwell interface outperformed the LIBER interface in terms of completion time (LIBER, mean 191.6sec, stdv 57.1sec; Longwell mean 96.5sec stdv 30.0s, p=0.000) and SUS score (LIBER, mean 37.63, stdv 18.11; Longwell mean 61.16, stdv 19.65 p=0.000). Subjects failed to complete tasks more often in LIBER (missing answers: LIBER, mean .47 stdv .62; Longwell mean .11, stdv .32, p = 0.015), but tended to provide more incorrect answers in Longwell (wrong answers: LIBER, mean .58 stdv 1.02; Longwell mean .84, stdev .90, p = 0.384). When asked to compare LIBER and Longwell directly, all but three users preferred Longwell; opinions on reliability were more divided but still in favour of Longwell (11 subjects).

Results: Screen Capture Analysis

We recorded screen captures and annotated the strategies that subjects employed in carrying out the querying task. Some videos did not record properly (N=16). Analysis of the data helped us to identify common errors, delaying factors and misunderstandings as reported below.

Strategies

A clear difference was found between the preferred strategy employed in subjects' initial use of the LIBER interface and the way in which subjects used LIBER over time. In answering the first question, the most frequently used strategy (7 subjects) was phrasing a query that when submitted retrieves the correct answer immediately, without need for further browsing. Five subjects used a different strategy, they formed a small query and used the LIBER browsing interface to find the final answer. From the second question onwards the "query then browse" strategy, dominated (used by 10, 8 and 7 subjects respectively).

With the Longwell interface the most popular strategy for finding answers to the questions was to use the provided descriptions rather than the filters. This preference was independent of the type of the question as well as independent of the experience with the interface that was built up during the task.

Errors

In general, subjects appeared to gain little understanding from the interfaces of how the data in the geographical ontology was modelled (e.g., classes, properties and values). For instance, in both interfaces subjects entered keywords such as 'largest city' (LIBER 4 subjects; Longwell 9 subjects). This shows the extent to which subjects are used to other types of search engines (e.g. a web search on 'largest city' will list the pages that include these search terms), and had difficulty adapting to search strategies suitable for RDF, which simply list population sizes, without comparing them. To search RDF you therefore need a different search strategy, a query that finds those population sizes and then compares them for you.

Compared to Longwell, in LIBER subjects made more mistakes that can be ascribed to minor issues in the interface, such as those caused by not moving values to boxes for inclusion in the query before confirming the query (18 subjects), and those caused by usage of the 'optional' checkbox (7 subjects). Most of these situations were catered for in that LIBER provided a warning or clarification, which brought subjects back on track. Still, in LIBER some errors seem to be specific to the natural language interface, like assigning a property or value to the wrong object (e.g. looking for lakes called 'Florida', rather than for 'lakes in a state called Florida') (4 subjects).

With Longwell fewer things could go wrong but, most likely due to the fact that subjects did not receive any feedback on what went wrong, the same errors were made repeatedly. Compared to LIBER, errors were of a different kind, such as selecting the wrong value for both filters (5 subjects) and descriptions (2 subjects), browsing through only one of multiple results (3 subjects), typos (5 subjects), and misinterpretations of descriptions (5 subjects).

Delays

With both interfaces, subjects appeared sometimes unsure whether all matches were found (Longwell, 5 subjects). In LIBER this happened, when the system stated the number of matches to the query without actually listing them (6 subjects), or when only one match was found (4 subjects). In contrast, it also happened that browsing was stopped after only a partial answer was found (LIBER, 5 subjects; Longwell, 4 subjects). In Longwell, subjects often clicked on links that did not lead them to anything useful, like the description of the ontology itself rather than the instances (10 subjects). In LIBER uncertainties appeared in the selection of menu items (8 subjects) and there were some interface issues that caused delays in task performance, for instance many subjects had trouble closing pop-up windows (11 subjects) or browsing windows (9 subjects). Many of them also experienced focus issues with pop-up windows; it was not understood that pop-up windows needed to be closed before a task could be continued (11 subjects).

DISCUSSION

From the experimental data, it is clear that subjects preferred Longwell over LIBER and they performed better with Longwell than with LIBER in almost all respects. It should be noted, however, that subjects felt that both interfaces were needlessly complicated. While the subject's preference for Longwell might help in choosing between the two applications at the current time, we are more interested in what the experiment tells us about the task of performing complex queries, and in how to improve interfaces to support this activity.

When contrasting the difficulties encountered in the LIBER interface with the comparatively fluid performance in Longwell, we see that with Longwell subjects generally used the same strategy in answering all four questions. In contrast, with LIBER subjects learned while working on the task that a browsing facility is available and that spending less time on a perfect query yielded better results. This indicates that novice users' initial expectations of the querying interface are incorrect. With LIBER many errors and delays can be attributed to minor usability issues in the interface, although some issues do appear to be related to the interface style. The analysis of the screen captures helped to identify areas where the LIBER interface might be improved such as clarification of the 'optional checkbox' and handling of pop-ups and browsing windows. Compared to LIBER, in Longwell fewer things can go wrong, users click on links and end up somewhere else (useful or not). Because of their familiarity with the web paradigm, users may explore the interface more confidently, as they can backtrack when they find themselves on an irrelevant page.

CONCLUSIONS

This paper described a study that was performed to help in the design and refinement of LIBER's interfaces for querying and browsing metadata. The study compares subjects' performance using LIBER with the existing Longwell interface, which provides a benchmark for performance. The study allows us to look at differences in interaction strategy, and to identify issues which may be associated with the interface style, including the use of natural language. The study has focused on initial use of tools for querying and browsing metadata by researchers with backgrounds in social science, yielding insight into the difficulties experienced by casual, non-technical users when operating an interface to an unknown database that nevertheless stored a general domain. A longer training time or a more longitudinal study could well yield different results, and could help to improve the system for use by more experienced users. Also, the use of a database that is less simple, as well as more relevant for the subjects, might make a difference in that subjects would have intuitions and expectations about the ontology used for representing the data, which would be more representative of real world use.

In general, it was found that subjects that do not have any knowledge of RDF data or SQL querying, seem to have difficulties recognizing and distinguishing concepts like classes, properties and values and the way in which they are defined in the ontology used in this study. Subjects seemed to rely on their methods for searching the internet, without realizing that different rules apply to metadata and the particular database that was used for the study. Neither LIBER nor Longwell provide the user with sufficient information about what type of input the system expects. Or in other terms, both LIBER and Longwell have not yet succeeded in providing an interface that supports users in efficiently constructing metadata-based queries.

We believe that the usability of LIBER and Longwell (and natural language interfaces and faceted browsers in general) depends on a number of factors that will vary between and even within domains, such as:

- The experience of users with ontologies and other metadata;
- The data described by the ontologies (for instance, a recipe is more usually described in natural language than geographical data);
- The type of interfaces that users normally utilise (those used to working with databases through e.g. Access would prefer Longwell);
- The size of the ontologies, and the number of individuals within them (large amounts of individuals might cause the generation of very long and therefore confusing descriptions in LIBER);
- The mix of tasks and goals which might have an effect on strategy (e.g. users may have a whole range of interaction types with a browsing system depending on their goals and mode of working.);
- The heterogeneity of the data (Longwell's filters work better if each individual has the same set of properties, while LIBER generates separate menus for each individual, and can thus deal better with heterogeneity).

Further studies should evaluate each of these factors separately in order to provide a better understanding of interfaces to support ontology-based queries.

ACKNOWLEDGMENTS

This research is funded by SFI as part of the CNGL project and the ESRC as part of the PolicyGrid project.

REFERENCES

1. J. Brooke, SUS: a "quick and dirty" usability scale, in: P. Jordan, B. Thomas, B. Weerdmeester, A. McClelland (eds.), Usability Evaluation in Industry, Taylor and Francis, London, 1996.
2. D. De Roure, N. Jennings, N. Shadbolt, The Semantic Grid: Past, Present and Future. In Proc. IEEE'05, 93(3), 2005.
3. P. Edwards, A. Chorley, F. Hielkema, E. Pignotti, A. Preece, C. Mellish, J. Farrington, Using the Grid to Support Evidence-Based Policy Assessment in Social Science. In Proc. UK e-Science All Hands Meeting, Nottingham, 2007.
4. C. Hallett, D. Scott, and R. Power. Composing Questions through Conceptual Authoring. Computational Linguistics, 33(1) (2007) 105–133.
5. C. Hallett. Generic Querying of Relational Databases using Natural Language Generation Techniques. In Proc. INLG'06, pages 88–95, Nottingham, UK, 2006.
6. S. Handschuh, S. Staab, A. Maedche, CREAM: creating relational metadata with a component-based, ontology-driven annotation framework. In Proc. K-CAP'01, ACM Press, Victoria, British Columbia, Canada, 2001.
7. E. Kaufmann, A. Bernstein, How Useful Are Natural Language Interfaces to the Semantic Web for Casual End-Users? In Proc. ISWC'07, vol. 4825 of LNCS, Springer Verlag, Busan, Korea, 2007.
8. Longwell. http://simile.mit.edu/wiki/Longwell
9. M. Petre, Why Looking isn't always Seeing: Readership Skills and Graphical Programming, Communications of the ACM 38 (6) (1995) 33-44.
10. P. Piwek, R. Evans, L. Cahil, and N. Tipper, Natural Language Generation in the MILE System. In Proc. of IMPACTS in NLG workshop, 33–42, Schloss Dagstuhl, Germany, 2000.
11. P. Piwek, Requirements Definition, Validation, Verification and Evaluation of the CLIME Interface and Language Processing Technology. Technical Report ITRI-02-03, ITRI, University of Brighton, 2002.
12. R. Power, D. Scott, and R. Evans. 1998. What You See Is What You Meant: Direct Knowledge Editing with Natural Language Feedback. In Proceedings of the Thirteenth European Conference on Artificial Intelligence, Brighton, UK.
13. J. Thoméré, K. Barker, V. Chaudhri, P. Clark, M. Eriksen, S. Mishra, B. Porter, A. Rodriguez, A Web-based Ontology Browsing and Editing System. In Proc. AAAI-02, Edmonton, Alberta, Canada, 2000.