Walking ThroughCS AKTive Space: a demonstration of an integrated Semantic Web Application

Nigel R. Shadbolt [*], Nicholas Gibbins, Hugh Glaser, Stephen Harris
m.c. schraefel

School of Electronics and Computer Science, University of Southampton, Southampton, United Kingdom

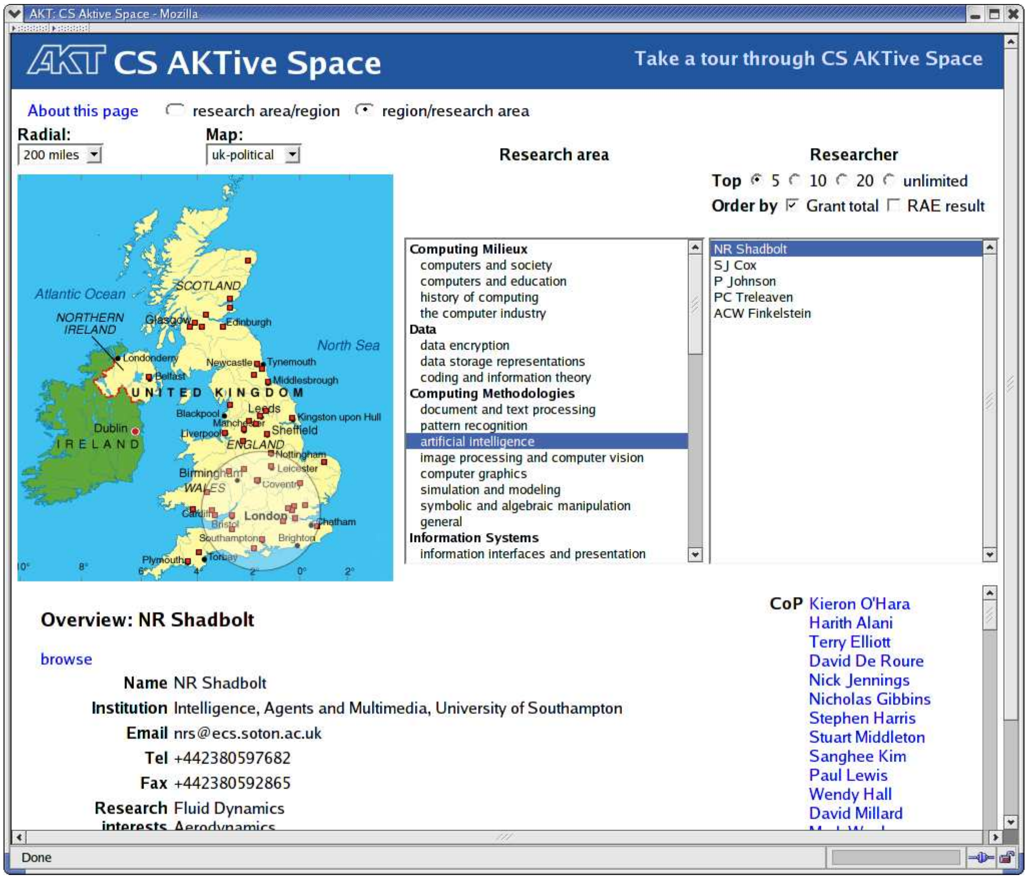
Abstract

We describe CS AKTive Space,an integrated Semantic Web application and winner of the 2003 Semantic Web Challenge [http://challenge.semanticweb.org/]. A demonstration of the application is available at http://cs.aktivespace.org. CS AKTive Space represents and integrates a wide range of heterogenous resources representing the Computer Science Domanin in the UK; it supports the exploration of patterns and implications inherent in the content and exploits a variety of services, visualisations and multidimensional representations to support questions like who is working with whom, where are there geographical concentrations in funding or research area, who are the most significant researchers in an area. We briefly show how this demonstration illustrates a number of substantial challenges for the Semantic Web. These include problems of referential integrity, tractable inference and interaction support. We review our approaches to these issues and discuss relevant related work.

1 Introduction

In this paper, we step through a demonstration of CS AKTive Space, a Semantic Web application lets users investigate the domain of UK University-based research in Computer Science. The application exploits a wide range of semantically heterogeneous and distributed content relating to Computer Science research in the UK. It provides services such as methods to explore research areas and institutions

[*] Corresponding author.
Email addresses: `nrs@ecs.soton.ac.uk` (Nigel R. Shadbolt), `nmg@ecs.soton.ac.uk` (Nicholas Gibbins), `hg@ecs.soton.ac.uk` (Hugh Glaser), `swh@ecs.soton.ac.uk` (Stephen Harris), `mc@ecs.soton.ac.uk` (m.c. schraefel).



○ research area/region ● region/research area

Radial: 200 miles

Map: uk-political

Research area

Researcher

Top ● 5 ○ 10 ○ 20 ○ unlimited

Order by ☑ Grant total □ RAE result

Computing Milieux
computers and society
computers and education
history of computing
the computer industry
Data
data encryption
data storage representations
coding and information theory
Computing Methodologies
document and text processing
pattern recognition
artificial intelligence
image processing and computer vision
computer graphics
simulation and modeling
symbolic and algebraic manipulation
general
Information Systems
information interfaces and presentation

NR Shadbolt
SJ Cox
P Johnson
PC Treleaven
ACW Finkelstein

Overview: NR Shadbolt

Name NR Shadbolt
Institution Intelligence, Agents and Multimedia, University of Southampton
Email nrs@ecs.soton.ac.uk
Tel +442380597682
Fax +442380592865
Research interests Fluid Dynamics
Aerodynamics

CoP Kieron O'Hara
Harith Alani
Terry Elliott
David De Roure
Nick Jennings
Nicholas Gibbins
Stephen Harris
Stuart Middleton
Sanghee Kim
Paul Lewis
Wendy Hall
David Millard

Fig. 1. CS AKTive Space UI featuring multicolumn queries and geographical visualizer and constraint controller

for researchers; it can show the geographic range and extent of where a topic is researched; it provides an estimation of "top" researchers in a topic and by geographic region, and it is able to calculate a researcher's Community of Practice. A longer description of CS AKTive Space is given in [1].

2 Interaction Design

CS AKTive Space presents a light-weight UI for exploring an ontology-modelled domain. By leveraging the semantic affordances of the domain as spatially presented visual queries, it disguises the complexity of the sources, services and queries running real-time beneath the interface. Hiding the complexity of the processes beneath a clear UI lets users explore both the relations and details of computer science research in the UK. If they wish, at any point, they can go "beneath the hood" to browse the sources and their provenance for any entity of interest presented in the UI. Thus, we provide multiple ways for users to engage the space to build up actionable knowledge. In order to contextualize the above interaction description, we present a brief walkthrough of the interaction pictured in Figure 1.

In the state pictured, the user has re-sorted the column views from the default arrangement of < Research Area | Region | Researcher > to < Region | Research Area | Researcher >. The Region column shows that a 200 mile reticule has been selected in the map view, and the reticule has been dragged over part of the map representing southern England. From the list of research areas that shows up in the Research Area column, artificial intelligence has been selected. With the constraints on Researcher set as Top 5, determined by Grant Total, the list of 5 people in AI with the greatest grant totals are represented. One of the researcher names has then been selected. The Detail View shows that that researcher is the current selection: it is populated with information about the researcher, including full name, contact information, URL and list of significant papers on the left two-thirds of the view. In the right remaining third, the list of the researcher's community of practice is generated automatically by the CoP service described in Section 4. Finally, at the bottom of the window the user can invoke the Armadillo service by the "Run Armadillo" service button. As described in Section 4, Armadillo will search for additional information about this researcher to supplement the information already present in the triplestore. Browsing the source data for the selected researcher will give the provenance of that data.

The final state of the UI shows both the context of discovery for the current selected information, as well as a view of the detailed information available about the researcher. The formal model underlying this interaction is described in [2].

3 Knowledge Acquisition and Maintenance

Even in a distributed environment such as the Semantic Web, the physical distribution of data is much less important than the semantic distribution of data brought about by the use of disparate ontologies for the same application domain. For the purposes of this application we use a single common ontology to express the data which drives the CS AKTive Space. We use this common ontology, the AKT Reference Ontology [3], to mediate and guide the integration of the different data sources. When expressed in terms of our ontology, the RDF data obtained from these sources is made publicly available through the hyphen.info web site (Hyphen being the name for the knowledge acquisition effort) and cached in an RDF triplestore [4] which provides the necessary query and inferential capabilities on which CS AKTive Space is built. The hyphen metadata gathered by these mediators consists of 430MB of RDF/XML files containing around 15 million RDF triples describing 800,000 instances of people, places, publications and other items of interest to the academic community.

Due to the wide variety of the data sources that we use, we have found it necessary to invest a degree of effort in developing individual mediators for each of our data sources that recast these sources in terms of our ontology. These mediators range from specialized database export scripts to XML transformation tools [5] that have been trained to extract the required content from semi-structured web pages. While the bulk translation of instance data by such a mediator is straightforward, our use of these mediators has shown that the mapping of existing structured and semi-structured data at the schema/ontology level is not a task that can be effectively automated in all cases; the investment of effort in building mediators for our common ontology is reflected in the consequent perceived value of the knowledge base to which they contribute.

The CS AKTive Space application requires that a range of content be available for use by the system. As it stands, some of this content already exists in suitable structured forms, while other content does not. We adopt a pragmatic attitude that reflects the fact that although the content that we are gathering is the prime mover that drives the interface, we should also be tolerant of inconsistencies in that content. We adopt a pragmatic approach in which we make the immediate best use of the available data sources, perhaps in an imperfect fashion, while anticipating that we will be able to make better use of them in future.

We employ both push and pull models of knowledge acquisition, where push and pull refer primarily to whether the publisher or consumer are responsible for translating the data into a form which is suitable for the consumer. We use the pull model predominantly for large, comparatively static data sources (for example, the list of countries and administrative regions given by ISO3166), and as an interim solution for high-value data sources that are of general interest to the community as a means to ’pump-prime’ the system with sufficient data to encourage other members of the community to participate.

The current development of knowledge services on the Semantic Web raises a number of issues which are not commonly encountered in existing knowledge based systems, and which pertain to the distribution of knowledge and the difficulty of obtaining agreement on a conceptualisation in a distributed environment when there is no ultimate authority. One such issue is that of coreference, which arises when more than one Uniform Resource Identifier is used to refer to a given resource, and which causes particular problems when statements from different knowledge bases are to be combined. We have three complementary approaches to this problem for CS AKTive Space. Firstly we support the simple social solution, where we allow the emerging knowledge base to be used as a gazetteer or name authority, so that new knowledge can be asserted with a common agreement on URIs. Secondly we have heuristic methods that conservatively coalesce appropriate entities [6] (using `owl:sameAs` assertions). Thirdly we have developed a coreference editor that builds on the second heuristic approach, but allows user intervention.

4 Key Services

The core of the CS AKTive Space system is a set of (mostly HTTP) services that collaborate to provide the knowledge capabilities required by the user interface. Currently, the services that comprise CS AKTive Space are manually composed a priori; we intend to migrate to a system in which the services are bound dynamically using service discovery techniques, to reduce brittleness and allow opportunistic service use. The current service portfolio is as follows:

3store — This application needs to be able to evaluate queries over a large volume of RDF data; to meet this requirement, we developed the scalable RDF(S) triplestore, 3store [4].

Geographic Visualizer — This service provides a graphical representation of the geospatial information within the ontology (the locations of institutions of interest) and permits the user to directly specify geographical constraints.

Armadillo — Armadillo [7] is a service for on-the-fly, user-determined, directed knowledge acquisition from web pages, which can be used to opportunistically expand the knowledge base. Pre-existing knowledge is used to inform natural language searches (over a variety of web sources) that extract further knowledge and assert it back into the triplestore.

Ontocopi — Ontocopi is a Community of Practice [8] analysis service that identifies the implicit communities that exist within a knowledge base. Ontocopi uses Ontological Network Analysis to discover connections between the objects that the ontology only implicitly represents. For example, the tool can discover that two people have similar patterns of interaction, work with similar people, go to the same conferences, or publish in the same journals.

5 Conclusions and Future Work

We have a strong proof of concept for an integrated Semantic Web application, but for CS AKTive Space to become a significant community resource, we need the content that we cannot get without participation. We need individuals and/or organizations to publish RDF content either directly against our ontology or else against an ontology we can translate. Participation cannot usually be enforced. Users have to see very strong benefits for the effort of publishing their content against our ontology.

We have started to see that happen within the UK CS community, and more recently, the eScience community. Interestingly, it is not simply the power of the back end tools that has provoked these enquiries and requests to apply CS AKTive Space to other domains; it has also been the user interaction. There is an immediate appeal to the fact that patterns and gestalts, particular and general content exploration can be so rapidly effected.

6 Acknowledgements

This work was supported by the Advanced Knowledge Technologies (AKT) Interdisciplinary Research Collaboration. AKT is sponsored by the UK Engineering and Physical Sciences Research Council under grant number GR/N15764/01 and comprises the Universities of Aberdeen, Edinburgh, Sheffield, Southampton and the Open University.

References

[1] m.c. schraefel, N. Gibbins, S. Harris, H. Glaser, N. Shadbolt, CS AKTive Space: Representing Computer Science in the Semantic Web, in: Thirteenth International World Wide Web Conference, 2004, pp. 384–392.

[2] N. Gibbins, S. Harris, m.c. schraefel, Applying mSpace interfaces to the Semantic Web, preprint: `http://eprints.ecs.soton.ac.uk/archive/00008639/` (2003).

[3] The AKT Reference Ontology, `http://www.aktors.org/publications/ontology/` (2002).

[4] S. Harris, N. Gibbins, 3store: Efficient bulk RDF storage, in: Proceedings of the 1st International Workshop on Practical and Scalable Semantic Systems (PSSS'03), 2003, pp. 1–20, `http://eprints.aktors.org/archive/00000273/`.

[5] T. Leonard, H. Glaser, Large scale acquisition and maintenance from the web without source access, in: Workshop 4, Knowledge Markup and Semantic Annotation, K-CAP 2001, 2001, pp. 97–101.

[6] H. Alani, S. Dasmahapatra, N. Gibbins, H. Glaser, S. Harris, Y. Kalfoglou, K. O'Hara, N. Shadbolt, Managing reference: Ensuring referential integrity of ontologies for the semantic web, in: Proceedings 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW'02), 2002, pp. 317–334.

[7] A. Dingli, F. Ciravegna, Y. Wilks, Automatic semantic annotation using unsupervised information extraction and integration, in: Proceedings of SemAnnot 2003 Workshop, 2003.

[8] H. Alani, S. Dasmahapatra, K. O'Hara, N. Shadbolt, Identifying communities of practice through ontology network analysis, IEEE Intelligent Systems 18(2) (2003) 18–25, `http://eprints.aktors.org/archive/00000172/`.