Monitoring Research Collaborations Using Semantic Web Technologies

Harith Alani Nicholas Gibbins Hugh Glaser Stephen Harris Nigel Shadbolt
{ha, nmg, hg, swh, nrs}@ecs.soton.ac.uk

Advanced Knowledge Technologies (AKT), School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK.

Abstract In the current research environment, funding agencies are increasingly required to demonstrate that the projects they fund represent value for money. When funds are disbursed in a speculative manner, in order to stimulate interdisciplinary collaboration, the determination of value for money relies on evidence that shows the generation of new collaborations. This paper summarises the work we carried out on behalf of the Engineering and Physical Sciences Research Council (EPSRC), in which we have implemented a set of applications to enable the research council to examine the existence and nature of collaborations between researchers. We have used Semantic Web technologies to construct a flexible application framework to provide multiple complementary visualisations of the data, while separating the issues of knowledge acquisition and curation from the more user-centric interface requirements.

1 Introduction

Organisations have the need and responsibility to review and analyse their activities. The need arises not only from internal review procedures, but also from external agencies (eg government) that are trying to ensure value for money. Current and forthcoming requirements from government are imposing increasing obligations in this respect on the Research Councils (RCs), with a particular focus on research outputs and citations.

For example, one of the questions that arises is the extent to which different groups, people and programmes collaborate with each other. This is a very complex issue, as it is not even necessarily clear what it means to collaborate, and certainly no general agreement of what would be evidence of collaboration. We should expect that good analysis would not only provide qualitative and quantitative data on collaboration, but also allow users to think about and explore the nature of collaboration itself. Such analysis is challenging, in particular when it requires analysis of data from a wide variety of sources, many of which are outside the direct control of the organisation.

We can roughly define two main stages in this work. The first stage is to integrate the distributed sources of data and store it in a format suitable for further use and analysis.

Integration of databases raises several well known challenges, such as resolving the conceptual differences between database schemas, identifying data duplications and inconsistencies, etc [4]. Ontologies have been widely proposed as a major role player in information integration [7][14][15]. They provide the mechanisms to establish a common understanding of heterogenous data domains and help to bridge multiple data source schemas. In the case of RCs, the ontology would have concepts such as person, funding

agency, grant value, etc. Data could then be gathered against those classes, and represented in a suitable form, such as RDF. It can then be kept in a suitable Knowledge-Base (KB) (3Store in our case, [9]), from which it can be queried in various fashions by other tools and applications.

The second stage of this work starts once the KB is set and populated with all the required data. This stage is concerned with building the suite of tools and applications needed to provide the type of data management and analysis of research collaborations required by the RCs. This involves the implementation of services for browsing the collecting data, and visualising research activities in interactive ways.

The following sections discuss the two stages described above. Section 2 describes the data gathering process. The architecture of the system is explained in section 3. Section 4 describes all the tools and applications we built for browsing, managing, and analysing the data. A discussion of the main issues and challenges that we came across during this work is given in section 5. Finally, section 6 concludes this work, and any major work to be done in the near future is highlighted in section 7.

2 Data Gathering

In the United Kingdom, there are a number of agencies and stakeholders who contribute towards the funding of research. The key initial activity in constructing a system which can provide an overview of this research is the gathering of appropriate data from the relevant participants. A production system that attempts to provide a full overview of the sector would need to embrace all required sources, including all RCs, and possibly publication data, academic staff data, and other funding agencies.

The sources from which data is gathered are heterogeneous, as would be expected from a group of organisations with distinct requirements and objectives, at least in terms of their research programmes. The integration of this data in a suitable manner for common browsers and visualisers requires that the data be mediated and cast into a common form. We use an ontology as the mediating construct, such that each of the heterogeneous sources is translated into the ontology.

The first requirement is to define the ontology. For this study we used an existing ontology, which was constructed in the AKT Project[1], and defined using OWL. The AKT ontology[2] represents general information about the academic research environment.

For the purposes of this study, we took data on projects and grants from RCs in three domains: engineering and physical sciences (EPSRC)[3], biotechnology and biological sciences (BBSRC)[4], and medicine (MRC)[5]. These RCs were chosen because their funding activities overlap in an area known as the *Life Sciences Interface*, which supports interdisciplinary research between engineering and physical sciences, and the life sciences. In addition to this, we provided a small amount of publication data for selected individuals who are active within the life sciences interface for demonstration purposes.

[1] Advanced Knowledge Technologies http://www.aktors.org
[2] http://www.aktors.org/ontology/
[3] http://www.epsrc.ac.uk/
[4] http://www.bbsrc.ac.uk/
[5] http://www.mrc.ac.uk/

We have adopted a hybrid approach [15] in our use of the AKT ontology, in which the ontology is used as a shared vocabulary to represent the data from each of the three RCs, with some local extensions to represent issues of local interest. These issues involved the representation of *research theme*. Each RC has its own notions of what constitute the different discrete areas of research which it funds. Such an approach allows us to easily integrate additional sources without any modifications to the rest of the system [15].

It is possible to gather data from institutions without their explicit cooperation, even when they have no intention to publish it in a machine-processable form. This is usually done by "screen-scraping" or extracting information from structured or semi-structured web pages, or even using Optical Character Recognition in extreme cases. In practice, such methods are far from satisfactory. They suffer from problems such as high error rates, and high maintenance (especially if web pages change), and are only sensible for initial experiments or for very valuable data that cannot be harvested any other way.

Far preferable is if the institution cooperates with the harvesting activity by publishing the data itself, either as web pages or other machine-readable form, against a well-specified structure. In our case, the EPSRC was able to supply us with the appropriate data from its own databases. The data was supplied in the form of tables (formatted as CSV files) which resulted from an agreed database query. We were then able to process the data to the form required for our activities (RDF, expressed in the AKT ontology) using simple scripts. EPSRC were also able to provide us with largely similar data from MRC, which was processed by using equivalent scripts. Some data from BBSRC was provided in the last minute. It is pleasing to note that the system was such that this data was incorporated within a few hours.

Resolving duplications is always a major task when integrating data from multiple sources [6][8]. We applied a set of heuristic techniques [1] for identifying duplicate entities and then consolidating them through the use of *owl:sameAs* assertions. By keeping the equality between entities as explicit assertions, rather than by making it implicit by rewriting gathered information to use canonical URIs for objects, we provide a means to roll back duplicate resolutions. In addition to that, we have also developed an editor which allows a user to vet the potential duplicates in a semi-automatic manner. Note that quality is very important in our context when it comes to resolving duplicates because any errors in doing so will almost definitely yield incorrect analysis results.

All this data is then stored in a 3Store [9] KB, ready for further action.

2.1 Statistics

In total, the information gathered from the RCs consists of some 3.1 million distinct statements (99.93% about instances), which when expressed in terms of the ontology and serialised as RDF/XML take up over 250Mb. The information is heavily weighted towards EPSRC-funded grants and postgraduate awards (and their associated investigators), which comprise a big part of the total data, the remainder being evenly divided between MRC and BBSRC. The data, both raw CSV and processed RDF, are available from triplestore.aktors.org/demo/EPSRC/data/.

3 Conceptual Architecture

Figure 1 gives a general summary of the architecture of our proof-of-concept system. The data sources (EPSRC, BBSRC and MRC) are gathered by dedicated programs that take the native data in its raw form from traditional relational databases, and express it in terms of the common ontology.

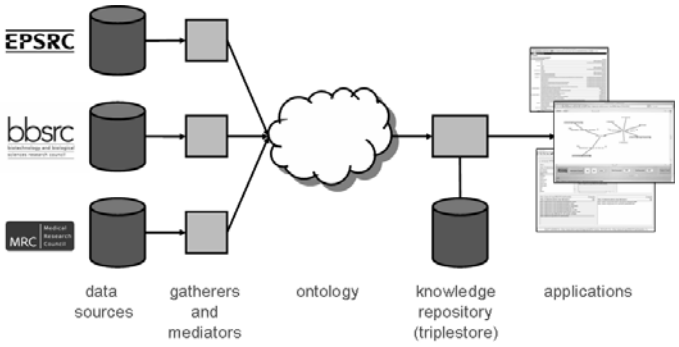

Fig. 1. Conceptual architecture.

The data describing the areas covered by this system are stored in the 3store KB. Data access is provided through a web service interface to the RDQL [10] query language. Applications can make HTTP requests to the server which returns query results in an XML format.

In addition to query processing, the triplestore also performs simple inference over its data, according to the formal semantics for the RDF and RDF Schema languages.

4 Data Presentation and Analysis

We identified a range of styles that would be interesting to explore and present for the study, and have implemented different points from the spectrum. Firstly, there is simply the ability to browse the data in its raw RDF format, as well as in a rendered fashion. Secondly, we provided two visualisation tools. One shows concepts from the system (people, grants, publications, etc), and the relationships between them, and the other presents a digest of total activity between funding sources, or activities by year.

It should be noted that, particularly in the case of the visualisations, these are intended to indicate the sort of tools that can be provided to explore the data. They represent our attempts to deliver interesting utilities as a result of a short study.

4.1 Browsers

Rendered Browsing Figure 2 shows a screenshot of the tool that allows the user to explore the data, rendered in HTML (tool located at http://triplestore.aktors.org/browse/epsrc/, browse for “JD Hirst”).

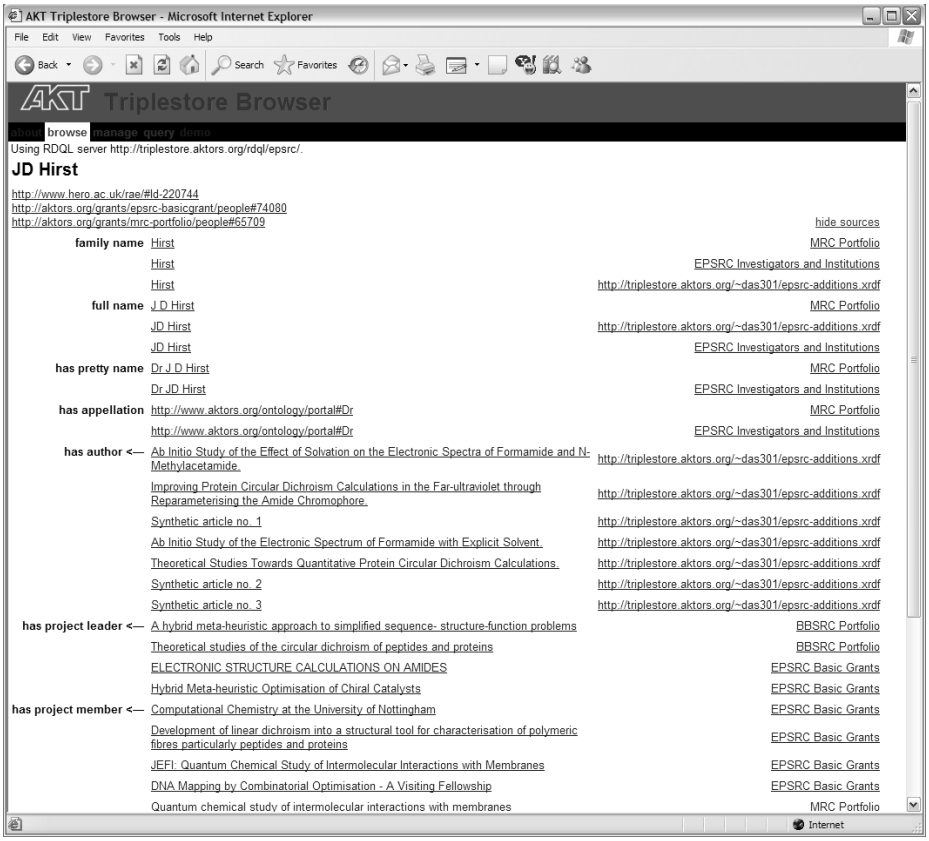

Fig. 2. Browser of HTML rendered data.

This is the data for Dr Hirst from the University of Nottingham, which we have chosen as the subject for our examples. This screenshot shows what it is like to browse the data, which is kept in a triplestore. Although not primarily intended for standard users, in this case we can identify some important issues.

The first is that this page offers a joint view of the data from the EPSRC, BBSRC, MRC, and also from the Research Assessment Exercise (RAE) submissions, which had already been gathered before this study. Looking at the full name data, note that the MRC knows this individual as “J D Hirst”, whereas the EPSRC knows the individual as “JD Hirst”. This is typical of the more simple variations that are seen.

Moving down the page, the sort of data we would expect for this exercise is then shown. Some publications are listed. As stated above, in this study we did not gather publication data; however, to demonstrate how such data would look, we have found publications for this individual, as well as manufacturing some synthetic articles.

Beneath this data are the basic details of projects from the RCs. The overlap on funding is of interest; we have chosen to leave the projects that are listed by more than

one council as separate projects. These projects could have been identified as the same, but the decision is one to be made in the light of the application requirements.

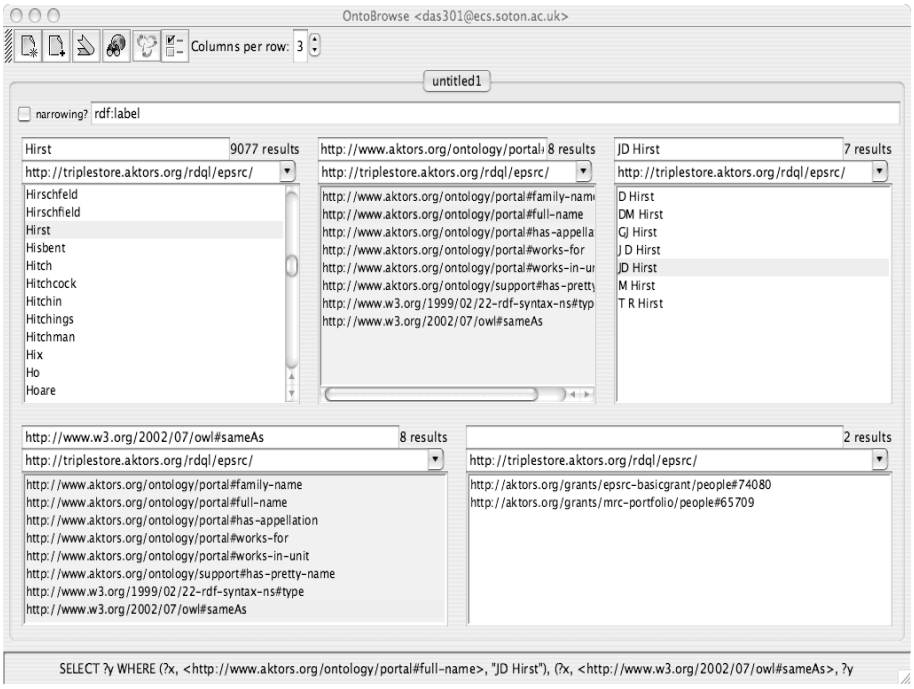

Fig. 3. Browser of raw data in triplestore.

Raw Browsing Figure 3 shows a prototype that gives a view that also exposes some of the more detailed workings of the system. This browsing tool is intended for the knowledge curator to get a detailed view of the data in its raw format in the triplestore.

The user has selected “Hirst” on the left, which has caused a column to appear showing the categories of knowledge the system has on people with that name. Selecting the “full-name” entry then displays the full names of all the Hirsts in the next column. Picking “JD Hirst” allows the user to find out that the data has been gathered from two sources, as the “sameAs” indicates this, and clicking on this shows the raw identifiers for the sources. The user could carry on clicking to find out details of the sources, institutions and so on. This style of interface is related to the mSpace user interfaces described in [12] and used in the application described in [13].

These interfaces are presented as sketches of different styles of low-level browsers, which gives the RCs a detailed view of their data once combined in one store.

4.2 Monitoring Collaborations

We built a tool to visualise the interactions between researchers, grants, research programs, etc. This visualisation tool (named EpsrcTGViz) runs as a Java applet from any browser. It makes use of a modified version of the TouchGraph[6] library, and it connects to a personalised ONTOCOPI web service.

ONTOCOPI (ONTOlogy-based Community Of Practice Identifier [2]) is a tool that finds sets of similar instances to a selected instance in a KB. If an ontology (i.e. both the classification structure and the KB of instantiations) represents the objects and relations in a domain, then connections between the objects can be analysed. The aim of ONTOCOPI is to extract patterns of relations that might help define a Community Of Practice (COP). COPs are informal groups of individuals interested in a particular job, procedure or work domain [16].

In the context of this work, ONTOCOPI is used to retrieve COPs of individuals or other type of object, and return the results to EpsrcTGViz. For example, the COP of the Life Science Interface research programme would include a number of individuals working on such grants, other related projects, institutions active in this research area, etc.

We can study the evolution of a scientist's collaborations by retrieving several COP sets for various dates, and monitor the rate of change of those collaborations. In other words, we can see when the scientist seises to interact with others, and when new interactions are born. We can also find out if, and how, interactions between scientists continued once a specific grant or a project has ended.

The idea is that it should be possible to gain a sense of how research and interactions have changed over time, while keeping an eye on some level of detail. We look here at Hirst's interactions, but it would be possible to focus on other things, such as a project or a programme. Using EpsrcTGViz, we can ask for the COPs of Hirst for the years from 1998 until 2008. This is achieved by clicking the Multiple Graph button, which retrieves the data from ONTOCOPI and constructs a set of graphs that represent them, one graph per selected year. The user can then browse those graphs by selecting the year of interest, or simply move backwards and forwards through the years to view how the graphs are evolving with time. Any change in the graph is displayed incrementally, dynamically, and slowly enough to help the user perceive any transformations.

The graphs show no information for Hirst before 2001, except for a couple of papers he published with Besley in 1999. It may be tempting to assume certain things, but it is always important to look at such data in relation to the sources. It may be that the sources did not go back so far. Also, with respect to publications, if there were entries for 2001, it might be sensible to represent them in earlier years, on the basis that collaborative work takes time, and publication can be very slow. As mentioned earlier, we did not collect any publication information for this system. However, we have created some publications as illustration.

When we reach the year 2001 (figure 4), significant funding activities start to show. Hirst is now funded on three projects from three different committees. Two are BBSRC, and one is EPSRC. Logos of RCs are shown as small icons to the top right corner of the nodes that represent research grants. Each grant node is linked to the nodes of people

[6] http://www.touchgraph.com

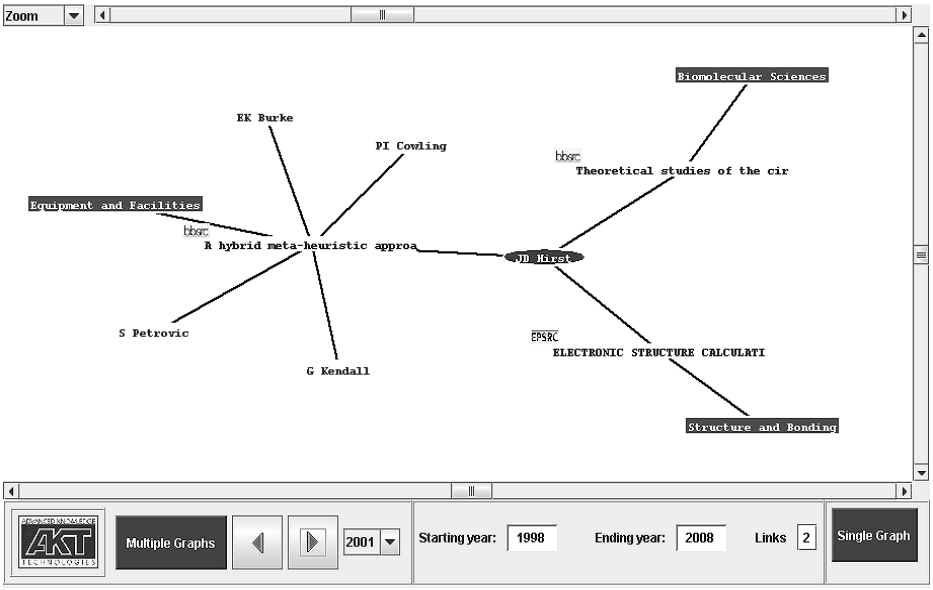

Fig. 4. Research activities of Hirst in 2001.

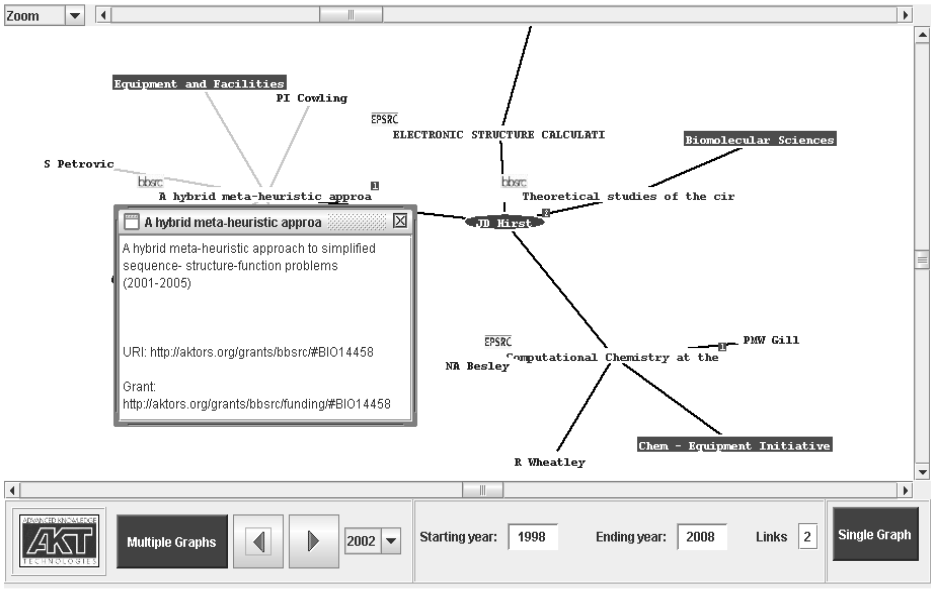

Fig. 5. Research activities of Hirst in 2002.

who have been identified as investigators in the RC's data, as well as to the research programme of the grant, which is displayed with dark background. As with the previous

year, 2002 has full RC data (figure 5), and we can see that Hirst has gained another EPSRC project, and in fact one of the co-investigators is Beasley.

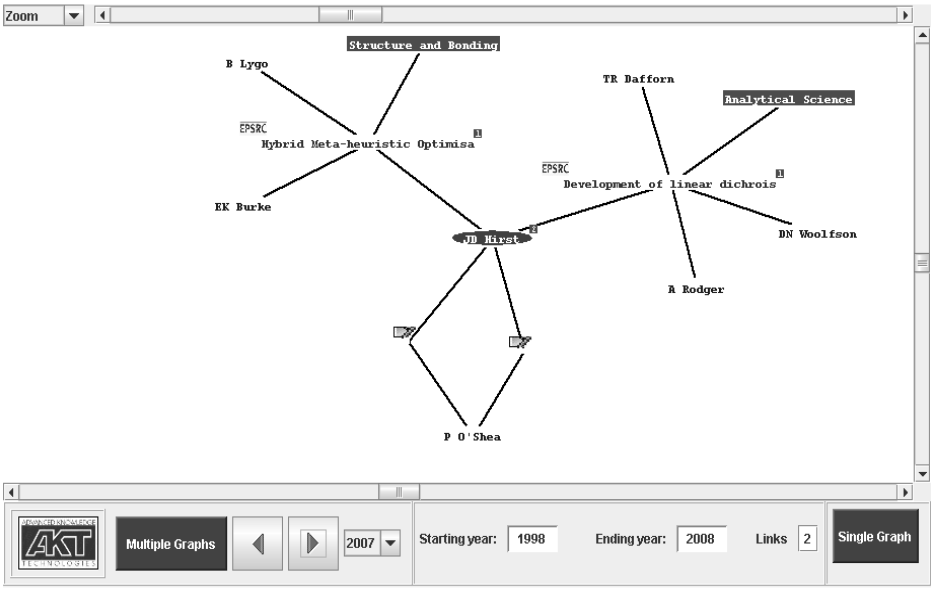

Fig. 6. Projection of Hirst's research activities in 2007.

Simply viewing in this way is very interesting, but clearly a user needs to be able to explore in more detail when they find things of interest. Figure 5 shows the pop-up window that appears when the mouse hovers over the project name. Any other details could be shown, and it is possible to right click any node in the graph and select to open the relevant page in the rendered browser described in section 4.1 for full details.

Because project data has fixed durations, in some way it is possible to look into the future. In 2007, some of the earlier awards begin to end (figure 6). If we have data on publications, then one would expect to see some paper writing collaborations emerging between Hirst and the other scientists with whom he shared some grants the years before (two publications are shown in the graph as illustration).

4.3 Summarised Information

Finally we explore a representation that abstracts away from the details of individuals, projects and publications. By counting numbers of projects or level of funding or other data, interactions between groups can be explored. In the next three figures we show such "Heat Charts".

Figure 7 looks at collaborations between different parts of the funding system, as it has changed over time. The brighter the colour, the more activity there is or was.

The metric for determining activity is the number of people who hold projects in the two areas in question, the value for each $(area1, area2)$ combination may be determined by the number of unique solutions to this query:

```
SELECT ?person        WHERE (?proj1 akt:has-project-leader ?person)
                            (?proj1 akt:has-research-interest area1)
                            (?proj2 akt:has-project-leader ?person)
                            (?proj2 akt:has-research-interest area2)
```

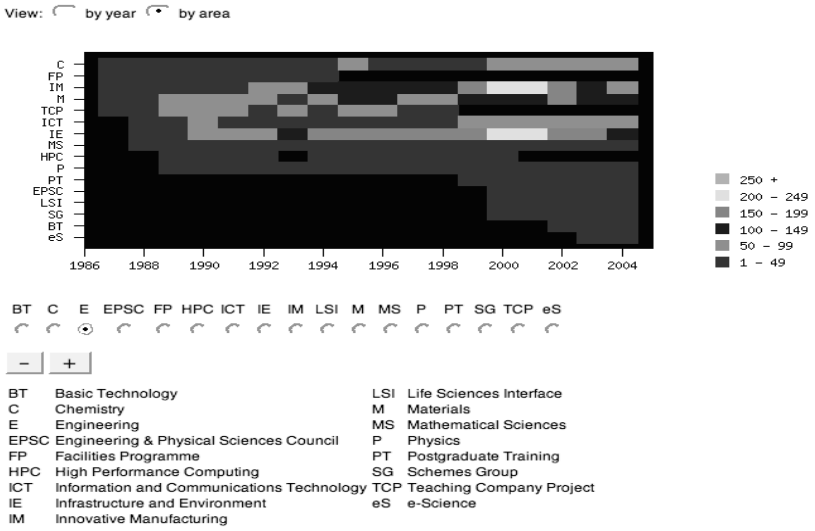

Fig. 7. Collaboration over time between Engineering and other disciplines.

The area we have focused on is "E" (Engineering), and as expected we see for example that Engineering makes great use of Infrastructure and Environment, and works closely over the years with Innovative Manufacturing, and sometimes with Materials.

Additionally, it is possible to move to a full view of the intersection of all the programmes by viewing the data on an area-area chart. Figure 8 shows the activities for 2004. The legends on the right in each figure shows the size of collaborations that each colour represents.

This tool is intended to give a flavour of what such a knowledge system might provide, once the data is stored in a triplestore and annotated according to an ontology.

The RDF characterisation, in terms of an agreed ontology was advantageous to the development in that it provided a relatively simple common data format for integration of the data from the disparate research funding councils, and the inferrential capabilities of the KB were exploited to allow general questions to be asked that would be answered using specific structures of the underlying data, without requiring the queries to be rewritten for each data source, or the data to be hand translated into a single common form.

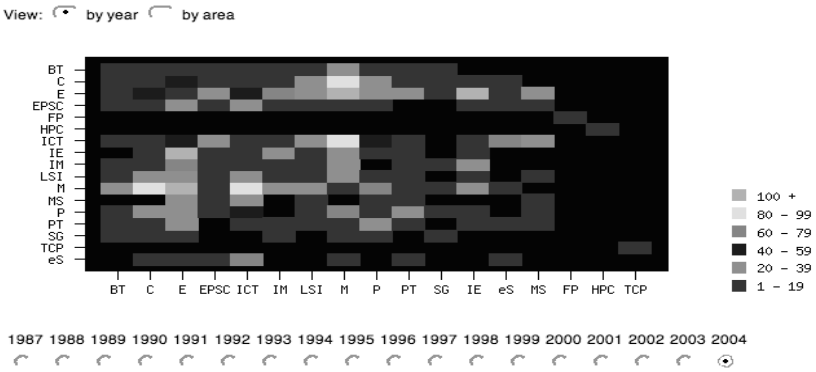

Fig. 8. Collaborations between all disciplines in 2004.

5 Challenges and Recommendations

This study has shown that an ontology-mediated KB system can be used to gather and process data from RCs at low cost. In this section, we discuss some of the lessons learned, and describe the recommendations we made to the RCs based on this study.

5.1 Data Publishing

We realised at the start of this study the importance of requiring as little effort as possible from the data providers (ie RCs). We believe that this was a major reason why this effort succeeded. Nevertheless, data providers could publish in a more convenient form (e.g. RDF), but the model should be that they simply publish the data they wish, and processors are able to collect the data over the Internet.

Output in RDF For this study, we received the output of database queries from the RCs and converted them to RDF with relatively simple scripts. Thus the publishing activity for the RCs is to take and maintain these small programs, or consider how they might perform the same function in other ways. The cost of doing this is very low.

Shared Ontology We used an existing ontology (the AKT ontology) for this study. Although it is possible that each RC could publish against a different ontology which we can map to each other, we recommend that some time should be spent considering an appropriate ontology that uses widely agreed concepts, such as those provided by the Dublin Core[7] metadata initiative.

Open Access The primary objective of this work was to build a management tool for RC use, and to provide information to stakeholders such as government. Once the data is published in a semantic web language, it can be put into many other usages and scenarios. For example, there is renewed interest in support for an “electronic CV” for researchers, which may be seen in part as an ontologically-informed view of RC information. The RCs may consider that the provision and publication of their data is an important part of their mission. Indeed, EPSRC and the RCs have the chance to lead the

[7] www.dublincore.org

knowledge-enabled society by publishing their own RDF for the Semantic Web. Doing so will encourage others to build additional tools and facilities.

5.2 Data Acquisition

We now move from the RCs publishing data, to how that original data is acquired. Data is obviously of vita importance to this work. If the RCs are to use our system to draw any reliable conclusions about research collaborations, then the underlying data must be highly complete and accurate.

Collecting Data in RDF The model for the RCs is that they should curate the data they need, publish it in RDF, along with mappings of their identifiers to those of other data providers. The same model should apply, in the longer term, to the institutions and any other organisations they deal with. Thus for example, instead of the RCs asking for and then curating the data for individuals from each institution, the institution should be encouraged and even required to provide the data in suitable RDF in public and private places (according to the confidentiality of the data).

Data on Publications During this study we found that it is of great importance to have access to data on publications when tracking research collaborations. RCs do not currently hold this type of data. Institutions are now setting up publication archives, and will soon be effectively mandated to do so. It is therefore timely that RCs should require all references to publications in documents submitted to include the URL of the document in the institutional archive. This will also provide a strong spur to initiatives that are currently underway, in particular the Open Archive Initiative (OAI[8]). Indeed, the House of Commons Select Committee on Science and Technology says [11]:

"*This Report recommends that all UK higher education institutions establish institutional repositories on which their published output can be stored and from which it can be read, free of charge, online.*"

In the short term, however, it is sensible for RCs to gather publications data by itself, where there are areas of particular interest. This will also enable the analysis tools to include publication data at an early stage.

5.3 Referential Integrity

A key issue that we have encountered in the course of this work is that of referential integrity, which is the problem of identifying that a pair of entities in two databases are actually referring to the same object.

Multi-Sourced Data We have taken information on projects and researchers from a number of disparate sources, as described earlier. In many cases, these sources discuss the same individuals and projects, but they each coin a different identifier to refer to these individuals and projects. For example, the subject of our earlier system description, JD Hirst, is identified by the key 74080 by EPSRC, the key 65709 by the MRC, and 220744 by the Higher Education Statistics Agency (HESA)[9] (as used in the RAE submissions). A substantial part of our work in adapting RC databases for our use was

[8] www.openarchives.org

[9] http://www.hesa.co.uk/

determining whether an identifier used by one source was coreferent with an identifier used by another source.

Imperfect Techniques While there are heuristic techniques that can be applied to the problem of referential integrity (eg [6][5][1]), these are frequently defeasible and often require a high degree of adaptation to a particular application domain. Moreover, these techniques require sufficient high-quality data to be able to judge whether individuals are coreferent; insufficient or inconsistent data (variant name forms, for example) increase the probability of incorrect judgements [3]. When applying these techniques, the cost of both false positives (incorrectly coalescing information on two distinct individuals) and false negatives (failing to identify two individual as coreferent) must be borne in mind. In other communities, such as the library and information science community, referential integrity is managed through social means, by using name authority files, gazetteers which list the correct form of authors' names. For a system of the sort we are developing, trust is of paramount importance, and so every step should be taken to avoid false positives or negatives of any kind.

Unique Identifier Authority The current situation in which each RC generates its own set of identifiers for referring to people, institutions, research programmes, etc., presents a significant legacy data issue, and the most appropriate way forward must take account of this. In addition, it is important to make best use of existing sources in order to minimise the duplication of effort in the alignment of these identifiers.

Our recommendation is that each RC continues to generate their own unique identifiers, but that they should also publish a mapping from their person and institution identifiers to those used by HESA, where such exist. HESA has good coverage of research staff and organisations across UK Higher Education (HE), which suggests that it is well placed to assume the role of identifier authority, but it has minimal coverage of non-HE entities. However, HESA has a high-quality dataset with properties that make it attractive for long-term use (HESA people identifiers are persistent, and do not change when personnel move between institutions). In the event that no coreferent identifier exists in HESA data, the RCs may publish pairwise mappings between their identifiers and those used by other councils.

5.4 Maintenance

Another important issue is to do with the "liveness" of the data. For this study, we chose to use data from a snapshot at a particular time, because the study did not need to keep it up to date. A production system would need to use the latest data when required. This can easily be achieved by ensuring that the RDF from the RCs is published at the required intervals (e.g. nightly), so that processors can ensure they are in synchronisation.

This approach can be practical and effective when using low cost automatic tools to format the data in RDF and publish it, with minimum or no human intervention. It is however of vital importance to couple this approach with good quality techniques and procedures for tackling the referential integrity problem discussed above.

5.5 Privacy

The RCs and others will need to have considerable regard for confidentiality and the requirements of the various Acts. We consider that the model whereby RCs choose what

to publish in RDF, which will be essentially the same data as provided in systems such as the EPSRC's Grants on the Web, gives effective control of this. By publishing the data through a single point, careful control can be maintained of what is published.

It may be that an individual RC will wish to analyse its activity using data that it does not make public, possibly using private RDF data from other sources.

6 Conclusions

Laying the proper semantic foundation for the data was the most important phase of our work. By doing so, we were able to implement a set of tools to help browsing and analysing this data. Previously, there was no joint access to this data which comes from multiple sources held in separate research agencies. There was a clear need for bridging these distributed data sources as a first step towards providing more comprehensive knowledge management tools.

We provided two tools for simple data browsing, and two other tools for analytical visualisations. These tools are meant to be simple demonstrators of what type of functionality we can add to this repository. Furthermore, they helped to gather more detailed user requirements from the RCs, some of which could be the focus of further work.

The criteria for importing data from traditional repositories into 3Store was of low cost and highly reusable. This can assure quick updates of the data whenever required. The RCs were quite pleased with the idea of keeping their traditional databases, while being able to fuse them together externally with minimum effort on their part.

The graph visualiser we developed offered the RCs a quick way to view if, and how, their grants are generating collaborations between the various scientists. This helps to make better decisions on when, and to whom, further grants should be awarded. Similarly, the heat charts provided an easy way to monitor collaborations between entire disciplines. Such a service is of great importance to RCs as it helps to quickly detect disconnections between research areas, which may feed into their future grant calls.

We believe this shows that an approach of this sort, on a wider scale, has the potential to provide EPSRC, other RCs and other stakeholders with the sort of information systems to deliver what they are now being expected to provide.

We made a set of recommendations to the RCs involved in this work to guide them through the process of building knowledge management systems. These general recommendations apply to any organisation with similar aims and requirements, and not strictly limited to any specific type of data, infrastructure, or application needs.

7 Future Work

The next phase of this study is already underway, focusing on two main objectives. The first objective is to collect data on publications to enable a finer grained analysis of collaborations. The second objective is to produce a set of data charts for the RCs, showing the rate of change of total collaborations between pairs of research programmes.

In the near future we might extend the system to display charts, similar to those presented in [2] to track changes in specific COPs over time. This type of chart can show the change in the level of n-order collaborations between researchers in relation to the duration of certain research programmes (e.g. Life Sciences Interface).

Acknowledgements. This work is supported under the Advanced Knowledge Technologies (AKT) Interdisciplinary Research Collaboration (IRC), which is sponsored by the UK Engineering and Physical Sciences Research Council under grant number GR/N15764/01. The AKT IRC comprises the Universities of Aberdeen, Edinburgh, Sheffield, Southampton and the Open University. Thanks to Daniel Smith for all his work. We are also grateful to EPSRC, and specifically to Elizabeth Hylton, Mark Hylton, and Gavin Salisbury for their time and support.

References

1. Alani H., Dasmahapatra S., Gibbins N., Glaser H., Harris S., Kalfoglou Y., O'Hara K., and Shadbolt N. *Managing Reference: Ensuring Referential Integrity of Ontologies for the Semantic Web*. Proc. 13th Int. Conf. on Knowledge Engineering and Knowledge Management, (EKAW02), Siguenza, Spain, LNAI, 317-334, 2002.
2. Alani H., Dasmahapatra S., O'Hara K., and Shadbolt, N. *ONTOCOPI - Using Ontology-Based Network Analysis to Identify Communities of Practice*. IEEE Intelligent Systems, 18(2), 18-25, 2003.
3. Alani H., Kim S., Millard D.E., Weal M.J., Hall W., Lewis P.H., and Shadbolt N. *Web based Knowledge Extraction and Consolidation for Automatic Ontology Instantiation*. Knowledge Capture (K-Cap'03), Workshop on Knowledge Markup and Semantic Annotation, Sanibel Island, FL, USA, 2003.
4. Batini C., Lenzerini M., and Navathe S.B. *A Comparative Analysis of Methodologies for Database Schema Integration*. ACM Computing Surveys, 18(4), 323-364, 1986.
5. Cohen W., Ravikumar P., and Fienberg S. *Adaptive Name Matching in Information Integration*. IEEE Intelligent Systems, Sept/Oct, 2-9, 2003.
6. Dey D., Sarkar S., and De P. *A Distance-Based Approach to Entity Reconciliation in Heterogeneous Databases*. IEEE Trans. on Knowledge And Data Eng., 14(3), 567-582, 2002.
7. Gruber T. R. *The Role of Common Ontology in Achieving Sharable, Reusabe Knowledge Bases*. Proc. 2nd Int. Conf. on Principles of Knowledge Representation and Reasoning, Cambridge, MA, Morgan Kaufmann, 1991.
8. Guha R. *Object Co-Identification on the Semantic Web* Proc. 13th World Wide Web Conf., New York, USA, 2004.
9. Harris S. and Gibbins N. *3Store: Efficient bulk RDF storage*. Proc. 1st Int. Workshop on Practical and Scalable Semantic Systems (PSSS'03), Sanibel Island, FL, USA, 1-20, 2003.
10. Hewlett-Packard Labs *RDQL - RDF Data Query Language*. http://www.hpl.hp.com/semweb/rdql.htm, 2003.
11. House of Commons Tenth Report, HC 399, July 2004. http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/39902.htm
12. schraefel m. c., Karam M. and Zhao S. *mSpace: interaction design for user-determined, adaptable domain exploration in hypermedia*. Proc. of AH 2003: Workshop on Adaptive Hypermedia and Adaptive Web Based Systems, 217-235, Nottingham, UK, 2003.
13. Shadbolt. N. R., Gibbins. N., Glaser. H., Harris. S., et. al. *CSAKTiveSpace or How we Learned to Stop Worrying and Love the Semantic Web*. IEEE Intelligent Systems, 2004.
14. Uschold M. and Gruninger M. *Ontologies: orinciples, methods and applications*. The Knowledge Engineering Review, 11(2), 93-136, 1996.
15. Wache H., Vögele T., Visser U., Stuckenschmidt H., et al. *Ontology-based Integration of Information - A Survey of Existing Approaches.* Proc. IJCAI-01 Workshop on Ontologies and Information Sharing, Seattle, WA, pp 108-177, 2001.
16. Wenger E., McDermott R., and Snyder W. *Cultivating Communities of Practice*. Harvard Business School Press, Cambridge, Mass, 2002