Evaluation of ResearchRevealed project

Author and contact details:

Nicky Ferguson	Managing Director, Clax Ltd	Email:	nicky at therightplace.net
10 Springhill		Text:	07770-262886
Stroud		Mobile:	07770-262886
GL5 1TN	Web: http://www.therightplace.net/nf-index.html		

Background

ResearchRevealed is a JISC-funded project based at the University of Bristol and running from August 2009 to May 2011. The general aims of the project were to facilitate customisable access to integrated research repositories and enable views of the real impact of research in its wider context (for example in industry, in society, and across the UK research community). Also to explore the use of semantic web technology to facilitate central repository information integration and to extend it with external sources of data such as data from research and funding councils.

I was asked to conduct a summative evaluation of ResearchRevealed towards the end of the project. I conducted interviews, face to face and by email, and also created an online questionnaire for this purpose. In addition I attended several team and steering group meetings as an observer and followed these up with conversations with some of those attending. I have also discussed issues with Phil Cross who is conducting the technical evaluation and whose report should be read alongside this one.

1 Methodology

I read the extensive internal wiki together with the project blog and followed email correspondence to inform myself of the project background, current issues and progress against objectives. I discussed the objectives, both of the project and of the evaluation with those involved, attending several team and steering group meetings as an observer and following those up with individual meetings with staff to clarify issues and fill in my understanding of the project. Together with Nikki Rogers, the project manager, and Phil Cross, the technical evaluator, we identified areas that the evaluation should cover.

I devised a questionnaire which was administered on the Web for those who could not make face to face interviews. I conducted interviews, both face to face and by phone and Skype – the interviews closely followed the structure of the questionnaire so that my notes and the questionnaire results were easily comparable. However, the face to face and phone interviews did not require or force the respondents to tick boxes or make selections but rather allowed them comment on and discuss the questions. For this reason I present the evaluation as a qualitative exercise and do not attempt a statistical or numeric presentation of the results. The respondents included the project team, members of the steering group, key contacts within the University of Bristol and key contacts outside of the university, including collaborators and workshop attendees.

2 Results

2.1 Software – process and product

Comments on the software process and product were all positive. Particular items highlighted included the small size of the team and the hands-on nature of the technical management. There was a good balance between clear division of tasks and discussion points involving the whole technical team.

The way the developers and designer worked together was mentioned both as a very positive outcome and also, several times, as a challenge with comments suggesting that having designers and developers located together would be desirable. It was generally recognised that designers and developers may have different priorities or expectations, especially in a research project, but although the challenges were discussed there was also a general recognition that the tension had been creative:

> *In general the process was easy, with a small team … it works best when co-located and looking over each other's shoulder … the challenge was working with the designer and maybe there should be a different approach to a research project than a service or product.*
>
> *the small development team made the process easier to handle. I think it's best if the designers & developers working on the project are seated close together.*
>
> *Although there were some difficulties* [working with the designer]*, it looks very good and that is very important*

I reproduce the designer's comments on this aspect in full because they may be useful to this and other technical development teams in future work:

> *The MVC approach allowed me as the visual designer to work directly on the code in the views, while the back-end developers did most of the heavy lifting in the model and controllers. Having direct access to the views during the development process as a designer is important because it makes it easier to maintain higher quality values in the front-end interface. In a project of this nature, it is necessary to make frequent changes to the interface. Having to make these changes by proxy, via the back-end developers, creates a bottle-neck and is highly inefficient. So this was a good solution and the quality of the interface has benefited. The only area where this broke down was in the more dynamic areas of the site where very involved Javascript was required. In these places, the MVC structure was breached and some of the "heavy lifting" logic bleeds into the views, or some of the view markup is hijacked by a Javascript file and mixed with logic. This makes it extremely difficult for me as a designer to build and adjust the user interface in these areas. It would be worth looking at ways to improve how this could be better structured in future.*

The project intended to have rapid, iterative software development, based on interactions with the users. Feedback from the questionnaire together with project updates and documentation suggests that this was successfully achieved.

In terms of the end product, this was well received both by potential users and collaborators:

> *Given it's a research project, the quality of the product is high*
>
> *As a non-techie, it looks fantastic, … Development seems to have been very thorough, well planned and managed*
>
> *The user interfaces look good and appear to work well.*

The software was regarded as "battle hardened" by the developers as a result of successfully dealing with the scalability issues of dealing with such large datasets. The developers are already using elements of the software on other JISC and university-funded projects and services. The faceted browser was particularly mentioned in this regard.

2.2 Data

It was recognised by all involved that problems with getting access to high quality data were a feature of the project. As well as getting access to data on research applications and data from the research repositories, the team had been hoping for access to data on research grants, holders etc direct from university's finance system. This was not forthcoming. They were also hoping for access to well-structured data from research councils – such data of sufficient quality was also not available in practice and the team were forced to obtain "screen-scraped" data from research council web sites and online databases. Interestingly, other collaborators reported similar or worse problems, for example at Oxford where it is reported that data is often "balkanised" within the college system. Although the team was disappointed by the measures they had to take to obtain data and in some cases by the quality of the data that they were dealing with, this did not affect the proof of concept

achieved by the system. Indeed, in this particular area (as in other areas of the project), although the team sometimes regarded their achievement as quite modest and

> *Not much really new, just putting things together and presenting it*

It was regarded as highly innovative and very important by the potential users/consumers of that data who rated it highly. It seems that just bringing together data which is already freely accessible and

> *Filling in the gaps, allowing it to speak for itself*

was a major and illuminating achievement – proving the potential for such data to make a difference to the working lives of its users.

> *Initial discussions with Research and Enterprise Development (RED) in the University indicated that there was limited scope in existing systems to draw together, integrate and gain different views of data relating to research, research outputs, research income and researchers themselves. Research Revealed has demonstrated how this could happen far more systematically and the potential benefits in research intelligence, benchmarking, promotion and management/enhancement of research performance. For a University that describes itself as 'research-intensive', access to and manipulation of these data at individual, organisational unit and university-wide levels is very important. This is further evidenced by the University now taking forward options analysis and business case for a university-wide system.*

In fact it was the developers themselves who were most acutely aware of quality problems with the data,

> *… there are large research-based issues around provenance of data in environments like this .. Also issues around being unable to capture everything which impacts design and its ultimate use.*

whereas users tended to regard it highly (presumably because of the undoubted excitement at the visibility and integration of data to which they had not previously had easy access)

> *The level of data integration that ResearchRevealed offers is unprecedented at the University - previously staff have not been able to explore data integrated from several underlying databases at the presentation layer, let alone integrated with external data.*

There appears to be a significant side benefit of casting a spotlight on the importance of the availability, openness and quality of data:

> *In terms of quality of data, this is more variable, with the data drawn from internal and external systems not necessarily designed for the purpose for which it is used in the demonstrator. ResearchRevealed has demonstrated the limitations of existing systems in capturing these kinds of data (internal systems particularly) and the need for consistency and rigour in data entry. Visualisations of the data make inconsistencies and variations in data quality very clear, so it is easier to explain to developers, policy makers and users alike the need for higher quality data.*

2.3 Collaboration

The project aimed to collaborate with two main constituencies:

- within the University of Bristol, with a diverse range of players, senior managers, RED, IT services, Marketing and Communications, the research repositories and active researchers;
- outside of the university with other institutions, particularly Oxford, Southampton and the Knowledge Media Institute (KMi) at the Open University through meetings, hands-on workshops and information sharing.

Interestingly, the RR team rated their success at collaboration with outside partners lower than the outside partners themselves who rated it highly. Others partners and colleagues within the university also gave the project a high rating for collaborating; while recognising the frustration of obtaining data, they admired the consultative nature of the project process within the university and the persistence with which the team pursued their mission, successfully obtaining requirements from the range of departments mentioned above and aligning with other strategic initiatives, such as the Research Information Systems (RIS) Project, the People Profiler (CMS project) and the Examiner (data mining project) at the University. One area of continued frustration was the failure to meaningfully share and use data between institutions. It may be that future data integration and collaboration projects should prioritise and incentivise this.

The workshops are dealt with in more detail below under Dissemination.

2.4 Impact and relevance

2.4.1 Impact on the team and within the university

The team reported a significant impact on current and planned work from the expertise and experience they gained working on the project.

> *It has increased our familiarity with Semantic Web / Linked Data approaches and helped us understand the issues and the landscape. Many RR components have been and will be reused within the university and there has been significant outside interest in the "bits and pieces".*
>
> *It's enabled me to develop things that we can re-use, components that others are interested in. I took stuff we already had and fixed it! Outside of the technical stuff it helped understanding and relationships in the university. There's a general level of people using our stuff.*

In addition, potential users within the university reported RR playing an important role in developing strategy:

> *It has also shown how research and research profiles can be promoted beyond the University, and how individual researchers can tailor those profiles for different applications. Beyond the immediate local impact, ResearchRevealed has also had a part to play in disseminating (and possibly co-creating) the [*national*] approach [*to data sharing*], particularly via workshops and the RIM group.*

Business analysts have recently completed a requirements analysis for the "Research Information Systems (RIS) Project", regarding the research data needs for Research Managers and Directors and Researchers at the University, in particular focussing on i) demands that the REF places on the institution, ii) on how we can provide a better solution through which academics' publications may be captured and iii) on discerning what our business intelligence needs are. The RIS project is now moving to the technical options analysis phase, evaluating how a range of supplier products compare to each other, whether we may be able to develop an in-house solution, or some combination of both. ResearchRevealed has helped enormously

2.4.2 Impact outside the university

Demonstrates the way forward for research information data - gets the data out of the box and starts to make it really useful.

It has been an important example for the community.

2.5 Advancing the technology

The technology which the project adopted at the outset of the project was only capable of dealing with a small subset of the University's data (for a couple of departments in fact). This was scaled to a solution that can integrate and offer a faceted browse over the entire University's data plus some external data drawn from funding councils. The project has also been able to publish the university's research data, publications data, and staff data as "linked data", offering a SPARQL endpoint to query it, which the team believe is possibly unique, certainly unusual, in UK HE. Development of different aspects of the demonstrator has involved novel application of existing technologies (for example, using a bookmarklet application as a tool to allow academics and active researchers to quickly record impact and include that record in the database, where that impact data becomes shared, searchable and browsable). As might be expected in a project of this nature, it has also raised a number of interesting issues relating to e.g. name matching and the current limitations of a linked data approach. There seems little doubt that the process of combining data from disparate sources (some well structured and some in the words of a team member "very messy") has created technology requirements which might not have been predicted before the project started. The team's work on querying and inputting the data was developing in parallel with and alongside W3C work in this area; and the specifics and mechanics of publishing of a large body of linked data was also a demonstrator for W3C standards. Once again the team tended to assess their own achievements perhaps more modestly

it's novelty is in merging existing datasets together. Aside from that, the technology isn't particularly revolutionary

than collaborators and users:

I think the importance here is how the technology is harnessed to create a service which I suspect has the potential to become strategically important once it's realised by significant individuals what the power of the technology provides.

2.6 Dissemination and engaging with users

The project's dissemination efforts were well attended and very well received, both inside and outside the university. In order to reach the very diverse community of stakeholders, a number of different mechanisms were adopted:

- Focussed technical meetings where collaborators and interested parties from other institutions came for an intensive day of presentations, discussions and goal-setting. Advantages: performed well in awareness-raising, clarifications of research agendas and avoiding duplication of effort; very worthwhile for project and technical managers. Disadvantage: some technical staff felt that there was a tendency towards talk as opposed to practical action/ hands-on work days.
- Linked data hackdays (see detailed account and videos from participants at http://tinyurl.com/3wemwwr) – brought together a wider collection of linked data activists/code-writers for two days (and nights in some cases) of working together on practical solutions. Advantages: highly popular with attendees, informal nature brought in new people (including several from non-academic sphere e.g. BBC and commercial enterprises) who had not previously worked with the project, productive in terms of practical results and forging new working relationships / collaborations. Disadvantage: focus on RR project objectives and deliverables was diluted, most of the work was on more general linked data issues and topics.
- Internal university-only briefing and strategy days where diverse stakeholders from the university came together to share project news, align project objectives and set priorities and agenda for current and future work in this area.
- Blog http://researchrevealed.ilrt.bris.ac.uk/ and mini-presentations at conferences and workshops. The blog was used extensively to update collaborators and potential users on progress and staff attended conferences and workshops giving mini-presentations were appropriate
- In particular, Nikki Rogers kept the Research Information Management group briefed on the project's work and contributed towards moving forward the debate on open data and data sharing within funding and research councils[1]

The workshops were highly rated by participants with external collaborators particularly appreciating the "outward facing" nature of the project.

ResearchRevealed was intended to provide a prototype interface for several groups within the university: end users (active researchers), research managers within departments, research managers within the administration and the Public Relations Office (PRO). There was some disappointment among the team that, despite positive feedback from end users during the user consultation, because of the difficulties encountered in involving a significant number of end users in the development/consultation process, the focus of the project appeared to

[1] See: http://www.jiscinfonet.ac.uk/infokits/research , http://tinyurl.com/3zoncsm http://www.ukoln.ac.uk/rim/dissemination/2010/rim-cerif.pdf ;

shift towards research managers – there was not a consensus on this, but it was expressed as a concern by some.

> *It's certainly engaged well with the dev community.*
>
> *The developer community and those involved in related areas of technical research in particular have engaged well*
>
> *There is always a fine balance between disseminating widely when you are dealing with a demonstrator - managing expectations particularly with users is difficult.*

2.7 Sustainability and longer term effects

It is uncertain whether the software itself will be adopted and developed by the university, but the project has certainly helped to set the agenda for future work is this field

> *The demonstrator continues to play a part in refining requirements for the business case for a university system within an over-arching approach to Enterprise Architecture*
>
> *This is a hot topic in universities. I'm sure the outcomes of this project will feed into others. The technology already has interested users outside the project so it likely to continue in some form.*
>
> *ResearchRevealed has shed much light on both data quality issues here at the University and also how Linked Data can be a solution in terms of integrating external data to great benefit.*
>
> *The technology will continue most likely as components, but also possibly being reused for other client needs. The system itself will live as a demonstration of how you can link, view and integrate multi-sourced research data.*
>
> *It provides data in a form that we can re-use it, it will enable the PRO to improve data's visibility and the university's visibility on the web*

2.8 Measuring against specific objectives

Within the university, the project aimed to make a contribution in two specific areas. I asked for feedback on these:

2.8.1 Address the need for improved cross-linking, discovery, manipulation, export and reuse of research repository data at the University of Bristol.

It was universally recognised that the project had been successful in this area, demonstrating the feasibility, benefits and desirability of tackling this issue and making major strides in analysing and addressing the issues involved. Of course the project did not provide a complete solution on this area:

> *Access to some data (eg financial) has been difficult, so whilst there are data from some internal data sources including publications, organisational and contact data, the complete picture has been difficult to draw.*

Nevertheless, the reaction to the novel views of data produced were a clear indication of the value of this approach.

2.8.2 Address the need for sustainable institutional policy for the privacy and access control of research data at the University of Bristol.

The project certainly addressed the need for a sustainable institutional policy and in a number of instances came up against the lack of it. Several institutional initiatives have been informed or affected by ResearchRevealed and the project manager has now moved on to a new role as Enterprise Architect for the university, guiding the "Research Information Systems (RIS) Project" http://estates.bris.ac.uk/ips/project/53/public. Business analysts have recently completed a requirements analysis for this project, regarding the research data needs for Research Managers and Directors and Researchers at the University. It will be interesting to see whether senior management recognise the advantages of a comprehensive policy which prioritises open access to appropriate research data. The danger is that a closed or agit-prop approach is taken, where data is made available only on a "need-to-know" basis and it is assumed that business intelligence needs are best served by allowing open access to datasets only when the business case is proven. It may be that the increasing importance of services such as Google Scholar and Microsoft Academic Search will demand a change of data strategy in the same way that Google demanded a change of information strategy some years ago.

> *… by exposing data so transparently (and as Linked Data) it has started to raise questions at the strategic level. But it has not pushed institutional policy high on to agendas ... it's generally a difficult area at the University*
>
> *Making data available seems to be regarded as more "dangerous" than publishing the same data on a web site*
>
> *We have started the conversation, but it is not within our power to provide a definitive solution- that is a political decision. We have shone a light on all the difficult questions. Once the political decisions are taken then we can definitely advise on the best practice for uncovering the data*

3 Conclusions

The project was successful in demonstrating the value of integrated views of research data. Technically, the solutions it produced were well-received and several components of the package have already been re-used in JISC and other projects. From a user point of view the main achievement was pulling together data to which access was already possible but fragmented and presenting it in a well-designed and easily usable interface. The project shone a light on problems with data accessibility and quality within and without the university. A particular success of the project was to demonstrate a quick and easy way for end users to note and store evidence of the impact of their research in such a way that others might use that impact evidence for future proposals or submissions.

Open data enthusiasts and people building technical solutions talk about the importance of making data freely available but this often does not translate into policy or institutional strategy. It seems that users within the university "got it", only when they were able to see and manipulate the real data that was of

concern to them, integrated and presented in ways they had not envisaged. It is not until people experience this in a hands-on way that they move from in principle approval to enthusiasm. It remains to be seen whether this enthusiasm can be communicated to senior management and outside bodies to affect institutional policy. Certainly it will be unlikely to happen widely without illustrative exemplars and "things to look at".

4 Recommendations

4.1 Presentation to Research Information Management group

The Research Information Management group is an informal group on which the funding and research councils are represented as well as JISC and other interested bodies. I recommend they invite the technical manager of ResearchRevealed to give a brief presentation and answer questions.

4.2 Presentation to Senior Management within University of Bristol

I recommend that the new Enterprise Architect be invited to give a presentation to senior management at the university on the lessons learnt from the project, including the difficulties in accessing data and the business advantages in adopting an open data policy with exceptions, rather than operating a need-to-know policy on data sharing.

4.3 Funders: Future projects and data sharing

I recommend that at least one strand of funding policy encourage projects involving multiple institutions, research councils, funding councils and other bodies to share real data on research information and make visible an integrated interface to it, and that a commitment to this sharing and visibility is a prerequisite for project funding in this strand.

5 Other projects and initiatives to note

Microsoft Academic Search http://academic.research.microsoft.com/

Google Scholar http://scholar.google.co.uk/

From the UKOLN RIM page http://www.ukoln.ac.uk/rim/ :

Relevant UK Projects:

- BRII (Building the Research Information Infrastructure) project (University of Oxford): http://brii.medsci.ox.ac.uk/
- BRUCE: Brunel Research Under a CERIF Environment (Brunel University): http://www.jisc.ac.uk/whatwedo/projects/bruce.aspx
- CERIFy (UKOLN, University of Bath): http://www.jisc.ac.uk/whatwedo/projects/cerify.aspx
- Content Integration Project CIP (University of Bristol): http://cip-blog.ilrt.bris.ac.uk/blog/

- dotAC: Exploring the UK research landscape project (University of Southampton): http://www.ecs.soton.ac.uk/research/projects/697
- Enrich project (University of Glasgow): http://www.gla.ac.uk/enrich/
- EVIE (Embedding a VRE in an Institutional Environment) project (University of Leeds): http://www.leeds.ac.uk/evie/
- EXRI-UK (Exchanging Research Information in the UK) project (University of Bristol): http://exri.ilrt.bris.ac.uk/
- IRIOS: Integrated Research Input and Output System (University of Sunderland): http://www.irios.sunderland.ac.uk/
- MICE: Measuring Impact under CERIF (Centre for e-Research, Kings College London): http://www.jisc.ac.uk/whatwedo/projects/mice.aspx
- RCUK Research Outcomes project: http://www.rcuk.ac.uk/aboutrcuk/efficiency/Researchoutcomes/default.htm
- Readiness4REF project (King's College London, University of Southampton): http://www.kcl.ac.uk/iss/cerch/projects/portfolio/r4r.html
- ResearchRevealed project (University of Bristol): http://researchrevealed.ilrt.bris.ac.uk/
- RMAS (Research Management and Administrative System) project (University of Exeter) [RMAS has developed a detailed set of requirements for a RIM system]: http://as.exeter.ac.uk/rmas/

ResearchRevealed Technical Evaluation

Author and contact details:
Phil Cross

philip.cross@manchester.ac.uk

Tel.: 01626 774269

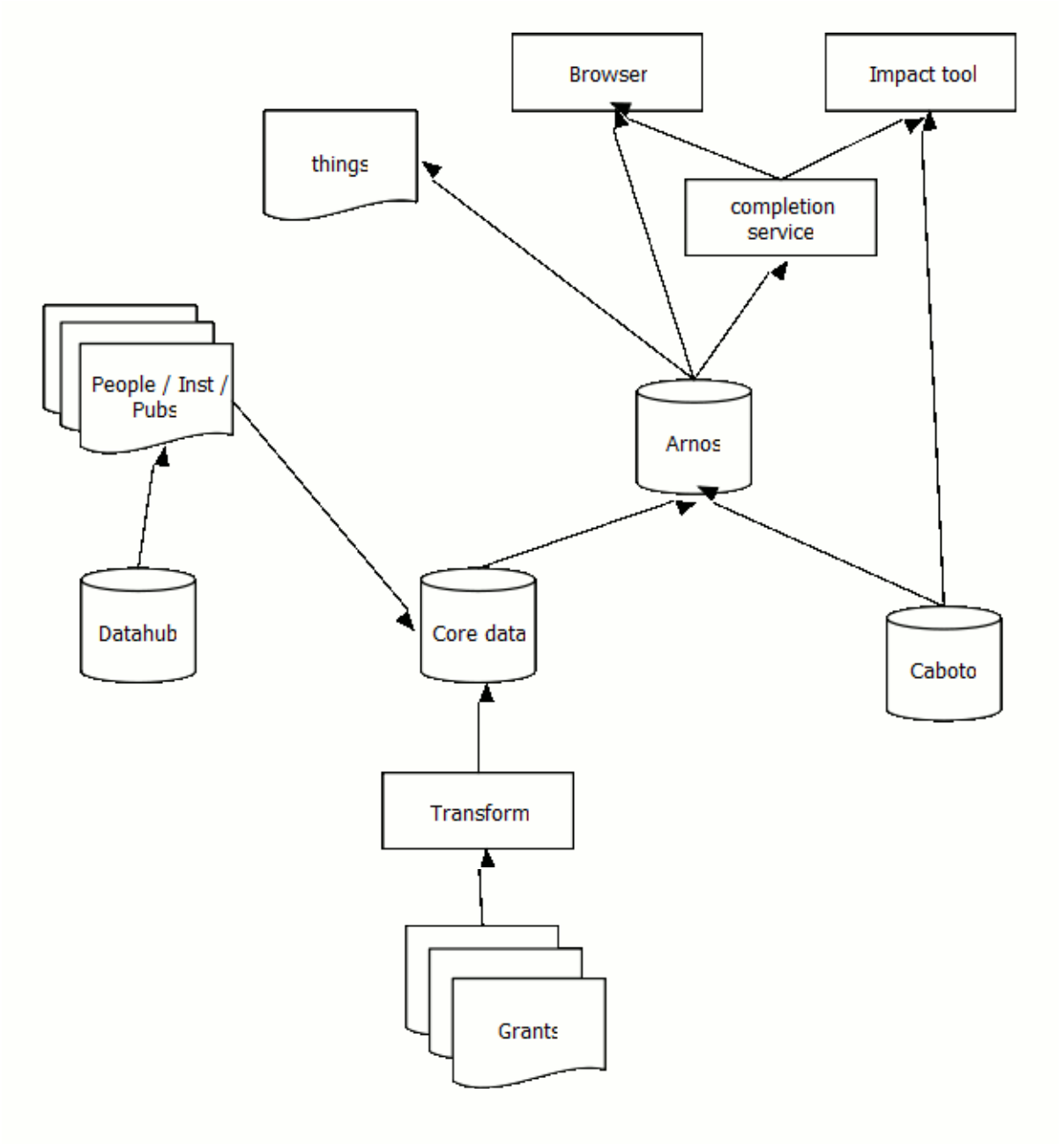
4 May 2011

Contents

1 Methodology

I used the project wiki as a starting point for understanding the technology used in this project, I had an initial meeting with the technical team and the project manager, and I obtained a copy of the software suite from the technical team. Due to time constraints it was not possible to install and run the system myself and in addition it would not have been possible for me to access the secure information from the Bristol University Data Hub or use another institution's data. I therefore simply attempted to install the software on my own Ubuntu Linux machine so that I could gain an overview of the technologies involved.

2 Overview



Architecture of ResearchRevealed software suite

This suite of software applications provides an integrated method for pulling in separate but interrelated data sets and allowing the links between the data items to be explored in an intuitive manner. This is accomplished by the use of Semantic Web technologies, specifically the use of RDF, and can be seen as an excellent demonstration of the purposes of the Semantic Web or Linked Data.

A major issue to achieving this objective was the lack of access to data in a readily usable format adhering to Semantic Web standards. Consequently, the system had to find solutions to obtaining data that may require authorisation to access and is found in a variety of formats, including using screen-scraping techniques to obtain data from funder web sites. Data is then mapped into namespaces chosen by the team and stored centrally in an RDF triplestore – the Core data.

The system then provides a faceted search over the harvested data. Faceted searching is an approach to navigating through information that has become prevalent over the last decade, particularly in commercial web sites. It combines textual searching with a progressive narrowing of search options based upon the common properties of the items in the data. This approach enables users to navigate through the underlying collection of RDF triples in a way that makes sense to them and usefully reveals the underlying linkages between data items. From the perspective of users, the faceted browser module might therefore be seen as the principle component of the ResearchRevealed system.

In addition, the system provides a mechanism for authenticated researchers to associate impact information with their own research data in the form of annotations, and allows this data to be surfaced via the faceted browser.

Under the hood is a system developed at ILRT for federating SPARQL queries between the Core data and researcher annotations, called *Arnos*.

The software suite may be viewed as consisting of three distinct parts:

- A faceted browser providing access to data stored in a distributed system of RDF triplestores, demonstrating the potential of the Linked Data aspect of the Semantic Web
- An annotation system for adding research impact information to an existing set of research data triples.
- A diverse set of mechanisms for accessing and converting the required data into a form that can be used by the other parts. This group of applications consists of some that are clearly temporary solutions to what are hoped to be temporary problems (for instance, accessing information from funder web sites), and some that provide innovative and useful approaches to issues that are likely to remain for the short to medium term (accessing data from institutional silos of protected data).

The impression of the architecture is hence not so much of a naturally integrated software suite that could be taken up as a whole and used in other academic institutions but of a set of complementary modules, put together to solve the particular problem space found at Bristol University, but which could be customised and developed further to provide solutions elsewhere.

The heterogeneous nature of the software suite is also apparent in the range of different programming technologies used.

Many of the modules will also clearly be useful in their own right in other projects.

One final issue that came to light at the end of the project was that ResearchRevealed has no mechanism for tracking movements of staff if they move between departments. According to the technical team:

*Essentially RR has no temporal semantics concerning people and departments. I.e. for any given paper or grant, we can't tell which department the author belonged to *at the time they wrote it*. If a research[er] switches from say, Biology to Chemistry, their entire back catalog of publications and grants moves with them and the RR browser would now show this material appearing under the department of Chemistry.*

Unless historic information about department memberships exist[s] in a university database somewhere, then we can't reliably assert anything about a department's grants or outputs in years gone by.

This is actually a problem of a lack of such information in the data provided by the university, and if the data was made available, the team assert that it would be fairly easy to change the inference rules used to generate the browser output.

3 Installing the software

As noted above, it was not possible to run the system on my test machine, due to time and logistical constraints. I simply installed each module separately into a Ubuntu Linux environment to see if there were any issues with the installation and to be able to view the code and configuration files.

In general, the various applications installed with no problems, several of them using Apache Maven to build the project and resolve dependencies, which makes installation straightforward. Others were Java web applications that could be directly installed into a web application server (I used Apache Tomcat). I had to install some extra packages such as Ruby for Rails for the ResearchRevealed Hub but otherwise had no problems. There is, however, very little documentation for installing or configuring the modules.

The only issue as noted above is the range of technologies used within the project and the fact that all the components of ResearchRevealed are separate applications. Installing and configuring the entire suite would therefore require knowledge of a broad range of technologies and be quite time consuming. However, this is partly due to the need to have several different solutions to harvesting and converting a disparate set of data formats into the central part of the system. But if ResearchRevealed was to be considered for marketing to other institutions, some work on integrating the entire system would be required. Since the majority of the software is based around Java, and much already uses Maven, this shouldn’t be a problem.

4 The ResearchRevealed modules

4.1 ResearchRevealed Hub

ResearchRevealed accesses information about Bristol researchers from the Bristol 'DataHub'[1]. This is an Oracle database that integrates data from a number of University systems and provides a single point of access for information about staff, students and the institution. This includes information about researchers and their publications.

ResearchRevealed Hub is an application developed by the technical team to access the information in the DataHub and convert it into RDFa encoded web pages containing information about the different research objects. These are then parsed by the RDFScraper module (see below) for inclusion in the Core data.

The approach has been to write an MVC web application (Ruby on Rails) whose data model is based on the existing DataHub database. The View configuration of Ruby on Rails outputs this information both as HTML and RDFa.

The team suggest this approach is solely due to the lack of availability of the required data either as usable RDF or already encoded in web pages in this manner. They hope in the future such an approach won't be necessary as universities hopefully move to more open data policies. However, it seems an innovative approach for interfacing an existing database system to an RDF store. The approach also generates end points for the browse nodes in the form of the web pages output by the application, which can display information that may be held in a secure system.

Configuring this application for a different database would be a matter of editing the configuration of the MVC model based on the structure of the database to access; assuming a basic set of data containing information on people, publications and organisations, this should in theory be fairly simple to achieve.

The team said that they used the MVC approach for converting the data since MVC frameworks are much more mature than relational database to RDF mappers. RDFa, used to encode the extracted data, is a microformat[2] and a W3C recommendation for encoding RDF triple data held in XHTML documents. According to rdfa.info:

"New research released by Yahoo! shows that RDFa demonstrated explosive growth in 2010. In fact, RDFa is the fastest growing data markup format on the web, and is used on more than 430 million web pages. It accounts for roughly 3.6% of the all of the web pages on the Internet".[3]

It would have been possible to use the RDF support provided by the Oracle database used by the University of Bristol directly. This was considered but there were a few issues. According to the team:

1. *Facetted browsing is a bit brutal on databases. I didn't think we [could reasonably] directly use the DataHub like that.*

[1] http://www.bristol.ac.uk/is/computing/applications/infosystems/datahub/overview.html
[2] http://microformats.org/
[3] http://rdfa.info/2011/01/26/rdfa-grows/

2. *Mapping the IRIS-derived[4] publication database was pretty hairy. Doing this in imperative ruby was much easier.*

3. *We needed a separate RDF store anyway for non-DataHub data*

4.2 RDFScraper

RDFScraper transforms the RDFa in the XHTML output from the ResearchRevealed Hub module using XSLT and stores it using SPARQL update into the Core RDF triple store. It is also used to scrape funder web sites for information about grants.

RDFScraper is intended to become a module for Heritrix[5], an open source web-crawler designed for web site capture by The Internet Archive[6]. This was chosen as it is a well-tested crawler, which behaves in a friendly manner and deals with the "sort of nasty web behaviour found in the wild". Heritrix is written in Java but only supported on the Linux platform. However, the Heritrix module has not yet been developed and, at present, RDFScraper makes use of a simple Ruby spider called Anemone[7]instead. RDFScraper requires Ivy and Ant to install.

In practice, the separation between the data storage implementation used by Bristol University (the DataHub) and the ResearchRevealed application achieved by the ResearchRevealed Hub module is broken by RDFScraper. This is because, in addition to harvesting the RDFa information directly from the ResearchRevealed Hub web pages, it is also set up to directly query the DataHub for item IDs. According to the team, this was originally to allow direct access to last modified dates to allow harvesting of changed or new records, information not stored in the RDFa. In practice this approach is not used as the last modified dates are not populated by DataHub. However, the direct access makes complete re-harvests of the data much easier, since it allows you to see the percentage progress of the import.

This makes the application specific to Bristol's set up so it would be useful to look for alternative ways to enable 'last modified' record updating of the RDFa metadata, should the ResearchRevealed Hub approach be continued, and to see how full harvesting of the data might be made more efficient - perhaps with the eventual move to using Heritrix.

The configuration files are written in Ruby, requiring some knowledge of this language, but the RDFa parsing is done by Jena[8], an open source Semantic web framework, written in Java and with a long track record within the Semantic Web community. Jena has a dedicated package for parsing RDFa within XHTML.

In addition to parsing the output from ResearchRevealed Hub, RDFScraper is used for scraping the web pages of funding bodies.

[4] The IRIS database contains details of research publications by Bristol University staff
[5] http://crawler.archive.org/
[6] http://www.archive.org/
[7] http://anemone.rubyforge.org/
[8] http://jena.sourceforge.net/

Each funding body web site has a spider defined for the pages that provide information on grants for Bristol University departments and staff. Whilst the spiders are defined using Ruby, the particular configuration for each uses XSL. These stylesheets can be quite complex and are obviously not ideal for extracting information from pages whose format could easily change. However, this is intended to be a temporary approach with the hope that, in the future, organisations such as funding bodies will populate their pages with microformats such as RDFa.

It is likely that if Heritrix is adopted for the crawler, the configuration of this module will become simpler and easier for other organisations to take up, as Heritrix is a widely used and well-supported application.

4.3 data-server

The central Core data application for storing the RDF triples produced by RDFScraper is run by a module called *data-server*. This uses Joseki[9], a SPARQL server developed for the Jena project. Joseki is a Java web application that supports the SPARQL protocol and RDF query language, allowing queries via HTTP and SOAP.

The data structure holds information about people, institutions, grants and publications and uses the following vocabularies:

rdfs: http://www.w3.org/2000/01/rdf-schema#
aiiso: http://purl.org/vocab/aiiso/schema#

- For organisational structure

foaf: http://xmlns.com/foaf/0.1/

- For people and organisations

proj: http://vocab.ouls.ox.ac.uk/projectfunding#

- For grant information

dc: http://purl.org/dc/terms/

- For title, abstract and source (provenance)

owl: http://www.w3.org/2002/07/owl#

- For *sameAs* relationship

closed: http://vocab.bris.ac.uk/rr/closed#

- Inferred hierarchical information

rr: http://vocab.bris.ac.uk/resrev#

- Space for web Future's 'made up' properties and some labels for local publication types

rel: http://purl.org/vocab/relationship/

- For the *collaborates* relationship

The technical team have consulted widely over the choice of vocabularies to use for describing research entities and their relationships, including Oxford and Southampton, who were developing similar applications. They also considered the use cases for a common Application Profile: such as the exchange of research information between institutions; opening up information to the Semantic Web/Linked Data; and reporting, such as to funders and the forthcoming Research Excellence Framework (REF) process. These issues will need to be kept under review, especially as it is still uncertain what the requirements will be for the REF, which comes in to force in 2014. However, the

[9] http://www.joseki.org/

team reported that there was little overlap between the three projects at Bristol, Oxford and Southampton, except for publications and people.

The team are presently developing CSV output from ResearchRevealed.

4.4 REFTool

REFTool is a bookmarklet[10] tool written using the Spring Framework for capturing impact data from researchers for the future REF process. It is designed to add value to the research information harvested from the university and funder web sites by associating RDF annotations to the research data statements.

The annotations essentially consist of the URLs of web pages that evidence an impact of the research, together with further metadata supplied by the researcher via a form, which describes the nature of the impact. The researcher is asked for a title for the impact evidence, a description, keywords, researcher involved, grant involved, publication involved, plus a classification for the impact.

REFTool makes use of two further applications developed by Web Futures: Caboto, an application developed in a previous project for managing annotations and providing authentication and authorisation; and Completor (see below), a tool for providing completion options for text typed into the form fields, based on the contents of the central Core data triplestore. Completor provides access to the researcher, grant and publication data held by ResearchRevealed, allowing the annotation to be linked to the correct data.

The annotations generated by Caboto are not added to the Core data but are stored in a separate Derby database which is then searched alongside the Core data using a federated SPARQL search provided by the Arnos module (see below). According to the technical team, the data for Caboto is kept separate as this operates as a system in itself, with its own API, SPARQL queries and authentication system. Arnos can then provide the mechanism for searching across these separate data sets.

REFTool is a clever, intuitive and potentially useful tool. Its value is enhanced since it makes use of existing and well-tried technologies in the form of bookmarklets and a RESTful interface to a SPARQL endpoint, together with Caboto, an application already tried and tested through three previous projects. This combination of a bookmarklet form with a "completor" system that taps into an existing SPARQL endpoint to generate annotations to that content clearly has widespread applications in the developing Linked Data world.

4.5 Arnos

Arnos is a Java web application that provides a federated query mechanism over SPARQL end-points. It is a Maven project built with the Spring Framework and Jena. Arnos was developed to enable searching across multiple triplestores or SPARQL endpoints, in particular the Core data of ResearchRevealed and the data store of annotations produced by the Caboto module, described above. It was developed as the team hadn't been able to find anything else *that suited [their] needs in an easily configurable way*. And was also developed with another project they were working on in mind.

[10] http://en.wikipedia.org/wiki/Bookmarklet

With Arnos, you can register all the endpoints an application is using via a RESTful interface. The application can then send all SPARQL queries to Arnos, which in turn issues the queries over all the registered endpoints. Results are then collated before being returned to the application.

Its benefits are that it:

- abstracts data sources farther from an application, so sources can be changed without affecting the code
- allows multiple sources to be hidden behind a single interface
- provides a query cache
- allows more security over connections

According to the team, the particularly nice approach with Arnos is that the client querying it doesn't need to have any knowledge of the back end stores being queried (although obviously in the real world that is never completely true), or need to use any specialist form of SPARQL.

According to the code web pages, there are some issues with Arnos since it is still under development to improve authentication through the interface and it also doesn't support full federated searching; instead taking a simplistic approach where a SPARQL query is simply forwarded on to the various end-points and the results merged.

However, this module also clearly has benefits beyond ResearchRevealed within the Semantic web.

4.6 Completor

This module provides a completion service for the browser and impact annotation tool - REFTool, described above. It is a fork of an auto-complete servlet developed for the STARS project[11]. Completor sits between a SPARQL endpoint and (typically) a web form to provide prefix completion for values entered into a field based on the contents of the target data store.

It is JavaScript driven, making use of the jQuery Autocomplete module, which poses some accessibility issues, but is another simple but useful tool which helps users to interact with the contents of a Semantic Web triplestore.

4.7 Facets

The Facets module, providing the browse and search interface to the Core data, might be seen as the heart of the system. It allows end users to navigate through complex sets of interlinked data stored as RDF triples using a faceted search, revealing information in an easy-to-understand manner.

The faceted search across the various objects held by ResearchRevealed is presented via two tabbed pages: *Research Outputs,* with the facets of Type, Department and Publication date; and *Grants and Funding*, with the facets of Funder, Grant values, Start year and End year. In addition, there are separate tabs for searching people and organisations without faceted browsing.

[11] http://stars.ilrt.bris.ac.uk/blog/

The Facets module is the most important component of ResearchRevealed as it allows users to explore the diverse but related data that has been harvested by the rest of the system. This is a vital element of the *Linked Data* approach to the extraction of knowledge from distributed data.

Linked Data is a component of the Semantic Web that uses URIs and RDF to pull together knowledge from disparate sets of data via machine-to-machine interfaces. A common approach would be to harvest data in the form of RDF from a number of separate sources and combine these in an RDF triplestore. Searches powered by the SPARQL query language can be made across the data using common or mapped vocabularies to connect related pieces of information. This is essentially what the ResearchRevealed software suite is enabling, where data about the University's staff and research publications can be combined with data harvested from publisher and grant funder web sites, combined with impact information added by the authors themselves via the browser plug-in REFTool - although in the case of ResearchRevealed, the data is not sourced using Linked Data approaches. The Faceted browser provides a graphical interface enabling the user to explore these sets of data.

Whilst Wikipedia lists a number of browsers designed for searching through Linked Data[12], these are generic and are limited by having no understanding of the domain of the information searched. This makes them difficult to use for those who have no knowledge of the nature of the data they are searching, i.e. that it consists of a collection of RDF triples. The Facets module is innovative in applying the familiar paradigm of faceted search across the data stored in a distributed set of triplestores, allowing users to navigate through the harvested data in an intuitive manner.

Depending on how easy it is to configure the browser to work with different sets of data using different namespaces for predicates and controlled vocabularies, Facets could have major re-use value for other data available through the Semantic Web.

The Facets module came from work originally carried out for the CREW[13] project, where it provided faceted browsing across events data. At this time, the objects forming the facets of the system existed as objects within the code. Since then, the team have further developed the application, so that customisation for browsing across different sets of data should require little or no need to edit Java code, although the team admit there are further areas for improvement here.

5 Conclusions

ResearchRevealed is not a homogenous suite of programs that could easily be installed and function "as is" in other institutions. Rather it is still a demonstrator, although one made from well-tested and robust components.

The central components*: Core data*, *Arnos, Impact tool*, *Completion service*, *Coboto* and the *faceted browser*, are a coherent set of applications but they depend upon the peripheral set of components that harvest and convert the source data into a form that the central components can use. This latter set of modules would need a lot of customisation allow them to work within other institutions, although not beyond the means of another technical team familiar with most of the current popular web technologies.

[12] http://en.wikipedia.org/wiki/Linked_Data#Browsers
[13] http://www.crew-vre.net/

However, it is to be hoped that the methods used to obtain the data that the central components of ResearchRevealed require will gradually become redundant as institutions become more willing to make their research data freely available and the Linked Data approach to information dissemination becomes more prevalent.