ourSpaces: Linking Provenance and Social Data in a Virtual Research Environment

Richard Reid
Computing Science
University of Aberdeen
United Kingdom
r.reid@abdn.ac.uk

Edoardo Pignotti
Computing Science
University of Aberdeen
United Kingdom
e.pignotti@abdn.ac.uk

Peter Edwards
Computing Science
University of Aberdeen
United Kingdom
p.edwards@abdn.ac.uk

Adele Laing
School of Geosciences
University of Aberdeen
United Kingdom
a.laing@abdn.ac.uk

ABSTRACT
Web-based Virtual Research Environments (VREs) have been proposed as one way in which e-Science tools can be deployed to support and enhance the research process. We are exploring the use of Linked Data in combination with the Open Provenance Model (OPM) and Social Web concepts to facilitate interactions between people and data in the context of a VRE. In this demo we present the *ourSpaces* VRE and outline the technologies used to link together provenance, research artefacts, projects, geographical locations and social data in the context of interdisciplinary research.

Categories and Subject Descriptors
I.2.4 [**Computing Methodologies**]: Knowledge Representation Formalisms and Methods; H.3.5 [**Information Systems**]: Online Information Services

General Terms
Documentation.

Keywords
VRE, e-Science, Linked Data, Provenance

1. INTRODUCTION
As scientific research becomes increasingly interdisciplinary in nature, the need for technologies that support collaboration and provide access to heterogeneous data and computational resources becomes ever more critical. The PolicyGrid [7] project, a collaboration between computer scientists and social scientists, is exploring how novel e-Science technologies can be used to support interdisciplinary research activities; in particular, the provision of support for evidence-based policy research. Evidence is used at various stages of policy making, from the design of new policies to the evaluation and review of existing policy. An evidence base supports transparency and accountability in the policy decision-making process. There is considerable potential for technologies to support such activities by providing tools to assist collaboration and interaction between researchers using the Web. De Roure [6] has argued that in order to accomplish this, it is crucial to develop solutions that facilitate the discovery and interpretation of knowledge generated by others and to allow connections between people, ideas and data.

Web-based Virtual Research Environments (VREs) [9] have been proposed as one way in which e-Science tools can be deployed to support and enhance the research process. Examples of existing VREs include myExperiment [5], which allows scientists to publish and share experimental workflows and provides additional social networking functionality such as tagging and commentaries; and SciSpace [1], a social networking site for scientists which incorporates many of the features one would expect such as tagging, blogging and maintaining contacts lists.

Semantic Web technologies [1] play an important role in facilitating such applications by providing a framework that enables linking of data. The principles of Linked Data [2] establish how RDF should be used to connect related (yet unconnected) data. Social Web technologies can be used in conjunction with Linked Data to allow people to comment and interact within the VRE, thus facilitating the interactions between people and data. Existing vocabularies such as FOAF [4] and SIOC [3] can be used to enable Semantic on Social Web applications. However, a gap still exists between the concept of Linked Data and the requirements of evidence-based research; it is difficult to give a detailed account of how a piece of data was derived (its provenance) just by linking data. To overcome this issue there are many existing approaches to representing provenance [12]. Most notably OPM [10] - an abstract model which defines a vocabulary to describe causation between artefacts, processes and agents (or controllers) and a set of rules to enable reasoning on provenance graphs.

This paper introduces *ourSpaces*: an interdisciplinary virtual research environment that facilitates the linking of research artefacts, provenance, projects, geographical locations and social data. We explain the rationale behind development of such an environment and summarise features of the system that we intend to demonstrate.

2. OURSPACES VRE
ourSpaces [11] promotes and enables online collaboration between researchers from various disciplines. Users within *ourSpaces* are encouraged to share their digital artefacts, download, tag and comment on other's work and form cohesive groups with other researchers.

Underpinning *ourSpaces* are a number of repositories and services. Each activity within the environment is enabled by a rich and pervasive RDF metadata infrastructure built upon a series of OWL ontologies (which are described more fully in Section 2.1). Information is retrieved from a metadata repository (running the Sesame[2] triple store) by a SPARQL query engine and forwarded

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

[1] www.scispace.org
[2] www.openrdf.org

to the interface. *ourSpaces* also makes use of folksonomies, allowing users to tag various events, digital artefacts and activities. A digital object repository is used for secure storage of digital artefacts and a number of web and grid services can be invoked from within the VRE. These include a natural language engine and interface [8] to support browsing and querying of data. Figure 1 presents the architecture of *ourSpaces* more fully.

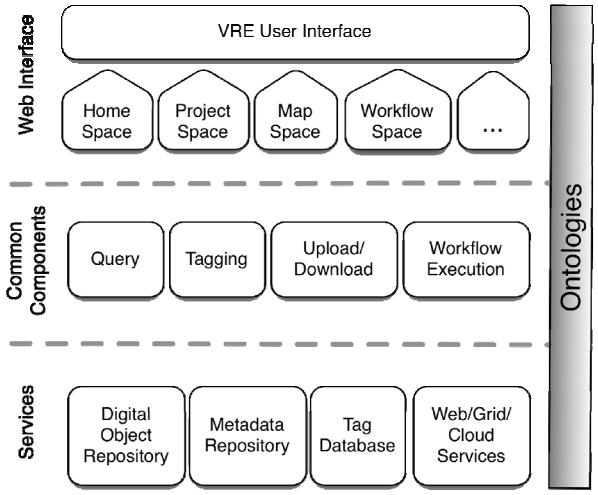

Figure 1 - ourSpaces Architecture.

2.1 Provenance Framework

In order to support linking of data, processes and people in *ourSpaces*, we have developed a provenance framework based on the Open Provenance Model. OPM provides a specification to express data provenance, process documentation and data derivation. It is based on three primary entities namely *Artefact*, *Process* and *Agent* and associated causal relationships namely *used*, *wasGeneratedBy*, *wasTriggeredBy*, *wasDerivedFrom* and *wasControlledBy*. Our framework consists of a generic provenance ontology developed in OWL, which defines the primary entities of OPM (*Artefact*, *Process*, *Agent* and *CausalRelationship*). Furthermore, our framework supports additional domain-specific provenance ontologies that can be created by extending the concepts defined in the OPM ontology with domain-specific classes. For example, in a Human Geography domain ontology, one might have a *Questionnaire* and a *Survey* as a type of *Artefact* and an *Interview* as a type of Process. To date we have developed a number of domain-specific provenance ontologies describing aspects of Human Geography, Social Simulation and Biology.

2.2 Creating Linked Data

A user in *ourSpaces* can upload and share different types of digital artefacts (e.g. documents, images, data-sets). During the upload process, a user is prompted to provide additional metadata describing specific properties of the artefact via a form. The form is generated dynamically depending on the domain ontology selected as well as the properties associated with the artefact type. For example, if an *Interview Transcript* is selected from the Human Geography ontology, it requires a *Title,* an *Interviewer* and a list of *Interviewees*. The artefact is stored into the digital object repository and an RDF instance describing the artefact is stored in the metadata repository.

We have created a utility ontology that can be used to annotate an *Artefact* (as defined by OPM) with geospatial information (e.g. *wasSubmittedAt*, *wasCollectedAt* and *wasDepositedAt*). Geographical locations are defined by the GeoNames ontology[3] by establishing the concept of a *Feature* as a means to represent any geographical entity. *Features* include properties relating to location names and postcodes, and can inherit properties relating to longitude and latitude from the W3C Basic Geo ontology[4]. In the context of *ourSpaces*, users can browse and create geospatial metadata within a map; an example is shown in Figure 2.

A user within *ourSpaces* can also create "Project Spaces"; an environment which provides various tools such as calendaring and resource aggregation. Project spaces allow research projects to display public information (such as their grant number, start-date, end-date, etc), while still offering a secure and private environment to project members as a means to govern their own research, collaborate within sub-projects or work-packages, create virtual meetings or events and manage artefacts. An uploaded artefact is automatically linked to a user, however, an artefact can also be linked to different aspects of a project space. For example, a data resource belonging to the project or a document produced as the result of a project event (such as a scheduled meeting or conference).

ourSpaces also provides a number of other "spaces", designed as a means to link, browse and share specific categories of data resources with other users. For example, the models space for handling simulation models and the maps space for browsing geo-spatial information.

Linking data in any online environment has the potential to raise issues of privacy. Resources become more accessible as users are able to browse and query through the RDF graph. Moreover, provenance provides greater transparency about the links between such resources. In order to overcome this issue, *ourSpaces* provides a mechanism to assign access rights to sensitive resources. Users can choose to make a particular resource private; alternatively, they can assign access to single users or groups of users (such as project members). Users can choose to create Read, Write and Edit rules for accessing data. For example, a user may have uploaded a *Paper* as part of a project within *ourSpaces*. This artefact may only be accessible to project members (read), however, the user may wish for other authors on the paper to create new and edit existing metadata (edit).

2.3 Linking Social Data

When users first join *ourSpaces*, they are presented with various means of establishing their social presence. *ourSpaces* utilises social networking concepts including tagging, collaboration invites, personal status updates, groups, etc. As *ourSpaces* is focused on improving collaboration among researchers, there are numerous tools dedicated to this such as an instant messaging (IM) service, internal messaging, emailing and comment facilities. Furthermore, these social interaction tools can establish links with digital artefacts and their associated provenance metadata. For

[3] http://www.geonames.org/ontology/

[4] http://www.w3.org/2003/01/geo/

example, a resource can be uploaded, then later tagged, commented upon or blogged about by another user.

To ensure that the link between social network activities and digital artefacts is established formally, we have extended our OPM-based framework to integrate social interaction as part of the provenance representation. The current OPM specification supports limited information about a person involved in the research process; there is also little regard for the wider social context. As a result, we have integrated existing social networking vocabularies including FOAF and SIOC with our provenance ontologies. The SIOC ontology (Semantically-Interlinked Online Communities) is designed to enable the integration of online community information by providing a model to express user-generated content. SIOC is used in conjunction with the FOAF vocabulary for expressing personal profiles and social networking activities such as blogging and commenting.

Figure 3 displays an example on how data provenance can be linked to social data; the data provenance of a set of interview notes has been enriched with a social context as a result of the introduction of the FOAF and SIOC ontologies. A set of interview notes has been produced, following an interview between Colin Hunter and John Farrington; Lorna Philip, who works with John Farrington, has posted a comment on the interview notes asking where the audio recording of that interview is. This example demonstrates how social context (e.g. *vre:workWith*) and user-generated content (e.g. *sioc:post*) can be integrated with the data provenance record. The effect of this is a far more definitive provenance representation of the digital artefact; the result of Lorna Philip's post may have inspired John Farrington to upload the audio recording of the interview. Alternatively, the lack of a vital piece of supporting information may have altered the level of trust in the interview transcript.

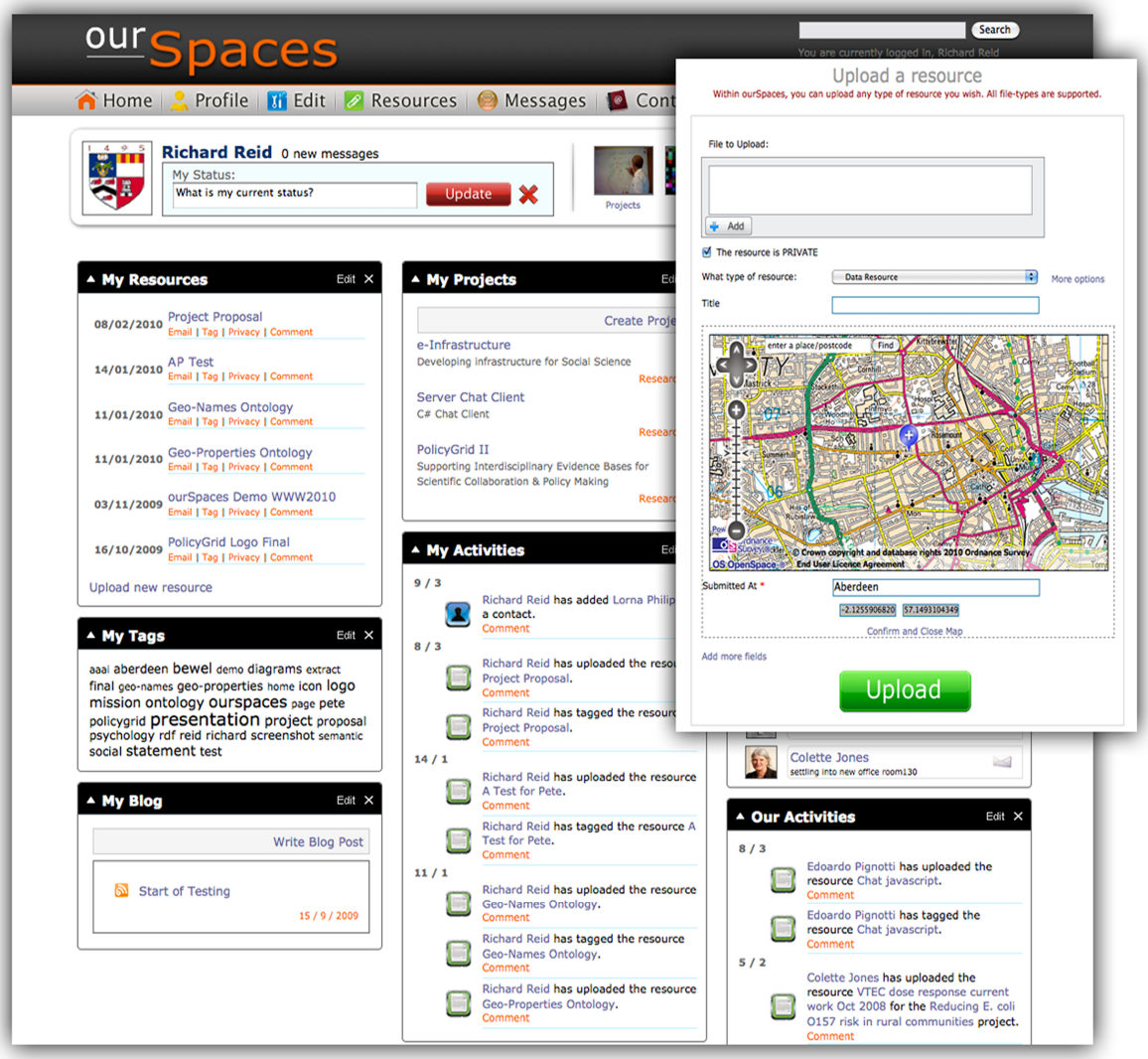

Figure 2 - ourSpaces Home Page with Upload Tool

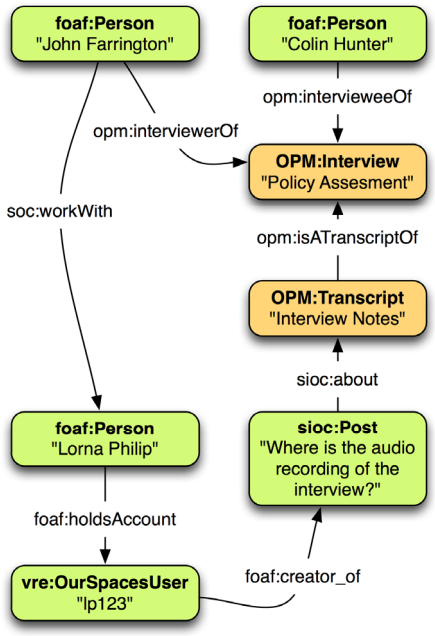

Figure 3 - Linking Provenance and Social Data

3. DEMONSTRATION CONTENT

During the demonstration, we will present the *ourSpaces* virtual research environment. Key features of the system will be demonstrated, as follows:

- Resource management including resource upload and description; this entails an overview of our ontology-driven (form-based) solution for metadata creation, enabling users, organisations, etc. to be connected with resources.
- Visualisation of provenance records. Displaying how data provenance is enriched by social context.
- Resource access controls and privacy. Methods for protecting sensitive data and providing individual (or group) access to particular resources.
- Group creation and maintenance features including management of a new project space (with focus on project security, event creation and group-resource aggregation).
- An overview of the social networking features such as blogging, tagging, maintaining contact lists, commenting, chatting, etc.
- A user's personal home page (Figure 2) and associated customisation features including component selection and appearance modification.
- Multiple approaches to metadata browsing including navigating through different spaces, visualising data within timelines and graphs.
- Different querying modes, including simple keyword searching, use of tag clouds and facet-browsing techniques.
- Use of mapping software to query and browse geo-spatial metadata regarding digital resources.

4. ACKNOWLEDGMENTS

The research described here was funded by the UK Economic & Social Research Council (ESRC) under the eSocial Science programme; award number: RES-149-25-1027.

We also acknowledge use of the OS OpenSpace development environment within *ourSpaces* to visualise geospatial metadata; see: http://openspace.ordnancesurvey.co.uk/openspace/

5. REFERENCES

[1] Berners-Lee, T., Hendler, J., Lassila, O. (2001) The Semantic Web. Scientific American, 284(5):34-43.

[2] Berners-Lee, T. (2006). Linked Data - Design Issues. Retrieved July 23, http://www.w3.org/DesignIssues/LinkedData.

[3] Bojars, U., Passant A., Cyganiak, R., Breslin J. (2008) Weaving SIOC into the Web of Linked Data. Proceedings of the WWW 2008 Workshop Linked Data on the Web LDOW2008, Beijing, China, Apr 2008.

[4] Brickley, D. and Miller, L. (2004) FOAF Vocabulary Specification. Namespace Doc- ument 2 Sept 2004, FOAF Project, 2004. http://xmlns.com/foaf/0.1/.

[5] De Roure, D., Goble, C. and Stevens, R. (2009) The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows. Future Generation Computer Systems 25, pp. 561-567.

[6] De Roure, D. 2008.The new e-Science presentation. http://www.slideshare.net/dder/the-new-science-bangalore-edution.

[7] Edwards P, Farrington J, Mellish C, Philip L, Chorley A, Hielkema F, Pignotti E, Reid R, Polhill J and Gotts N (2009) eSocial Science and Evidence-Based Policy Assessment: Challenges and Solutions. Social Science Computer Review 2009;27:553.

[8] Hielkema F, Mellish C and Edwards P (2008) Evaluating an Ontology-Driven WYSIWYM Interface. In Proceedings of the Fifth International Conference on Natural Language Generation (INLG2008) 2008, Columbus, Ohio, USA.

[9] Hodson, S. (2007) .The use of Virtual Research Environments and eScience to enhance use of online resources: the History of Political Discourse VRE Project. UK eScience All Hands Meeting 2007.

[10] Moreau L, Freire J, Futrelle J, McGrath R, Myers J and Paulson P. (2007). The open provenance model. Technical report, University of Southampton.

[11] Reid, R. and Edwards P. (2009). ourSpaces - A Social Semantic Web Environment for eScience. Proceedings of the AAAI Spring Symposium on the Social Semantic Web, 2009, 22-25 March, Stanford, USA.

[12] Simmhan Y, Pale B, and Gannon D (2005). A Survey of Data Provenance in eScience. SIGMOD Record, 34(3): 31-36, Sept. 2005.