Web Semantics: Science, Services and Agents on the World Wide Web 27–28 (2014) 70–77

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

Lessons learnt from the deployment of a semantic virtual research environment

Peter Edwards [a,*], Edoardo Pignotti [a], Chris Mellish [a], Alan Eckhardt [a], Kapila Ponnamperuma [a], Thomas Bouttaz [a], Lorna Philip [b], Kate Pangbourne [b], Gary Polhill [c], Nick Gotts [c]

[a] Computing Science, University of Aberdeen, United Kingdom
[b] School of Geosciences, University of Aberdeen, United Kingdom
[c] The James Hutton Institute, Aberdeen, United Kingdom

ARTICLE INFO

Article history:
Received 28 September 2013
Received in revised form
6 May 2014
Accepted 24 July 2014
Available online 23 August 2014

Keywords:
Provenance
Policies
Natural language
OPM
PROV-O

ABSTRACT

The *ourSpaces* Virtual Research Environment makes use of Semantic Web technologies to create a platform to support multi-disciplinary research groups. This paper introduces the main semantic components of the system: a framework to capture the provenance of the research process, a collection of services to create and visualise metadata and a policy reasoning service. We also describe different approaches to authoring and accessing metadata within the VRE. Using evidence gathered from data provided by the users of the system we discuss the lessons learnt from deployment with three case study groups.

© 2014 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/3.0/).

1. Introduction

Many of the contemporary challenges facing society such as climate change, require researchers from a range of disciplines to work together. Underpinning the scientific process is the transfer of ideas, knowledge and resources, and in recent years, the Web has drastically altered both the nature and speed of this exchange.

Web-based Virtual Research Environments (VREs) [1] have been proposed as one way to help researchers across all disciplines to manage the increasingly complex range of tasks involved in carrying out research. In the UK, the Joint Information Systems Committee (JISC) VRE programme[1] explored the virtual research environment collaborative landscape. JISC recognised that a major shift in research practice will occur through the formation of common taxonomies, data standards and metadata as researchers collaborate with others across disciplinary, institutional and national boundaries [1]. Semantic web technologies [2] are seen as crucial in this context in order to provide a common framework to allow the creation of intelligent applications and services which can be integrated with data resources, people and other objects in a VRE.

Some of these issues have been explored by the PolicyGrid[2] project, a collaboration between human geographers and computer scientists as part of the UK Digital Social Research initiative. As part of the project we developed *ourSpaces*,[3] a virtual research environment that allows researchers to collaborate online. A screenshot of the *ourSpaces* web interface is presented in Fig. 1.

The system was co-developed with three interdisciplinary case study groups: a research team investigating *E. coli* O157 risk in communities, members and affiliates of the Aberdeen Centre for Environmental Sustainability, and a group of agent based social simulation modellers. Based upon interactions with these groups

* Corresponding author. Tel.: +44 1224274065.
E-mail addresses: p.edwards@abdn.ac.uk (P. Edwards), e.pignotti@abdn.ac.uk (E. Pignotti), c.mellish@abdn.ac.uk (C. Mellish), a.e@centrum.cz (A. Eckhardt), k.ponnamperuma@abdn.ac.uk (K. Ponnamperuma), thomas.bouttaz@googlemail.com (T. Bouttaz), l.philip@abdn.ac.uk (L. Philip), k.pangbourne@abdn.ac.uk (K. Pangbourne), Gary.Polhill@hutton.ac.uk (G. Polhill), ngotts@gn.acp.org (N. Gotts).

[1] http://www.jisc.ac.uk/whatwedo/programmes/vre.aspx.

[2] http://www.purl.org/policygrid.

[3] http://www.purl.org/policygrid/ourspaces.

http://dx.doi.org/10.1016/j.websem.2014.07.008
1570-8268/© 2014 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/3.0/).

Fig. 1. A screenshot of the *ourSpaces* VRE showing a user's personal page, a natural language representation of project metadata, and a provenance graph associated with a report.

the core requirements for the *ourSpaces* VRE were identified as follows: (a) It should be possible to describe and uniquely identify a range of entities: artefacts (digital and physical); processes (both computational services and human activities); people; organisational structures and membership; social networks; (b) The system should incorporate online communication (e.g. instant messaging, blog entries, email) into the research record; (c) It should be possible to define relationships (e.g. causal, social, organisational) between entities; (d) It should be possible to define access control and documentation policies.

To satisfy these requirements the *ourSpaces* architecture implements a number of core and Web services for creating, editing and querying data, metadata and digital artefacts. These include a service used to upload and access digital artefacts, a natural language service to support browsing and querying data, and a policy reasoning service [3]. The diagram in Fig. 2 illustrates the basic components of the system architecture. While *ourSpaces* is currently accessible online, it is no longer supported (as the PolicyGrid project ended in Summer 2012). The source code is available for download via GitHub[4] under the GNU LGPL version 2.1. The system requires a linux-based environment capable of running Java EE 1.6.0, apache tomcat 6.18 or above, MySQL version 14.2 or above and openrdf-sesame 2.0.

The system has been designed in order to encourage users to share their digital artefacts, download and comment on each other's work and form cohesive groups with other researchers. In order to support the formation of interdisciplinary groups in *ourSpaces*, users are presented with various means of establishing their social presence, e.g. tagging, blogging, personal status updates, instant messaging.

In a collaborative environment such as *ourSpaces*, understanding the provenance of scientific data and other research artefacts is crucial in order to understand and verify their authenticity and completeness [4]. *ourSpaces* is thus underpinned by a provenance framework capable of capturing the derivation history of research artefacts, including the original sources, intermediate products and processes involved.

In this paper we revisit the design, implementation and deployment of the *ourSpaces* VRE introduced in a previous publication [5]. We introduce a new Personal Lexicon Service (PLS) designed to create a set of vocabulary mappings and to utilise them during search and online communication. We also introduce a possible solution for adapting the provenance framework used within *ourSpaces* to the recent W3C provenance recommendations.

The remainder of this paper is organised as follows. In Section 2 we describe the design of the ontological framework required to support provenance in the VRE; Section 3 presents a number of tools that we have developed in order to support interaction with semantic metadata. In Section 4 we discuss lessons learnt from deployment of *ourSpaces* with our case-study communities. In Section 5 we introduce a set of mappings and rules to convert an *ourSpaces* provenance record to PROV-O. Finally, we discuss related work, our conclusions and future directions.

2. The *ourSpaces* ontological framework

We have developed an extensible ontological framework for capturing the provenance of the research process based on the requirements highlighted in Section 1. In order to describe and uniquely identify entities (such as artefacts, people, locations) and to make explicit relations between entities we follow the linked data principles [6]. At the heart of *ourSpaces* (and thus, our provenance framework) is an OWL representation of the Open

[4] https://github.com/policygrid/ourSpaces.

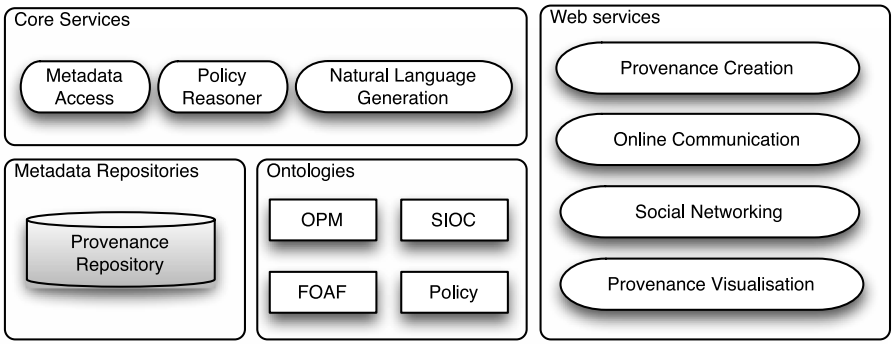

Fig. 2. The *ourSpaces* system architecture.

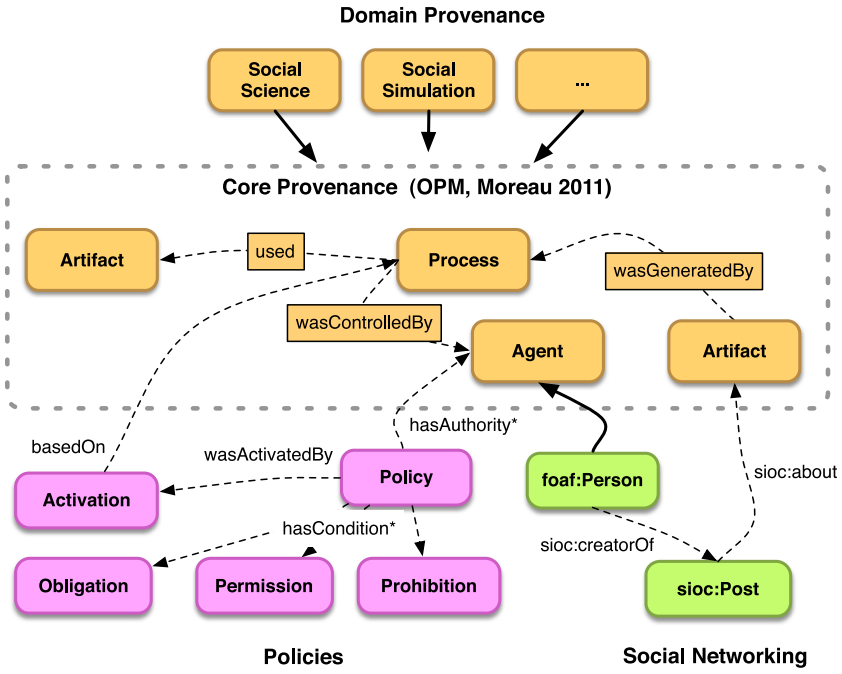

Fig. 3. *ourSpaces* ontological framework.

Provenance Model (OPM) [7]. This ontology defines the primary entities of OPM as well as the causal relationships that link them (see Fig. 3, *Core Provenance*). OPM is a generic solution and as a result, our framework supports additional domain-specific provenance ontologies that are created by extending the concepts defined in the OPM ontology with domain-specific classes. To date we have developed a number of such provenance ontologies describing aspects of Human Geography[5] and Social Simulation.[6] Using these ontologies it is possible, for example, to describe a physical research activity (e.g. an interview) as an `opm:Process`, and how such an activity causes an `opm:Artifact` to be generated (interview notes).

Based on the requirements from our case study groups, the provenance framework should not only capture information regarding artefacts and processes, but must be able to situate these alongside people and their associated organisational structures. Friend-of-a-Friend[7] (FOAF) is an established RDF vocabulary for describing people and their social networks and we have opted to utilise this within our framework; a `foaf:Profile` is thus a subclass of `opm:Agent`. Several FOAF profiles are visible in Fig. 1, as contacts of the user (My Contacts). Organisational structures such as projects or employer institutions can also be defined, and users within *ourSpaces* may belong to several projects or groups.

Another requirement was to capture the provenance of online communication within the social network. However, the OPM specification supports limited information about the relationship between a person (`opm:Agent`) and the research process (`opm:Process`). As a result, we have integrated the social networking vocabulary SIOC[8] (Semantically-Interlinked Online Communities) within our provenance framework. Using this vocabulary, traditional provenance can be extended to incorporate social data. For example, a collaborator (defined with `foaf:worksWith`) could post a comment (`sioc:Post`) about some artefact (e.g. `opm:Paper`) uploaded by a colleague asking for some clarification about the method used to generate the data.

Another important requirement was the ability to manage users and their behaviours to ensure compliance with certain policies. For example, a user uploading a digital artefact into *ourSpaces* may be obliged (by the project PI) to provide certain information such as a geographical location. We have thus extended our provenance

[5] http://purl.org/policygrid/ontologies/provenance-generic.
[6] http://purl.org/policygrid/ontologies/provenance-simulation.
[7] http://www.foaf-project.org/.
[8] http://sioc-project.org/.

framework to define such policies as a combination of obligations, prohibitions or permissions.

We have combined the existing OWL binding of the Open Provenance Model with an OWL ontology (inspired by the work of Sensoy et al. [8]) defining the concepts introduced above. An extract of the provenance policy ontology is shown in Fig. 3 (Policies). Moreover, we make use of the SPIN ontology[9] to support the use of the SPARQL query language to specify rules and logical constraints necessary to reason about policies.

In our ontology a policy is a combination of `PolicyCondition` instances described by the property `hasCondition`. Each condition can be defined as an Obligation, Prohibition or Permission depending on the nature of the policy. We define a condition as a `spin:Construct` query describing its logic in the form of an *if-then* statement where *if* is represented by the WHERE block of the query and *then* by the CONSTRUCT block of the query (see Fig. 4). Once processed by the SPIN reasoner a *spin:Construct* can assert a new `ActionRequest` instance which is constructed as part of the query, such as the `InformationRequest` in Fig. 4. A policy in our ontology also has one or more `ActivationCondition` instances describing the activation condition of the policy via a `spin:Construct` query. As a result of an activation, the `spin:Construct` query asserts a new `PolicyActivation` instance. A `PolicyActivation` links a specific policy instance to the event (`opm:Process`) that activated the policy, e.g. a resource action `UploadResource`.

In order to reason about obligation, permission or prohibition conditions we require a reasoning mechanism able to check conditions over a provenance graph. This is done by evaluating each condition defined as a `spin:rule`. For an obligation, conditions have to be met; for a prohibition, the condition cannot be met; and for a permission, the condition might (or might not) be met.

Using this approach in *ourSpaces* we were able to implement a policy for use by the *E. coli* O157 Risk project team as illustrated in Fig. 4. The policy specifies the kind of metadata required for artefacts that will eventually be archived to the UK social science data archive—UKDA.[10] More specifically, the policy is created by the PI of the project and it is addressed to its members. The policy is activated when a person uploads an artefact.

Activation Rule:

```
CONSTRUCT {
  _:b0 a pol:PolicyActivation .
  _:b0 pol:activePolicy :UKDADocPolicy .
  _:b0 pol:basedOnEvent ?event . }
WHERE {
  ?up a vre:UploadResource .
  ?edge opm:cause ?up .
  ?edge opm:effect ?resource .
  ?resource a pggen:TabulatedData
  ?resource pggen:producedInProject ?project.
```

Obligation Rule:

```
CONSTRUCT {
  _:b0 a pol:InformationRequest .
  _:b0 pol:onDate ?date .
  _:b0 pol:requestAboutResource ?this .
  _:b0 pol:requireProperty pggen:geoCoverage.
}
WHERE {
  ?policy pol:basedOnEvent ?up .
  ?policy pol:activePolicy :UKDADocPolicy .
  ?up a vre:UploadResource .
  ?edge opm:cause ?up .
  ?edge opm:effect ?this .
  NOT EXISTS {
    ?this pggen:geoCoverage ?x .
  } .
}
```

Fig. 4. Example of activation rule and obligation rule associated with the UKDA documentation policy.

3. Authoring and accessing metadata

We have developed a web interface to make creation of metadata by the users of the VRE as intuitive as possible, allowing them to utilise a traditional web form and to create metadata automatically where possible. The form is dynamically generated based on the current user context. For example, a user uploading a paper in the context of a research project, might be required to provide additional information about the paper specific to that project. This is achieved by a background service that continuously reasons about the user context (e.g. user uploading a project resource). Inferences generated by the reasoner are used to dynamically populate and remove fields in the form. This includes determining what fields are mandatory or optional depending on the type of metadata being generated by the user.

We have also developed methods to support the creation of semantic links within communication items such as messages, comments and posts. For instance, when writing a message to a colleague, a user can refer to a person or an artefact in the system, by using @ (for people) or # (for artefacts) in combination with an autocomplete search function which returns instances from the repository. This allows users to make fine grain semantic references within unstructured artefacts such as blogs.

In *ourSpaces* a "space" acts as a container to provide access to information about a specific resource or a category of resources. In order to generate a space a number of SPARQL queries are performed over the provenance repository, extracting relevant sections of the RDF graph. In order to allow users to access metadata, we have incorporated a number of visualisation modalities at different levels. A space-based interface acts as a container and provides high-level access to information on specific resources or resource types, while a text-based interface and a graph-based interface are used to provide detailed access to the information on a specific resource. These interfaces are controlled by preferences of users and projects, which are formulated as policies according to the policy framework introduced in Section 2. Fig. 1 (bottom right) illustrates the graphical interface used to visualise provenance metadata relating to an artefact. In this example the provenance of the document *People Living with Diabetes in Grampian* is presented. By clicking the + button, the user can expand the graph in order to explore additional provenance information. Moreover, by hovering the mouse pointer over processes or artefacts the user is presented with additional information which is rendered in plain text by a Natural Language Generation (NLG) service. This service translates RDF statements into English sentences, based on the approach described by Hielkema [9]. To generate the description of a particular RDF resource, this service queries the metadata repository with the ID of that resource to retrieve all related statements. A local model is then built from that list of statements, representing the information about that resource. This model is subsequently used by the NLG service to convert the axioms into plain text, using the appropriate *language specification* files. These files (encoded in XML) describe how axioms should be translated to English. Each file represents a particular property in the ontology and contains

[9] http://spinrdf.org/spin.html.
[10] http://www.data-archive.ac.uk/.

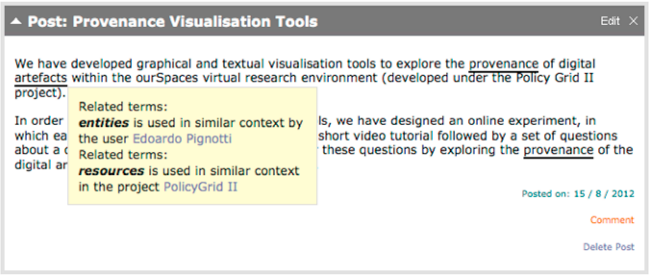

Fig. 5. Example of vocabulary mappings used in a blog post.

linguistic information about how to structure the sentence corresponding to that property (e.g. syntactic category, verb, source and target). For each file we define a dependency tree structure to represent the relationships between the different syntactic units of the sentence. This allows properties with similar syntactical structures to be aggregated together in the text. The final stage of linguistic realisation is carried out using the SimpleNLG realiser [10], which converts abstract representations of sentences into actual text using rules of grammar (morphology and syntax).

ourSpaces is designed to support collaboration within multidisciplinary research groups and users from our case study teams often stressed that people from different backgrounds tend to have different information presentation preferences. Empirical evidence also suggests that there is a need to adapt information interfaces to users and their context [1]. To address this issue, we have used policies to prioritise between data presentation strategies (e.g. graphical or textual visualisation to explore the provenance graph), as well as to control the content of the text generated by the NLG service [11]. The latter is a kind of rule-based content determination [12]. For example, the principal investigator of a project might want to protect the identity of the person who transcribed an artefact from users who are not members of that project. Such a preference can be expressed by constructing a policy that triggers an action request to remove the relevant property *transcribedBy* from the extracted RDF graph, if the user visualising the description of a *Transcript* is not a member of the project which produced that artefact. In this manner, the NLG service combined with the policy framework allows the system to generate descriptions aligned to the user's context and preferences.

Our case study groups also raised the use of discipline specific terms in communications (emails, blogs, comments) as a major issue faced by multidisciplinary teams. However, creating and maintaining personal ontologies for each user containing these terms is a difficult task due to the number of users, and the dynamic nature of such personal vocabularies. The discipline specific vocabulary of an individual is reflected in his or her writing, especially within documents produced as part of their research. Therefore, we used a corpus based Distributional Similarity (DS) [13] approach to create mappings between different personal vocabularies. Users without a personal corpus of documents were assigned mappings based on the documents available in their projects. We have studied many DS algorithms [14] and implemented a Personal Lexicon Service (PLS) in *ourSpaces* to create a set of vocabulary mappings for use during search and online communication.

The search function within *ourSpaces* is key word based and uses SPARQL to query resources in the triple-store. The key words within the search query are checked against the lexical mappings and if a match is found, the SPARQL query is re-written by adding more filters corresponding to each similar term identified in the vocabulary mappings.

The use of discipline specific terms is also common in online communication such as emails, blogs and comments. In order to indicate possible similar terms for such words, each communication item is parsed and each word is checked against the user's vocabulary mappings. As shown in Fig. 5, if a match is found the term is underlined, allowing a user to see similar terms and their associations in the system.

4. Lessons learnt

In this section we illustrate some of the lessons learnt during the deployment of *ourSpaces* with different case-study groups, using evidence gathered from real data provided by the users of the system. The sources of data used to inform development of the system were: (a) metadata about resources, people, events, projects, etc. stored in the repository; (b) the MySQL database containing user account information and system logs; and (c) interviews and focus groups with *ourSpaces* users. Throughout the development of the system, user feedback contributed directly to the introduction of new features, and to changes in existing system functionality.

Between 2009 and 2012 the *ourSpaces* VRE went through two major software revisions. As of August 1st 2012 there were 254 `foaf:profiles` defined in *ourSpaces* of which 183 were registered users.[11] The social network in the VRE contained 204 links (`foaf:knows`) between user accounts. Users had created 49 projects and sub-projects, with 92% of the accounts in *ourSpaces* being a member of at least one project. A total of 435 research artefacts had been uploaded, with the metadata repository containing 14 680 triples describing 4388 entities. 63 distinct classes and 105 distinct properties had been used to describe entities in the repository, utilising 33% of the classes and 40% of the properties defined in the supporting ontologies.

In the early stages of *ourSpaces*, users were required to provide a great deal of metadata about research artefacts in order to guarantee a detailed provenance record. The result was that few users went through the effort of providing such metadata and only a small number of artefacts were uploaded. Following feedback from users we adopted a more relaxed approach, where very few mandatory fields were required and the users themselves had the option to choose which metadata to add to the artefact. This resulted in more artefacts being uploaded, but with a much sparser metadata record. To illustrate this issue we now present some summary data collected from our metadata repository. While selecting the type of artefact to upload, a user is presented with a list of mandatory fields depending on the class selected. Types of artefact are shown on a tree-like structure, where `Artefact` is the root class and more specialised types are presented up to two levels down the class hierarchy. From our analysis, 20% of the artefacts in *ourSpaces* have been associated with the root class, 71% with subclasses of `Artefact` and 9% with classes at the next level in the hierarchy. The classes nearest to the root class have significantly less mandatory properties than others at lower levels in the hierarchy.

We aimed to solve the problem of sparse metadata by allowing people (with the right authority) to define policies in *ourSpaces* to specify the mandatory information required when uploading a research artefact. In this way, the request for additional information originated with a person rather than the system, e.g. the principal investigator of a project. After introducing policies into the *E. coli* O157 Risk project, the average number of RDF triples used to describe research artefacts increased from 9 to 32. However, policies and particularly the SPIN reasoner require additional computational resources, resulting in a delay when a user is using a web form.

[11] FOAF profiles without an *ourSpaces* account are created by the system when a user specifies authors of documents.

We have analysed the logs from the policy reasoning service in order to assess the performance of the reasoner based on the policies generated by the users in the system. The hardware used for the deployment of the *ourSpaces* services and repositories consists of three Sun Fire X4100 M2 servers with two dual-core AMD Opteron 2218 processors and 32 GB of memory. Based on 2895 runs of the reasoner logged by the system, the average time to run a policy was 1.9 s. Miller [15] and Card [16] argue that system response times of less than 10 s do not compromise the user's attention on the current task. Based on analysis of the logs we determined that time taken to reason about a policy was acceptable. A similar analysis has also been conducted for the use of policies by the Natural Language Generation service. In spite of the overhead associated with the use of the policy reasoner, text is generated and appears within an average of 200 ms.

Based on feedback from users we discovered that the graphical visualisation of provenance metadata served a useful function as a means to validate the metadata uploaded via the form based interface. This was especially useful as the UKDA policy required users to provide detailed information about the methods used to generate a research artefact. The graphical interface was also used by representatives of the UKDA to review the data (and metadata) uploaded by project members. Screen-grabs of the provenance graphs generated by our system have been used by the UKDA as part of their internal documentation describing the project archive.

As a part of a wider *ourSpaces* evaluation, we carried out an experiment to evaluate different distribution similarity algorithms that could be used in creating automatic vocabulary mappings in the personal lexicon service [14]. This study showed that the precision and recall of vocabulary mappings increased with the size of the personal corpus, and that the quality of the mappings fall below useable levels if the personal corpus contains less than 6500 dependency relations with nouns.

A crucial part of maintenance of the system is to take into account the requirements of new user groups. When a new group joins *ourSpaces*, it is normal to expect that they might have their own way to describe research artefacts and processes. The *ourSpaces* provenance framework can easily be extended in order to accommodate new domain-specific provenance concepts. This issue was detected very early during the development of the system and we therefore designed it in such a way that new domain ontologies could be integrated into the system without the need to change the underlying source code. Our implementation of the policy framework also allows new policies to be integrated without the need for alterations to the system. Although we do not yet have a specific tool for designing policies, a standard ontology editor can be used.

5. Adapting *ourSpaces* to PROV-O

Since the *ourSpaces* system was developed and deployed, a new provenance specification (W3C PROV) has emerged.[12] The PROV model is similar to OPM as it describes an *Entity* (physical, digital, conceptual); an *Activity* (something that occurs over a period of time and acts upon or with entities); and an *Agent* (something that bears some form of responsibility for an activity). Adapting the *ourSpaces* implementation to support the PROV specification was beyond the scope of the original project. However, we have since investigated a solution to support automatic generation of PROV compatible provenance from the *ourSpaces* record. In order to do this we make use of the OPM profile specification [7]. An OPM profile is intended to define a specialisation of OPM while maintaining the compatibility with the semantics of the OPM provenance model. Following the OPM profile conventions it is therefore possible to define an OWL based extension of our provenance framework that contains logical rules that can be used to generate W3C PROV-O[13] compatible provenance records. This can be done by introducing ontology mappings and rules that a reasoner can apply over an *ourSpaces* provenance graph to infer additional PROV-O properties. A set of such mappings and rules are summarised below:

- An instance of `opm:Artifact` is also an instance of `prov:Entity`;
- An instance of `opm:Process` is also an instance of `prov:Activity`;
- An instance of `opm:Agent` is also an instance of `prov:Agent`;
- An instance of `opm:WasGeneratedBy` is also an instance of `prov:Generation`.
  The statement `[?wgb opm:causeWasGeneratedBy ?ar; ?wgb opm:effectWasGeneratedBy ?p]` is also described as `[?ar prov:wasGeneratedBy ?p; ?ar prov:qualifiedGeneration ?wgb; ?wgb prov:activity ?p]`;
- An instance of `opm:Used` is also an instance of `prov:Usage`.
  The statement `[?u opm:causeUsed ?ar; ?u opm:effectUsed ?p]` is also described as `[?p prov:used ?ar; ?p prov:qualifiedUsage ?u; ?u prov:entity ?ar]`;
- An instance of `opm:WasControlledBy` is also an instance of `prov:Association`.
  The statement `[?wcb opm:causeWasControlledBy ?ag; ?wcb opm:effectWasControlledBy ?p]` is also described as `[?p prov:wasAssociatedWith ?ag; ?p prov:qualifiedAssociation ?wcb; ?wcb prov:agent ?ag]`;
- An instance of `opm:WasDerivedFrom` is also an instance of `prov:Derivation`.
  The statement `[?wdf opm:causeWasDerivedFrom ?a1; ?wdf opm:effectWasDerivedFrom ?a2]` is also described as `[?a2 prov:wasDerivedFrom ?a1; ?a2 prov:qualifiedDerivation ?wdf; ?wdf prov:entity ?a1]`;
- An instance of `opm:WasTriggeredBy` is also an instance of `prov:Communication`.
  The statement `[?wtb opm:causeWasTriggeredBy ?p1; ?wdf opm:effectWasTriggeredBy ?p2]` is also described as `[?p2 prov:wasInformedBy ?p1; ?p2 prov:qualifiedCommunication ?wdf; ?wdf prov:activity ?p1]`.

We have created an OWL ontology[14] implementing the mappings and rules described above in the form of SPIN-SPARQL rules. We have tested the ontology using the SPIN reasoner with a sample provenance graph extracted from the *ourSpaces* repository. The resulting provenance graph was determined to be valid PROV-O by manually testing it against the PROV-O specifications and constraints.

6. Related work

Semantic Web technologies have been applied to the development of a number of virtual research environments. The *myExperiment* system [17] enables people to share digital objects associated with their research. The notion of *research objects* is used in *myExperiment* to provide a container for semantic aggregation of

[12] http://www.w3.org/TR/prov-overview/.

[13] http://www.w3.org/TR/prov-o/.

[14] http://www.purl.org/policygrid/ontologies/prov-mappings.

resources produced and consumed by common services. We have investigated the role of policies in *myExperiment* and have concluded that their use is limited to access control via the Simple Network Access Rights Management (SNARM) ontology.[15] VIVO [18] is an open source application designed to manage metadata about scholarly activities from different institutions for the purpose of information discovery. The system supports semantic linking of resources across different disciplines and institutions and utilises web-based graphical tools to visualise linked data about research networks, papers and grants. *VIVO* defines a custom policy framework to implement role-level authorisation rules but as with *myExperiment* this appears to be limited to access control.

Semantic Web approaches have also been used in enterprise knowledge management tools [19]. For example, the IBM *WebSphere* Portal [20] uses ontologies to support different aspects of document management such as tagging and searching. All these systems (including *ourSpaces*) employ Semantic Web technologies to provide a representational framework that can be used across different domain applications. However, the main difference between *ourSpaces* and the environments discussed above is that it utilises policy reasoning to control the behaviour of users and services. This allows us to adapt the environment to meet domain-specific requirements without changing the logic behind services.

One of the other core aspects of *ourSpaces* is the support for capturing the provenance of research artefacts and processes, inspired by similar approaches in use in other application domains [21]. Groth et al. [22] discuss general requirements for provenance on the Web, focusing on three key aspects: the content of provenance, the management of provenance records, and the uses of provenance information. We argue that the provenance framework in *ourSpaces* aligns with many of the provenance dimensions discussed by Groth et al. such as object, attribution, process, versioning, entailment, publication, access, dissemination, understanding, interoperability, comparison, trust, imperfections and accountability. Most notably *ourSpaces* addresses the issue of incomplete provenance (imperfections) using its policy reasoning approach.

7. Conclusions

In this paper we have introduced the *ourSpaces* virtual research environment focusing on three elements which makes use of Semantic Web technologies: a provenance framework, a collection of services for creating and visualising metadata, and a policy reasoning service.

The process of designing *ourSpaces* with real user groups gave us the opportunity to learn directly about the advantages and disadvantages of using an ontology-based approach for representing and managing research metadata. The use of linked data makes certain aspects of information discovery (e.g. identifying related resources) and information presentation possible within the *ourSpaces* environment. Linked data also allows components such as the natural language visualisation service to exploit this model to allow users to explore the provenance graph. Users in *ourSpaces* told us that they found this aspect of the system useful (e.g. UKDA staff were able to identify the methods that were used to create a research artefact).

However, in a such a complex semantic environment we discovered that there is a trade-off between flexibility and performance. For example, we discovered that storing large amounts of text (e.g. blogs) in RDF impacted on the performance of the semantic components of the system. We thus adopted an approach which combined lightweight RDF metadata with a relational database.

Another disadvantage we have identified is that users are often not prepared to go through the effort required to provide the metadata required by the system. While we designed the ontologies with very few mandatory properties, we had to introduce a policy reasoning component to enforce certain policies (e.g. to enforce a request by a project P.I. for mandatory information about a research artefact). Of course, depending on the nature of the policy, such reasoning does require additional computational resources which can impact on overall system responsiveness.

We are currently working with the Scottish Environmental Protection Agency to transfer some of the semantic technology developed as part of *ourSpaces* to extend their existing SEWeb portal.[16]

Acknowledgement

The work described here was supported by the UK Economic and Social Research Council (ESRC) under the Digital Social Research programme; award RES-149-25-1075.

References

[1] A. Butterworth, T. Reimer, Virtual research environment collaborative landscape study, Tech. Rep., JISC, 2010.
[2] T. Berners-Lee, J. Hendler, O. Lassila, The semantic web, Sci. Am. 284 (5) (2001) 34–43.
[3] E. Pignotti, P. Edwards, Using web services and policies within a social platform to support collaborative research, in: Working Notes of AAAI 2012 Stanford Spring Symposium on Intelligent Web Services Meet Social Computing, 2012.
[4] T. Heinis, G. Alonso, Efficient lineage tracking for scientific workflows, in: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, (SIGMOD'08), ACM, 2008, pp. 1007–1018.
[5] P. Edwards, E. Pignotti, A. Eckhardt, K. Ponnamperuma, C. Mellish, T. Bouttaz, Ourspaces—design and deployment of a semantic virtual research environment, in: P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, J. Euzenat, M. Hauswirth, J.X. Parreira, J. Hendler, G. Schreiber, A. Bernstein, E. Blomqvist (Eds.), International Semantic Web Conference, in: Lecture Notes in Computer Science, vol. 7650, Springer, 2012, pp. 50–65.
[6] C. Bizer, T. Heath, T. Berners-Lee, Linked data—the story so far, Int. J. Semant. Web Inf. Syst. (IJSWIS) 5 (3) (2009) 1–22.
[7] L. Moreau, J. Freire, J. Futrelle, R.E. McGrath, J. Myers, P. Paulson, The open provenance model: An overview, in: J. Freire, D. Koop, L. Moreau (Eds.), IPAW, in: Lecture Notes in Computer Science, vol. 5272, Springer, 2008, pp. 323–326.
[8] M. Sensoy, T.J. Norman, W. Vasconcelos, K. Sycara, Owl-polar: semantic policies for agent reasoning, in: International Semantic Web Conference, 2010.
[9] F. Hielkema, Using natural language generation to provide access to semantic metadata (Ph.D. thesis), University of Aberdeen, 2010.
[10] A. Gatt, E. Reiter, Simplenlg: a realisation engine for practical applications, in: Proceedings of the 12th European Workshop on Natural Language Generation, ENLG'09, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 90–93.
[11] T. Bouttaz, E. Pignotti, C. Mellish, P. Edwards, A policy-based approach to context dependent natural language generation, in: Proceedings of the 13th European Workshop on Natural Language Generation, Association for Computational Linguistics, Nancy, France, 2011, pp. 151–157.
[12] E. Reiter, R. Dale, Building Natural Language Generation Systems, Cambridge University Press, New York, NY, USA, 2000.
[13] L. Lee, Measures of distributional similarity, in: 37th Annual Meeting of the ACL, Stroudsburg, PA, USA, 1999, p. 25.
[14] K. Ponnamperuma, C. Mellish, P. Edwards, Using distributional similarity for identifying vocabulary differences between individuals, in: Workshop on Computational Approaches to the Study of Dialectal and Typological Variation, ESSLLI 2012, Opole Poland, 2012.
[15] R.B. Miller, Response time in man–computer conversational transactions, in: Proceedings of the Joint Computer Conference 1968, ACM, New York, NY, USA, 1968, pp. 267–277.
[16] S.K. Card, G.G. Robertson, J.D. Mackinlay, The information visualizer, an information workspace, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Reaching Through Technology, CHI'91, ACM, New York, NY, USA, 1991, pp. 181–186.
[17] D.D. Roure, C. Goble, R. Stevens, The design and realisation of the virtual research environment for social sharing of workflows, Future Gener. Comput. Syst. 25 (5) (2009) 561–567. http://dx.doi.org/10.1016/j.future.2008.06.010.

15 http://rdf.myexperiment.org/ontologies/snarm/.

16 http://www.environment.scotland.gov.uk/.

[18] K. Börner, M. Conlon, J. Corson-Rikert, Y. Ding, VIVO: A semantic approach to scholarly networking and discovery, Synth. Lect. Semant. Web Theory Technol. 7 (1) (2012) 1–178.

[19] G. Aastrand, R. Celebi, L. Sauermann, Using linked open data to bootstrap corporate knowledge management in the organik project, in: Proceedings of the 6th International Conference on Semantic Systems, I-SEMANTICS'10, ACM, New York, NY, USA, 2010, pp. 18:1–18:8. http://dx.doi.org/10.1145/1839707.1839730.

[20] A. Kreiser, A. Naurex, F. Bakalov, A web 3.0 approach for improving tagging systems, in: Proceedings of the International Workshop on Web 3.0: Merging Semantic Web and Social Web (in conjunction with the 20th International Conference on Hypertext and Hypermedia 2009), vol. 467, Torino, Italy, 2009.

[21] Y. Gil, J. CVheney, P. Groth, O. Hartig, S. Miles, L. Moreau, P. Pinheiro da Silva, Provenance XG final report, Tech. Rep., World Wide Web Consortium, 2010.

[22] P. Groth, Y. Gil, J. Cheney, S. Miles, Requirements for provenance on the web, Int. J. Digit. Curation 7 (1) (2012) 39–56.