mSpace meets EPrints: a Case Study in Creating Dynamic Digital Collections

m.c. schraefel, Daniel A. Smith, Leslie A. Carr
IAM Group, Electronics and Computer Science
University of Southampton, UK
[mc, ds, lac] @ ecs.soton.ac.uk

ABSTRACT

In this case study we look at issues involved in (a) generating dynamic digital libraries that are on a particular topic but span heterogeneous collections at distinct sites, (b) supplementing the artefacts in that collection with additional information available either from databases at the artefact's home or from the Web at large, and (c) providing an interaction paradigm that will support effective exploration of this new resource. We describe how we used two available frameworks, mSpace and EPrints to support this kind of collection building. The result of the study is a set of recommendations to improve the connectivity of remote resources both to one another and to related Web resources, and that will also reduce problems like co-referencing in order to enable the creation of new collections on demand.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**] Collection, Dissemination, Standards, User issues; H.5.2 [**User Interfaces**] Ergonomics, Evaluation/methodology, User-centered design.

General Terms

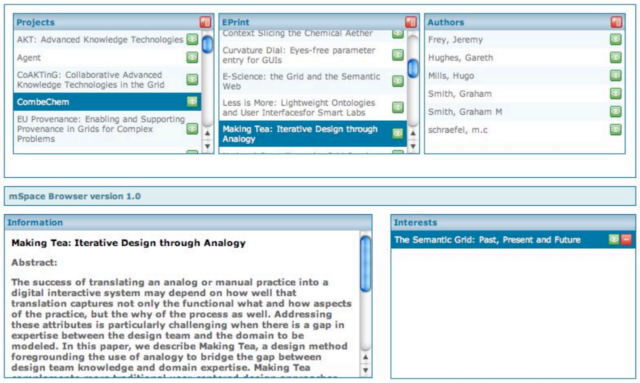
Design, Experimentation, Human Factors

Keywords

mSpace, EPrints, digital libraries, association, user interface, social factors

1. INTRODUCTION

In the UK, there is a new requirement being proposed by the main Research Council (RCUK) that all organizations make research publications resulting from RCUK funding publicly available. The Council is responsible for a number of discipline-specific granting bodies such as the EPSRC in the physical sciences and the MSRC in medicine. A call for proposals may be from a programme for one granting body, or may be programme that spans councils. One such programme involving multiple councils is the e-Science call, (similar to the Cyber Infrastructure programme in the NSF in the the US). There was interest in the e-Science program to promote the research outputs of the program; it wanted to be able to promote the papers from all projects funded by the call. To represent this output, a system would need to gather information from all e-Science projects across all councils. As the projects involved multiple universities, this would also mean gathering data across institutions.


Projects
AKT: Advanced Knowledge Technologies
Agent
CoAKTinG: Collaborative Advanced Knowledge Technologies in the Grid
CombeChem
EU Provenance: Enabling and Supporting Provenance in Grids for Complex Problems
EPrint
Context Slicing the Chemical Aether
Curvature Dial: Eyes-free parameter entry for GUIs
E-Science: the Grid and the Semantic Web
Less is More: Lightweight Ontologies and User Interfacesfor Smart Labs
Making Tea: Iterative Design through Analogy
Authors
Frey, Jeremy
Hughes, Gareth
Mills, Hugo
Smith, Graham
Smith, Graham M
schraefel, m.c
mSpace Browser version 1.0
Information
Interests
The Semantic Grid: Past, Present and Future

Making Tea: Iterative Design through Analogy

Abstract:

The success of translating an analog or manual practice into a digital interactive system may depend on how well that translation captures not only the functional what and how aspects of the practice, but the why of the process as well. Addressing these attributes is particularly challenging when there is a gap in expertise between the design team and the domain to be modeled. In this paper, we describe Making Tea, a design method foregrounding the use of analogy to bridge the gap between design team knowledge and domain expertise. Making Tea

Figure 1. mSpace interface on EPrints e-Science collection

This past year, our group ran a pilot project with the RCUK to investigate how to bring together these diverse outputs into one virtual location, such that people coming to the collection could easily explore the projects, the people associated with them, and especially, the papers produced in a project. Information on authors would lead to the author's Web site; publication info to the publisher, and so on. In other words, we wanted the patrons of the collection to be able to explore the associated digital geography of the collection. To this end, we combined the mSpace software framework and interaction model for exploring heterogeneous Web sources [19] with EPrints open access repository software [9]. mSpace provides both a software framework to correlate heterogeneous Web sources and an interface (see Figure 1) to explore them. EPrints is digital repository software, the primary objective of which is to provide a deposit-and-view interface for a local context (an institution or a subject discipline) that will participate in global collections and services through an OAI interface (we describe OAI below).

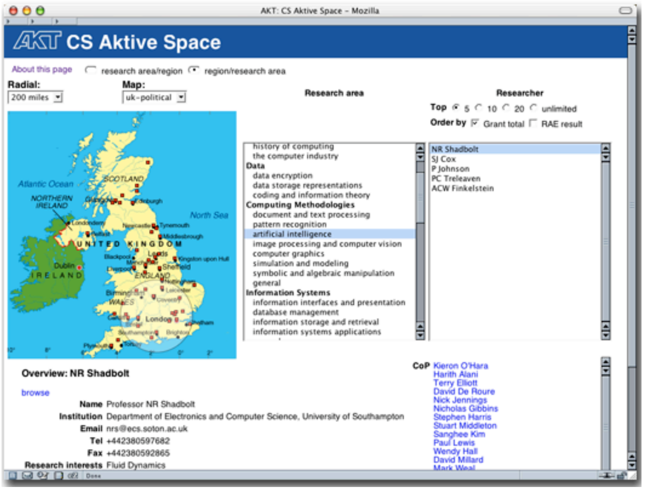
In the following case study, we present our efforts to generate a cross-institution/cross-council digital collection using the mSpace/EPrints approach. We overview related work, describe the rationale for choosing these components, describe how mSpace and EPrints interoperate, and overview the specific implementation, its strengths and it limitations. We present the largely social, rather than technical gaps that these approaches exposed in our efforts to deliver on demand collections. Based on the lessons learned from this project, we propose a set of recommendations to enable on demand, richly explorable, dynamic, digital collections.

JCDL 2006, June, 2006, City, State, Country.

2. RELATED WORK

In our prototype, we focused on sites that used EPrints digital repository software because the software would give us metadata stream about the papers. While many universities' individual Schools and Departments provided lists of papers published in their groups, not all used mechanisms that could produce harvestable metadata about the publication, in particular OAI metadata [12]. The purpose of OAI is to promote archive interoperability rather than individual archive functionality. Currently, the interoperability is principally based on Dublin Core metadata [24]. Dublin Core is a standardized XML schema for expression and sharing document metadata. It specifies predicates such as "name", "title", "description" for creating XML descriptions of documents. These descriptions have the potential to be filtered against a criteria and amalgamated into a new collection.

EPrints is not the only software to provide harvestable OAI output. DSpace [3] and Greenstone [25] are also exemplars in the digital archive space. Greenstone is designed specifically as a Digital Library: trained librarians deposit artifacts and create collections around those artificats. EPrints and DSpace are both digital repositories: with EPrints, authors deposit their own artefacts and the software makes collections based on the values of the metadata. DSpace is likewise a self-archive style repository with more of an emphasis on collections. Unlike digital libraries, repository software may well have a particular integration with authors' workflows and tasks such as maintaining up-to-date CVs or providing administrative form filling for research audits, where these tasks are facilitated by an OAI service to provide simple federated search [11] or more involved citation analysis functionality [7].


AKT: CS Aktive Space – Mozilla

AKT CS Aktive Space

About this page ○ research area/region ● region/research area

Radial: 200 miles

Map: uk-political

Research area

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

Researcher

Top ● 5 ○ 10 ○ 20 ○ unlimited

Order by ☑ Grant total □ RAE result

NR Shadbolt
SJ Cox
P Johnson
PC Treleaven
ACW Finkelstein

Overview: NR Shadbolt

browse

Name	Professor NR Shadbolt
Institution	Department of Electronics and Computer Science, University of Southampton
Email	nrs@ecs.soton.ac.uk
Tel	+442380597682
Fax	+442380592865
Research interests	Fluid Dynamics

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

Done

Figure 2. csAKTiveSpace (cs.aktivespace.org) showing highest grant earners in Artificial Intelligence in the south of the UK. Author selected, shows contact and community of practice information of the author.

Dublin Core and OAI-PMH [12] provide standard ways to format information, but the standard XML formatting of the information lacks a mechanism to properly apply semantics to the information, that is, it lacks the mechanisms to describe *the relations* possible in the metadata information: how information about a paper (held in an archive) relates to information about an author (held by an institution; published to the Web). One approach to creating such connections is to take advantage of new Web protocols which fall under the umbrella of the Semantic Web [5]. In the CSAKTiveSpace project [21], for example, a precursor to mSpace, the project used Semantic Web-enabled approaches to harvest data from databases, Web sources and resources which publish data in a native Semantic Web format in order to be able to explore associated questions about research activities in the UK. As shown in Figure 2, one can choose a region on a map of the UK, for instance, and see who the researchers are in that region, in a given area of computer science. They can then filter the results based on highest funding total. When a researcher is selected in the interface, one also sees the researcher's contact information, as well as their community of practice, derived from their research collaborations. A related project, AKTive Futures [8] used a similar approach to harvest resources on oil production in order to be able to explore both stories about and analysis of oil production over time, represented graphically. If one saw a dip in a region's oil production at a particular year, selecting that point would connect to stories about that region at that time. Likewise, multiple regions' production could be compared over time.

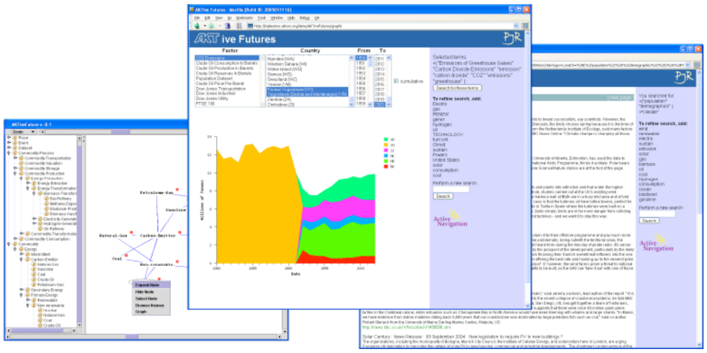
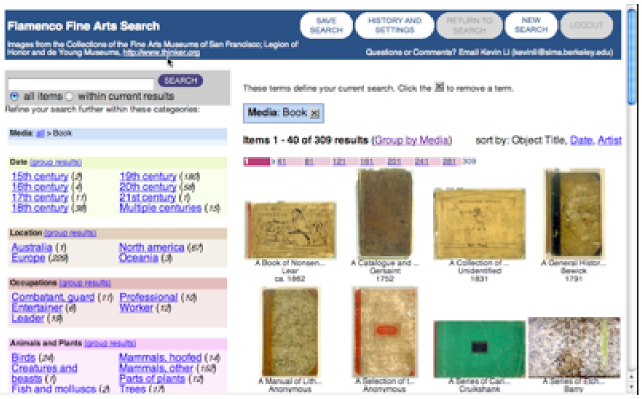


Figure 3. AKTive Futures, showing graphing of energy data for several regions over a selected period connected with news stories semantically associated against any point selected in that period.

In the same way that Semantic Web technologies enables exploration over the association of data about researchers and energy production, it can facilitate statements to be made about a much larger universe of information than what is captured in the OAI data, including information about the authors, institutions, journals, conferences and other actors or stakeholders in the scientific communication and publication process. In our prototype, therefore, we planned to use a Semantic Web layer, provided by the mSpace framework, to integrate supplementary sources of information with the OAI data.

One of the critical components for exploring these richly associated resources, however, is an effective user interface. Raw keyword search interfaces are unquestionably effective when a person knows what they want to get with a degree of certainty. When material goes digital, however, it often goes invisible. When a person is less certain of the data they want, or when they are interested in other kinds of search, such as exploring relations of one thing with another, the common keyword interfaces of most digital archive systems are not as effective for these kinds of explorations.

While open access archives have focused on the mechanisms of collection and storage for institutional repositories, complementary research has looked at new paradigms for improving access to making the treasures within a collection accessible. One paradigm in particular, "faceted browsing" [http://user-experience.org/uefiles/facetedbrowse/], is proving considerably effective. Faceted browsing presents categories in a domain for selection. Each selection acts as a filter organizing and limiting the sub-categories displayed.



Images from the Collections of the Fine Arts Museums of San Francisco; Legion of Honor and de Young Museums, http://www.thinker.org

Questions or Comments? Email Kevin Li (kevinli@sims.berkeley.edu)

● all items ○ within current results

Refine your search further within these categories:

Media: all > Book

Date (group results)
15th century (2) 19th century (180)
16th century (4) 20th century (58)
17th century (11) 21st century (1)
18th century (38) Multiple centuries (15)

Location (group results)
Australia (1) North america (87)
Europe (220) Oceania (3)

Occupations (group results)
Combatant, guard (11) Professional (10)
Entertainer (8) Worker (12)
Leader (10)

Animals and Plants (group results)
Birds (24) Mammals, hoofed (14)
Creatures and beasts (1) Mammals, other (162)
Parts of plants (12)
Fish and molluscs (4) Trees (17)

These terms define your current search. Click the ☒ to remove a term.

Media: Book ☒

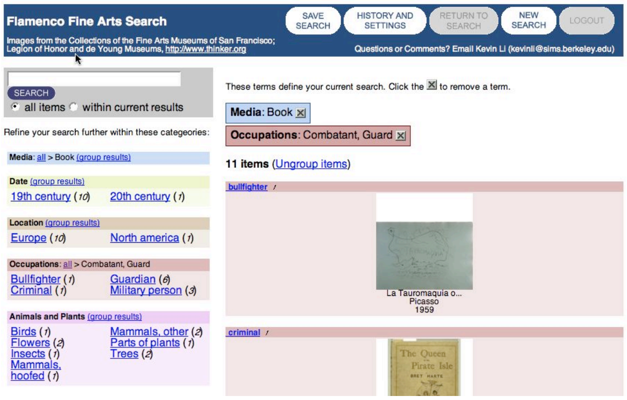
Items 1 - 40 of 309 results (Group by Media) sort by: Object Title, Date, Artist

1 41 81 121 161 201 241 281 309

A Book of Nonsen... Lear ca. 1862
A Catalogue and ... Gersaint 1752
A Collection of ... Unidentified 1831
A General Histor... Bewick 1791
A Manual of Lith... Anonymous
A Selection of t... Anonymous
A Series of Cari... Cruikshank
A Series of Etch... Barry

Figure 4. Flamenco Browser.

In the Flamenco project, digital library patrons are presented first with a table of categories, subcategories and the number of instances in a subcategory. The next page presents a further sub-categorization (Figure 4, left pane) of the selected category as well as a sample visual representation of the main category. Further selections in the subcategories act as filters on the objects displayed. This approach is similar to the Topia project [18] which enables criteria to be selected and then matching artifacts to be determined. A key difference between Topia and Flamenco is that Topia uses Semantic Web protocols to enable inferences across the museum's data. For instance, Vermeer's painting a Kitchen Maid will show up in the categories for which it has been explicitly categorized, such as Domestic Interiors and Women, but it will also be available in another category, People, for which it has not explicitly been tagged: the ontology underpinning the categories of this collection enables the other categories to be inferred automatically, and thus the artifacts can themselves appear in a richer variety of contexts. Endeca (endeca.com), Mercedes "select a model" [17] and Yahoo's camera selector [26] each use a similar approach in its dynamic generation of facets.



Images from the Collections of the Fine Arts Museums of San Francisco; Legion of Honor and de Young Museums, http://www.thinker.org

Questions or Comments? Email Kevin Li (kevinli@sims.berkeley.edu)

● all items ○ within current results

Refine your search further within these categeories:

Media: all > Book (group results)

Date (group results)
19th century (10) 20th century (1)

Location (group results)
Europe (10) North america (1)

Occupations: all > Combatant, Guard
Bullfighter (1) Guardian (6)
Criminal (1) Military person (3)

Animals and Plants (group results)
Birds (1) Mammals, other (2)
Flowers (2) Parts of plants (1)
Insects (1) Trees (2)
Mammals, hoofed (1)

These terms define your current search. Click the ☒ to remove a term.

Media: Book ☒

Occupations: Combatant, Guard ☒

11 items (Ungroup items)

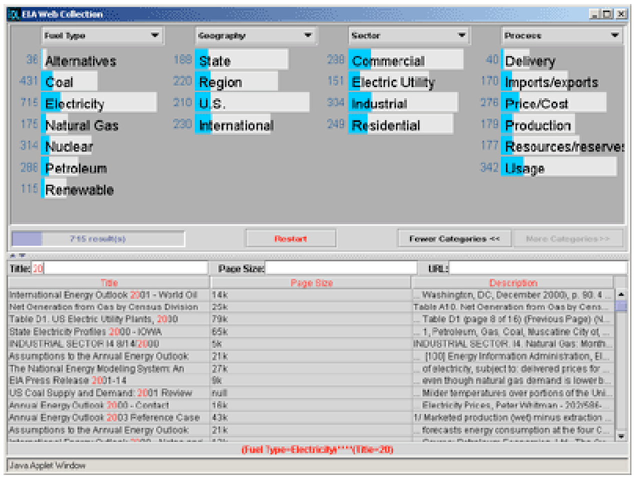
bullfighter 1

La Tauromaquia o... Picasso 1959

criminal 1

Figure 5: Refined selection in Flamenco Browser

These visualizations have several advantages over keyword search, not the least of these is that rather than confronting a person with an empty text box, they present a series of attributes (categories) that make up the domain to help people orient their explorations. One of the disadvantages of the above implementations of the facet approach is that the previous context is usually erased from view when a new facet is selected as the current focus. As shown in Figure 5, a selection from elements in Figure 4 (a selection of books on a particular topic in the database) removes the previous context from view. The effects of losing context means increased cognitive load: rather than seeing the previous context and thus being able to recognize it, one has to work to remember its detail. Likewise, with interfaces that rely on the web to make single click selections, a time delay is introduced as a call out to the network must be answered. Such a delay can mean the difference between a gesture, like scanning multiple open books on a desk, and having to get up, grab a new book, open it, look at it, put it away each time a new book is requested.



Fuel Type		Geography		Sector		Process	
36	Alternatives	188	State	238	Commercial	40	Delivery
431	Coal	220	Region	151	Electric Utility	170	Imports/exports
715	Electricity	210	U.S.	304	Industrial	276	Price/Cost
175	Natural Gas	230	International	249	Residential	179	Production
314	Nuclear					177	Resources/reserve:
288	Petroleum					342	Usage
115	Renewable						

715 result(s)

Title: 20 Page Size: URL:

Title	Page Size	Description
International Energy Outlook 2001 - World Oil	14k	... Washington, DC, December 2000), p. 90. 4 ...
Net Generation from Gas by Census Division	25k	Table A10. Net Generation from Gas by Cens...
Table D1. US Electric Utility Plants, 2000	79k	... Table D1 (page 8 of 16) (Previous Page) (N...
State Electricity Profiles 2000 - IOWA	65k	... 1, Petroleum, Gas, Coal, Muscatine City of, ...
INDUSTRIAL SECTOR I4 8/14/2000	5k	INDUSTRIAL SECTOR. I4. Natural Gas: Month...
Assumptions to the Annual Energy Outlook	21k	... [100] Energy Information Administration, EI...
The National Energy Modeling System: An	27k	... of electricity, subject to: delivered prices for ...
EIA Press Release 2001-14	9k	... even though natural gas demand is lower b...
US Coal Supply and Demand: 2001 Review	null	... Milder temperatures over portions of the Uni...
Annual Energy Outlook 2000 - Contact	16k	... Electricity Prices, Peter Whitman - 202/586-...
Annual Energy Outlook 2003 Reference Case	43k	1/ Marketed production (wet) minus extraction ...
Assumptions to the Annual Energy Outlook	21k	... forecasts energy consumption at the four C...

(Fuel Type=Electricity)***(Title=20)

Figure 6. Rave Browser

In the Relational Browser [27] (aka the Rave Browser) a different approach is taken to artefact exploration and facet representation. Here, relations among artifacts are exposed dynamically while context is preserved. The Rave browser is implemented as a Java ap, enabling it to use more sophisticated UI components than basic HTML.

The Rave Browser works on a dataset with a small number of relations (typically 3-5) and takes the approach of showing all of the possible values of that relation at all times, and showing the effect of one filter on the number of returned documents matching the other values. In Figure 6, the Rave browser has four relations, "Fuel Type", "Geography", "Sector" and "Process", laid out as columns. The values of those relations are shown as the rows in these columns, e.g. for "Geography", the possible values are "State", "Region", "US" and "International". Each one of these values has a variable-length bar associated with it, which represents the proportion of documents that exhibit this relationship. A mouse hover over a particular bar limits the bars in the other columns as if this filter were applied, showing the reduced number of matches that would remain if that filter were fixed. Clicking on the bar fixes the filter on. As many filters as desired can be set over as many of the relations as are required. The system will then fetch a listing of documents matching the criteria. Figure 6 shows the browser in use. A selection has been made and the bars show what proportion of each value the matching documents meet that selection, compared to the entire library's documents that match that value.

The Rave browser demonstrates a means of maintaining context while rapidly being able to shift focus. Its limitation currently is the small number of fixed relations it can make available before the approach ceases to be as effective. **As well, i**n each of the above cases, the interfaces have been used on single, well-controlled collections. In order to support dynamically generated cross-archive collections, our approach needs to span multiple, heterogeneous sources in a visually effective manner and be able to support the inclusion of new data sources effectively.

3. MSPACE MEETS EPRINTS

In our prototype, we wanted to take advantage of OAI metadata across a range of sources, supplement this data from non-OAI sources, and present this related information for effective exploration. mSpace and EPrints gave us the tools to explore this space: (1) mSpace provides an infrastructure for managing the metadata representing the holdings of diverse collections made available via OAI services/gateways and other metadata sources, (2) EPrints acts as the OAI data provider, and (3) mSpace also provides an interaction model that can be wrapped over the resulting data to enable it to be explored by collection patrons. In this section we describe each component in turn.

3.1 EPrints Repository

EPrints as we have stated, is tailored for self-archiving of research papers by their authors. Self-archiving is the technique supported by proponents of Open Access (OA) [10], in order to make research open and available. EPrints is a system that can be used by institutions to allow their researchers to disseminate their research papers on the institutions' Web sites, for example http://eprints.ecs.soton.ac.uk.

EPrints we discovered is particularly useful for the requirements set out in this project as it allows for the EPrints software administrators to optionally hook-in another data source, such as an institutional database, to the repository so that a greater level of metadata can be associated with the papers in the library. Data about projects and people for example, which many institutions already have in their databases, can therefore be associated with paper submissions. While this information is not available through the OAI-PMH gateway, it is available in a public-facing XML dump from the repository that provides the information in a standardized way using EPrints own schema.

3.2 mSpace and the Semantic Web

In the case of our prototype, we are associating projects from one domain – a funding council – with papers held in another domain. As a result, we want to be able to facilitate a variety of ways to expose and explore the relationships between the data associated with these domains. These associations are made possible via specific properties of Semantic Web protocols. Before getting to a description of the architecture, it is worth taking a close look at one of these specific properties, the URI.

3.2.1 The URI

The Semantic Web provides a data-centric approach to information and data query, compared to the schema-centric approach of conventional databases. This difference allows us to combine information from different sources, expanding our schema, or in the case of the Semantic Web, swapping which ontologies we use when we harvest data that may utilize relationships we have yet to consider. The Semantic Web approach represents all resources, such as people, papers, journals, horses with a Universal Resource Identifier (URI) [4]. This URI refers to a specific resource, such as a researcher. When data is marked up in the Semantic Web's Resource Description Framework (RDF), relationships between different URIs are defined in triples of subject predicate object, to makes links, like author – has a – paper; paper-has an – author, publisher, etc.

Using URIs instead of strings (freeform text) when referring to resources facilitates the accuracy of the import process: URIs ensure that when information is imported from a variety of sources, information that refers to the same things is identified correctly and unambiguously. When importing information from many sources, the certainty of referencing resources is extremely important to data integrity. This issue is important for digital library metadata harvesting, as when using, for example, OAI-PMH, one library may refer to an author as "Daniel Smith", another as "Daniel A. Smith."

The Semantic Web approach does not require that there be only one URI to denote the unique person Daniel Smith. There is provision in the Semantic Web ontology language (OWL) [13] that allows for the marking up of the fact that URI-A used by one resource represents the same D.A. Smith as pointed to by URI-B. We will come back to the issue of managing multiple URIs for the same resource later in the paper.

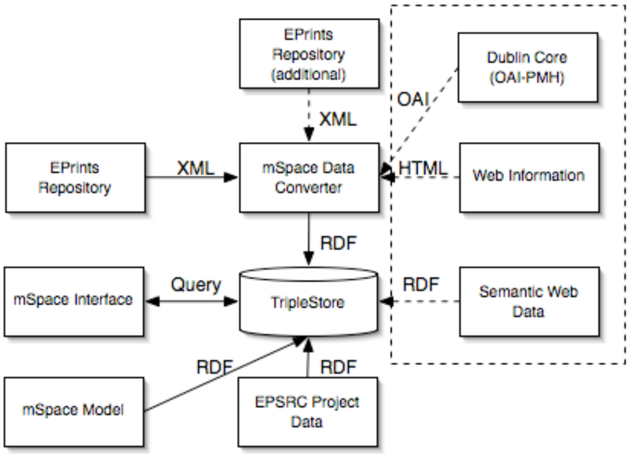


Figure 7. mSpace meets EPrints architecture

3.2.2 Architecture

mSpace utilizes Semantic Web protocols within an architecture similar to that of OAI-PMH harvesters. The difference is that the OAI harvesters can only rely on combining information when the same form for names is used, whereas the Semantic Web architecture matches on identifiers. The architecture diagram (Figure 7) shows the XML metadata dump coming from the EPrints repositories. These streams are converted into RDF at regular intervals, and asserted into a TripleStore. A TripleStore is a particular kind of database that holds Semantic Web data, allowing it to be queried using a language similar to SQL, called RDQL. RDQL differs from SQL in that it applies constraints on triples in a similar way that prolog applies constraints to variables. In RDQL, triples are given with of required form, with variables returned as the result, whereas SQL queries on tables, constraining on the contents of the specific columns in those tables. mSpace gets its information directly from the TripleStore as users are browsing. The mSpace model describing the dimensions in the domain that will be represented in the interface is also asserted into the TripleStore. Effectively, the model specifies which parts of the data streams coming into the system are to be collected to be available to be queried by the mSpace system. Specifically, the mSpace model, describes how the properties relate to each other and how to show them in the mSpace. This specification is required as mSpace is a generic system that can be used to browse any information, not just digital libraries. Once the shape of the data is determined – what the attributes of the data are that the collection wishes to make available for exploration – this needs to be described in the mSpace format, which uses RDF, and is also asserted into the TripleStore as shown in Figure 7. More formally, the mSpace model describes the links between resources, specifying which predicates, from which ontologies, should be used in the queries (modeled by the selection of slice, arrangement of the slice, and selection within the dimensions (columns) of the slice)_ which gather the data to populate the interface. The mSpace was then configured to use this knowledge base, the final part of configuring the mSpace explorer.

The right of the diagram shows how future data sources can feed into the architecture. Dublin Core metadata, combined with "screen-scraped" HTML web pages would be input into the mSpace Data Converter, before making up part of the TripleStore's knowledge base. Related information and metadata from the Semantic Web can also be asserting directly into the TripleStore, creating relationships between metadata from repositories.

3.3 mSpace Interface Model

The main attributes of the mSpace interface have been described elsewhere [16] [19]. Suffice it to say for our purposes here that the model for the interface is of a domain with n-dimensions. For instance, in the case of our prototype, the domain is research in the UK e-Science program. Dimensions include author, investigator, project, paper, publisher, granting council and so on. To manage the visualization of a high-dimensional space, we take a projection onto a plane. This flattens the space and creates temporary hierarchies of the dimensions in the projection. We call these projections "slices." A slice is currently represented in the interface as columns in a spatial layout (Figure 8). The slice shown there is Project | Eprint | Author.

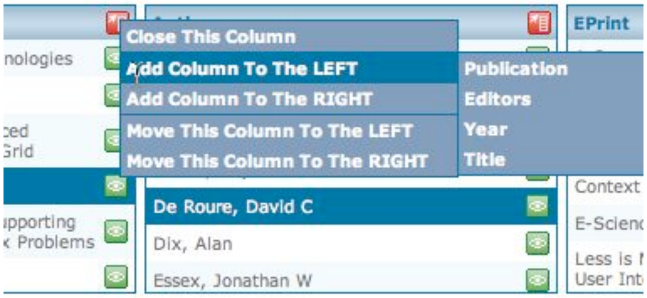
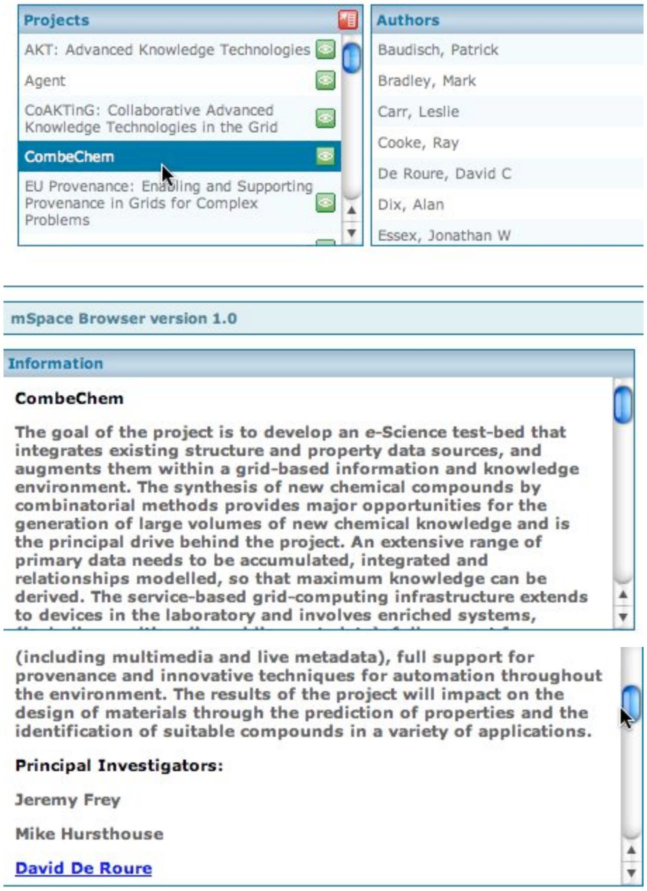


Figure 8. Choosing an operation on a slice.

Several operations are provided on a slice: sorting, swapping, adding and subtracting. Sorting means that the dimensions in a slice can be rearranged; swapping means that one dimension is traded for another dimension that is available but is not part of the current slice; adding and subtracting means that dimensions can be added to the slice or taken away. There are several operations within a slice: selecting an entity within a dimension (such as Daniel Smith under Author) causes the next dimension/column to be populated. If the next column is papers, a listing of Daniel Smith's papers will be presented. As well, a pane below the columns, called the Info View, will present information about the current selection. In this case, information about the author Daniel Smith will appear. Questions like what papers from the e-Science program have been published in JDCL, or who is involved in a variety of e-Science projects, can be asked readily.


Projects
AKT: Advanced Knowledge Technologies
Agent
CoAKTinG: Collaborative Advanced Knowledge Technologies in the Grid
CombeChem
EU Provenance: Enabling and Supporting Provenance in Grids for Complex Problems
Authors
Baudisch, Patrick
Bradley, Mark
Carr, Leslie
Cooke, Ray
De Roure, David C
Dix, Alan
Essex, Jonathan W
mSpace Browser version 1.0
Information

CombeChem

The goal of the project is to develop an e-Science test-bed that integrates existing structure and property data sources, and augments them within a grid-based information and knowledge environment. The synthesis of new chemical compounds by combinatorial methods provides major opportunities for the generation of large volumes of new chemical knowledge and is the principal drive behind the project. An extensive range of primary data needs to be accumulated, integrated and relationships modelled, so that maximum knowledge can be derived. The service-based grid-computing infrastructure extends to devices in the laboratory and involves enriched systems,

(including multimedia and live metadata), full support for provenance and innovative techniques for automation throughout the environment. The results of the project will impact on the design of materials through the prediction of properties and the identification of suitable compounds in a variety of applications.

Principal Investigators:

Jeremy Frey

Mike Hursthouse

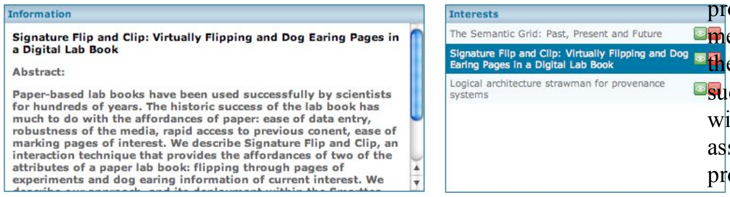
David De Roure

Figure 9. Info View associated with selected entity

The benefit of the Information box is that a person gets descriptive information about an entity in the same view as the facets selected. Based on this information, people exploring the collection can better decide whether they wish to click out of this context to another site, or if they simply wish to switch their current focus and move to a distinct entity.

The interface makes it easy to move rapidly among entries in the interface to carry out what Marshall calls "information triage. " [15] On finding something of note, a simple double click adds the item to an "of interest" list. Selecting that item again in the Interests list brings up the associated information about that item (Figure 10).

This ability to rapidly peruse and thumbnail information also leverages some of the navigation approaches observed in paper based reading strategies [14] thus offering an effective paradigm for rapid exploration and then deeper exploration of a space. The use of the spatial layout also provides a persistent context for current information focus while making relationships among entities clear.


Information

Signature Flip and Clip: Virtually Flipping and Dog Earing Pages in a Digital Lab Book

Abstract:

Paper-based lab books have been used successfully by scientists for hundreds of years. The historic success of the lab book has much to do with the affordances of paper: ease of data entry, robustness of the media, rapid access to previous conent, ease of marking pages of interest. We describe Signature Flip and Clip, an interaction technique that provides the affordances of two of the attributes of a paper lab book: flipping through pages of experiments and dog earing information of current interest. We

Interests

The Semantic Grid: Past, Present and Future

Signature Flip and Clip: Virtually Flipping and Dog Earing Pages in a Digital Lab Book

Logical architecture strawman for provenance systems

Figure 10. Selecting a collected item in the interest list shows info about it.

4. BUILDING THE COLLECTION

With the approach in place to coordinate heterogeneous sources, the next phase of the prototype was to collect the data. The specification to gather scan several EPrints repositories meant we we would need information from the granting councils about which projects were funded under a particular call, and from the institutions, we would need to know which papers were associated with which projects. We would then need the associated OAI data as well as information about the projects and the authors to create the associated spaces. By using the RDF of the Semantic Web protocols in the mSpace architecture, we could then associate these components to be displayed in the mSpace interface. The final result in progress can be explored at http://cortex.ecs.soton.ac.uk/mspace/printspace. The following section describes the successes and hurdles of this build process.

4.1 Scoping the Data Space

4.1.1 Initial Repositories

For the initial prototype, we looked at two EPrints services, both held at our home institution: the University-wide repository and our School's repository. We used our home repositories (a) because we only had a short time to build the prototype and (b) we know the people managing these archives and they were keen to help, should direct contact be necessary. This latter point was an asset in project development. Thus, based on our findings with EPrints repositories at home, we could generalize to the 100+ other EPrint repositories in the UK.

4.1.2 The OAI Data – Complete but Not

The first thing we learned was that the OAI data produced by the University's and the School's OAI-PMH gateways represents all the publication metadata entered about a paper. Other information that is relevant to the paper, and critical to us, such as the project that the paper was associated with, or the program that funded the paper, or the author's home institution, is not part of this source. This meant that in order to determine the funding source of a paper, we would potentially have to create a list of papers associated with an author (This would be the easy part) and then beg their authors to identify on a form whether the papers were part of an e-Science project or not. That later point would be the hard part: why should the academics – or at least enough of them - take the time? We could not ask the granting councils for such a list. After all, they did not know themselves: that is why, in part, they wanted this project.

4.1.3 The EPrints Repositories Differences

Fortunately, the second thing we learned is that, as open source software, not all EPrints repositories are implemented in the same way. The ECS EPrints service differs from standard installs such as the University's because, it has made use of the EPrints facility to augmented the EPrint deposit with the School's people and projects information from the various School database. That means that authors who are members of the school (identified by their email addresses) have associated information about them, such as contact data, Web page and associated projects, connected with their entry. Likewise for projects, the repository records the associated project's name, which in turn connects it with the project's description, the people on the project. (The space in the project page for the granting council and grant number seems rarely to be filled in). While this extra data on people and projects is not piped via the OAI-PMH gateway, it is accessible from the separate, publicly available, EPrints-defined XML dump. Interestingly, in the ECS EPrints example, these associations of author and project are not made to augment the view of a deposit in EPrints, but so that an EPrint deposit can augment the view of a project on the School Web site. The list of EPrints-deposited papers associated with a project or author is automatically associated with the project's or academic's official Web page. For instance, the project page on mSpace lists associated links to investigators and to papers on the project: http://www.ecs.soton.ac.uk/research/projects/mspace/. This additional harvestable metadata in the ECS EPrints archive became crucial for our prototype, and has lead to interesting questions about the role of an archive which we address at the end of this paper.

In the case of our data requirement, none of the project data specified the call by which it had been funded, so we could not simply filter only those projects funded under the e-Science program call. We did hope, however, that the funding councils would publish a list of funded projects for a given program call. It was at this point, that for the purposes of the prototype and in the interest of time, we decided to focus on using the School repository alone, rather than trying to combine its results with the University repository. It seemed there would be ample to learn from the effort of connecting call information and papers with a single repository that had at least, at a minimum, some project markers that could potentially be associated with a call, rather than one that had none.

4.1.4 Matching Projects to Programs

While the School's project database did not give us Program information, we did learn that one of the granting councils in the RCUK, the EPSRC, provides a web site called "Grants on the Web", which classifies the projects they have funded according to their own taxonomy. E-Science, a major program, exists in its own part of their taxonomy. This site provided a list of projects at all participating Universities that were funded by the EPSRC under e-Science. Our first task was to filter out from the EPSRC data only those projects involving our School/University. With this list in hand (literally), we were then able to filter the EPrints information dump for only the projects which matched our e-Science EPSRC program list. This matching/filtering was carried out by a custom built script to compare the list and the output.

4.2 The Build

From the list we created of local projects associated with e-Science, we could then collect all the data in EPrints associated with this list. There are two components designed after we understand what sources we have available: the mSpace model and the information box.

4.2.1 The mSpace model

With the data collected, it was a matter of marking it up into RDF, the standard format for describing knowledge on the Semantic Web (see Figure 2). Once data was in this format, which is required for the mSpace interface to be able to explore it, we developed the mSpace model, as described above, for the collection. This model let us define the specific dimensions that would be available for slicing. Based on the data we had, we could readily provide the dimensions Project | Author | Paper | Publisher (the journal or conference for the publication) | Year. What we could not model is the research council, since we only had data for one council's projects.

4.2.2 The Information for the Info Box

One of the key concepts to the mSpace interface is the Information Box (Figure 9), which provides information about the currently selected item. This is true for both documents that are selected, as well as anything else that can be selected, such as authors, projects and institutions. The approach taken for this project was to show the linked information from School's database on these projects and people, as it is published to the Web. This information would be published directly to the Info Box area of the interface. This linking and repurposing is facilitated by the EPrints XML feed that provides the email usernames of the authors. Thus hooking up with the ECS website via author identifier to show information was relatively simple. Without that information (not provided by the OAI gateway), it would not have been so easy.

The ability to populate the information box from the above simple linkage of available resources highlights the usefulness of EPrints' built-in ability to connect with departmental databases, so that it can provide information beyond a library of research papers, and allow for a richer exploration experience by showing such information to the user in-context at the time of exploration.

5. ISSUES

One might think that by focusing on effectively a single source of information, already well-associated with additional local resources (about people and projects) the resulting prototype would go forward without incident. After all, we are effectively only adding a new interface to the collection to improve interaction. While what we call slinging an mSpace interface over a model is now straightforward, that very exercise exposed some critical issues for deploying even this highly constrained prototype of an on-demand collections. Two key issues are co-referencing and non-perceive(d/able) data holes.

5.1.1 Co-Referencing

In the final collection representation, it is not unusual to see many authors listed on a paper who seem to have remarkably similar names: M. Luck, Michael Luck, M.R. Luck, for instance. It may seem that these are the same person, but in the case like this, of similar names, other information has to be taken into account before one can say for certain that they are the same person, and even then it's hard to be sure. This problem of co-reference is one that is becoming increasingly foregrounded as the remit of digital libraries starts to grow.

Similar projects to ours, but outside the digital archive space, [1], [21] have also come across co-reference problems as critical. The complementary problem also occurs, and is much harder to identify: where data apparently on one person actually was the combination of data on two people, which has been incorrectly merged at some indeterminate point in the past. The use of unique identifiers such as URIs, and corresponding information when two URIs refer to the same resource help to alleviate this problem, as it provides the confidence and machine-readability that two pieces of data refer to the same resource. In our case that M. Luck is Michael Luck is M.R. Luck.

5.1.2 Incomplete Data Looks Complete

The connection of program data with projects and associated publications means that projects that should be represented in the collection have likely fallen on the floor. There are three issues which have contributed to this. First, of all the RCUK includes a number of granting councils (for Arts, Medicine, Social Sciences, etc) in addition to the EPSRC that also have e-Science projects,. Only the EPSRC produces a public-facing list of projects associated with a call. This lack of information means obviously that people participating in work with other councils would potentially go unrepresented in the collection.

Second, the EPrints repository itself relies on people manually typing in the name of the project to which a publication is associated. It is entirely possible for a person, on submitting a paper, either to skip the step for entering the project, or to use a slightly different representation of the project name than that which is in the database. Thus, there are likely papers in the EPrints archive that are e-Science funded, but not included in the final collection For example, many of the projects encountered were acronyms, such as PASOA, the Provenance-Aware Service-Oriented Architecture project. Papers in the EPrints archive list papers as being part of the project "PASOA", "pasoa", "PASOA: Provenance-Aware Service-Oriented Architecture", and even "Provenance". These variances had to be manually adjusted; such adjustment increases the risk that some papers will not be picked up, because some version of the name is missed, causing the associated papers to stay indivisible, and thus missed in the new collection.

Third, we lost an entire repository, the University one, because there was *no* connective tissue available at all to show whether or not a paper had been produced as part of a particular project. Hiring someone to follow up with all the authors of a collection individually in order to determine, post-hoc, the project associated with a publication was not within the means of the project. There is no requirement under either OAI or EPrints institutional archives to list the projects associated with a paper. The University archive does have a free form text field into which a depositor can write the name of a project, but (a) this field is rarely used and (b) it is not labeled with any identifier, so it is next to useless in automated efforts to integrate collections.

6. RECOMMENDATIONS

One of the questions that our work on the prototype raises is how best to integrate relevant external sources with the repositories to create the kinds of dynamic collections we have investigated developing. Our recommendations are both technical and social. In other words there are relatively straightforward technical solutions that require social enactment, rather than engineering, to make these types of collections straightforward to deploy.

6.1 URI's vs Invisibility

Co-referencing, and its opposite harm of mis-referencing can largely be addressed by the adoption of URIs and a method for resolving them. We strongly recommend that URIs become integrated components of digital archives. The mechanics of incorporating URIs within Dublin Core, for instance, is highly tractable: in addition to specifying the textual value of a relation i.e. for an author "Daniel Smith", the archive would also specify the universal resource identifier (URI) for this person e.g. http://ecs.soton.ac.uk/people/ds.

One of the benefits of URI use beyond co-referencing is implicit linking of associated resources. When merging together multiple sources of data, or when shared URIs are used, or when mappings of URIs are provided (i.e. that two URIs represent the same resource), the link between one piece of data about that resource (such as paper authorship) from one source is implicitly linked to the resource from the second source of information. This merging of Semantic Web data (known as smushing) makes exploring data about the same things, but from multiple sources, possible. This multiple referencing is of course the key idea behind mSpace meeting EPrints: the use of one interface to browse multiple archives for particular values. .For example, one Web site or Web service may give information about a researcher's articles in a particular journal, and another about papers in conferences. If they both used the same URI for the things they hold in common, or a URI-URI mapping stating they were the same, then the smushing system would know that the researcher behind the journal articles also wrote the conference papers from the second website.

6.1.1 Generating and Managing URIs

The knock on effect of adopting URIs to represent attributes in an archive means that attributes like authors and publishers would themselves have URIs readily harvestable in order to be associated with the data submitted to an archive about a deposit. When it comes to generating URIs, the approache regarded as current best practice is a policy of using a domain name that the URI-maker owns and controls , and of generating the URI such that it will not change in the future [http://www.w3.org/Consortium/Persistence], otherwise the URI can be any string that conforms to the URL schema, which, to the layperson means that it looks like a world wide web address. For example, image the "example" company a URI for "Daniel Smith" may be: http://example.com/people/Daniel_Smith, or someone may decide to use the primary key from an institution's database instead: http://example.com/people/3341, or the person's email username: http://example.com/people/dsmith. If an identifier already exists for the resource, using that in the URI will help to maintain the uniqueness: the primary key of a database or username is a good example for people, while an asset tag number might be appropriate for identifying a workstation.

If all archives represented identical resources, such as People, Projects, Institutions, with the same URI, the data from multiple institutions could be merged together without having to deal with problems of whether two resources are the same person, or that some of the metadata is missing. We would suggest that there are organizations which are best candidates for creating and managing core URI's. For instance, in the UK, all institutions which and all academics whom receive funding from the RCUK (or wish to receive funding from the RCUK) are registered on the Joint Electronic Submission (Je-S) system. Funding applications now have data about universities, schools and investigators filled in by keyword search for the person, and clicking on a list of possible matches. This is a source ideally situated to generate URIs, which institutions themselves can adopt as their URI's for themselves and their staff, or they can link them to existing URI's via sameAs. While large publishers like the ACM can generate their own URIs for themselves and their conferences, smaller conferences and workshops, using the framework for URI generation, can create and publish their own. Services for conferences may even spring up, as exist for people [6] [22] and files [2] [22], to act as pointers for these events. One could imagine the presence or absence of a URI becoming a mark of credibility: if an event doesn't have one, it doesn't take itself seriously.

6.1.2 Social Issues for Technical Solutions

For URI's to be useful, they must be supported and adopted. Just as the RCUK is considering a policy to make research outputs publicly available, a similar policy would likely need to be enacted to ensure critical data representing the institutions, projects and people involved in that research must be likewise identifiable, so that the producers can be clearly associated with what is produced. Thus, while there may need to be policy to insist that URI's are generated for use, there likewise would need to be policy to make use of URIs either directly or via sameAs services. For example, if Je-S generated URIs for institutions and people, it could provide those URIs via a Web Service so that EPrints could make use of them directly to represent its own institution's authors, *as well as* its collaborators from other UK institutions. Likewise URI's for Conferences or Journals could be used to ensure the correct data is harvested for an "in submission" paper, which can be updated by publisher data once the submission is published.

The adoption of URI's by publishers does not mean that self-archivists or librarians will now have to enter one complex URI rather than several lines of text strings: there are a number of models that could be used to make URI use tractable in even a largely manual deposit process. Once the local system knows about an entry, it can store it in human readable/searchable form. A person can check the archive to see if anyone else has already entered the URI for instance for JCDL06. If they have not, a person could trackback it. Trackback[23], made popular in blogging, means that a person can point from one site to another site. A person could point their digital archive software at the appropriate conference URI. The trackback would provide the appropriate link to the resource to the system - in this case a URI to the conference.

Once URIs are in use, a follow up question becomes, to what do the URIs refer? The URI is used to determine if one text string relates to the same entity as another text string. In the case of an enriched digital archive, we may want to know more than if two similar names represent the same author. Technically, however, that is the core function of the URI: to disambiguate references. They do not have to "resolve" to a Web page with information about that entity. It may therefore be another policy decision to say official URIs *must* be resolvable, and must resolve/point to a minimum criteria for a referenced entity so that there is a clear path to the essential elements of an artefact's provenance: where publication data tells of where the deposit was published, an author URI would enable tracing of the associated provenance of the writer. An author's minimum information may be institution, Web page, contact details, and dates of service (time becomes an important marker). But it may not. For example, some may argue that it is basic to include a person's institutional email address in any URI result; others may say that that opens the person up too easily to abuse like spam. Institutions, whether the Universities themselves or the granting councils, however, need not wait for complete policies to be developed in order to take advantage of URIs. They can construct their own based on the pattern in section

6.1.1 models. As research within the OAI community has shown, papers available online are cited 20-40% more [10] than those that are not available digitally. Our prototype has clearly demonstrated is that if the information is not there on the public Web to be gathered, that work is effectively invisible. As open archives become more prevelent, and as the OAI metadata is increasingly used to create these kinds of dynamic collections, in order simply to be visible, it will become critical to make sure information structures are in place to enable this kind of dynamic connectivity. URIs are one tractable method to help insure visibility.

6.2 Integration with the (Semantic) Web

To further integrate with the Semantic Web, and thereby extend the possible virtual geography associated with a collection, digital libraries could export their metadata directly into RDF, in addition to continuing to output the OAI-PMH they support at present. There is a standard ontology for describing the Dublin Core schema used by OAI-PMH, and as such it is likely that this integration would not constitute a significant development effort. Combined with the use of institution-neutral shared URIs, data harvesting using Semantic Web exported data would make exploring the data as a whole, without omissions, and with confidence on accuracy, a tractable concept.

6.3 Interfaces for Exploratory Search

Interfaces which support exploration of the relationships in a collection are critical for taking these collections beyond what a keyword search engine offers, and therefore should not be overlooked in creating cross-archive collections. We have looked at the exploration attributes enabled in interfaces that support exploratory rather than keyword-only search. In the case of a dynamic collection, mSpace as an exemplar of exploratory interfaces, foreground rapid triage of the domain being highlighted by the collection. This domain exploration is part of the critical value add of creating a collection in the first place: to facilitate ready exploration of the domain geography at least somewhat beyond the immediate range of the artifact itself. Keyword search is target oriented, missing the context of the publication. Exploratory search interfaces like mSpace, wrapped over a collection foreground a variety of contexts.

6.4 Aside: Self Archiving Deposit Interfaces

One of the social issues impacting the success of repositories is the mechanism for deposit. In the case of the self-archive, few depositors would claim to enjoy the experience of the requisite form filling associated with the deposit. In our School one of the reasons motivating the success of its repository is that a complete listing of publications for an academic member of staff must be available in the archive for annual review, as well as for the national RAE institutional assessment on which funding levels rest. There is an implicit "or else" associated with making the deposit, as so much rests on it. That said, there is also now a more visible and immediate payoff for making the effort to deposit a paper: the most recent deposits are broadcast to numerous sites: public displays in the school which great visitors; the School and Group Web pages, and the ePrints main Web interface itself. "Heh, that's my paper" can be heard increasingly as people walk past one of the displays in the main hall. Despite these carrots and sticks, depositors are frequently delinquent in submitting papers. This is in no small part because most the service relies largely on manual effort: the archive presents a series of forms that must be filled in, in order to associate the appropriate metadata with a deposit. This is a tedious and error-prone process. The results, as we have seen, of a deposit suffering from missing or incorrectly entered data is the deposit's potential absence from a dynamic collection which requires the presence of that particular marker.

One might suggest that a solution would be for a form to prohibit submission until a field is filled in. That is no guarantee that an open text box will be filled in correctly; it is also irritating, especially when this may be the umpteenth time a depositor has had to fill in the same information. There are likewise numerous instances in the deposit process where a depositor is required to fill in a text field with information that is already available in another source, either a in local database (eg. project names) or remote Web sources (conference information). The use of URIs as proposed above has the potential not only to improve quality of data entry, but to increase the data entry by, in no small part, reducing the number of repetitive steps in the deposit process.

7. OPEN QUESTIONS

In our prototype, we were able to create a dynamic collection of e-Science publications that could be explored readily across a variety of criteria, from project to author to publication to year of publication. We were able to take the public-facing output of EPrints, an OAI archive service and, using the mSpace framework, process and select a subset of publications against particular criteria and wrap the new mSpace interface around this collection to facilitate exploration of the collection domain. Crucially, however, the collection represents only those papers available from one EPrints resource. The block to connecting archives is the lack of the connective tissue within the archive to make connection to external sources possible *automatically.* While it is possible to blend collections in general, based on what is in common to these collections (all papers across institutional collections which have papers with a Daniel Smith as a co-author), our collection mix required knowledge of data outside the various collections' usual holdings: projects associated with the authors. Without this information, we could not connect the data about papers with the projects listings held by the RCUK. In the one EPrints collection where this hook was possible, we were able to go to work with relative ease. The question that arises from this connectivity issue is whether or not it is the mission, provenance, duty of the digital archive to be the maintainer of such hooks, and as we have proposed, such hooks as URIs.

It has been historical bibliographic practice to list the author of a work, and so providing an author URI (and implicitly the institution URI) may seem harmless at worst. It has not, however, been the provenance of a library to hold information in an entry about the source of funding that enabled that deposit to be generated. There may be an argument – and we think there is – that a digital repository is the right place to hold that information by way of exposing the full provenance of the work. We note that there seem to be counter arguments to this association: one of the funding bodies under the RCUK, we learned, deliberately does not make the projects it funds publicly available. The best and not official answer we have received about this is fear of retaliation upon the funded institutions from anti-vivisectionists.

Our initial prototype has suggested that a minimum set of URIs an archive would hold is: the event or source of submission/publication, the author and institution, the project which enables linking to funding call and council, and the subject, either in terms of ACM type classification or Library of Congress subjects or both. These URIs-as-hooks would make a range of interconnections possible. It is the absence of these hooks that scuttled automatically generating a cross-institution collection of RCUK e-Science project publications.

8. CONCLUSION

In this paper, we have presented a case study exploring an approach to deliver cross-repository, dynamic collections defined against semi-external criteria. In this case, the external criteria is a funding council wishing to present a view of publications associated with projects it has funded nationally, across universities. We have shown how to build these collections by combining the OAI metadata and additional XML data stream of EPrints with the Semantic Web technologies and exploratory interface available in the mSpace software framework. With these combined frameworks it is possible to define the scope on an on-demand collection, gather the data and provide the explorable interface. We have also shown where the automation afforded by this technology breaks down without effective hooks within the archive to relevant external sources. As a result of this case study, we have made a series of recommendations, in particular the adoption of URIs for referencing entities like authors, publications and funding sources within the archive. While we have shown that the technical challenges are largely surmountable, social policy decisions and questions about the nature and responsibilities of institutional archives need to be addressed before technologies can be deployed to make automatic, dynamic generation of collections across archives possible.

9. ACKNOWLEDGMENTS

Thank you to the JCSR for its support of this project. Thank you to Tony Hey and Anne Trefethen.

10. REFERENCES

[1] H. Alani, Glaser, H., Gibbins, N., Harris, S. and Shadbolt, N., *Organising and Integrating Knowledge from Research Councils to Monitor Scientific Collaboration*, The New Review of Hypermedia and Multimedia (2006).

[2] W. Arms, *Uniform resource names: handles, PURLs, and digital object identifiers*, *Communications of the ACM*, 2001, pp. 68.

[3] M. J. Bass and M. Branschofsky, *DSpace at MIT: meeting the challenges*, *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, ACM Press, Roanoke, Virginia, United States, 2001.

[4] T. Berners-Lee, R. Fielding and L. Masinter, *RFC3986: Uniform Resource Identifier (URI): Generic Syntax*, 2005.

[5] T. Berners-Lee, J. Hendler and O. Lassila, *The Semantic Web*, Scientific American (2001).

[6] D. Brickley and L. Miller, *FOAF Vocabulary Specification*, RDFWeb Namespace Document, 2004.

[7] T. Brody, *Citebase Search: Autonomous Citation Database for e-Print Archives SINN03*, Germany, 2003.

[8] N. Gibbins, H. Alani, S. Harris and N. Shadbolt, *AKTive Futures: Supporting strategic decision-making (Demo)*, *ESWC 2005*, Heraklion, Greece, 2005.

[9] C. Gutteridge, *GNU EPrints 2 Overview*, *11th Panhellenic Academic Libraries Conference*, Greece, 2002.

[10] S. Harnad and T. Brody, *Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals*, *D-Lib Magazine*, 2004.

[11] K. Hegedorn, *OAIster: a ‘‘no dead ends’’ OAI service provider*, *Library Hi Tech*, 2003, pp. 170-181.

[12] P. Hochstenbach, H. Jerez and H. Van de Sompel, *The OAI-PMH static repository and static repository gateway*, *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, IEEE Computer Society, Houston, Texas, 2003.

[13] H. Horrocks, P. F. Patel-Schneider and F. van Harmelen, *From SHIQ and RDF to OWL: The Making of a Web Ontology Language*, Journal of Web Semantics, 1 (2003).

[14] C. C. Marshall and S. Bly, *Turning the page on navigation*, *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, ACM Press, Denver, CO, USA, 2005.

[15] C. C. Marshall, F. M. Shipman and J. H. Coombs, *VIKI: spatial hypertext supporting emergent structure*, *Proceedings of the 1994 ACM European conference on Hypermedia technology*, ACM Press, Edinburgh, Scotland, 1994.

[16] M. J. McGuffin and m. c. schraefel, *A comparison of hyperstructures: zzstructures, mSpaces, and polyarchies*, *Proceedings of the fifteenth ACM conference on Hypertext and hypermedia*, ACM Press, Santa Cruz, CA, USA, 2004.

[17] Mercedes, *Select a model* - http://www.mbusa.com/brand/selector/controller.jsp.

[18] L. Rutledge, M. Alberink, R. Brussee, S. Pokraev, W. van Dieten and M. Veenstra, *Finding the story: broader applicability of semantics and discourse for hypermedia generation*, *Proceedings of the fourteenth ACM conference on Hypertext and hypermedia*, ACM Press, Nottingham, UK, 2003.

[19] m. c. schraefel, D. A. Smith, A. Owens, A. Russell, C. Harris and M. Wilson, *The evolving mSpace platform: leveraging the semantic web on the trail of the memex*, *Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*, ACM Press, Salzburg, Austria, 2005.

[20] P. Scrimshaw, *Text completion programs*, *Language, classrooms and computers*, Routledge, 1993, pp. 130-144.

[21] N. R. Shadbolt, N. Gibbins, H. Glaser, S. Harris and m. c. schraefel, *CS AKTive Space or how we stopped worrying and learned to love the Semantic Web*, IEEE Intelligent Systems (2004).

[22] K. Shafer, S. Weibel, E. Jul and J. Fausey, *Introduction to Persistent Uniform Resource Locators*, *INET1996*, 1996.

[23] B. Trott and M. Trott, *TrackBack Technical Specification* - http://www.sixapart.com/pronet/docs/trackback_spec, 2001.

[24] S. Weibel, J. Godby, E. Miller and R. Daniel, *OCLC/NCSA Metadata Workshop Report*, *OCLC/NCSA Metadata Workshop*, Office of Research, OCLC Online Computer Library Center, Inc, 1995.

[25] I. H. Witten, S. J. Boddie, D. Bainbridge and R. J. McNab, *Greenstone: a comprehensive open-source digital library software system*, *Proceedings of the fifth ACM conference on Digital libraries*, ACM Press, San Antonio, Texas, United States, 2000.

[26] Yahoo, *Camera Selector* - http://shopping.yahoo.com/sort_tool_digicamera.

[27] J. Zhang and G. Marchionini, *Evaluation and evolution of a browse and search interface: relation browser*, *Proceedings of the 2005 national conference on Digital government research*, Digital Government Research Center o, Atlanta, Georgia, 2005.