JISC DEVELOPMENT PROGRAMMES

Project Document Cover Sheet

FINAL PROJECT REPORT: May 2008

Project

Project Acronym	RIDIR	Project ID	
Project Title	Resourcing IDentifier Interoperability for Repositories		
Start Date	April 2007	End Date	May 2008
Lead Institution	University of Hull		
Project Director	Ian Dolphin		
Project Manager & contact details	Richard Green r.green@hull.ac.uk		
Partner Institutions	Rightscom Ltd, Linton House, 164/180 Union Street, London SE1 0LH		
Project Web URL	www.hull.ac.uk/ridir		
Programme Name (and number)	Digital Repositories Programme: Discovery to Delivery		
Programme Manager	Balviar Notay		

Document

Document Title	Final project report		
Reporting Period	April 2007 - May 2008		
Author(s) & project role	Richard Green Project Manager		
Date	2008-07-09	Filename	final-rep-v10.doc
URL			
Access	☐ Project and JISC internal		☑ General dissemination

Document History

Version	Date	Comments
1.0	09/07/2008	Version submitted to JISC

THE UNIVERSITY OF HULL

with

rightscom

RIDIR Project

Final Project Report

April 2007 - May 2008

Richard Green, Chris Awre, Steve Bayliss, Ian Dolphin, Martin Dow, and Hugh Look

June 2008

The RIDIR Project

Project Director:	Ian Dolphin, Head of e-Strategy, University of Hull	(i.dolphin@hull.ac.uk)
Project Manager:	Richard Green	(r.green@hull.ac.uk)
Repository Domain Specialist:	Chris Awre	(c.awre@hull.ac.uk)
Workshops management and coordination for Rightscom:	Hugh Look	(hugh.look@rightscom.com)
Identifier consultant:	Mark Bide	(mark.bide@rightscom.com)
Lead software architect and developer :	Martin Dow	(martin.dow@acuityunlimited.co.uk)
Software developer:	Steve Bayliss	(steve.bayliss@acuityunlimited.co.uk)

The RIDIR Project was undertaken by the e-Services Integration Group at the University of Hull and Rightscom Limited in London. It is funded by the JISC Repositories and Preservation Programme.

Table of contents

Acknowledgements

The Resourcing IDentifier Interoperability for Repositories (RIDIR) Project was funded by the 'Discovery to Delivery' strand of the JISC Digital Repositories Programme. The technical development was carried out by Rightscom Limited of London in conjunction with staff of the e-Services Integration Group at the University of Hull. The Project Manager was Richard Green, an independent IT consultant, working as a sub-contractor.

We acknowledge with thanks the help and assistance given us by the digital repositories community world-wide. In particular we should like to acknowledge the contributions made by members of the ARROW Project in Australia (Project Manager, David Groenewegen) and the PILIN Project in Australia (Dennis Macnamara and Nick Nicholas); and in the UK by members of the Version Identification Framework team based at the LSE (Jenny Brace, Catherine Jones and Dave Puplett), who contributed to a number of thought-provoking discussions; to staff of Spoken Word Services at Glasgow Caledonian University (Caroline Noakes, Iain Wallace, and Graeme West) who were extraordinarily helpful in providing insight into a range of identifier issues and for providing a range of test data to work with; to staff of the Intute Repository Search team at UKOLN (Paul Walk and Monica Duke) for helping us to establish direct access to their search facilities; to members of the Fedora UK & Ireland User Group; and to UK-based members of the OAI-ORE advisory group.

Last, but certainly not least, we acknowledge the patient assistance given to us by our JISC Programme Manager, Balviar Notay.

Executive summary

The RIDIR Project set out to investigate how the appropriate use of identifiers for digital objects might aid interoperability between repositories and to build a self-contained software demonstrator that would illustrate the findings.

The project started by holding two Focus Group Meetings with repository practitioners to explore the (then) current range of issues around the topic and to map out what RIDIR might do. The result of the two meetings was an emerging understanding that new repository managers did not see identifier interoperability, whatever it might mean, as a burning issue and that experts in the field had a wide range of problems in the field and an equally wide range of potential solutions.

The RIDIR project team narrowed down the issues and expressed them as five scenarios which it was felt could be addressed within the lifetime and resources of the project. However it became clear that there were two approaches that could be taken to the work.

The first would demonstrate the value of interoperability: In this case, the role of the demonstrator would be to show some of these scenarios in action. The team would also aim to describe the impact of failure, in terms of what could not be accomplished or the cost of an alternative approach (for example, manual creation or editing of large volumes of repository metadata). The demonstrator would not attempt the creation of identifiers: it would focus on offering a clear demonstration of value unconstrained by the contingent factors of current practice, which we knew to be very limited. The work would focus on cases where there was a unique, *accessible and actionable* identifier of some form. It would also thereby establish what conditions and changes in practice would be necessary to make feasible the use of identifiers to support interoperability.

The second approach would look at the cost of interoperability: In this approach, using the scenarios to give context, we would concentrate the demonstrator on showing *how* identifiers can be created, mediated and therefore managed cost-effectively on the assumption that the value of so doing is axiomatic. The benefit of the 'cost' or 'how' approach would be to assist the community by demonstrating approaches that will facilitate the achievement of interoperability, and so enable more rapid adoption of technology and working practices. It would also be likely to reveal issues that have not been encountered by other projects to date.

It was accepted that there were elements of work common to both approaches. In the event, the JISC asked the project team to follow the second approach. This has resulted in a demonstrator that addresses most of the issues raised in our scenarios by providing two related services allowing identifiers to be used as the means to build up and record potentially rich relationships between objects and identifiers and between digital object and other digital objects. More explicitly, it shows how such techniques can be used to locate and record the whereabouts of objects that have moved outside their original curation space and become 'lost', and how rich networks of relationships can be built up between related objects in disparate locations enabling a user who discovers one immediately to be aware of and investigate the others.

During the work done to move from the outcomes of the focus group meetings to a proposal for demonstrator development, a significant amount of research was done concerning the general

aspects of identification and interoperability, to evaluate the various approaches that the RIDIR demonstrator might take with respect to existing work, and to ensure value was added rather than duplicating extant work. This included analysis of existing identifier schemes and services, existing services for enabling interoperability. We anticipate that this research will also be of use to the repository community.

The totality of the RIDIR work has enabled the team to make a range of recommendations which we trust will be given due consideration by the repositories community.

1. Introduction

The continuing growth in the number of repositories in research and teaching institutions worldwide is resulting in an increasing number of digital objects that are openly available. Increasingly flexible discovery mechanisms are becoming required to facilitate access to these resources, particularly where content is being moved around between repositories and in a state of flux. The adoption of standards for repository-related activities offers an opportunity to build such services on a firm base and support interoperability, though there is a need to be conscious of working across standards as well as working with them for the best outcome. One area of interoperability that has attracted limited attention is that between identifiers, and particularly those identifiers intended to be persistent. Attention has focused predominantly around the adoption of standard identifier schemes, such as Handles[1] or DOI.[2] In a diverse repository landscape, however, multiple identifiers are available and are being used, presenting a real interoperability challenge.

The RIDIR project was set up to investigate and demonstrate the links that can be established between digital objects via identifiers. In this way, the project has sought to contribute towards, as stated in the original JISC call, "… better understanding and practice relating to rich search and discovery of content within repositories."

1.1 Identifiers and persistent identifiers

Throughout the remainder of this document it is important to distinguish between our use of the terms 'identifier' and 'persistent identifier'. In the digital repository community, it is common for the term 'persistent identifier' to be associated with the URL or URI of a digital object; typing it into the address bar of a user's web browser software should retrieve the object or, at least, a splash page about it. The adjective 'persistent' refers to the notion that this URL or URI should be stable over time. We have used 'persistent identifier' in this way. By contrast, in this report we use the term 'identifier' to refer to any label that can legitimately be associated with the content of a digital object but which may not necessarily be resolvable in a browser. An object's persistent identifier would be one such, but - depending on type - so might the name of its author or creator, an ISSN for the serial journal in which the text appears, the catalogue number for the broadcast in a particular system, the research contract details of the project which created the data, and so on.

1.2 Project Aims

To engage with the identifier and repository communities to understand better their requirements and highlight the benefits of the clear use of persistent identifiers in order to facilitate interoperability where required.
To develop and build a fully working demonstrator to showcase the findings of this engagement and demonstrate potential means for addressing the issues raised.

---

[1] See: http://www.handle.net
[2] Digital Object Identifier System, see: http://www.doi.org

1.3 Project Objectives

To raise awareness of persistent identifier interoperability issues within the Higher and Further Education community, influencing repository practices to incorporate these issues and contributing to the understanding of the governance procedures around identifier management
To provide a clear way of demonstrating issues relating to persistent identifier interoperability and potential solutions for addressing a range of use cases

1.4 Non-aims and non-objectives

It is perhaps worth noting here that whilst the brief for the RIDIR Project was quite broad there was a specific area of work that was not in scope. The RIDIR team, in submitting the Project Proposal, and the JISC, in funding it, were both aware of the DEST-funded PILIN Project[3] in Australia which was set up to pilot a shared, standards-based, persistent identifier management infrastructure. There was concern in both quarters that RIDIR should not duplicate the work being done by PILIN but that the team should take cognisance of the work being carried out there concurrently.

1.5 Project outcomes

The RIDIR Project set out to investigate how the appropriate use of identifiers for digital objects might aid interoperability between repositories and to build a self-contained software demonstrator that would illustrate the findings.

The project started by holding two Focus Group Meetings with repository practitioners to explore the (then) current range of issues around the topic and to map out what RIDIR might do. Following these, the RIDIR project team narrowed down the issues and expressed them as five scenarios which it was felt could be addressed within the lifetime and resources of the project.

It became clear that there were two approaches that could be taken to the work. The first would demonstrate the *value* of interoperability whilst the second approach would look at the *cost* of interoperability. It was accepted that there were elements of work common to both approaches.

In the event, the JISC asked the project team to follow the second approach. Subsequent work to determine the scope of the demonstrator indicated that:[4]

- RIDIR should not focus on identifier schemes and resolution mechanisms *per se*
- Even when appropriate persistent identifier schemes and services are implemented, there will still be 'corner cases' which RIDIR could address
- RIDIR should focus on how relationships are created, described and navigated to aid interoperability

[3] PILIN Project see: https://www.pilin.net.au
[4] These points are presented in rather more detail on page 37

- The RIDIR demonstrator should not be a 'black box' but should make apparent to a user what was happening

- RIDIR should primarily focus on user-driven discovery of relationships between identifiers, and the persistence and usage of these relationships, whilst recognising that it may be possible in the future to have machine to machine discovery of relationships.

This has resulted in a demonstrator that addresses most of the issues raised in our scenarios by providing two related services allowing identifiers to be used as the means to build up and record potentially rich relationships between objects and identifiers and between a digital object and other digital objects. More explicitly, it shows how such techniques can be used to locate and record the whereabouts of objects that have moved outside their original curation space and become 'lost', and how rich networks of relationships can be built up between related objects in disparate locations enabling a user who discovers one immediately to be aware of and investigate the others.

During the work done to move from the Workshop Report outcomes to a proposal for demonstrator development, a significant amount of research was done concerning the general aspects of identification and interoperability, to evaluate the various approaches that the RIDIR demonstrator might take with respect to existing work, and to ensure value was added rather than duplicating extant work. This included analysis of existing identifier schemes and services, and existing services for enabling interoperability. We anticipate that this research which is reported here, will also be of use to the repository community.

The totality of the RIDIR work has enabled the team to make a range of recommendations which are dealt with in Section 8 and Appendix A of this report.

2. Background

The RIDIR Project Plan set out the following background to the work that the team originally intended to undertake:

What does an International Standard Book Number (ISBN) identify? This may seem a naïve question – isn't the answer that it identifies "a book"? – but a more thoughtful answer to the question goes to the heart of the issue of identifier interoperability.

Because in reality an ISBN doesn't identify "a book"; it identifies a class of books, all of which (for the purpose of the ISBN) are regarded as being "the same" – or at least directly substitutable one for another. Two books are never of course completely identical – but for the purpose for which the ISBN was developed, as a *publisher's product identifier* to use in the book supply chain, they can be treated as identical. What is more, the ISBN is extensively used to identify books in contexts other than the one for which it was devised – with greater or lesser problems as a result. In the physical world, where carrier and content are so intimately bound together, the challenges of resource identity are less obvious and identifiers developed for one purpose can often act as proxies for another purpose. However, as the primary mechanism for the management and dissemination of content resources migrates from the physical to the digital environment, the challenges of coherent models of resource identity become much more pressing. There is an extensive discussion of many of the issues involved in the introduction to the RIVER project report[5], which it is not necessary to repeat here. In précis, when it is so easy to create and disseminate copies of resources, the exact identity of those resources becomes critical for users in many different contexts – although the ones with which we are particularly concerned for the purpose of this demonstrator are those within the discovery to delivery chain, and in particular those contexts post-discovery, in academic (institutional or subject based) digital repositories. The importance of interoperability between identifiers in supporting preservation is also key.

The challenge of identifier interoperability is being taken up within ISO TC46/SC9, the part of ISO which is responsible for a familiar group of standard identifiers – including ISBN, ISSN, ISRC, ISWC, ISTC, ISAN; the work of this group has been reported extensively in a recent D-Lib article by Norman Paskin.[6] The first part of this article, which is based (with acknowledgement) on work undertaken by Mark Bide for TC46/SC9, describes the nature of interoperability from the perspective of the organisations that are responsible for the management of international standards. In particular, it proposes that there are three different areas which merit attention in exploring what "identifier interoperability" means:

- **Metadata interoperability**, using different identifier metadata schemes
- The creation of standard mechanisms for the **expression of relationships between the referents** of different standard identifiers

[5] http://www.jisc.ac.uk/uploaded_documents/RIVER%20Final%20Report.pdf

[6] Paskin N (April 2006) "Identifier Interoperability: A Report on Two Recent ISO Activities" *D-Lib Magazine* 12.4 http://www.dlib.org/dlib/april06/paskin/04paskin.html

- The creation of **common services** which give consistent user experiences using different identifiers

It is our contention that these three areas of interoperability are at least as appropriate for exploration in the context of academic digital repositories as they are in the context of international standards.

The RIDIR team embarked on the project by attempting to organise two practitioner workshop focus groups. It was the intention that these workshops be attended by representatives from communities working on identifiers and digital repositories. The invitees to the first workshop would deliberately be practitioners in the field who deal with day-to-day issues; to the second, more expert individuals. The first workshop would be used to develop a set of views to help inform the construction of the RIDIR demonstrator software, whilst the second would be used to validate the project team's understanding of the issues from the first workshop and to help chart the course that it planned to take.

In the event it proved almost impossible to recruit for the first workshop focus group. Feedback from potential participants suggested that the project might be exploring areas that were not yet perceived as problems by most repository managers and that therefore they had done little thinking about the issue. Some of the invitees noted that they generated identifiers internally and had not yet considered interoperability. This was the project team's first firm indication that it was working in an area poorly understood, or even yet considered, by day-to-day practitioners and that many elements of the project as originally conceived would need to be revisited and rethought.

3. Methodology

RIDIR's intended methodology was straightforward. The project would start with a period of desk research followed by two User Focus Groups to scope the range of the demonstrator that was to be implemented. The first group would comprise repository staff who were relatively new to the work and who thus might be expected to have fresh opinions which may or may not reflect the 'orthodox view' in the field. A second meeting of more established practitioners would then be convened to discuss the output from the first meeting, add in their own perceptions and then help develop a first outline for RIDIR's development work. Following consultation with the JISC to establish their views about the outline, iterative development work on the demonstrator would then begin.

The actual experiences of holding the Focus Groups are discussed in the next section. Suffice it to say here that all did not go to plan, but ultimately the workshop participants identified four examples that they felt the RIDIR team should take forward; after developing these further the team presented them as five scenarios.

The RIDIR Project Plan set out the team's intention to consult with the Programme Manager, following the two focus workshop meetings, to agree the best way forward. Accordingly a Business Plan was produced for discussion at a meeting which took place during August 2007 in London. The Plan set out the five use cases and argued that RIDIR could take two different approaches to the issues identified:

- dealing with versions of objects where there might be a long chain of connecting events
- locating related objects to one just discovered (potentially both parents and children)
- dealing with issues encountered when an object becomes lost outside the curation boundary of its home repository (typically manifested to the user as '404' errors)

One the one hand the RIDIR Project could demonstrate how the ability to work with identifiers yields value, whilst on the other it could concentrate the demonstrator on showing how identifiers can be created, mediated and therefore the cost of managing them effectively on the assumption that the value of so doing is axiomatic. It was understood that the two approaches had elements of common ground.

The Business Plan was considered at some length in the meeting and, subsequently, by others that the JISC consulted. The result was a request that RIDIR concentrate on the how or cost approach.

Following this clarification the RIDIR team spent a considerable period of time analysing the five scenarios and turning them into more detailed use cases; these each identified the 'pain points' in terms of identifiers and suggested potential solutions in terms of manual processes and/or policies and in terms of possible RIDIR processes and services. After extensive discussion an abstract architecture for the RIDIR demonstrator was drawn up taking care that the principles underlying its workings were potentially applicable in a wider, real-world context. The extensive background research done at this stage of the project forms an important part of the project's output and is detailed further in Section 4 and Appendix C.

Work then started on developing the demonstrator which is now available to the JISC on data DVD. The software components are integrated in a VM-ware package that can, with relative ease, be deployed on a PC. Basic documentation is provided. During this period of development the team held weekly conference calls to assess progress and inform the next period of work.

The totality of the work undertaken during the project has enabled the project team to produce, not only a demonstrator, but also a complementary set of related recommendations for the repositories community.

4. Implementation

4.1 Workshop process and outcomes

As noted in Section 1, it was intended that the project should begin by holding two focus workshop groups. The first to help us understand better the requirements of the identifier and repository communities and the second, some time later, to validate the approach that we were taking to address the issues identified.

The first workshop proved almost impossible to recruit for. Repository practitioners seemed to have an understanding that objects in their repositories needed identifiers of some sort but did not seem to have considered the further implications of this and certainly not what interoperability between such identifiers might imply.

As a result, the first workshop as originally conceived was cancelled and, instead, a meeting of five well established repository practitioners was convened in London to brainstorm some of the problems. Desk research was used to provide that meeting with a range of possible scenarios in which identifiers would be important to repository and cross-repository tasks.

Following this meeting the team re-affirmed its view that RIDIR should be developed as an adjunct to the work of the PILIN project and not attempt to be an alternative to it. With this in mind an overall scope for RIDIR was determined, and an abstract systems architecture was developed to embody the scope.

A second focus workshop was held just more than two weeks after the first, this time in Manchester. There seven acknowledged leaders in the field discussed the outcomes from the first meeting, their own thoughts on the subject and, again, some of the scenarios that had been presented in London.

The second group identified four use cases that they felt were the most useful to pursue; the use cases covered the following:

- finding further digital objects related to a 'known' digital object
- locating the original version of discovered content
- establishing identifier chains between objects
- migrating repository content

These were ultimately reworked somewhat and presented to the JISC as the five use cases that are discussed below in section 4.2.3.

In addition, the group was invited to provide a wish-list of things that the finished RIDIR demonstrator might show in the absence of time or resource constraints:

- functional use of identifiers to support de-duplicating and explicit grouping of objects

- that identification of groups should be capable of representing multiple levels of granularity
- that relationships between (classes of) objects should be identified, for instance the 'hasPart' relationship
- ensuring machine-to-machine services are built
- an 'identifier cloud' interface, similar to the 'tag cloud' concept
- a 'crawler' to provide support in the user's maintenance of a semantic map, as a means to relate together identified objects (referents) and their classifications or types
- a 'push' mechanism to enable explicit updates to the semantic map from existing maps, from, say, chosen authoritative sources, and for user-specified semantic categories/concepts

The project deliverable 'RIDIR Focus Groups report'[7] sets out the detail of the workshop discussions

The RIDIR Project Plan set out the team's intention to consult with the Programme Manager, following the two focus workshop meetings, to agree the best way forward. Accordingly a Business Plan was produced for discussion at a meeting which took place during August 2007 in London. The Plan set out the five use cases and argued that RIDIR could take two different approaches to the three issues identified:

- dealing with versions of objects where there might be a long chain of connecting events
- locating related objects to one just discovered (potentially both parents and children)
- dealing with issues encountered when an object becomes lost outside the curation boundary of its home repository (typically manifested to the user as '404' errors)

On the one hand the RIDIR Project could demonstrate how the ability to work with identifiers yields *value*, whilst on the other it could concentrate the demonstrator on showing *how* identifiers can be created, mediated and therefore the *cost* of managing them effectively on the assumption that the value of so doing is axiomatic. It was understood that the two approaches had elements of common ground.

The Business Plan was considered at some length in the meeting and, subsequently, by others that the JISC consulted. The result was a request that RIDIR concentrate on the *how* or *cost* approach.

4.2 Requirements analysis and issue identification

4.2.1 Approach

Following the analysis of workshop outcomes and determination of project scope with the JISC, the approach taken on the project to arrive at a proposal that could serve as the requirements specification for the physical architecture and demonstrator software was as follows:

- development of an abstract architecture

[7] Available at: Http://www.hull.ac.uk/ridir/documents/index.html

- development of final use cases
- determination of the scope of the demonstrator software via Process Maps
- identification and analysis of issues:
    - draft formulation of some outcomes (best practices/recommendations)
    - research and draft foundational model for RIDIR functionality
- definition of requirements from a user perspective via 'narratives'
- detailed analysis of use cases and production of a demonstrator development proposal for team review

This final development proposal stage was not reached until near the end of February 2008 due to the difficulty in synthesising systems software requirements from the workshop outcomes. In agreeing the demonstrator proposal, the JISC Programme Manager also approved a short, unfunded, extension of RIDIR's software development phase to mid-May 2008.

4.2.2 Development of an abstract architecture

To meet the needs outlined in the workshop, it was decided RIDIR should demonstrate an *architecture* addressing the underlying identifier interoperability issues. Components of the architecture should be modular so that those principles could be rolled out in a production context within the JISC Information Environment in the future, should they be found to be applicable, by modifying or adding modules, but without substantially altering the architecture. This emphasis meant that no one component would be 'complete', but development activities should focus on completing enough to demonstrate that future projects could build upon the work without needing to revisit the fundamental approach.

Therefore an *abstract RIDIR architecture* was developed according to the following principles:

- RIDIR should not 're-invent the wheel', but build upon existing work in the field such as OAI-ORE[8] and FRBR.
- Metadata should be explicit, machine readable and interpretable

The treatment of metadata within the abstract architecture was by definition critical to the project, since metadata interoperability was an assumed integral part of identifier interoperability according to the project proposal. The principle that metadata must be “explicit, machine-readable and interpretable” led to the following assumptions and requirements of any system conforming to the abstract architecture:

[8] Open Archives Initiative Protocol - Object Exchange and Reuse See: http://www.openarchives.org/ore/

- to meet the needs of the community regarding the use of identifiers going forward, it is not enough to have the basic categories offered by Dublin Core,[9] and free text data for human interpretation of identifiers implicit in text

- there is not yet any usable framework or standard impacting (and possibly governing) the interoperability of identifiers for digital resources and their agents, although there are efforts underway and levels of consensus suitable for RIDIR to incorporate, notably the outcomes of the PILIN project, OAI-ORE and FRBR.

- explicit descriptions of resources, users/agents, and software services, are essential in order to build the rules and algorithms to resolve between different identifiers that originate in differing contexts

- metadata for these purposes must be comprised of abstract concepts, each of which has an addressable identifier

- concepts must be related to each other, using relations which are independently identified

- no concept can also be a relation at the same time; concepts and relations are disjoint - ie, two concepts must be linked by an identified binary relation which itself is not a concept. This is similar to, or identical to, the concepts underlying RDF and Topic Maps. For example, the concepts of Book and Author can be related by the relation 'hasAuthor', resulting in the 'triple': Book hasAuthor Author.

- the network of concepts and relations form a *semantic map*. The ability of the user to build, maintain and use a semantic map for the purposes of resolving identifiers has become a key requirement of the project across the three use cases identified as most significant.

[9] Dublin Core Metadata Initiative See: http://dublincore.org/

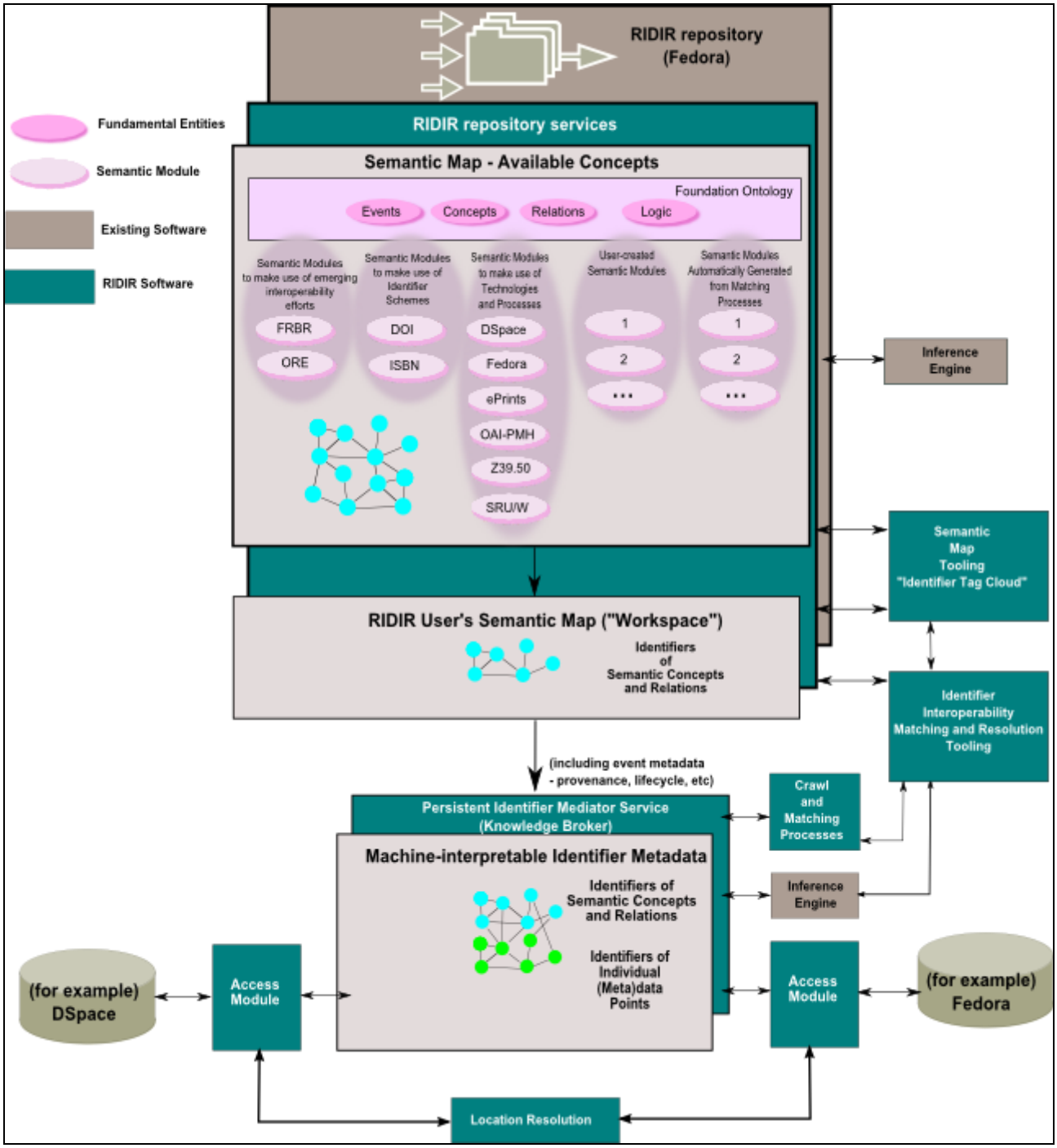

Figure 1: Abstract Architecture

The key areas of functionality covered by the architecture are:

- construction and maintenance of a semantic map to configure relationships between, and precise definitions of, any item having an identifier

- deployment of the semantic map into the Persistent Identifier Mediation Service, or PIMS (a RIDIR term). The PIMS was seen as a kind of 'knowledge broker' for a range of configurable services related to identifier resolution.

The architecture defined the following key components:

Semantic Map

- *Ontology and inference*: The idea of 'semantics' within the context of RIDIR refers to those meanings that can be interpreted unambiguously and systematically, which is a prerequisite of providing tool support. This essentially means that meanings ultimately must be expressed as collections of inference rules.

- *Framework (or Foundation) Ontology*: The framework, or foundation, ontology, consists of core concepts that must be considered axiomatic, ie 'givens' that are 'taken as read', for systematic processing to occur.

- *RIDIR User's Semantic Map ('Workspace')*: The semantic map would consist of an individual, user-specific semantic map. A user could build and maintain this not unlike building and maintaining a collection of tags found in modern web applications.

- *Semantic modules*: Built on top of the foundation ontology, explicit modules representing different views of what a digital object 'is' (so as to define what an identifier refers to) can be developed and mapped together, via the semantic map. A semantic module is where each 'view' is made explicit and is localised. Modularisation of semantic definitions would ensure identifier mappings remain generic, flexible, quick to construct and easy to maintain.

Persistent Identifier Mediator Service (PIMS)

The PIMS (or 'knowledge broker') sits at the heart the demonstration of identifier interoperability. Its function is to take the semantic map as its basis for deriving conclusions relevant to identifier interoperability, such as finding candidate objects which could be considered 'the same' or related in a certain way in the user's context.

The following functional components were identified:

- *Matching Subsystem*: The matching subcomponent is a key component of discovery of resource identifiers, in that it has the responsibility for executing rules that derive candidate matches from data and metadata presented to it, based on the identity modules; its 'intelligence' is derived ultimately from axioms within the foundation ontology.

- *Access subcomponents*: Access components should ideally be 'pluggable', encapsulating the underlying mechanisms for source repository and service access
- *Location Resolution*: The 'Location Resolution' subsystem is seen as an independent module to allow for variation in location identifier schemes, such as Handle and URL, and their location resolution methods.

- *User tool support*: Candidate identity matches would be under user control; ie, identified items would not be related through a single, global domain-specific scheme such as FRBR. In order to drive future adoption and therefore the success of the RIDIR component, the

system would instead put the user in control of the identification of relationships between identified items, but also allow them to adopt identities derived from controlled vocabularies or the classifications of other users.

- *Identity Crawler:* The default model for using the PIMS in an implementation would be as part of a discovery-driven process. A crawler implementation would process raw source data and the matching plug-ins propose any candidate identifiers implicit in that data. A key potential feature of the demonstrator that building a crawler would demonstrate is the automated generation of candidate semantic categories based on free text, rather than on a user's categories defined by mapping to established vocabularies or other users. This was considered potentially important as most metadata currently exists in free text form alone, for instance in Dublin Core fields (especially considering the importance of OAI-PMH[10] in the institutional repository (IR) context). In general, though, the implementation of an identity crawler within the demonstrator was considered a lower priority than a discovery-driven demonstration.

4.2.3 Developing the final use-cases

Following the development of a potential abstract architecture the team spent a considerable amount of time determining concrete scenarios to serve as requirements suitable for realising the abstract architecture in demonstrator software. The starting point for this was to identify candidate real-world use-cases (at that stage not much more than high-level scenarios) and look at the issues involved. In very brief summary, the use cases were:

EThOSnet

The EThOSnet Project, based at the British Library (BL), is handling e-theses that may exist in a BL repository or the originating institution's repository or both. There is a further possibility that copies of a thesis and/or its associated metadata may exist in repositories elsewhere. How can identifiers help a user make informed choices within this complexity?

The Depot

The Depot exists to provide a repository for research papers that must be made available by mandate but that cannot be deposited at the 'home' institution which does not yet have a repository of its own. Ultimately these papers will be transferred back to an institutional repository. The Depot does not currently use an identifier system that can be remapped to reflect the move. Potentially this results in broken links; how could RIDIR help?

Repository migration

There will be cases where digital objects need to be migrated from one repository to another, either within an institution or across institutions. This use case shares many problems with the Depot case outlined above.

Spoken Word Services

[10] Open Archives Initiative - Protocol for Metadata Handling See: http://www.openarchives.org

Spoken Word Services at Glasgow Caledonian University operate a large repository of audio-visual materials. Their objects have a range of relationships with objects in other repositories world-wide. How could a RIDIR tool facilitate the discovery and re-use of the many possible relationships that exist?

Locate related version

This is a more abstracted version of the Spoken Word scenario. How can a user, discovering a potentially useful digital object, find out about related objects that may potentially be of use. How can this information be recorded for re-use?

By mid-October 2007, the team had reached the following provisional conclusions:

- Firstly, RIDIR is a project to produce a demonstrator; it is not a system, or a service
- Secondly, it is unlikely that a single system or service could deliver the full range of functionality needed to solve the problems identified by the project
- Thirdly, the need for such functionality is at present immature. It is also highly specialised – the people who need it really know they need it, but no-one else cares very much. There isn't a middle ground to speak of.
- Finally, for any given case where interoperability would be facilitated by identifiers, there will be a number of people, systems and entities involved. These could be a different collection in each case.

The overall conclusion was the need to centre the project on defining a process (and perhaps significant variants of that process) rather than just a system.

The team had also developed detailed scenarios from the high-level summaries to achieve the following stated purposes:

- Identify what a RIDIR process would constitute, by considering the five use cases presented in the business requirements document
- Outline what practices would be required to make RIDIR processes work successfully
- Serve to demonstrate that RIDIR is a set of processes and practices, rather than a discrete systems deliverable, ie, an ongoing service (although this was not precluded)
- Illustrate the human factors comprising RIDIR
- Identify factors that demonstrate the importance of defining policy (since tools created in the absence of policy create a policy).
- Provide an emphasis on the intimate relationship between identity and metadata in the context of RIDIR – that identifier interoperability can be seen as the metadata from one identifier in the context of another one.

- That the function of the RIDIR demonstrator is not to identify correlations with certainty; rather, it should make a record that an individual (person) considers there to be sufficient evidence that they themselves make the decision that there is a match between two identified things (primarily, 'manifestations' or 'works' in FRBR terms).

By way of example, two diagrammatic depictions from the document are given below (pertaining to the use cases eventually implemented in the demonstrator). The overall role of a 'RIDIR process' and the issues the RIDIR project would address were also proposed for each use case.

Example 1: The Depot use case

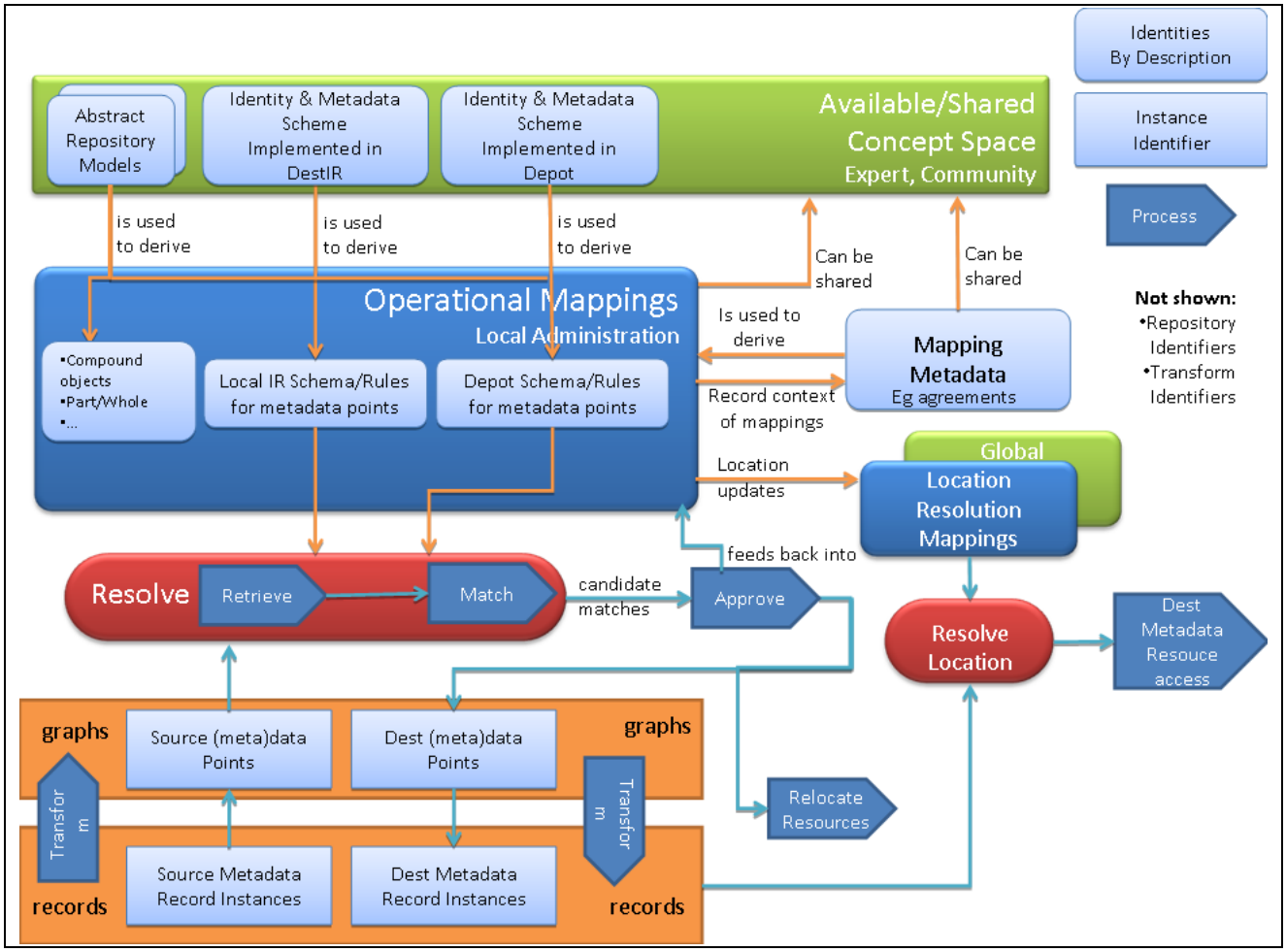

Figure 2: Depot Use Case Scenario

RIDIR process:

- maintain historical information concerning previous resource location (assuming no other system is present to record this information)
- maintain registry of identity and metadata schemes
- maintain relationship between identity scheme and metadata record structures
- maintain mapping between identifiers
- maintain explicit identifiers for expressing semantics of compound objects.

Issues considered by RIDIR:

- how to represent objects with different granularity
- how to assign identifiers to represent (ie, make explicit) the precise semantics of Resources, including Compound Resources, metadata records and their identifiers,

within Depot and Local IRs. (This is required to enable the system to suggest identifier mappings, and for the user then to authorise them.)

- maintaining a record of who authorised the mapping of identifiers (to typed objects and locations) as part of data pertaining to the mapping event.

- decision criteria for the IR Administrator, concerning the construction, management and application of rules and practices, that could be based in part upon the mapping event data, as well as the mapping of the static Resource entities, and their inter-relationships.

- should RIDIR represent guideline models for IR content structure, protocol, semantics, eg OAIS[11] or OAI-ORE, to assist mapping?

- when there are opportunities for the IR Administrator to execute transformations between schemes using the capabilities using the Source or Destination IR itself, could these be modelled and made explicit as 'repository capabilities' that would be of use to RIDIR?

- under what conditions would a researcher be presented with the location details of a resource?
- would such location details be restricted to the last location, or should a more comprehensive history of locations be maintained?

Example 2: The Spoken Word Services use case

[11] See, for instance, http://en.wikipedia.org/wiki/Open_Archival_Information_System

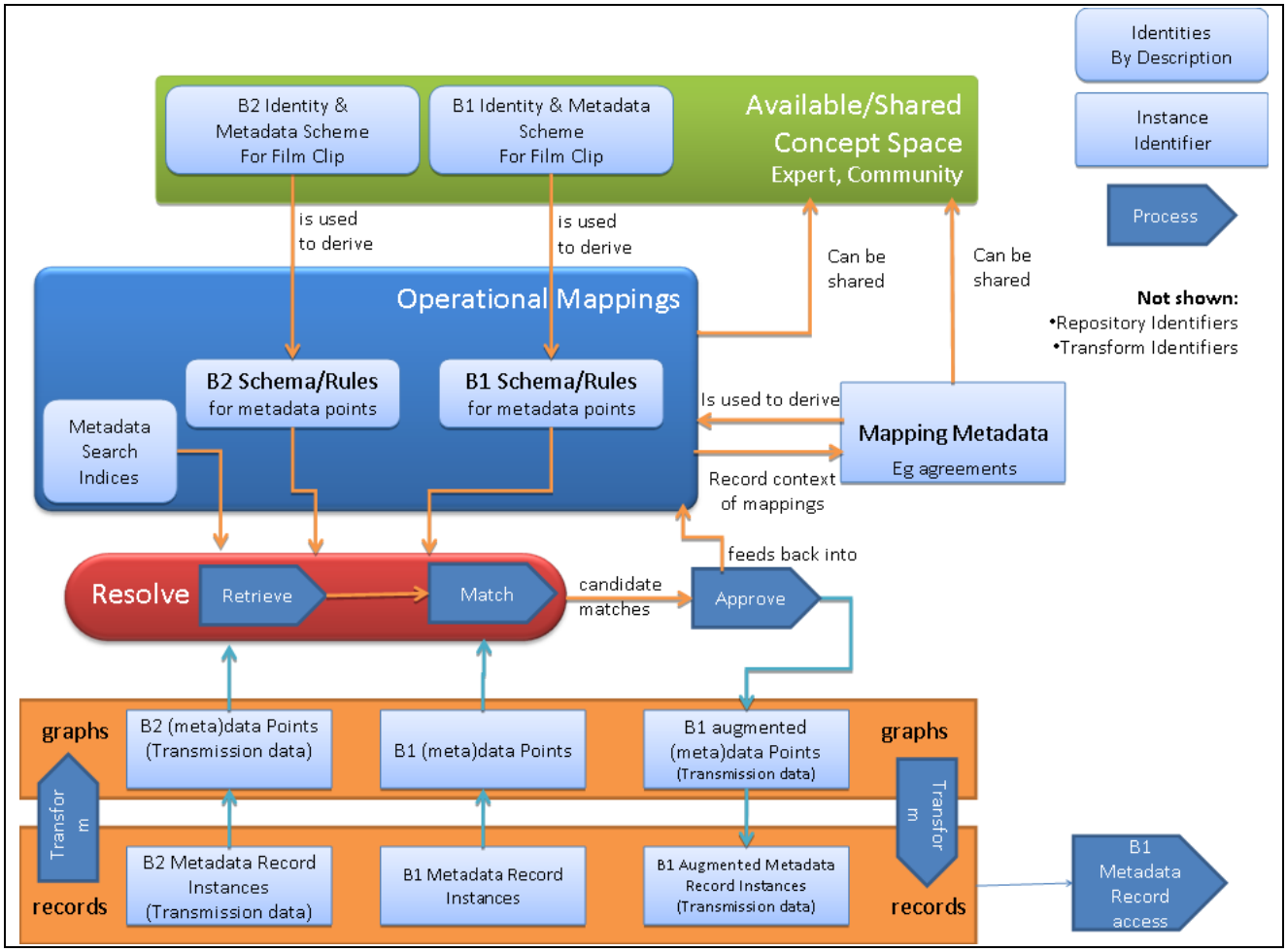

Figure 3: Spoken Word Services Scenario

RIDIR Process:

- maintain registry of identity and metadata schemes
- maintain relationship between identity scheme and metadata record structures
- maintain mapping between identifiers

Issues to be considered by RIDIR

- how are identities issued and propagated?
- how are relationships between identifiers maintained?
- what is the relationship between an identifier and a metadata record?
- how is a metadata record structured, managed, maintained, especially with respect to the identifiers contained within it?
- how is the resolution mechanism managed and maintained?
- how does a comprehensive history of locations come to be maintained?

4.2.4 Determination of the scope of the demonstrator software

4.2.4.1 Approach

To move from the use cases, whose proposed solutions reference the abstract architecture, to defining a set of concrete requirements to satisfy in the demonstrator, in October 2007 the team proposed the following steps be then taken:

- RIDIR 'process maps' would be developed to include all the steps made by different actors in the use cases to ensure that identifiers can be used reliably to facilitate interoperability. The processes require actions to be performed by both humans and machines. We do not believe that all steps can be performed by machines (certainly at the moment; perhaps ever).

- The 'RIDIR process' would be mapped for each use case, and identify all the components and steps that would be needed to make it work. Many of these would be outside the scope of this project.

For each RIDIR process, the approach would be to identify the following 'views':

- The *abstract process* showing the steps and entities that would comprise the process under ideal circumstances.

- The *real-world process* showing the steps and entities that would comprise the process under current conditions: not all those required by the abstract process might be present, and some might require multiple steps of components to achieve the same as a single step or component in the abstract process.

- The *risk/quality process* identifying where the real world process needs to be strengthened or enhanced to bring it closer to the abstract process. Risk points occur when a chain of identification might be broken; quality points occur where the chain could become less reliable or robust.

By the end of 2007 the team had performed the requisite analysis and developed process map documentation for all five use cases, and was in a position to review them in early January 2008.

In the event the presentation of risk/quality process was simplified in terms of 'pain or "cost" points' (in recognition of the development direction identified for the project by the JISC), and potential solutions proposed in each case, both in terms of manual RIDIR processes and policies and opportunities for automated RIDIR processes services. The process map analysis also identified a number of assumptions requiring clarification prior to firming up system requirements for the RIDIR demonstrator software.

4.2.4.2 Results of analysis

By way of an illustrative example, the following summarises the pain/cost points from the Spoken Word use case:[12]

[12] The full set of use cases are to be found in the document *Report from National Workshops including use cases and development plan* available at http://www.hull.ac.uk/ridir/documents/index.html

Pain Points/Cost Points:

- No relationships between identifiers when clip is copied, relationships not persisted
    - relationship to original metadata not present; not possible to determine rights and descriptive metadata
- Difficult to mediate across different identifier schemes
    - different identifier schemes in use in different archives and repositories, difficult to determine the identifier of the original clip
- Potentially multiple part/whole identifier and metadata schemes
    - different schemes in use for describing excerpts of a clip, difficult to determine where an excerpt was sourced from
- Objects may be part of collections, metadata may be associated with the collection rather than the object
    - semantics not made explicit, difficult to identify the collection and the associated metadata
- Disambiguation of free-text metadata
    - Usage of free text to identify people's names (for example) makes it difficult to use this information to relate between versions of the clip
- Knowing which repositories to search
    - If no 'registry of repositories' is provided, difficult to know which repositories to search for related versions and their metadata.

Potential solutions:

Manual processes and policies

- Persist identifiers
    - When a clip is copied to another repository, ensure that the identifier of the original is persisted in the metadata of the new copy
    - Will require usage of appropriately scoped identifier schemes

- Document metadata and identifier schemes
    - including part/whole relationships expressed in identifiers or metadata
    - including collection membership information
- Document repositories
    - Provide list of repositories and the type of resources they contain

Potential RIDIR processes and services

- Maps of metadata and identity schemes
    - including cross-scheme metadata relationships
    - including semantics of identifier schemes
    - including collection and part-whole schemes (compound and complex objects)
- Discovering relationships between resources and persisting those relationships (crawling metadata and proposing candidate matches)
    - including information on who asserted relationships so authority/veracity of the relationships can be assessed by other users
    - including usage of any disambiguation services available, so relationships to resources (eg authors, creators) identified in these can be persisted

The following list of issues and pain/cost points were identified as a result of this analysis:

- ***Identifiers change (resource moved): resource cannot be located***

  The identifiers used are not 'true' persistent identifiers, in the sense that a guarantee that an identifier will resolve to the intended resource for all time cannot be absolutely guaranteed. Commonly, identifiers may be URLs that are used to indicate the current location of a resource. When objects are moved from one repository to another, the URLs change, as the URL syntax identifies the location of the resource, the system used to serve the resource, but not the resource itself.

- ***Resource deleted; identifier refers to non-existent resource***

  Resources are deleted after identifiers have been published for the resource.

- ***More than one copy of resource, cannot determine appropriate copy***

  Unable to resolve to the most appropriate copy of the resource for the user accessing the resource. User may not be able to access the resource as a license only allows access to the institution's local copy, which cannot be resolved to.

- ***Not clear what identifier referent is (eg, 'raw' resource, splash page, metadata)***

  Identifiers created that refer to resources, to splash pages and to metadata for resources; no consistent usage of these different identifiers so that it is clear what is being identified in a particular context. Particularly important for machine-machine interactions (eg metadata crawling and discovery, discovery of related versions).

- ***Free-text metadata difficult to disambiguate***

  Expression of people's names (for example) in free text makes it difficult to identify when one has found the 'right' John Smith.

- ***Relationships between objects not persisted (objects, metadata enrichment)***

  No mechanisms for persisting relationships between objects once they have been discovered leads to duplication of effort in rediscovering these relationships.

- ***Mapping between metadata schemes (Mediation)***

  Requirements to map between metadata schemes. This could also include usage of metadata (what gets indexed, what gets presented), and the syntax/packaging of the metadata.

- ***Mapping/translation of taxonomies, thesauri, controlled vocabularies***

  Different repositories may use different semantics and different mechanisms for controlled vocabularies, taxonomies and thesauri that need mapping

- ***Mapping between identifier schemes***

  Mapping between different identifier schemes, including dealing with syntactic restrictions in different schemes, dealing with semantics implicit in the identifier syntax.

- ***Mapping between different object/content models***

  Mapping between both the content models implemented in different repositories and the content models implicit from the repository software and different way the repository software chooses to model digital objects.

- ***Mapping/translation between object packaging and ingest schemes***

  Mapping to and from schemes for packaging and describing objects ready for ingest.

- ***Mapping/translation between different ownership and security models***

  Mapping between different repositories' models for handling object ownership and between repository-specific security model implementations.

- ***Need to handle complex objects and collections***

  Ability is needed to deal with part/whole relationships and collections.

- ***Location of appropriate repositories, where to search***

  A list or registry of repositories with information on what resources are contained in each and details of how to access the repositories is required

- ***Information on assertion of relationships is required***

  It is necessary to know who claimed that a particular relationship between objects or metadata items is present to make an assessment of the authority and/or veracity of the relationship for other users, ultimately as a means to facilitate the formation of mechanisms for establishing trust.

- ***Mapping/translation between different versioning schemes***

  Mapping between different schemes of representing versions is required.

- ***Reintegration issues***

  Joining up with other services, eg integration with persistent identifier infrastructure, integration with harvesting services

- ***Implicit metadata that needs making explicit***

  There is implicit information about objects in a repository that is not explicitly stated in metadata; for instance migrating a repository known to contain MPEG-2 clips to a general multimedia repository; the MPEG-2 repository does not explicitly state that its contents are MPEG-2; all of the users of the repository are aware that the repository is there to hold MPEG-2 objects.

- ***Other repository-specific and technical issues***

    - Managing stateful repository constraints
        - eg constraints on the order of ingest of objects – children must be ingested before relationships to parents can be created or vice versa.
    - Catering for duplicate objects found in source repository
    - Handling any orphaned objects discovered
    - Recording of provenance information about the mapping
    - Validation of the migration and test procedures
    - Resolution of object requests based on location identifiers of source repository after migration
    - Processes involved in migration through pain points listed above are not easy to automate.

One objective of this analysis piece was to show that the potential scope for building a demonstrator was vast, and that an exercise was needed to help the project focus down on some specific functionality which was relevant in terms of value (pain/cost), which could be related to a real-world, demonstrable use case, and which was possible to implement in a demonstrator within the diminishing amount of time available.

To this end the list of issues and pain/cost points was cross-referenced against both use cases and functionality, shown in the following tables:

Table 1: Issues and pain/cost points related to scenarios

	Scenarios				
	Depot	EThOSnet	Locate Related	Spoken Word	Migrate Repository
Issue					
Identifiers change (resource moved): resource cannot be located	H	H		?	L*
Resource deleted; identifier refers to non-existent resource		H		?	L*
More than one copy of resource, cannot determine appropriate copy		H		?	
Not clear what identifier referent is (resource, splash page, metadata)	M			?	L*
Free-text metadata difficult to disambiguate	L	M	M	M	L*
Relationships between objects not persisted (objects, metadata enrichment)			H	H	
Mapping between metadata schemes (Mediation)			H	H	H
Mapping/translation of taxonomies, thesauri, controlled vocabularies					H
Mapping/translation between identifier schemes			H	M	H
Mapping between different object/content models					H
Mapping/translation between object packaging and ingest schemes					H
Mapping/translation between different ownership and security models					H
Need to handle complex objects and collections		?	H	H	H
Location of appropriate repositories, where to search			M	M	
'Which is the best, which is the original'			L		
Mapping/translation between versioning scheme					H
Reintegration issues					H
Implicit metadata that needs making explicit (eg technical metadata)					H
Other repository-specific and technical issues					H

Notes:
The relationships between issues and scenarios have been rated H (high), M (medium) and L (low) to indicate how important we believe the issue to be to each scenario.
L* indicates a low priority for repository migration, but a high priority for repository customers.

Table 2: Issues and pain/cost points related to potential RIDIR functionality

	Service/Functionality							
Issue	Persistent Identifiers	Resolution	Contextual Resolution	Defined ID semantics	Semantic maps	Identity crawling	Relationship persistence	Repository registry
Identifiers change (resource moved): resource cannot be located	X	X						
Resource deleted; identifier refers to non-existent resource	X	X						
More than one copy of resource, cannot determine appropriate copy	X		X					
Not clear what identifier referent is (resource, splash page, metadata)				X				
Free-text metadata difficult to disambiguate						X	X	
Relationships between objects not persisted (objects, metadata enrichment)							X	
Mapping between metadata schemes (Mediation)					X			
Mapping/translation of taxonomies, thesauri, controlled vocabularies					X			
Mapping/translation between identifier schemes					X			
Mapping between different object/content models					X			
Mapping/translation between object packaging and ingest schemes					X			
Mapping/translation between different ownership and security models					X			
Need to handle complex objects and collections					X			
Embedded identifier semantics					X			
Location of appropriate repositories, where to search								X
'Which is the best, which is the original'								
Mapping/translation between versioning scheme					X			
Reintegration issues								
Implicit metadata that needs making explicit (eg technical metadata)								
Other repository-specific and technical issues								

To simplify the demonstrator scoping exercise further, three main directions were identified and proposed as options for demonstrator implementation, of which only one would be achievable within the remaining time frame. Each option represented a grouping of the potential functionality and services identified within each use case:

1. Focus on usage of persistent identifiers and resolution mechanisms

   Implementation of services to provide persistent identifiers and their resolution would go a long way to resolve the issues in the Depot and EThOSnet scenarios. It was proposed that these services should be designed and implemented after the semantic requirements for other scenarios have been evaluated in greater depth (potentially outside the RIDIR project). The PILIN project was recognised as a useful source of information in the context of a service for persistent identifiers and their resolution.

2. Focus on discovering related resources with support from Semantic Maps for metadata mediation

   Good balance between reasonable 'spread' across proposed abstract architecture components and use cases. A risk was identified whereby this option would not be seen to be specifically tackling 'Persistent Identifiers' without provision of sufficient context of the RIDIR analysis and findings.

3. Focus on Migrate Repository

   Though the work would undoubtedly be useful, there was a potential for a large amount of effort being spent on technical issues specific to each repository implementation, with less effort being spent on more generic RIDIR services and functionality that would have wider usage outside of this use case.

It became clear in this analysis that a number of the 'pain points' could largely be alleviated by implementation of an appropriately scoped JISC-wide, shared persistent identifier management scheme and resolution service, such as that scoped in Australia by the PILIN Project. PILIN considered in depth the use of identifiers and resolution mechanisms and in particular the benefits to be had from a centrally managed, but shared, set of policies and services. It was clear to the RIDIR team that duplicating this effort would not be productive and, as noted at section 1.4, there was an agreement that it should not do so, thus Option 1 was rejected. RIDIR would add little value by simply demonstrating identifier schemes and services.

The focus on 'migrate repository' functionality was likewise rejected on the grounds above; the value of the RIDIR approach within this context would be difficult to demonstrate within the limited time available, and was felt likely to overlap with other related initiatives, most notably the, at that time yet-to-be-released, OAI-ORE specification.

Instead it was agreed within the team that RIDIR should focus on functionality falling within Option 2, firstly examining the potential relationships between identifiers in the broadest sense and how these relationships could be created, described and navigated to aid discovery and interoperability in general. This functionality corresponds to the 'Locate Related (or Original)

Versions' use case, looking at identifying and persisting of loose chains of identifiers. In adopting this focus, the RIDIR work becomes complementary to a PILIN-like approach by offering a " 'value added' identifier enabled service" as conceived in this diagram which is taken from page 29 of the PILIN Closure Report[13] and which shows a possible infrastructure for a centrally managed, shared identifier management system. The work of the PILIN Project is further considered in our Recommendations in Section 5.

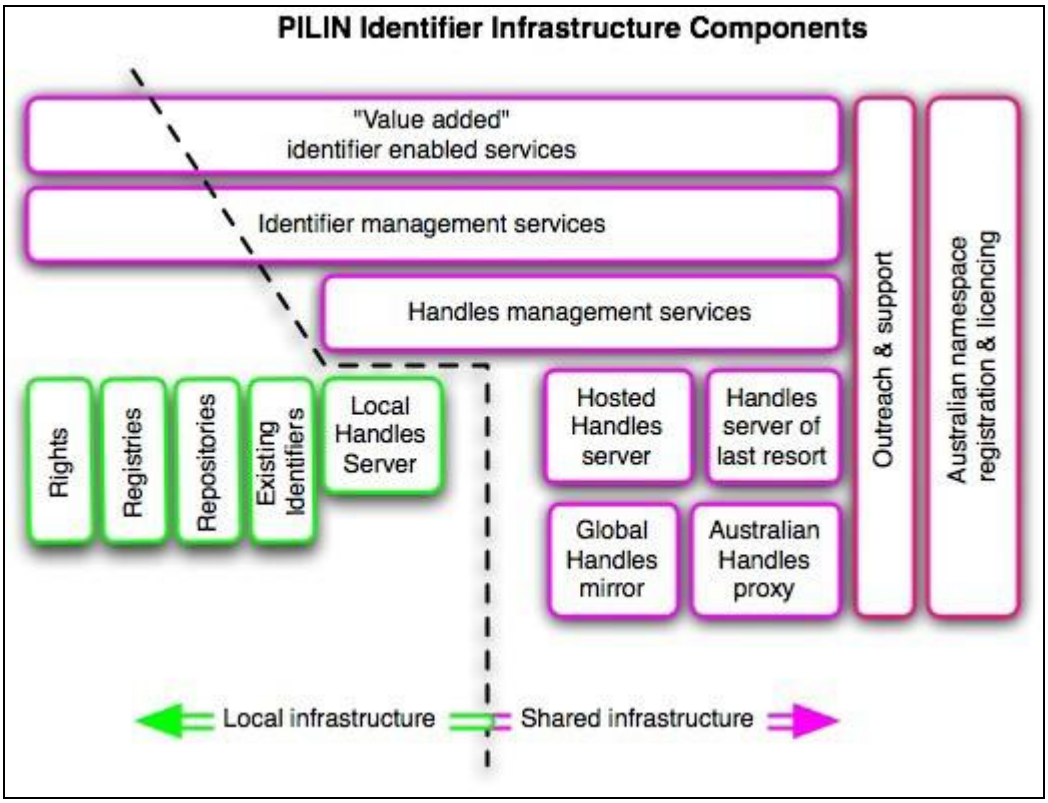

Figure 4: PILIN Identifier Infrastructure Components

As noted above, the RIDIR team decided that its work would most usefully be deployed in exploring shared, 'value added' identifier enabled services, as shown in the top box of the diagram. A key aspect of this thinking was that even with policies and services akin to those covered by PILIN's scope in place, there will always be the 'corner cases' that RIDIR would address, meaning those situations which have not implemented such policies and do not adhere to shared service usage models, which will require attention. The degree to which this will become an issue cannot be ascertained at present, but clearly with the increasing adoption of institutional repositories with potentially divergent management policies, resource and identifier resolution curation boundaries and differing technical implementations, the issue is likely to be increasingly significant.

Specifically, in terms of building demonstrator software, RIDIR could explore how to deal with digital objects that have somehow become 'lost', which is to say that their URL no longer resolves appropriately, and how identifiers could be used in the creation of a network of relationships between one object and others in such a way that a user discovering one object in the network is made aware of other, related objects.

In summary the outcome of the analysis was that:

[13] PILIN Closure Report. See: http://resolver.net.au/hdl/102.100.272/RPG891PQH

- RIDIR should not focus on identifier schemes and resolution mechanisms *per se*, as this would be duplication of other efforts such as PILIN and would add little value. Similarly the, at the time forthcoming, work of the Open Archives Initiative's Object Reuse and Exchange (OAI-ORE) project should not be duplicated if possible.

- Even when appropriate persistent identifier schemes and services are implemented, there will still be 'corner cases' which RIDIR could address, arising from objects moving outside of the curation boundaries/scopes of these identifier schemes; deficient design of the identifier schemes; poor or no implementation of the schemes

- RIDIR should focus on how relationships between referents of various identifiers are created, described, themselves identified and navigated to aid interoperability, and develop an understanding of the significance of these relationships.

- The RIDIR demonstrator should not be a 'black box', but should make visible to the demonstrator audience how relationships are created and what these relationships are

- RIDIR should primarily focus on user-driven discovery of relationships between identifiers, and the persistence and usage of these relationships, whilst recognising that it may be possible in the future to have machine to machine discovery of relationships

In terms of practicality it was determined that the five workshop scenarios could be combined into two more generic ones: the first based around potential issues identified with objects passing outside a curation boundary becoming 'lost' and the other based around the problems of locating related versions of an object:

- A resource is relocated, and the existing identifier for that resource ceases to function in terms of locating the resource for access
    - Mediation to determine if RIDIR contains a new location for the resource
    - If no new location is present in RIDIR, discovery to seek out a new resource location
    - Persistence of discovered relations
- A resource is found for which it is desired that related resources are located
    - Mediation to determine what relationships (if any) have already been identified to related resources
    - Discovery to seek out related resources and their identifiers
    - Persistence of the discovered relationships

Where the scenarios differ is in the mediation and discovery aspects, the 'types' of relationships persisted and the user interface. The semantics used, for instance, in The Depot example to identify the relationships between identified things and their locations would be different to those used in the Spoken Word one. Similarly, the relationships used in the Spoken Word

example between different versions of resources are likely to be different to those used in Locate Related Version. The mechanisms used for discovery of new relationships would be different across the scenarios. That said, the process for asserting, recording and retrieving all these relationships would be very similar.

The RIDIR demonstrator's implementation architecture would be based upon the concepts represented in the abstract architecture, divided into 'core services' which represent automated RIDIR functionality, and 'application services', which are use-case specific and serve as a setting for the demonstration only.

4.2.5 Identification and analysis of issues

During the work done to move from the Workshop report outcomes to a proposal for demonstrator development, a significant amount of research was done concerning the general aspects of identification and interoperability, to evaluate the various approaches that the RIDIR demonstrator might take with respect to existing work, and to ensure value was added rather than duplicating extant work.

This phase of the project included analysis of existing identifier schemes and services, existing services for enabling interoperability, and led to some best practice recommendations.

Analysis was also conducted at a more theoretical level to address the area of metadata interoperability, identified in the background to the project as meriting attention in terms of clarifying what is meant by "identifier interoperability", and is described in Appendix D.

Presented in a later section are some overall recommendations concerning identifiers and interoperability. We believe that presenting these, then setting the outcomes of the demonstrator software development and resulting applications in that context, allows validation of some of these recommendations; it effectively demonstrates 'how to make it work' by giving real-world situations where adoption of these recommendations would 'make it work'.

Regarding the outcomes of the theoretical investigation into identifier and metadata interoperability, there is evidence to suggest that a commitment to a foundation model, expressed in terms of an ontology, is a important component of future activity.

What, then, did RIDIR find in its research?

4.2.5.1 What do we mean by interoperability?

From Paskin [14]

- Metadata interoperability, using different identifier metadata schemes
- The creation of standard mechanisms for the expression of relationships between the referents of different standard identifiers

[14] Paskin N (April 2006) "Identifier Interoperability: A Report on Two Recent ISO Activities" *D-Lib Magazine* 12.4 http://www.dlib.org/dlib/april06/paskin/04paskin.html

- The creation of common services which give consistent user experiences using different identifiers

From PILIN [15]

Interoperable

"A component is interoperable if an action can operate on the component from outside the curation boundary of the identifier management system. The action must follow a well-defined interface, which is known outside the curation boundary. If a component is not interoperable, then only the identifier management system's own infrastructure can be used to operate on it. If the action uses a publicly documented interface through an open protocol such as Web services, it is interoperable."

From Paul Miller [16]

"to be interoperable, one should actively be engaged in the ongoing process of ensuring that the systems, procedures and culture of an organisation are managed in such a way as to maximise opportunities for exchange and re-use of information, whether internally or externally."

From the JISC [17]

"Interoperability requires commonly agreed standards and protocols. Standards exist at different levels and types of interoperability. The prospect is emerging for a broad set of standards across different aspects of terminology services [TS] - persistent identifiers, representation of vocabularies, protocols for programmatic access, vocabulary-level metadata in repositories. Such standards are an infrastructure upon which future TS will rest but it is not feasible to wait for international agreements; international consensus will be influenced by operational experience. Pilot TS projects should orient to existing potential standards (in persistent identifiers, representations, protocols for programmatic access) and help to evaluate and evolve them."

4.2.5.2 The Institutional Repository context

The issues surrounding identifiers and interoperability need to be set in the context of the range of services and responsibilities of institutional repositories.

- **preservation** of a resource – ensuring that the resource is adequately preserved over archival time spans

[15] See the PILIN 'Glossary' at: https://www.pilin.net.au/Project_Documents/Glossary.htm
[16] Interoperability - what is it and why should I want it? http://www.ariadne.ac.uk/issue24/interoperability/

[17] JISC Terminology Services report (2006)
http://www.jisc.ac.uk/media/documents/programmes/capital/terminology_services_and_technology_review_sep_06.pdf

- **description** of a resource – providing descriptive metadata about the resource and making that metadata available, to *facilitate discovery* of the resource, and to lend credence that the preserved resource is the correct resource.

- **identification** of a resource – creation of (usually) a text string identifier and *association of that identifier with the resource*

- **resolution** of the identifier – ensuring that the identifier, when actioned, can be used to *access* the resource.

These services and responsibilities are necessarily tightly-coupled, for instance description of a resource is necessary to give strength to the claim that the identified resource is the one served when the identifier is resolved. Of particular importance in the RIDIR project is identifiers and their resolution with respect to providing continued interoperability. Several of the use-cases examined by the project (EThOSnet, Depot, Migrate Repository) involved the potential movement of resources from one location to another. This can lead to changing responsibilities for the above services, and in particular in the desired scenario where identifiers are to be persistent, can lead to different organisational units being responsible for the preservation and resolution responsibilities.

The PILIN project identifies as important the concept of a *Curation Boundary*[18] and in particular identifies that there are curation boundaries for both the resources and for their identifiers. These two curation boundaries are particularly important with regard to persistent identification schemes and services and their governance. For instance, if an institution operates a repository and maintains identifier resolution for its resources, and a resource is then moved to a different institution's repository, the original institution has the continued responsibility for resolution, but no longer has the responsibility for preservation. In itself this will not necessarily lead to a breakdown in interoperability, but there is an increased risk of this happening – the original institution does not have the same level of motivation to provide resolution services for resources in other institutional repositories as it does for its own resources. Therefore selection of the appropriate curation boundary for identifiers needs to encompass the likelihood of movement of resources between different organisations.

4.2.5.3 Identifiers and persistence

There is generally a trade-off between the persistence of an identifier and its continued actionability (ie the continuing - persistent - ability to resolve to whatever is defined to be the correct resource). (There may be differing definitions of "correct" depending on the situation; for example, the PILIN project's FRBR tool will resolve to the most recent manifestation of a resource at the FRBR 'work' level, whereas if an identifier identifies a particular representation of some content, an exact copy at the bit-stream level might be more appropriately considered "correct"). Truly persistent identifiers would have no information encoded (syntactically) within them that may be required to change during the lifetime of the resource. This would include information about where the resource is located and the location of the service responsible for resolving the resource – as these pieces of information may change during the lifetime of the resource, resulting in the identifier no longer performing as an actionable identifier.

[18] See: https://www.pilin.net.au/Project_Documents/Community_Guidelines/ID_Association_Guidelines.htm#id233)

Identifier schemes that have identifiers which encode no information regarding resolution responsibility imply a single point of resolution (since no decision can be made based on the identifier on where to go to in order to resolve the identifier); and furthermore a single point of resolution cannot then examine the identifier to delegate the resolution to another service. With a large number of identifiers, this is likely to lead to issues regarding performance and scalability (and there are other issues such as having a single point of failure). Examples of such schemes are UUIDs[19] and ISBNs.

Identifiers which do encode a resolution service responsibility or location with them do not suffer from this, especially with devolved resolution models such as Handle and DNS,[20] which are designed to address these scalability and performance issues. In this case, appropriate choice of the identifier curation boundary is essential, to avoid the identifier not being persistent, as discussed in the previous section. A solution here is to choose larger and larger curation boundaries. However this leads to larger and larger numbers of identifiers that must be resolved by a single service, and this results in the same issues of scalability and performance.

In general, there is no perfect solution to these issues, and RIDIR considers a layered approach to be the best solution, including:

- having a reasonable degree of uniqueness in some part of the identifier;
- choosing a curation boundary appropriate to the specific environment[21] (especially the set of policies and agreements surrounding resource curation) and the anticipated lifetime of a resource;
- carrying descriptive metadata with the identifier.

These points are considered in detail as part of the Best Practice Recommendations summarised below and dealt with in detail at Appendix C.

In general it must be emphasised that there is no single solution to these issues, and systems (and practices) should be designed to cope with failure. Were a national, shared identifier infrastructure service to be implemented, certain procedures could be included to cope with failure. For example, the PILIN project has identified that offering an “identifier of last resort” would be a key role for a national service. Here, metadata would record the last known provider of the resource, explicitly recording any curation boundary changes, so that a physical address for the resource in its last known accessible state can be determined and the institution responsible contacted.

The RIDIR team have considered these issues and produced a summary of best practice recommendations. These are presented in detail at Appendix C and listed in summary here:

A. Minting identifiers for resources

1. Mint resolvable persistent identifiers
2. Identifier structural semantics should have the same lifetime as the resource
3. Use semantically opaque identifiers

[19] See for instance: http://www.iso.ch/cate/d2229.html, ISO/IEC 11578:1996 and http://tools.ietf.org/html/rfc4122 RFC 4122
[20] See, for instance, http://en.wikipedia.org/wiki/Domain_Name_System
[21] The PILIN project identifies a trade-off between the size of the curation boundary and the performance and scalability of the associated identifier resolution system

4. Mint universally unique semantically opaque identifiers
5. Use universally unique identifiers within resolvable identifier schemes
6. Consider human communication factors
7. Generate identifiers early in the origination process
8. Provide semantically-precise descriptions of what is being identified
9. Combine preservation and resolution responsibilities
10. Maintain a registry of identifier syntax

B. Publishing and citing resources

11. Include descriptive metadata in resolution services
12. Include descriptive metadata when citing resources
13. Carry old identifiers in metadata when moving objects
14. Use disambiguation services
15. Provide capabilities for user-generated metadata
16. Use metadata standards and provide clarification and best practices for usage of standards

C. Resource discovery

17. Implement automated resource rediscovery mechanisms
18. Don't rely on identifiers being persistent

D. Linking of resources

19. Provide resource linking capabilities with semantics, publish relationships

Regarding the adoption or development of identifier schemes themselves, the question of metadata associated with each scheme must be considered. Given metadata interoperability is a critical component of ensuring identifier interoperability, it is evident that a commitment to an extensible foundational model, expressed in terms of an ontology, is an important component of future activity. Specifically, the IRE ontology[22] was analysed for reasons given in more detail in Appendix D.

It seems clear from this work that such an ontology must clarify the relationship between an identifier and its referent(s), which may or may not be a 'real-world' object, and that the four-layer model of reference should be considered in future work. This work is especially important given that the work conducted by the W3C Technical Architecture Group[23] towards formalising the web architecture[24] is not specifically targeted at the needs of the institutional repository community; more specifically, its definitions of “information resource” and “non-information resource” as the referents of (resolvable) HTTP URI identifiers alone are unlikely to suffice.[25] Such issues are dealt with to some extent by the PILIN ontology[26], and suggestions have been made as to the application of FRBR definitions to the web architecture.

[22] IRE ontology. See http://wiki.loa-cnr.it/index.php/LoaWiki:IRE
[23] W3C Technical Architecture Group See: http://www.w3.org/2001/tag/
[24] Web architecture See: http://www.w3.org/TR/webarch/
[25] A discussion of the W3C TAG findings and information resources can be found in the paper “URIs and the Myth of Resource Identity” See http://dbooth.org/2006/identity/
[26] PILIN Ontology for Identifiers and Identifier Services: See https://www.pilin.net.au/Project_Documents/PILIN_Ontology/Ontology.htm

Whilst the OAI-ORE efforts go a long way towards defining mechanisms to allow repositories to describe and communicate their contents in a way that addresses the web architecture, the project does not have within its scope the kind of foundational theory of identity and reference that the RIDIR work suggests is required to achieve interoperability at a semantic level. OAI-ORE allows any aggregated resource to be given a type in RDF, but it is also important to adopt a means to integrate the types and their definitions themselves. The RIDIR project felt it important to consider agreement over types, facilitated in two ways, the first is being through adoption of an 'already agreed' controlled vocabulary or ontology designed to support a community need, as in the case of FRBR, the second through 'ad-hoc' agreement on terminology, exemplified by tag-based classifications popular on the web. A foundational ontology must carefully consider supporting both cases in order that the choice of identifier scheme and associated software remains robust over time.

The RIDIR analysis illustrated that behaviour of software services required to manage and interact identifiers and metadata could be modelled in terms compatible with the foundational ontology. The benefit of this approach is in terms of identifier interoperability across systems and differing software implementations. This part of the work could only be conducted to a very preliminary stage, but suffices to illustrate that the approach should be considered in future efforts, alongside appropriate adoption of related standards, primarily OAI-ORE.

A specific case arises where the size of the curation boundary over archival time spans may require consideration; for instance, in the case of a national identifier service, agreements between the national body and the provider of resolution services and even the resolution mechanism itself may be subject to change. Even if infrequent, the foundational model governing the handling of identifies must be resilient to such change, and explicitly identify resolution mechanism for example, in order that the persistence requirement be met.

4.2.5.4 Barriers to interoperability

Factors which represent barriers to implementing a persistent identification service for institutional repositories were identified, and include:

- Integration of such a service within institutional repository software

- Diverse persistent identifier implementations within popular institutional repository systems. DSpace, EPrints and Fedora were examined to varying degrees of detail: all have persistent identifier capabilities in their own terms, but are far from interoperable 'out of the box'. To illustrate this point around the Handle system, at the time of writing, DSpace implements Handle as standard, with institutions commonly registering themselves as Handle naming authorities. Should a shared infrastructure service based on Handle be adopted further customisation (and migration of identifiers) would still be required to take account of the naming authority scheme used for the shared infrastructure service. In this circumstance, EPrints and Fedora would require integration with a Local Handle Server at a technical level as well as a policy level to ensure the native identifier implementation works in concert with the handle identifiers required.

- Further technical issues may arise if a non-standard Handle implementation is required for the JISC IE[27] for a national service, requiring integration modules for each software to be custom.

- Lack of a standard 'fallback' or failure policy set and associated services, eg at a national level.

4.2.6 Research and draft foundational model for issues within RIDIR scope

The work outlined above provided a basis for the development team to research, identify and further develop a foundational model to clarify those issues that were determined should fall within the scope of the RIDIR project. The overall objectives were as follows:

1. To clarify the meanings associated with terms, and further develop a vocabulary for RIDIR as issues became clarified
2. To represent these meanings, and the issues they cover, with formal semantics if possible, as a formal ontology[28], so that the model possesses unambiguous semantics and which are therefore machine-interpretable, thereby forming a technical basis for semantic interoperability (for such metadata associated with an identifier expressed in a form amenable to 'semantic' processing, such as RDF)
3. To model software interactions with the foundation model in terms of formal ontology, in order to provide future directions for RIDIR work and to help inform development of the simpler ontology implemented within the demonstrator software
4. To evaluate and potentially engage with any communities and practitioners covering similar or related scope
5. To meet a requirement of the RIDIR abstract model developed early in the RIDIR project.

To summarise, a custom-built version of the IRE ontology based on DOLCE[29] upper ontology and expressed in the Web Ontology Language OWL[30], was identified early as the most promising candidate within the limited development schedule available. All further work on the foundational model was based upon this version of IRE and, and related ontology models were developed based upon existing DOLCE-Lite ontology modules. Note that although the foundational model was implemented and tested within various ontology tools, the task of integrating it with the RIDIR demonstrator software would require a separate project phase that could incorporate other factors (chiefly, performance and reliability), and was not attempted. Rather, the model served to inform the more basic semantic model implemented to focus more directly on the demonstrator requirements within the time frame available.

Overall, this was a substantial piece of work which is described in full at Appendix D. Here it will suffice to list the issues addressed by the functionality covered in the model; note that not all this functionality is implemented in the RIDIR demonstrator:

---

[27] JISC Information Environment See: http://www.jisc.ac.uk/whatwedo/themes/information_environment.aspx
[28] Formal ontology See: http://en.wikipedia.org/wiki/Formal_ontology
[29] Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) is the first module of the WonderWeb foundational ontologies library. Project homepage See: http://www.loa-cnr.it/DOLCE.html; DOLCE within the context of other upper ontologies See: http://wonderweb.semanticweb.org/deliverables/documents/D18.pdf
[30] OWL See: http://www.w3.org/TR/owl-ref/

- Identifiers change (resource moved): resource cannot be located
- Resource deleted; identifier refers to non-existent resource
- More than one copy of resource, cannot determine appropriate copy
- Not clear what identifier referent is (resource, splash page, metadata)
- Free-text metadata difficult to disambiguate
- Relationships between objects not persisted (objects, metadata enrichment)
- Mapping between metadata schemes (Mediation)
- Mapping/translation of taxonomies, thesauri, controlled vocabularies
- Mapping between identifier schemes
- Mapping between different object/content models
- Mapping/translation between object packaging and ingest schemes
- Mapping/translation between different ownership and security models
- Need to handle complex objects and collections
- Information on assertion of relationships is required
- Mapping/translation between different versioning schemes
- Reintegration issues
- Implicit metadata that needs making explicit

4.2.7 Definition of requirements from a user perspective via 'narratives'

Narratives were provided by the University of Hull to assist in the final aspect of the requirements definition process for the demonstrator: from the point of view of RIDIR users in the institutional repository domain, what would be the primary roles, the behaviour and the expectations in each case?

In addition, over the course of early 2008 institutions were found which agreed to help fill the necessary roles to ensure these narratives could be realised within the demonstrator development.

4.2.8 Development of demonstrator proposal specification

4.2.8.1 Overview

Once defined, the scope of the demonstrator was analysed in terms of interactions informed by the narratives, some investigation and prototyping was carried out, and the specification firmed up. By the end of February 2008 the team was in a position to issue a proposal for the demonstrator development phase.

The proposal described a demonstrator consisting of two example web applications, one for each scenario being demonstrated: 'lost objects' and 'locate related versions'. The web applications serve as exemplary third party services developed to demonstrate the use of underlying RIDIR services, built upon some common underlying services representing core RIDIR functionality.

4.2.8.2 'Lost objects'

This is essentially a 'broken link resolver' service.

When a URL ceases to function in a subscribing repository (as might happen, for instance, when a Depot object is moved to an institutional repository), this would be detected (by a browser plug-in, or potentially by a custom 404 page).

The user would be redirected to the 'broken link resolver' service.

If this service is able to find a matching authoritative[31] record pointing elsewhere, then the user would be redirected to the new URL, via a 'splash' page informing them that they should use/bookmark this new URL for future use.

If non-authoritative[32] matching records are found, the user would be presented with a list of these together with associated metadata against the new matches. The user would be able to navigate to the resources against these matches, and be able to indicate if they believe the match is correct or not - this information would be captured in RIDIR for reuse when presenting matches to subsequent users. The user would also be able to perform a search for a new resource at this point. Locating these potential matching objects depends to an extent on the quality of the metadata associated with them and it is in this regard that RIDIR will make use of a range of identifiers as part of the search process

If no matches are already provided by the RIDIR service, the user would be able to search for a new resource, based through a range of available systems, for example the Intute Repository Search, keying in any identifiers or other metadata they know about the resource. Once they have found a candidate match they will be able to navigate to this resource, and, if appropriate, indicate that they believe this is the new URL for this resource. This information would be captured for subsequent users.

In this way a network of non-resolvable identifiers and their new identifiers will be built up, with the relationships between them indicating whether users of the system believe the relationships to be certain or not. This relationship information will be used by the system in presenting candidate matches for subsequent users.

In the context of the RIDIR demonstrator, it was proposed to reveal some detailed information about the resources and relationships between them in the user interface so that those viewing or using the demonstrator could understand what was happening 'behind the scenes'.

A flowchart was developed to help visualise the process:

[31] Authoritative: A relationship added by eg a resource owner or the Depot system manager, therefore to be treated as the "correct" replacement location identifier

[32] Non-authoritative: A relationship proposed by a user of the system based on discovery of a replacement resource through Intute Repository Search.

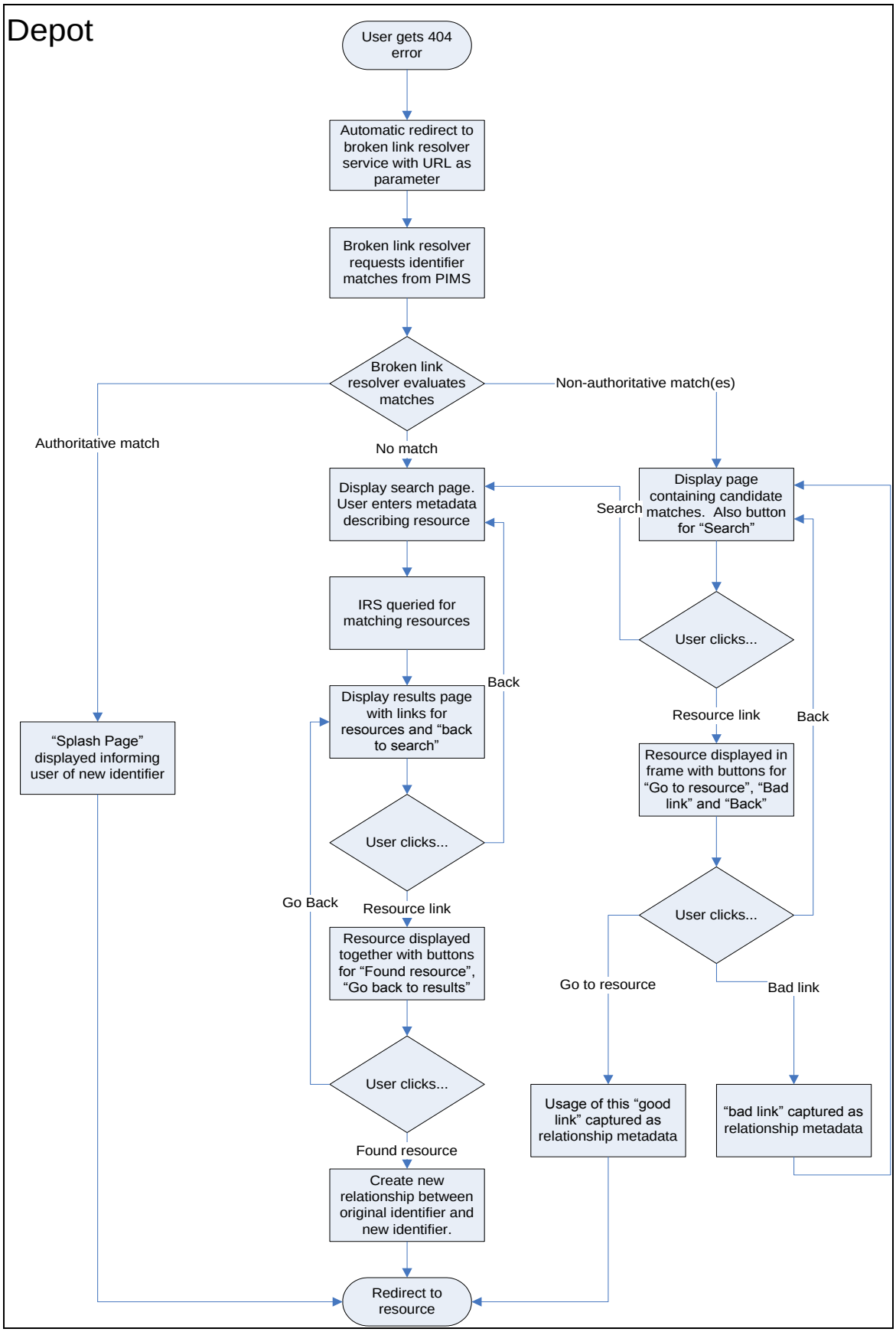

Figure 5: Flowchart for dealing with a 'lost' resource

4.2.8.3 Locate Related Version

'Locate Related Version' would be a service that allowed users to navigate previously-captured relationships between resources, view associated metadata and navigate to the resources displayed, coupled with a user-driven discovery of new resources and the ability to categorise the relationship between new and existing resources for reuse by subsequent users of the system.

The user of a RIDIR-enabled system would be able to determine if it contained information about digital objects related to one which they had accessed and they would be able to discover and record the whereabouts of (possibly further) related materials. The resource would have a 'find related' link next to it.

The user would click on the 'find related' link, and be able to view previously-captured relationships with other resources with their associated metadata.

The user would be able to navigate to these related resources, and be able to add to relationship metadata.

If none of the resources displayed are appropriate to the user's needs, the user would be able to navigate through to discovery of new resources.

Discovery of new resources would be user-driven, based on existing discovery services, such as Intute Repository Search, The European Library, xISBN and others. The search would be capable of using a range of identifiers

Once a user has located a new resource of interest, the user would be able to categorise the relationship of this new resource to the existing resource. This new relationship, plus metadata about the resource, would then be available to subsequent users of the system when attempting to locate related resources.

Again, this process is visualised in a flowchart:

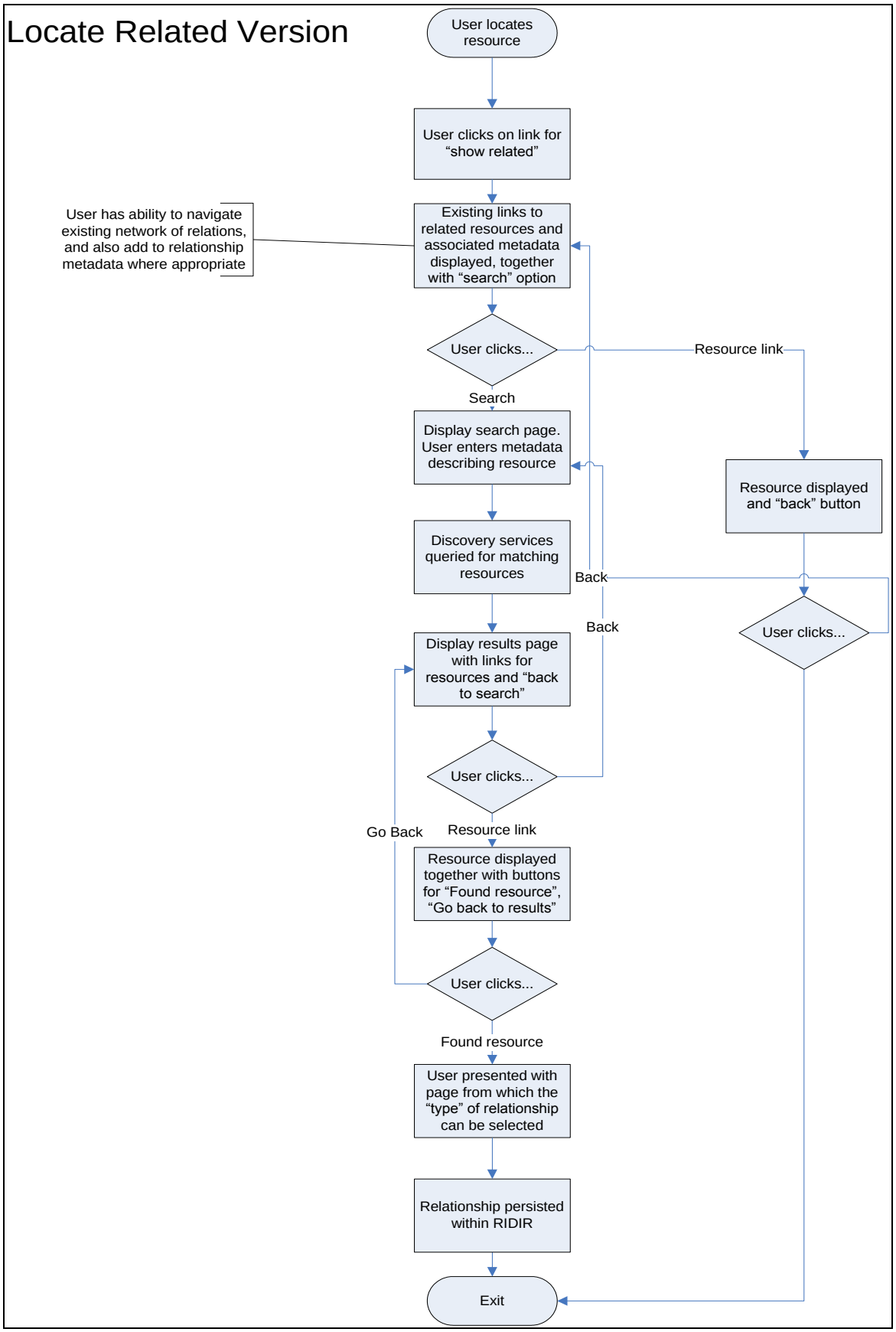

Figure 6: Flowchart for dealing with related versions of a resource

4.2.8.4 Architecture: RIDIR demonstrator scope

Consideration of the various requirements allowed the Rightscom team to develop a plan for the core scope of the demonstrator ("core" here in the sense that the team also proposed a set of optional features which could be phased into the development according to priority and available time remaining).

The web applications would be built according to the RIDIR architecture, upon a set of common services and repository core. Certain services were considered to pertain to a class of applications requiring services around 'lost resources', others to a class requiring those around 'locate related'. Both application classes call upon the common RIDIR services, representing the Persistent Identifier Mediator Service and RIDIR Repository Services.

As indicated earlier, some of the scope of the demonstrator could be satisfied by the provision of a more fully-scoped shared infrastructure service such as that provided by the PILIN project. For instance, wherever the RIDIR Repository mints a new identifier, it currently uses a simple naming convention and the Fedora Commons services, whereas in practice it would be desirable to devolve such cross-institutional responsibilities to the overall governance procedures offered by a shared infrastructure service such as PILIN.

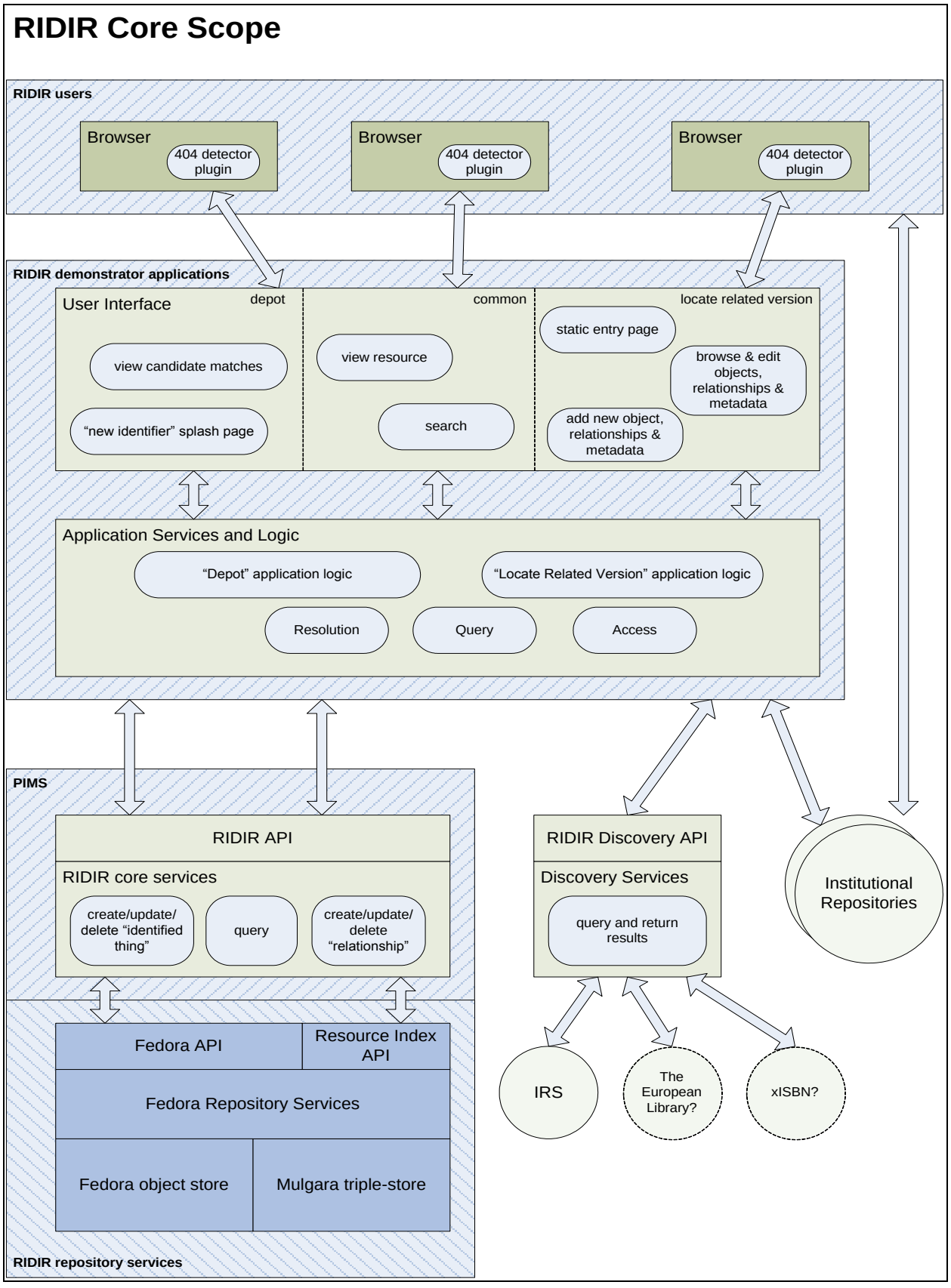

Figure 7: The RIDIR demonstrator's physical architecture

This architecture and its relevance to a possible RIDIR service available nationally is discussed further in Section 5. Development of the demonstrator proceeded on this basis, beginning in March, and completing in mid-May 2008.

5. Outputs and Results

5.1 The RIDIR Demonstrator

The demonstrator contains explanatory text designed to give the demonstrator user an overview of the issues being addressed, without having to resort to extracting the relevant parts of a separate report.

The demonstrator is entered through a 'welcome' page that introduces the overall requirements addressed by RIDIR and which provides links to the two areas of demonstration:

Figure 8: The Demonstrator 'welcome' page

5.1.1 Using the 'lost object' process

As noted above, this first RIDIR process is essentially a mechanism for dealing with broken links. It could, in theory, be initiated by a server-side action - a modified '404 page' that forwarded the broken URL to the RIDIR service, or a client-side action - for instance a browser plug-in - again to forward the broken URL. In view of the limited time available for completion of the work no additional development was carried out on externally hosted systems (‘404’) or a browser plug-in

environment, rather it offers hard coded links from the RIDIR home page to the resolver, each of which specifies the 'broken' URL as a parameter, for instance:

`http://ridir.ac.uk/link-resolver?url=http%3A%2F%2Fdeposit.depot.edina.ac.uk%2F99902%2F` [33]

The demonstrator provides a number of links to explore based on fabricated broken links in 'The Depot'. (Readers will recall that The Depot was one of RIDIR's use cases because its content is, in a sense, designed to be moved elsewhere.)

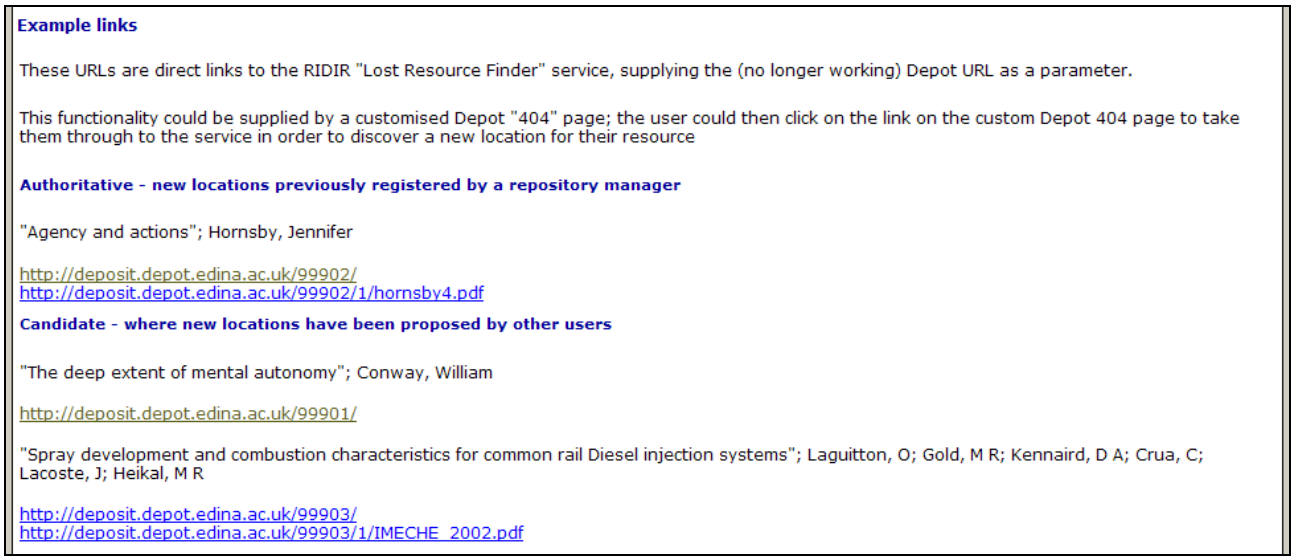

Figure 9: Part of the RIDIR 'Lost Resource Finder' home page showing links for the resolver service

Access to any of the links shown will result in the user being asked to log in (if they have not already logged in); this allows changes to the RIDIR database to be tracked. Note that whilst the starting points for these 'missing' objects have been pre-programmed, the demonstrator uses a live internet link to search for and retrieve candidate resources. For the purposes of the demonstrator the Intute Repository Search[34] is used in search routines; in a possible RIDIR service a range of search services might be made available to choose from.

5.1.1.1 Authoritative links

The first pair of links shown are so-called 'authoritative links'. These take the user to a splash page and a document respectively in the Birkbeck College EPrints repository which are the new 'homes' of the imaginary Depot equivalents; in other words, the curation boundary of the resource has changed. An authoritative link is one provided by a repository manager where the new location of the resource is exactly known and so the user gets no options to search for the 'missing' material, rather a simple page explaining what has happened and suggesting that the new URL should be used:

[33] Note that the domain name 'ridir.ac.uk' does not exist; it is used for illustration only
[34] Intute Repository Search website at: http://www.intute.ac.uk/irs/

Figure 10: The result of following an authoritative link

5.1.1.2 Non-authoritative links

The next set of links are 'non-authoritative links', which is to say that there is no authoritative information about where the missing resource might now be. Rather, when the user follows the RIDIR links (simulating the referral process that would occur with a real service) they are presented with candidate matches that other users of the system have suggested.

The first link, looking for the lost "The deep extent of mental autonomy" by William Conway results in this page:

Figure 11: The result of following a non-authoritative link

Users have looked for this resource before and have identified two possibilities, one in the Edinburgh Research Archive and one at the STFC ePublication Archive. Three people have visited the first candidate link of whom two have felt it to be the missing resource, resulting in a level of confidence of 67%. One person has looked at the STFC material and suggested that it is the missing resource, from that point of view and at that point in time could be said to have a 100% confident assertion metric.

Our current user will probably want to look at these before making a decision. For instance, following the Edinburgh link results in the following:

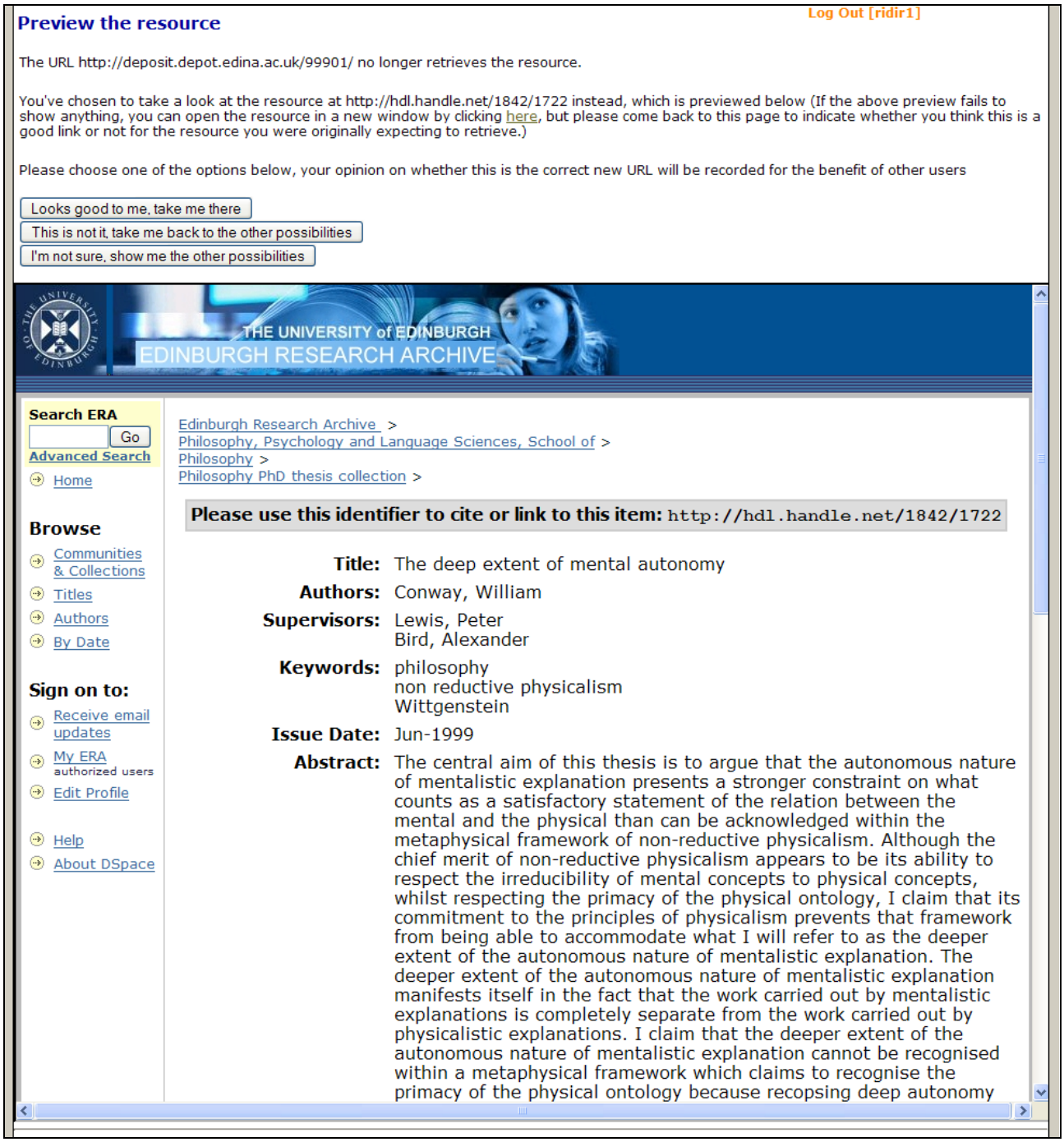

Figure 12: A RIDIR candidate match

The first button "looks good to me, take me there" will take the user to the resource and records in the RIDIR database that another RIDIR user thinks this is the correct material; the confidence would now be 75% based on four users.

The second button "this is not it, take me back to the other possibilities" will take the user back to the previous page recording their negative comment so that the confidence would become 50% based on two users.

The third button "I'm not sure, show me the other possibilities" takes the user back to the previous page leaving the RIDIR database of relationships unaltered.

In this way users will normally update the database and the score, the level of confidence in a resource, reflects their views. Because the user is logged in, the RIDIR system can manage the database sensibly. Thus, for instance, a single user cannot make multiple entries against a single resource, although they can change their entry should they have cause to change their opinion.

Using the "not sure" button takes the user back to the page shown in Figure 11 where they might now choose to search for the missing resource at an alternative URL. If they do so, they are presented with a search page:

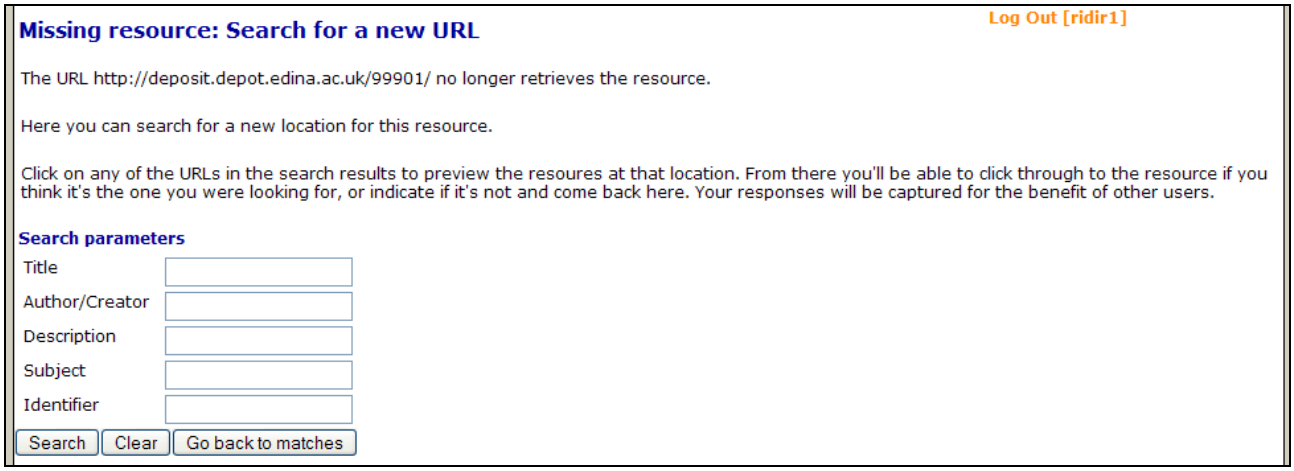

Figure 13: Searching for a missing resource

Searching here for 'mental autonomy' in a description (see figures 11 and 12) results in three matches from the repository searches available to RIDIR. One is a document already in the RIDIR system from the Edinburgh Research Archive, another is from Durham Research Online, whilst the third is a document in the Cognitive Sciences ePrint Archive. Following the links for this third match gives the user a preview similar to Figure 12:

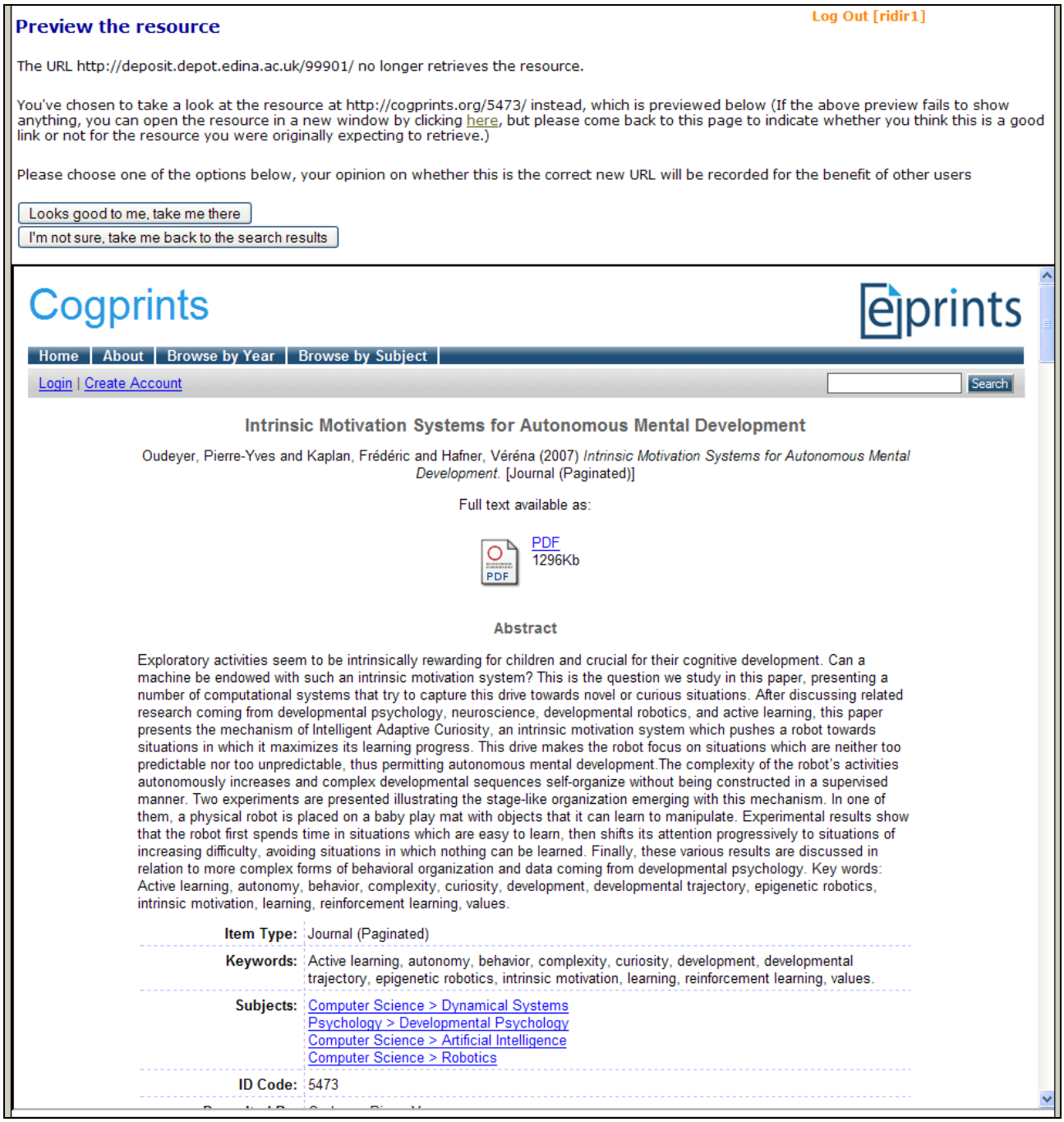

Figure 14: The preview page for the new candidate match

Using the "looks good to me" button takes the user to the resource and records it in the RIDIR database as a new candidate match for the missing resource having a confidence of 100% based on a single user. "Not sure" takes the user back to their search page.

The search page shown at Figure 13 deserves further comment. It is here, in the underlying process, that the importance of using identifiers in object metadata becomes clear. Rather than a vague search on keywords of some sort, a user armed with some firm identifiers - perhaps an author name, an ISSN, a catalogue entry, even an 'old' URL - could use them in a search in the 'identifier' field and have confidence that the results should have a good level of relevance. However this will only be so if the creators of object metadata conscientiously record all the identifiers that might reasonably be associated with a given object.

5.1.1.3 Unknown location

The final link in this section of the RIDIR 'Lost Resource Finder' home page, headed 'unknown', allows a user to search for a missing resource 'from scratch'. In other words, on a clean install of the demonstrator there are no candidate matches already in place.

5.1.2 Using the 'locate related version' process

In order to demonstrate the possibilities of a 'locate related version' service, the contents of two real repositories have been used: the Spoken Word Services Archive at Glasgow Caledonian University,[35] and TRILT - the Television and Radio Index for Learning and Teaching.[36]

[35] Spoken Word Services website at: http://www.spokenword.ac.uk/
[36] TRILT website at: www.trilt.ac.uk

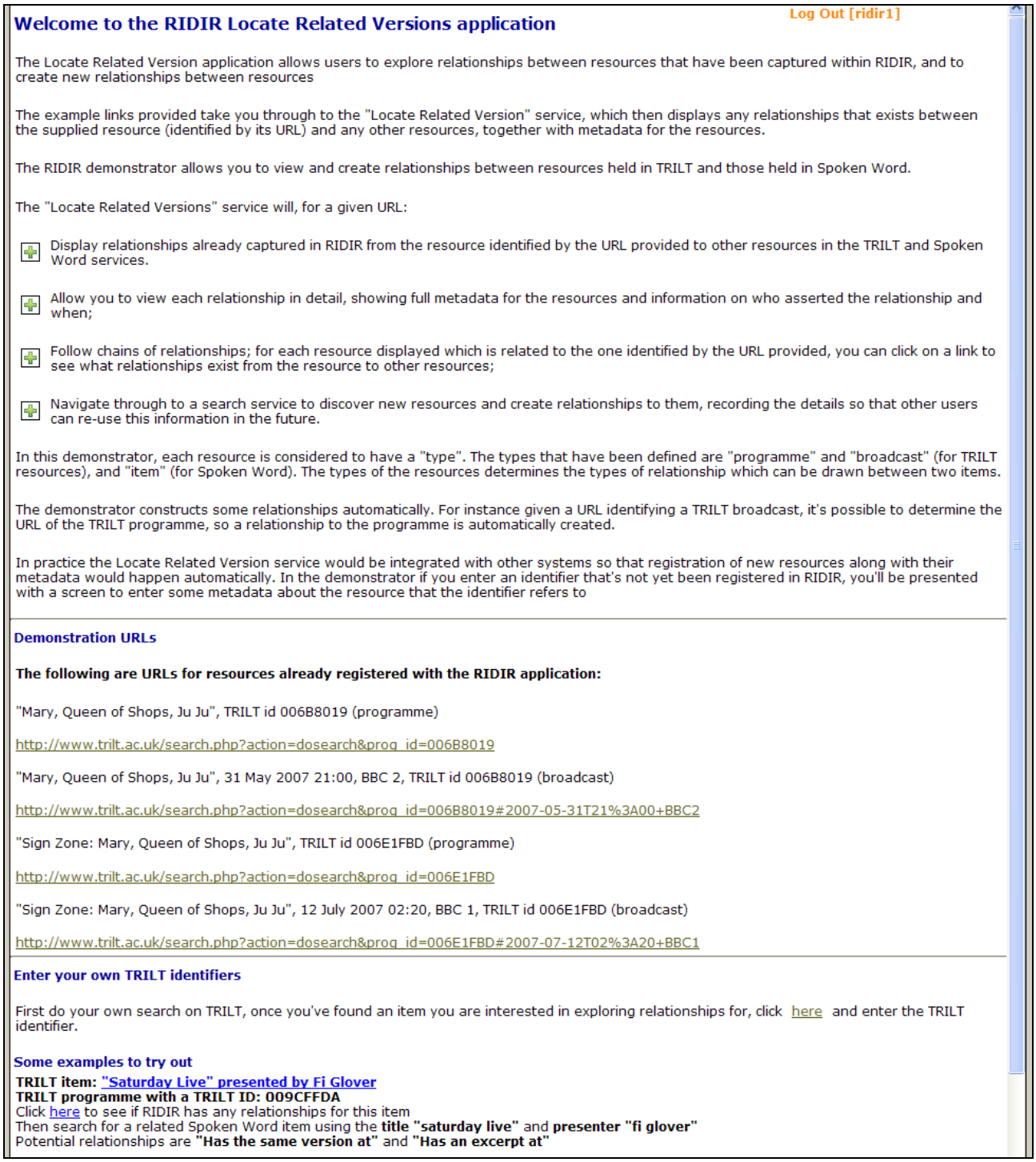

Figure 15: Part of the homepage for 'Locate related resources' service

The home page for the 'Locate related version' service has two sections. The first provides a set of 'Demonstration URLs'. These relate to a set of related broadcasts in the TRILT system and between them show how a network of named, resolvable relationships can be built up through normal usage of the application. Following any one of these links (simulating a user finding one of these resources in a RIDIR-enabled system) reveals details about the resource and also provides links to related resources.

The second section of the home page allows users to use a newly provided TRILT resource as the basis for building a network of relationships.

The first time during a browser session that a user attempts to access any of the items on this page they will be asked to log in. Logging in allows the RIDIR system to track and manage changes to its relationship store.

5.1.2.1 Test URLs

The first test URL relates to TRILT item 006B8019, a programme called “Mary, Queen of Shops, Ju-Ju". Following the link gives the following screen:

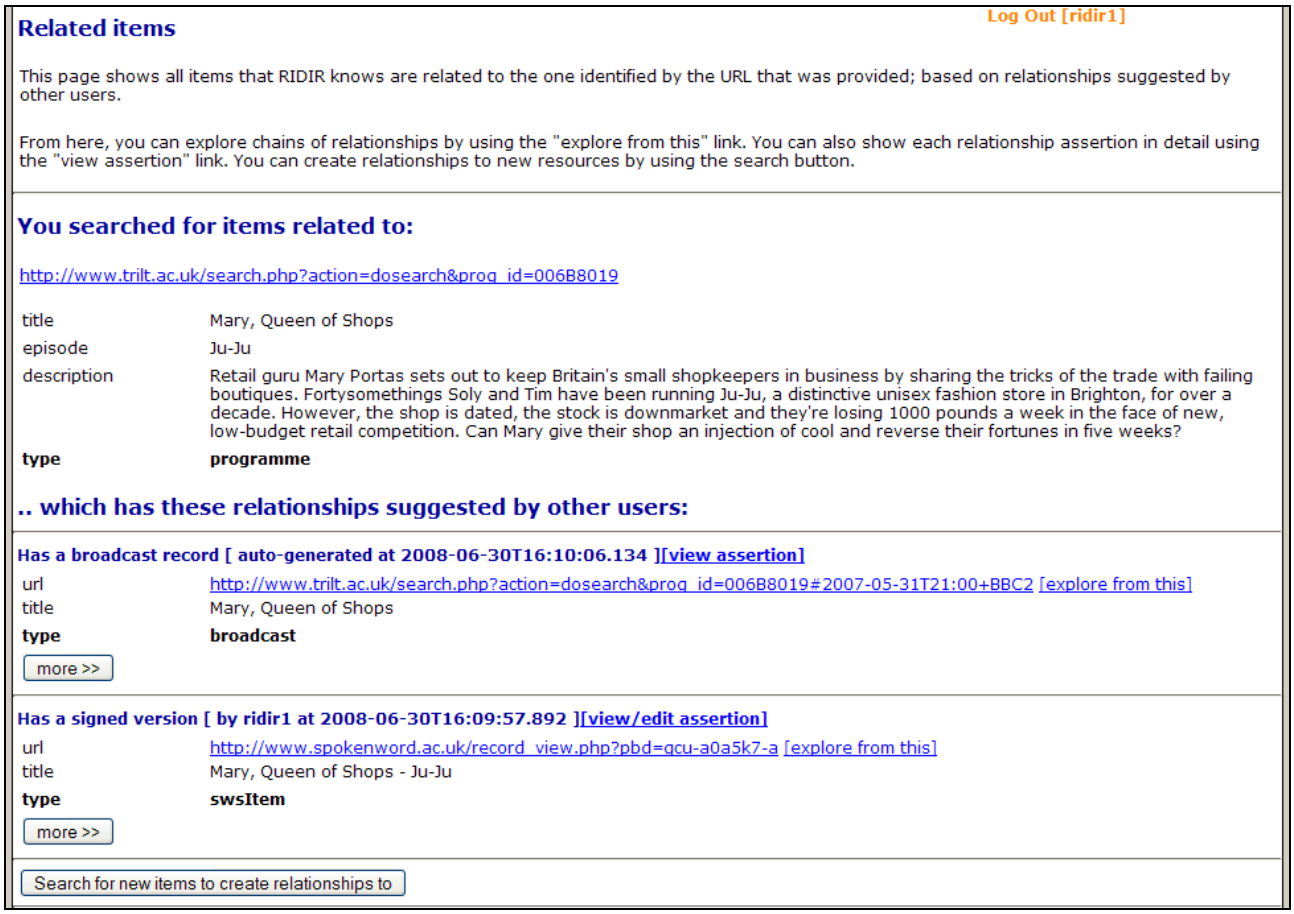

Figure 16: RIDIR screen for TRILT 006B8019

The programme has been added to the RIDIR database by a user and, because RIDIR knows something about TRILT, the system has automatically located a broadcast of the programme. (TRILT distinguished between a generic 'programme' and specific broadcasts of the same.) The 'related items' panel says that the programme 'has a broadcast' and gives details of it. A user has also asserted that there is a signed version of the programme (for the hard of hearing) in the Spoken Word Services repository. An expansion button [more >>>] can be used to show further details of the broadcast:

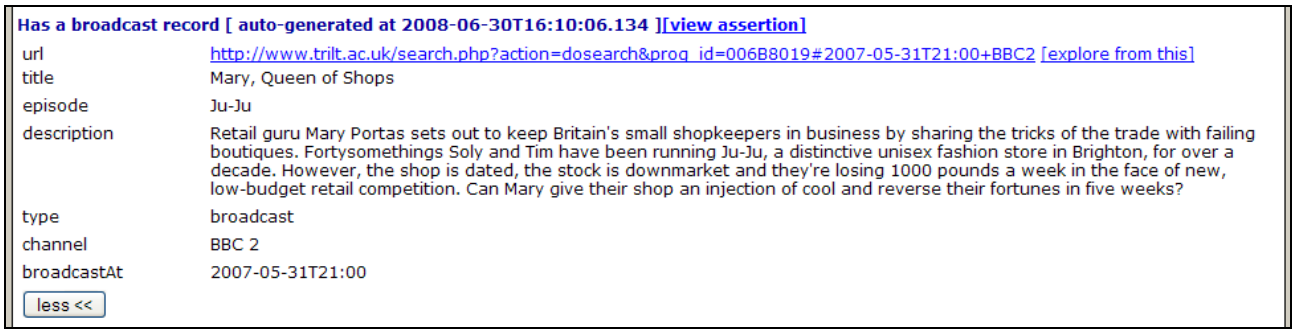

Figure 17: Part of the previous diagram expanded to show broadcast details

Alongside the line showing the 'has broadcast' relationship is a link: [view assertion] which allows users of the demonstrator to see the internal format of the assertion within the RIDIR system. This would not appear in a production version.

Alongside the URL of the broadcast is a link: [explore using this] which, if followed, will show items related to the broadcast:

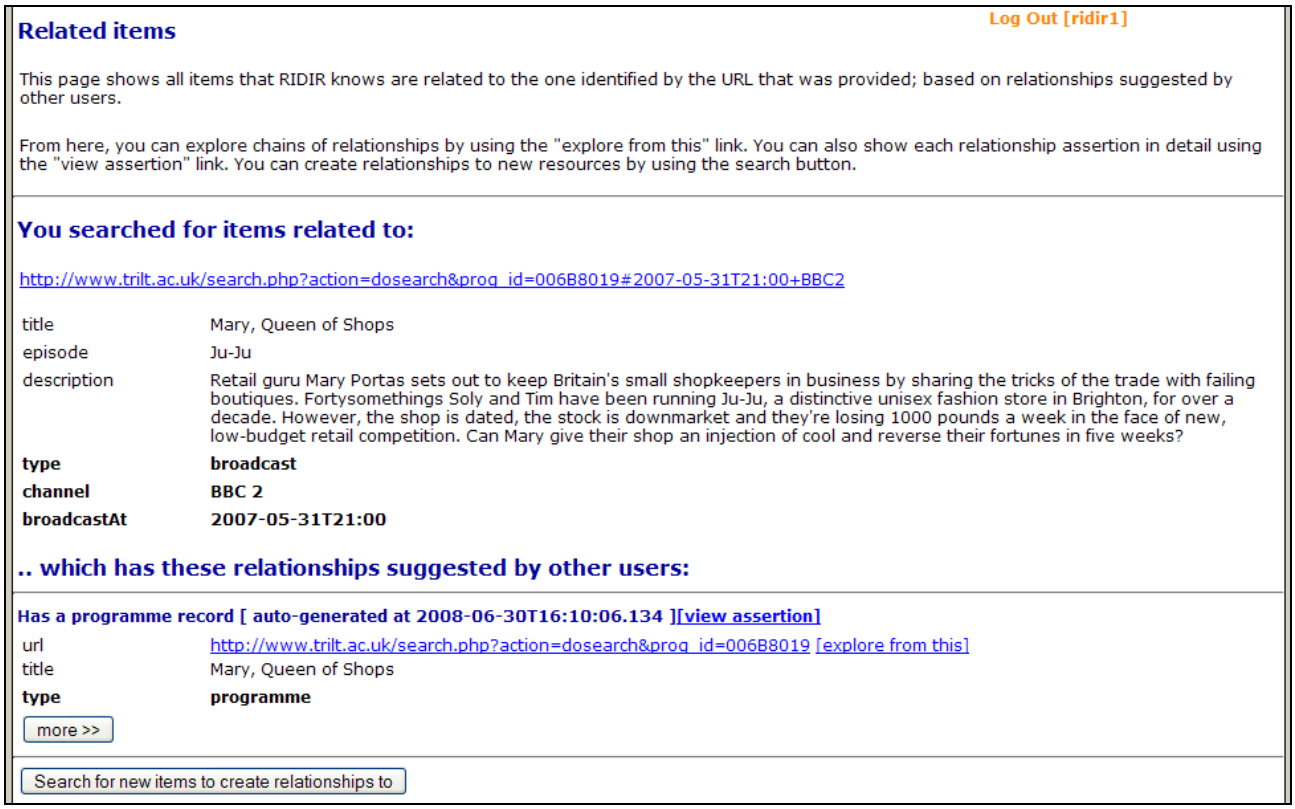

Figure 18: Items related to the TRILT broadcast

Unsurprisingly, this page shows the reverse relationship: the specific broadcast is related to the general TRILT programme entry. Again, this relationship was system generated when the TRILT item was added to the RIDIR database. However it would be quite possible that there may be other resources shown which relate specifically to the broadcast rather than to the programme.

The second link on the 'locate related' homepage shows these same relationships but starting from the discovery of the broadcast.

The third link on the 'locate related' home page simulates a user identifying a different TRILT programme entry. In this case an episode of 'The Sign Zone'

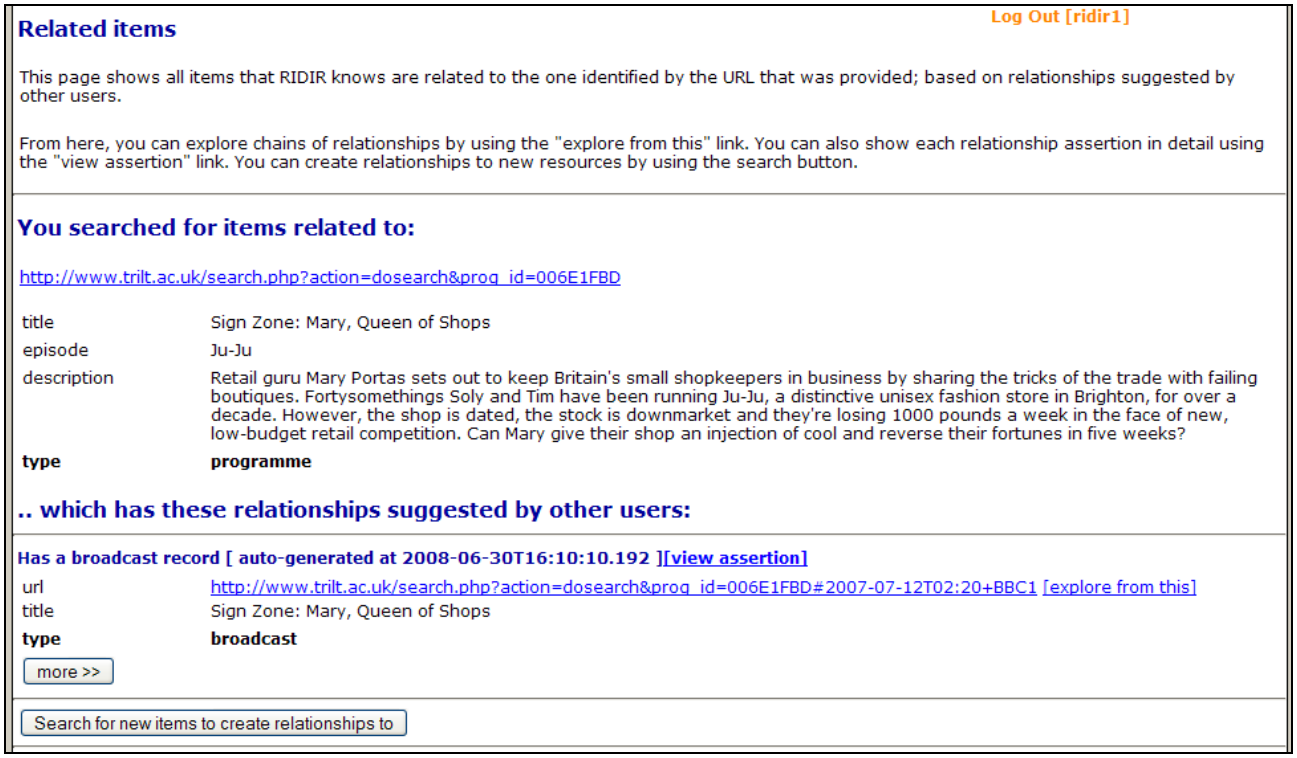

Figure 19: Details and related items for a Sign Zone programme

This and the fourth link provide similar functionality to those already described above.

At the bottom of each of the pages described in this section is a button [Search for new items to create relationships to]. This allows a user to search for related items in another repository of similar material, in the demonstrator this is the Spoken Word Services repository, in a RIDIR service it would search a number of related repositories.

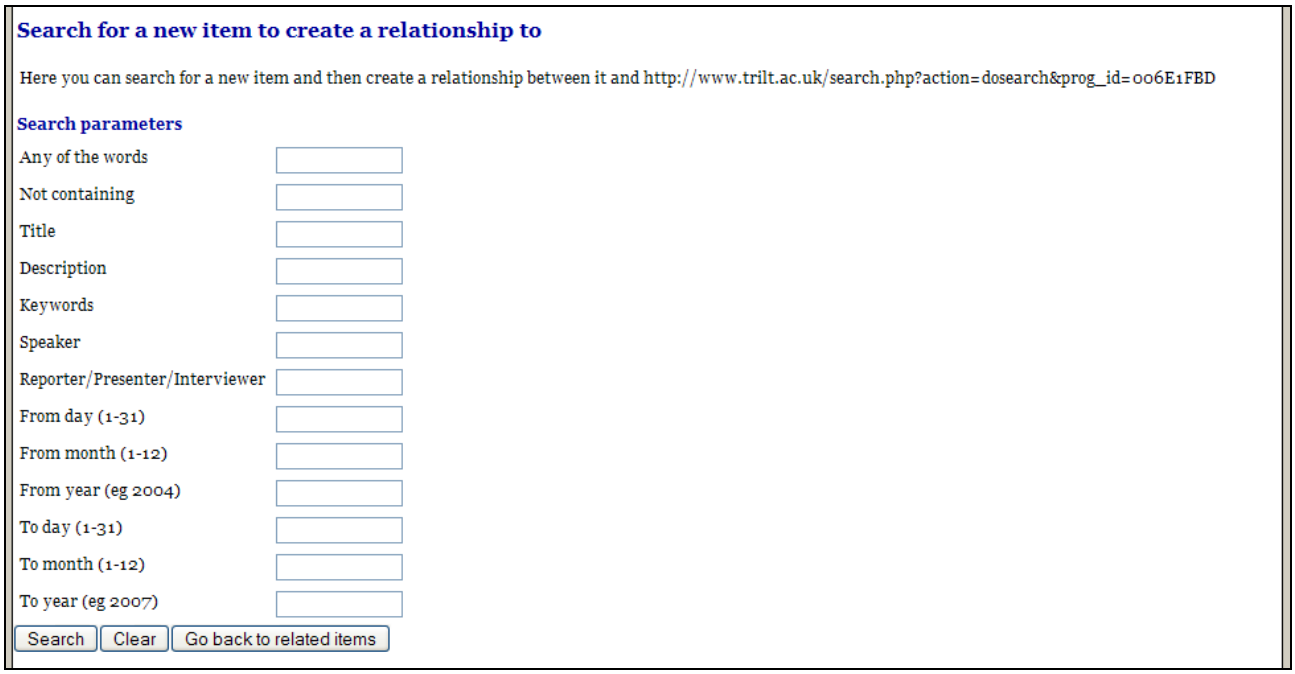

Figure 20: The search screen for finding related items elsewhere

In fact, the Spoken Word Services repository does not appear to contain obviously related material. This search will be dealt with further in the next section.

5.1.2.2 Enter your own TRILT identifier

This section of the interface simulates the situation in which a user has found a potentially useful item used in a resource and wishes to use RIDIR to search for related items. In a production system one might hope to provide an intelligent transfer between the TRILT (or other) search page and the RIDIR tools.

Suppose that a user has discovered a radio programme from which they would like to use an extract in some teaching materials; its TRILT identifier is 0018111D. They enter this into the RIDIR search page. No further information is needed for this search:

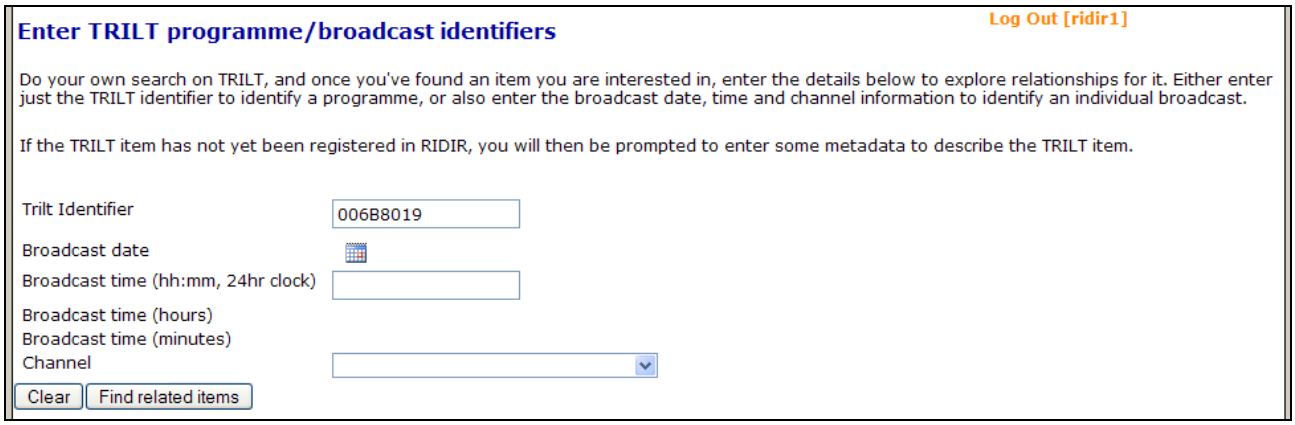

Figure 21: The search page to find TRILT items in RIDIR

In this case the programme does already exist in the system:

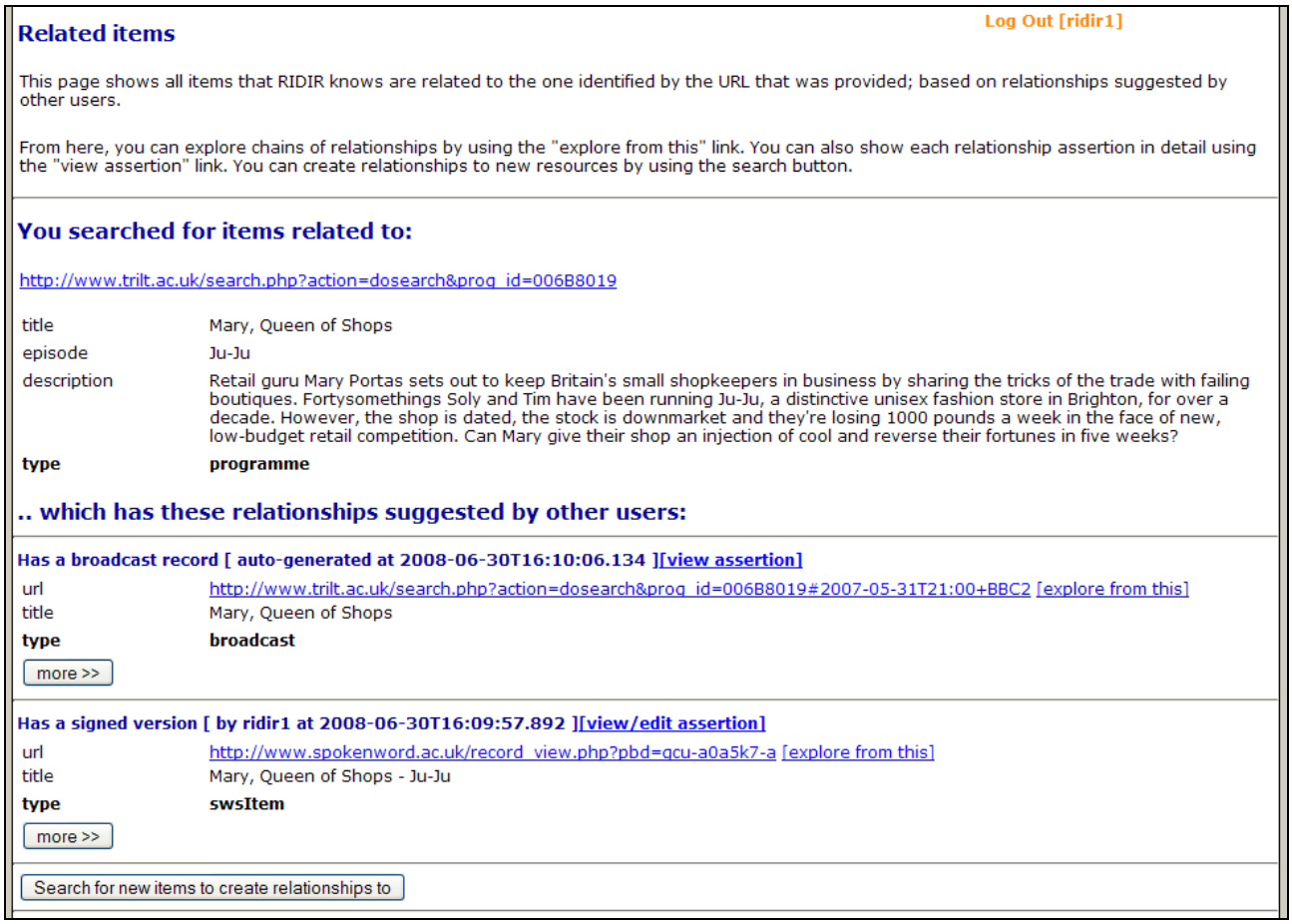

Figure 22: Related items for a known TRILT programme ID

RIDIR has retrieved the details from its database and notes that there are related sources of further information available too.

Entering a TRILT identifier for a programme that does not yet exist in the RIDIR system gives a different page:

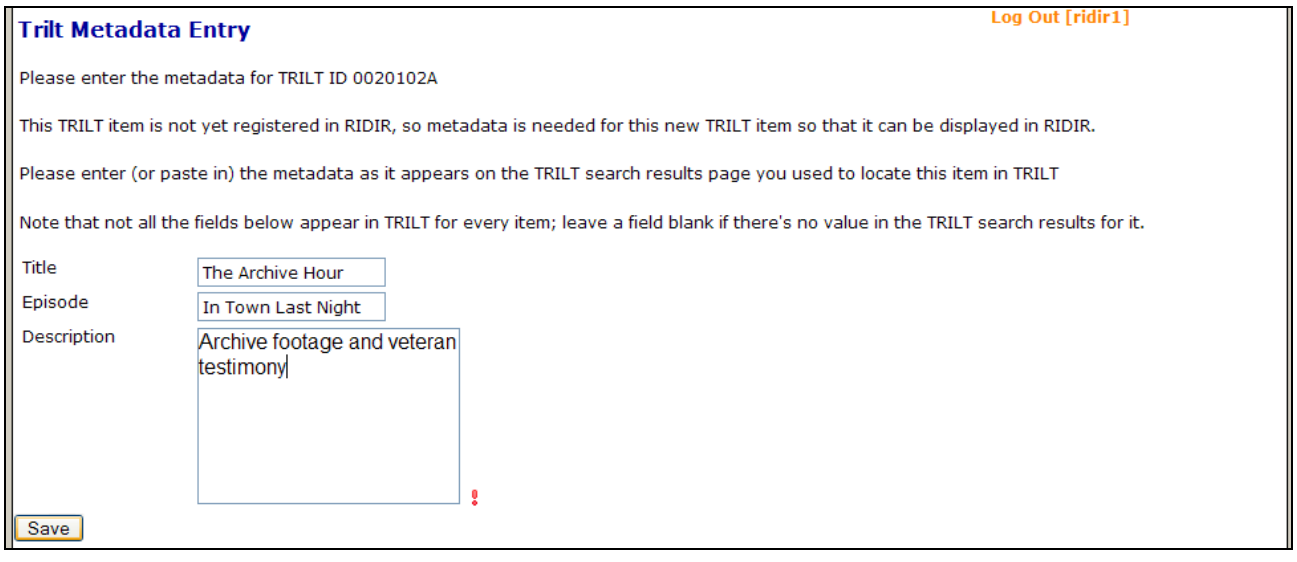

Figure 23: Metadata request for an unknown TRILT Programme ID

The user is asked to enter basic metadata for the new item. Clearly this is an undesirable step, unfortunately the protocols used in the TRILT standard search interface make it difficult to transfer this information to RIDIR and it was not possible to resolve this issue in the lifetime of the project. Note that the description of the programme retrieved from TRILT is minimal.

Clicking the [Save] button transfers the information to the system:

Figure 24: The entry for the 'new' RIDIR item

Subsequent users might search for related material in another repository, for instance using the search routine shown in Figure 25:

Figure 25: Searching for related material elsewhere

and find amongst the search results:

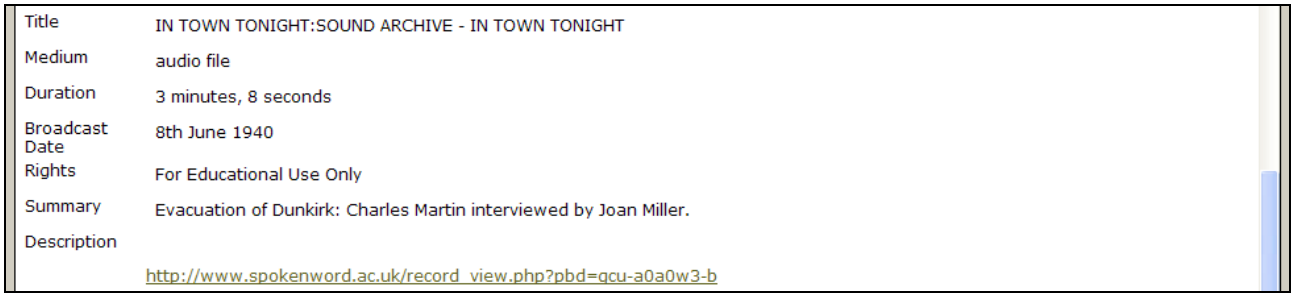

Figure 26: Candidate related material

The user decides that this is an example of the programmes described in 'The Archive Hour' broadcast and decides to record the fact by clicking the link at the bottom of the search result:

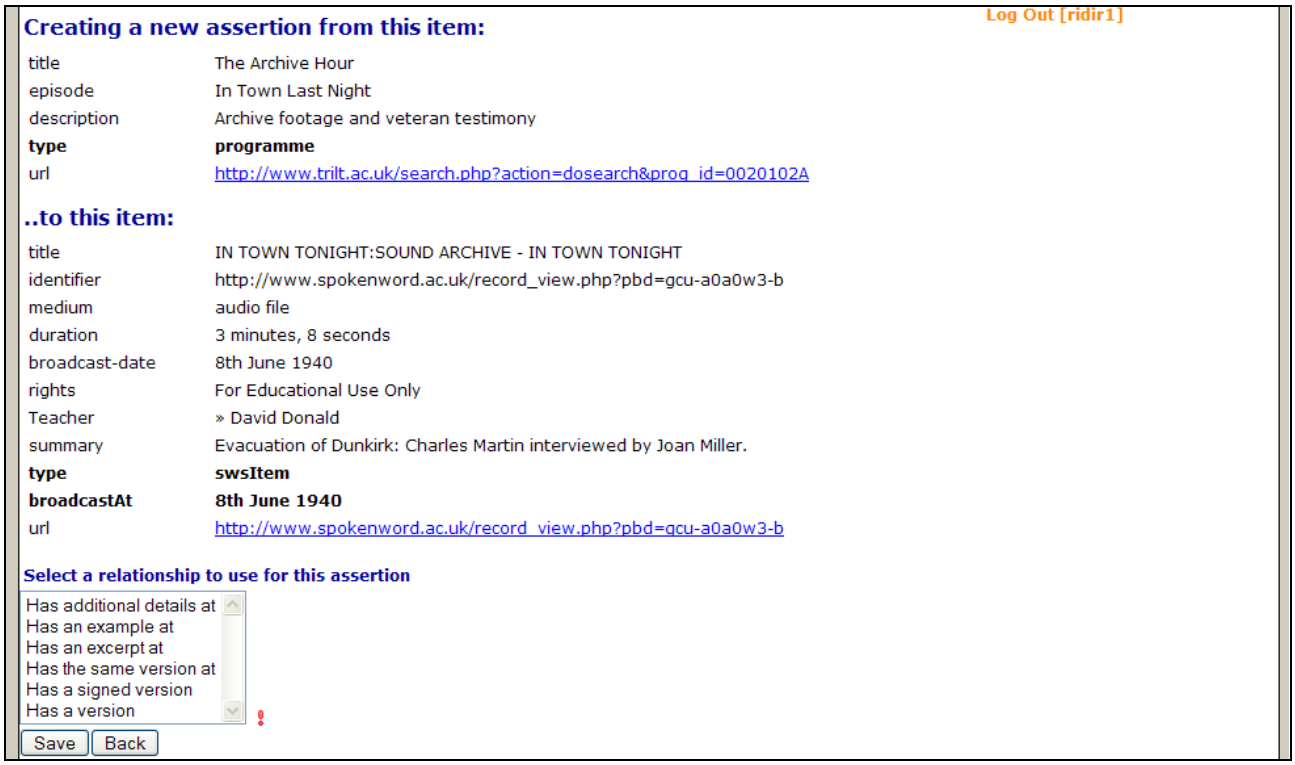

Figure 27: Additional metadata retrieved from the candidate material

and then uses the relationship "Has an example at" from the list of options to do so.

The RIDIR page shown at Figure 24 now has an additional section:

Figure 28: The supplemented RIDIR page

Future users coming to RIDIR for information on this broadcast will immediately be given access to this additional material. They do not even have to follow the link. Expanding the 'has additional details' section provides the majority of the Spoken Word metadata.

Figure 29: The expanded page

5.2 The RIDIR API

The functionality described above is provided through application services according to the class of functionality, 'lost object' or 'located related', both of which communicate with an underlying RIDIR API which wraps the PIMS (“persistent identifier mediation services”) component.

The RIDIR API is not fully implemented, in terms of the abstract architecture and foundational model and ontology developed during the analysis phase, and in terms of the full range of functionality listed below (for instance the application services provided had no need of the 'Delete' part of CRUD operations). The focus of the software development phase was to implement enough to support the use case functionality according to priority. The implementation of the underlying foundational ontology was limited in a similar fashion.

The RIDIR API offers the following services:

- Registration of existing 'persistent identifiers'

    - *Mints a PIMS 'mediation identifier'* – whose scope is that of the RIDIR system and whose lifetime is under the governance of the RIDIR system, so as such may be considered 'persistent'. In the 'lost resource' use case, RIDIR occupies a shared infrastructure service role whose function is to mint and resolve in perpetuity identifiers from the community across the UK; however as the project was not to duplicate work done by PILIN it does this by minting RIDIR demonstrator repository-scoped identifiers. It also fulfils the role of registering identifiers that are not registered as part of a national identification service.

    - *Associates human-readable descriptive metadata with the mediation identifier*. Such descriptive metadata is important because in terms of 'identifier interoperability' they represent 'truth claims' as to the veracity of the relationship between a resource and the identifier itself (which is typically an alphanumeric string). (In the demonstrator the metadata is sourced from existing search services, and mediation between metadata is done through human interpretation of search results; longer-term we would anticipate such mediation being done automatically via the foundational ontology to fulfil the overall functional requirement for metadata interoperability.)

    - *Associates the identifier with a machine-readable semantic description ('semantic metadata') of what type of thing the identifier refers to* (ie, a classification for the referent; the main expected types are sorts of 'Resources' and 'Representations'). The classifications are not rigid, global or prescribed in origin, any URI may be used to specify a classification. The identity of the user is recorded, and as such the classifications are localised to the context of the registration. The classification used may be user-specific, may be local to RIDIR (entered into the RIDIR system by another user), or it may a predefined vocabulary which the user chooses to adopt, such as FRBR. In the 'locate related versions' application, a set of types are provided for demonstration purposes, such as 'broadcast' and 'Spoken Word Services item'. The localisation aspect of the semantic metadata facility is a mechanism for providing machine-readable metadata interoperability. In other words, there is the facility for many views to be asserted as to the classification of the referent of an identifier.

    - *Overall, CRUD (create, update, delete) services for administration* of the identifiers and associated metadata (human and machine-readable).

- Identification and registration of relationships between registered 'persistent identifiers'

    - *Association of referents* In contrast to an approach which simply draws a one-to-one 'crosswalk' between identifiers (rather than examining the semantics surrounding the referent, such as its type definition in terms of other identified referents), the RIDIR model allows the relationships to be drawn between the referents which themselves are assigned named and resolvable identifiers, expressed internally using RDF. The existence of these associations form meaningful, machine-interpretable metadata associated with the identifier.

- *Classification of referents* Referents may be assigned a type; for interoperability using identifiers, it is important to classify what sort of identified thing an identifier refers to. *Use of ontology within semantic classification Types may be linked together* as part of the semantic classification such that further conclusions may be drawn by inference. For example, if 'Is Part Of' is asserted to always be the inverse relation to 'Has Part', then further conclusions may be deduced when retrieving all referents of the 'Is Part Of' relation even when only the 'Has Part' relation was defined between two referents. In this way, various ontologies may be constructed within RIDIR and 'overlaid' onto a set of identifiers held within RIDIR's repository that allow 'domain models' of referents of persistent identifiers to be constructed by users. Within the demonstrator, simple 'models' have been created by way of demonstration: one for 'items' stored within Spoken Word Services items, one for 'resources' identified by the TRILT service, and one for modelling how a 'splash page' relates to a 'resource' within institutional repositories, in the context of a specific user.

- *Context is preserved with each semantic classification.* As for classification of referents, all classifications are treated as 'assertions', where semantic metadata describing the context of the classification is recorded. So here, 'localisation' specifically means 'the classification holds only within the circumstances (context) local to the act of classification'; in other words, no classification is applied globally. RDF metadata is created to store those circumstances. The only exception to this is for the semantics governed by the foundational model.

- Mediation of identifiers

  - *Determining related versions of the referent of a persistent identifier (a 'resource')* Given a persistent identifier, finding identifiers of referents each of which have been (or could potentially be) deemed a 'related version' of its referent. This feature is exercised within the demonstrator within the 'Located Related Versions' application.

  - *Determining alternative addresses (locations) for accessing the referent of a persistent identifier* Given an identifier whose referent is a location of a resource, finding alternative location identifiers for that resource. This feature is exercised within the demonstrator within the 'Lost Resources' application.

  - *RIDIR-governed mediation* The above forms of application-specific mediation implemented to support the 'Lost resource finder' and 'Locate Related' services. We would anticipate that this would be replaced with a more generic mediation service in a full implementation, which would offer services taking full advantage of the foundation ontology and those classifications linked to it by users.

Implementation of Semantic Model for RIDIR Demonstrator

Due to the time pressures placed on the development part of the project described above, a decision was taken not to separate out the various aspects of the ontology into 'foundational' or 'domain-specific' modules for the early iterations of the development (those leading up to the test release), but to explore the possibility of including it in the release in later iterations. The primary focus of the iterations prior to release was to issue a working end-to-end system capable of meeting the proposal requirements for a *demonstrator* for RIDIR, rather than attempting to build a more robust pilot system whose software assets would be designed for reuse. An additional factor was the degree of risk attached to developing and implementing a system based on a rigorous formal ontology, given the time constraints.

Therefore the model implemented for the RIDIR demonstrator was not implemented in the web ontology language, OWL, but expressed using a specific RDF vocabulary defined for the demonstrator. Using OWL was considered unfeasible in the short-term due to the time frames involved in considerations over the formal logic involved when dealing with OWL, and moreover with the complexity of the explicitly reified assertions required for RIDIR's preservation of assertion context that is not addressed directly in OWL; for the longer term, investigating development in OWL based upon the foundational ontology work would however be a recommendation. One issue to consider would be the management of the impact of release cycles of OWL itself; OWL is approaching its second major public release within five years.[37] Implementation issues would also require exploration to commit to OWL, specifically performance, scalability and robustness of implementation of ontology reasoner software, together with the trade-offs in terms of dealing with large framework ontologies during the development cycle (which were found to require very significant computational resources in the case of developments with IRE and DOLCE).

Therefore the demonstration-specific RDF term vocabulary defines an 'OWL-like' schema vocabulary which was suitable for rapid, iterative development, using the less strict constraint definitions of RDF query languages. Consequently, the model itself is implemented using the RDF vocabulary, and the constraints defined in terms of the RDF query language iTQL supported by the Fedora repository.

Those 'OWL-like' relations defined to be axiomatic (always hold true for the RIDIR demonstrator) are expressed in iTQL and applied to the triples comprising the 'ontology' in RDF used in the demonstrator, by virtue of fully-expanding RDF triples of an ontology digital object stored in the repository. Certain axioms are expressed within the ontology itself. Examples are:

- in the 'ontology':

```
(X ridir:inverseAssertion Y) and (X ridir:domain D) and (X ridir:range R) =>
(Y ridir:range D) and (Y ridir:domain R)
```

- in queries, the queries search for inverses explicitly; so

```
(A ridir:subject S) and (A ridir:object O) and (A ridir:assertionType X) and
(X ridir:inverseAssertion Y) => (A ridir:object S) and (A ridir:subject O)
and (A ridir:assertionType Y)
```

[37] See news item at: http://www.w3.org/News/2008#item71

For the short term, a different vocabulary was used to ensure that terms from the foundational ontology (or any models built upon it) would not be confused with the model used to build the demonstrator. For future developments, a recommendation would be to base developments on the foundational ontology. Specifically, the prefix "proxy" was defined: creating a proxy in RIDIR is effectively registering (external) resource and representation identifiers (URLs) and associated metadata and descriptions (types) with RIDIR, and in doing so minting RIDIR identifiers. RIDIR identifiers are considered 'internal', as the project is not intended to replicate a shared identifier infrastructure in the manner addressed by PILIN; however, it needs to be RIDIR identifiers that form the basis of any relationships, descriptions, etc used to define the behaviour of the application services. Therefore, in the longer term, it would be expected that the RIDIR identifiers are considered 'persistent' via a shared infrastructure service whose identifier curation boundary creates longevity expectations outside those of the computational machinery provided at the RIDIR system-level.

Implementation of Semantic Relation Browser

In order to make explicit in the demonstrator the network of relationships constructed, a basic enhancement was introduced which converts the RDF triples and feeds them into visualisation software[38]. The visualisation software allows a user very simply to see and navigate chains of relationships created within RIDIR. Any node can be clicked upon, which places the node into the centre of the viewing area. The information displayed for each node gives its type, and its RIDIR identifier is shown in brackets. The right-hand bar shows key information about the central node: Name (here, type and identifier), Location URL (URI), and Description (textual content gleaned by RIDIR).

Illustrative screenshots are shown below that show the basic relation chains implemented:

[38] Relation browser See: http://der-mo.net/relationBrowser/

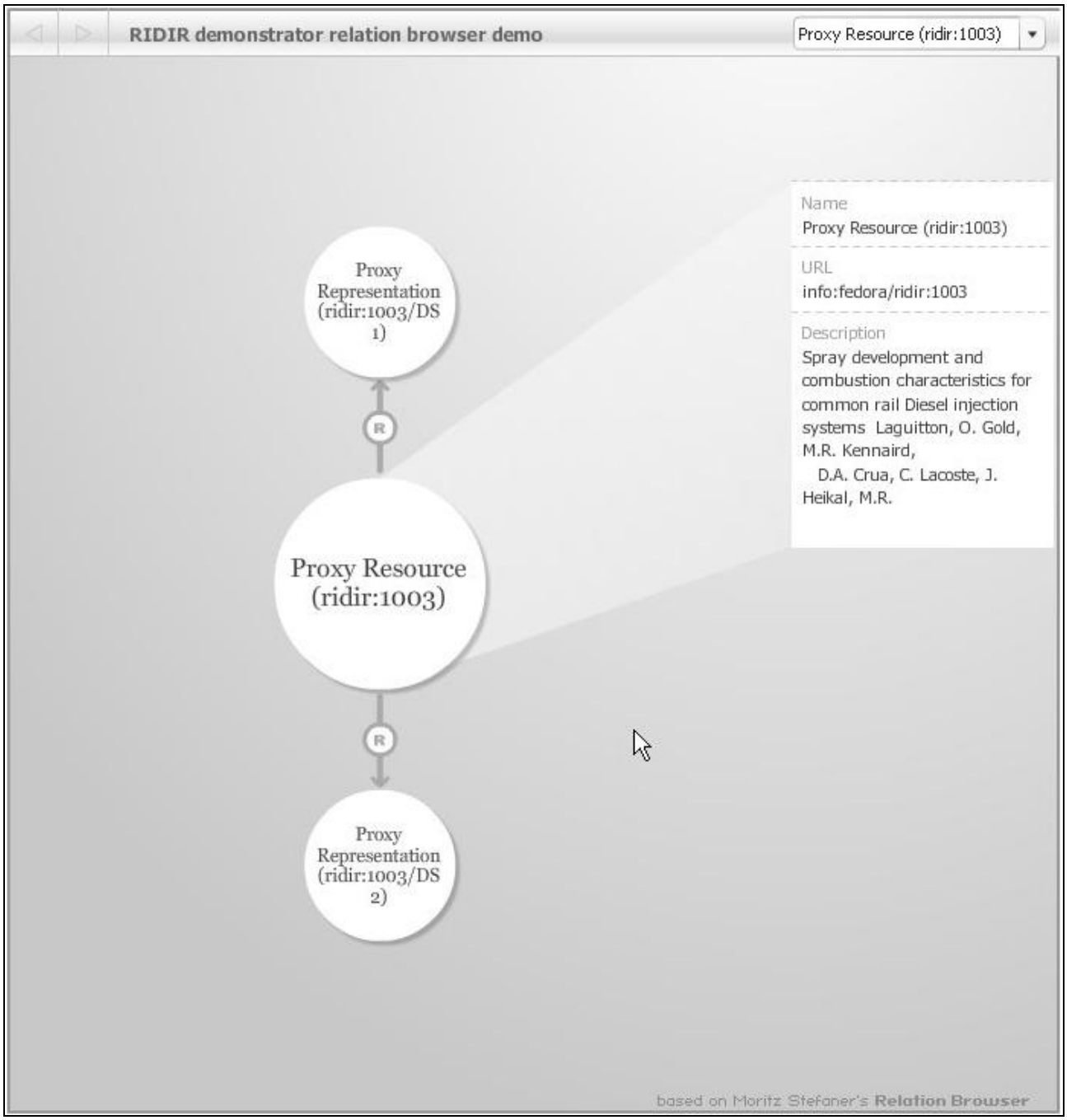

Figure 30: Semantic relation browser, screenshot #1

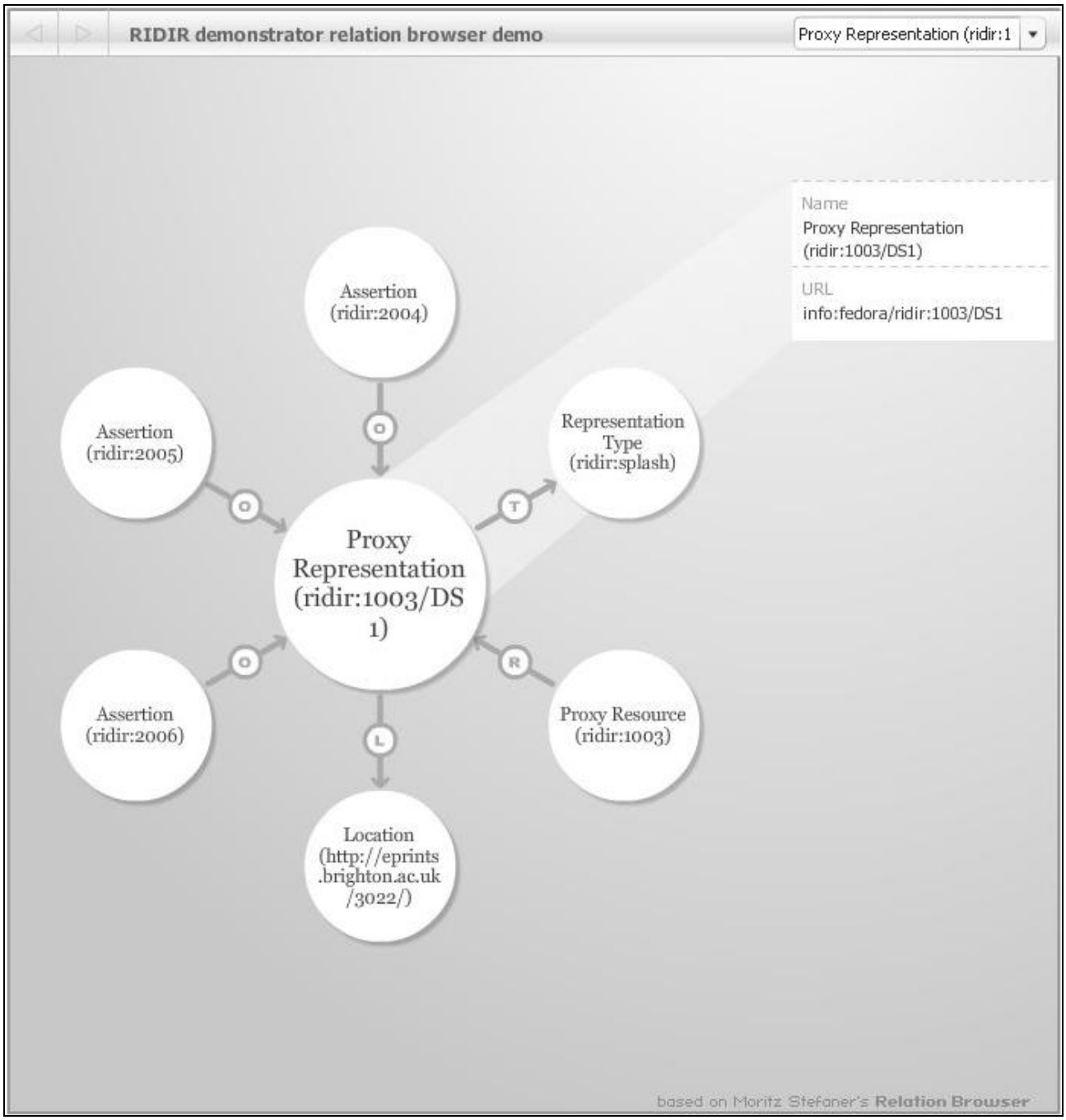

Figure 31: Semantic relation browser, screenshot #2

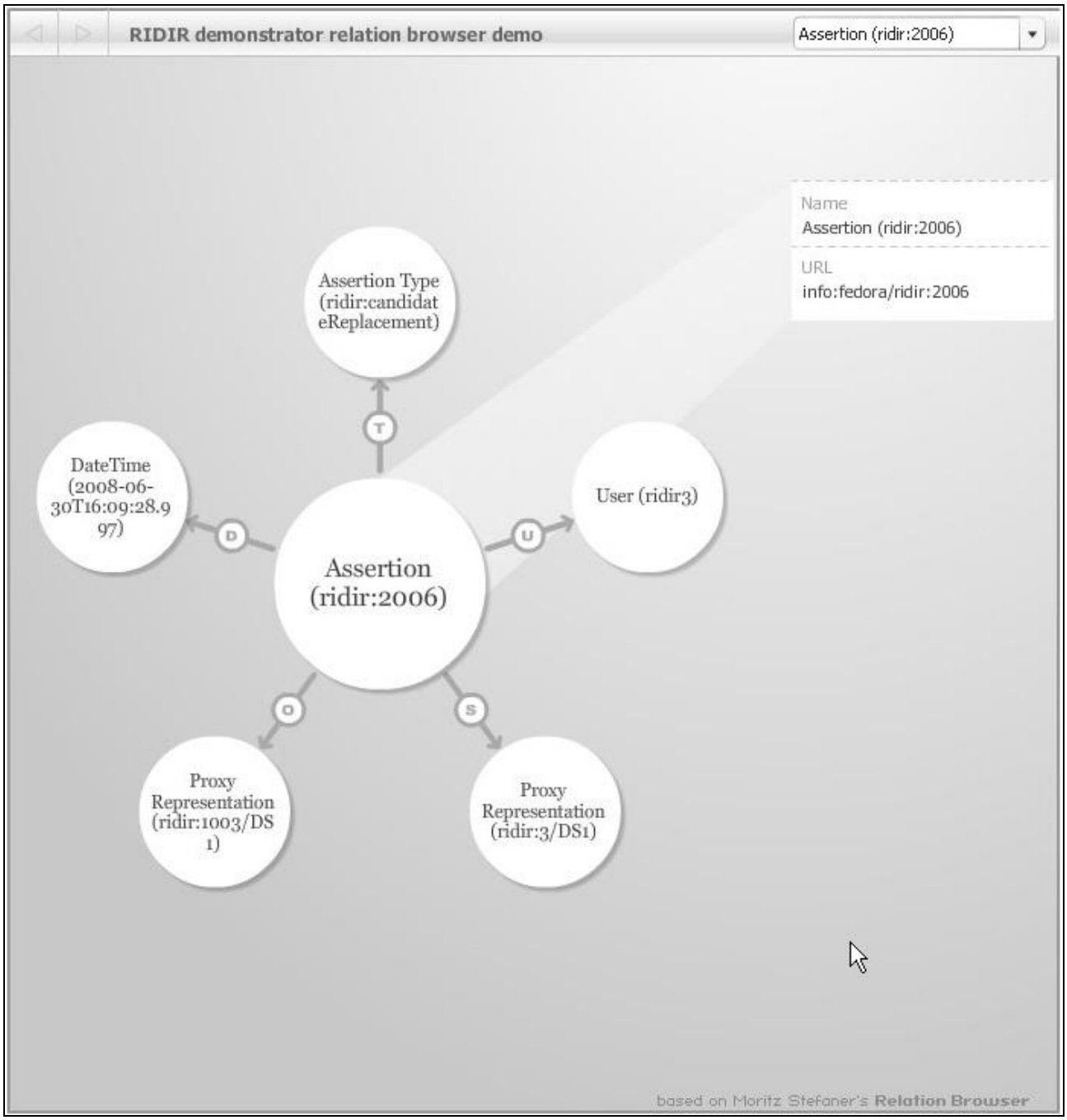

Figure 32: Semantic relation browser, screenshot #3

5.3 Demonstrator outcomes

5.3.1 Lost Resource Finder

Identifier and Interoperability landscape

RIDIR's work here was based on dealing with resources that moved from one repository to another (such as could happen when digital objects are moved from the JISC's Depot repository to a new institutional home). However the work has application to other scenarios where resources

are moved from one repository (or place) to another, such as within EThOSnet, migrating to new repository software, merging and combining of IRs and indeed institutions themselves.
The demonstrator dealt with trapping situations where URLs for resources no longer resolve (giving an HTML '404' error), and then attempting to find the new location (URL) for the resource by using the metadata captured by Intute Repository Search (IRS).

Observations and recommendations (recommendations taken from Appendix A)

Observation	Recommendation	Impact (of observation)
No persistent identifier scheme is in place with an appropriate curation boundary to handle migration of resources between IRs (treating Depot as an IR)	1. Mint resolvable persistent identifiers 9. Combine preservation and resolution responsibilities	Potential for a large number of broken links over time; requirement for a 'lost resource finder' service or for continued provision of resolution for non-curated resources; user frustration with broken links
Some implementations of persistent identifier schemes are scoped to institutions. Around 30% of the IRs harvested by IRS were DSpace implementations that published Handle identifiers for their resources; each institution has its own Handle Naming Authority. As these implementations were scoped to the institution this means that if resources are migrated outside the institution, the institution has a commitment to maintain resolution, but this represents a decoupling of preservation and resolution responsibilities, so in time is likely to lead to 'link rot'	1. Mint resolvable persistent identifiers 9. Combine preservation and resolution responsibilities	Potential for a large number of broken links over time; requirement for a 'lost resource finder' service or for continued provision of resolution for non-curated resources; user frustration with broken links
Identifier schemes with embedded semantics Software packages identified as part of the URL, eg http://eprints.institution.ac.uk	1. Mint resolvable persistent identifiers 2. Identifier structural semantics should have the same lifetime as the resource 3. Use semantically opaque identifiers	Likelihood of identifiers changing when repository software is changed leading to broken links

Software-specific identifier schemes; the identifier syntax is dependent on the repository software used; in some cases different versions of the same software use different identifier syntax (though there were some mechanisms in place to resolve between the different syntaxes)	1. Mint resolvable persistent identifiers 2. Identifier structural semantics should have the same lifetime as the resource 3. Use semantically opaque identifiers	Likelihood of identifiers changing when repository software is changed leading to broken links; requirement to provide services to translate between different identifier syntaxes
Separation of resolution and persistence responsibilities - intent of Depot is that resources move, but published identifier includes depot'; therefore there is a requirement either for Depot to provide ongoing resolution once the resource is moved, or to provide services to cope with URLs that no longer resolve.	1. Mint resolvable persistent identifiers 9. Combine preservation and resolution responsibilities	Need to provide continuing resolution services for non-curated resources or to provide 'lost resource finder' service
No indication of what the identifiers actually identify, eg a splash page, the resource itself, a different version of the resource. Both dc:identifier and dc:relation are variously used for URLs for resources and splash pages.	8. Provide semantically-precise descriptions of what is being identified (ontology) 19. Provide resource linking capabilities with semantics, publish relationships 10. Maintain a registry of identifier syntax, (and any embedded semantics in the identifier, eg whether a splash page or resource is referred to).	Code written to parse identifiers to determine at a basic level what they refer to. Provision of information about what identifiers refer to would allow more automated matching between identifiers.
No standard/published registration of identifier scheme syntax and semantics. Work was done to parse' the DSpace (handle implementation) and EPrints identifiers to derive where an identifier was for a splash page and where an identifier was for a representation.	10. Maintain a registry of identifier syntax	Code written to parse identifiers; 'best guess' based on multiple search results on what identifier schemes were in use in different repositories, time-consuming and potential for errors.
No indications of relationships between identified things. Some relationships (just that there is a relationship) can be inferred where there are several identifiers present in one IRS record; some inferred	19. Provide resource linking capabilities with semantics, publish relationships 10. Maintain a registry of identifier syntax	Code written to create basic relationships between representations and splash pages. Improved service could be provided if better discovery of relationships was possible.

relationships (between splash and representation) where institution had an identifier scheme that could be parsed (DSpace, EPrints).		
No universally-unique identifiers; in general limited alternative identifiers (however some examples of ISBNs were found); no unique syntactical components of identifiers.	4. Mint universally unique semantically opaque identifiers 5. Use universally unique identifiers within resolvable identifier schemes 13. Carry old identifiers in metadata when moving objects	Automated linking of resource identifiers could be provided based on unique identifiers. If URLs included some unique component automated linking could be achieved in this way.
No disambiguation of, for example, author names	14. Use disambiguation services	Automated matching of resources based on metadata could be provided; the accuracy of this would be improved with disambiguated names

Potential uses/deployments for the work done

- General observation: if a persistent identification service with an appropriate curation boundary is in place, there would be very few broken links; so the work to be done is to select and implement such a service, and provide the appropriate modifications/additions to repository software to automatically keep the service up-to-date

- Given that there will still be some cases of broken links, there could therefore be a general 'lost resource finder' service available to all IRs (through 404 page links) to cope with this situation.

- Given the potential for automated lost resource discovery through the use of metadata (and identifiers, where possible), the logical place for such a service is within a service that already harvests the metadata for IRs, ie within Intute Repository Search itself.

- Recommendations for IRS would be:

    - Provision of a 'lost resource finder' service, redirected to by IR 404 pages

    - User searching and discovery to find new locations for resources

    - Archiving of OAI-PMH metadata for automatic discovery

    - System-generation of suggested alternative links based on metadata

        - The scenario here is that the relocated Depot resource, whilst it was still in Depot, will have had its metadata harvested by IRS; realising this was not feasible within the demonstrator

- That harvested old OAI-PMH record can then be used as the starting point (when the user hits the service through a 404) of rediscovery

5.3.2 Locate Related Version

Observations and recommendations (recommendations taken from Appendix A)

Observation	Recommendation	Impact
Lack of URL identifiers for resources - SWS: SWS does provide URLs (from a search) that can be used to link to a particular metadata record; but these URLs include parameters indicating (search) words to highlight, so they are not canonical (there can be two URLs for the same item, with different words highlighted).	1. Mint resolvable persistent identifiers	Code was written to generate canonical URL identifiers for SWS items from those returned in search results
Lack of URL identifiers for resources - TRILT: TRILT identifiers are an 8 character string. There seem to be no citable and resolvable identifiers (ie, URLs on the web) for TRILT resources.	1. Mint resolvable persistent identifiers	Code was written to (a) generate URL identifiers for TRILT programmes (which are in fact a search URL) and (b) formulate our own URL scheme for TRILT broadcasts, using the # fragment identifier to append broadcast information
Knowing what is identified - SWS: The SWS URLs used resolve to a metadata splash page which also embeds a 'player' to view/play the resource. It's not clear what the URL identifier is actually identifying.	8. Provide semantically-precise descriptions of what is being identified	Imprecise semantics when linking items within TRILT and SWS: Without knowing exactly what the identifiers refer to it's not possible to come up with semantically precise relationships between the items, though candidate relationships are used in the demonstrator they are of limited utility due to this.

Knowing what is identified - TRILT: Again it is not defined what the TRILT identifier refers to (the programme itself, the metadata record of the programme); additionally there are cases where a series has its own identifier and other cases where each programme in a series has an identifier.	8. Provide semantically-precise descriptions of what is being identified	Imprecise semantics when linking items within TRILT and SWS: Without knowng exactly what the identifiers refer to it's not possible to come up with semantically precise relationships between the items, though candidate relationships are used in the demonstrator they are of limited utility due to this.
Searching and harvesting interfaces - SWS: SWS provides XHTML, Atom and RSS versions of search results, with varying metadata coverage.		A more sophisticated (eg SOLR) interface would have proved useful. Ability to harvest metadata (eg OAI-PMH, believed not to be present) would be useful to provide integrated discovery services.
Searching and harvesting interfaces - TRILT. TRILT provides no (known) machine search or harvesting interface.		Both a search and a harvesting interface would be useful

Potential uses/deployments/integration of the work done

Could provide a 'show related' type of service in-line with the demonstrator

- Have TRILT, SWS (and others) produce OAI-PMH
- Integrate the resources into existing discovery services (or new ones) such as Intute, IRS
- Build 'show related' service similar to ours following requirements analysis, include things like semantically precise relationships and also investigate user tagging as a mechanism for deriving relationships.
- Provide an OAI-ORE Resource Map of RIDIR’s semantic network of identifier referents and identifiers themselves

6. Outcomes

The RIDIR Project Plan identified a number of aims and objectives which are reproduced below with commentary:

Aims

- *To engage with the identifier and repository communities to understand better their requirements and highlight the benefits of the clear use of persistent identifiers in order to facilitate interoperability where required.*

  The RIDIR Project set out by 'engaging with the ... communities' and, as noted at length elsewhere, quickly discovered that the majority in those communities did not yet share our understanding of how important identifiers would be to the process of interoperability between repositories. We hope that by producing the RIDIR Demonstrator and associated extensive documentation and recommendations the dissemination process that will now follow will help raise awareness of the potential importance of this area and how it can be addressed. Even during the timespan of the project, dissemination opportunities that we have had, interacting with other project teams or with JISC staff, have clearly contributed to such awareness raising, and been welcomed.

- *To develop and build a fully working demonstrator to showcase the findings of this engagement and demonstrate potential means for addressing the issues raised.*

  The Demonstrator that RIDIR has produced shows, we hope, two major benefits that identifier interoperability could bring: on the one hand an approach to dealing with 'lost' digital resources (resources that have somehow strayed outside their original curation boundary so that their URL no longer resolves), and the other an approach that allows the construction of potentially rich semantic maps recording relationships between objects, possibly in widely separated and differing repositories, and which allows those relationships to be persisted and made available for others to use. If implemented as services at a national level these would be a significant contribution to interoperability.

Objectives

- *To raise awareness of persistent identifier interoperability issues within the Higher and Further Education community, influencing repository practices to incorporate these issues and contributing to the understanding of the governance procedures around identifier management*

  The RIDIR Project has produced an extensive set of recommendations ranging from suggestions for national strategy down to practical suggestions for working with digital objects in repositories. We hope that these will be considered seriously be the JISC and by the repositories community. In particular we hope that the JISC will give due consideration to the possibility of a UK persistent identifier management infrastructure, similar to that proposed for Australia, and to the inclusion of complementary RIDIR-like services within it.

- *To provide a clear way of demonstrating issues relating to persistent identifier interoperability and potential solutions for addressing a range of use cases*

  RIDIR set out to address five possible use cases which were grounded firmly in repository practice and needs. The Demonstrator, as delivered, addresses most of those needs and, we hope, will be used to show the possibilities that exist for interoperability when identifiers are used properly and in particular the benefits to ensuring that all identifiers that might reasonably be associated with a digital object are, in fact, represented - thus greatly raising its discovery potential.

Stakeholders

The original RIDIR Project Plan listed six groups of 'stakeholders' in the project's outcomes. Whilst the project has changed somewhat from that originally envisaged it is worth considering the potential benefits of what RIDIR has done to these same groups. For the purposes of this review, RIDIR's outputs will be considered in isolation, which is to say not as part of a potential, broader national infrastructure. Clearly, the comments below assume that the stakeholder(s) have access to a RIDIR-enabled system. The stakeholders listed were:

- repository managers
- repository users
- content owners
- content aggregators
- repository search services
- linking services

The first RIDIR service, to address the problem of persistent identifiers that no longer resolve, could be of clear benefit to the first three groups.

Where a resource is deliberately moved and it is not possible to continue the successful resolution of the old persistent identifier repository managers could use the service to make an authoritative relationship between the old and the new. Users are then not faced with an annoying '404 error' and the possibly long process of trying to find the resource elsewhere. Users of repositories and repository search services who, during a search, come across a broken persistent identifier would be able to use the RIDIR service: the most helpful case would be one where the 'missing' resource has already been discovered and linked by others, even were this not the case a full RIDIR service would provide an efficient multi-target search system to try and find the missing material and record that discovery for others. Content owners too would benefit by knowing that, by these mechanisms, the RIDIR service will help to maintain the discoverability of their work even if a particular persistent identifier fails to resolve to the 'correct' resource.

The second RIDIR service, to locate and record related items, is of potential use to the first four identified stakeholders.

It should be clear from extensive comment earlier in this report how the ability to retrieve relationships between related resources could be of benefit to repository users, content owners and content aggregators. They themselves have the ability to add to the network of these relationships by recording new ones. Whilst we have identified 'linking services' as a stakeholder,

the second RIDIR offering is effectively a linking service which extends the linkages available to an end user. Potentially the user of a RIDIR-enabled system starting out using a non-RIDIR linking service may discover useful material some part of which occurs in the RIDIR system. The additional availability of RIDIR may then extend the linkages available to the user.

In general, by explicit separation of the concerns of resolution and naming allows curation boundaries to be redefined in accordance with policies governing agreements between parties responsible for the curation of both identifiers and resources.

Methodology

Many of the lessons leaned from the methodology undertaken by the project have been discussed under the appropriate sections earlier in this report. From the inauspicious start of discovering a lack of awareness and, in many cases, interest, in the area of identifiers within the repository community the project has built upon the principles that did come out of the focus groups to reach a point where subsequent investigations can be made on a more informed basis.

Specific lessons that other projects could benefit from when undertaking similar studies include:

- The benefit of regular team meetings to share ideas, progress and address problems
- The need to be adaptable and flexible in project planning
- The benefit of clear communication, both with external bodies we have dependencies on and with the JISC

These have all contributed to the outputs the project has delivered.

7. Conclusions

We hope and believe that the RIDIR Project will make a lasting contribution to understanding of many issues surrounding the notion of identifier interoperability. We were more than a little surprised, at the outset of our work, to find that many in the repositories community did not seem to understand the questions that we were asking nor, indeed, our purpose in asking them. We hope that, over the year-long timespan of the project, our discussions with people up and down the country have gone some way to raise awareness of the issues and why our questions are of importance.

The project has also reinforced the originally taken view that identifiers are key to the ongoing and long-term management of digital objects, in repositories and elsewhere. Having said that, the nature of what can be understood to be an identifier needs to be broader than those metadata elements that are usually associated with this term. Many characteristics of a digital object can identify it, and specific schemas that are useful for processing of objects, e.g., Handles, are one of these. Looking at the identity of digital objects in this broader way can, we believe, assist with their management and interoperability.

As the work of the project has demonstrated, the value of the Semantic Web in enabling the relationships that lead to interoperability is vital. The Semantic Web has always promised much, but has not caught the public eye perhaps due to its perceived complexity. By opening up the Semantic Web to the end-users, through allowing them to establish the relationships involved, it is hoped that the RIDIR project can help to demonstrate how the Semantic Web can be of wide use and value to the repository community. As mentioned elsewhere, the OAI-ORE initiative is another approach to bringing the Semantic Web into the repository sphere, and any further RIDIR developments would seek to take full account of this emerging specification.

This lack of initial input from the community at large made for some early problems in scoping the work that RIDIR should do. We were fortunate to have a Programme Manager who was prepared to listen to these difficulties and discuss with us our suggestions for ways forward. Whilst this took up a considerable amount of time, there is no doubt in our minds that the outcomes of the Project are the better for it. The Demonstrator that we have produced addresses real-world issues and goes beyond them to point up possible developments for the future; the extensive research that was carried out by the development team has been documented and, we hope, represents a firm foundation on which that further development might be based; finally, we have been able to make a range of recommendations ranging from the macro to the micro which, if taken up, should ease the way to wider and easier interoperability.

8. Implications and Recommendations

Achievements and recommendations

As noted in Section 2, the RIDIR team was asked not to duplicate any of the work of the PILIN Project in Australia, rather to stay cognisant of the work that they were doing and to produce outputs that were complementary to it. To understand better what RIDIR produced and why, it is first necessary to make some comments about PILIN.

The 15-month PILIN Project in Australia was funded by the then Department of Education, Science and Technology with somewhat more than a million Australian dollars. It was tasked with building and piloting a shared, standards-based, persistent identifier management infrastructure. It was believed that such a service infrastructure would assist with finding digital resources as they move around during their lifecycle and would bring central governance and policy to identifiers associated with Australian repositories. It was recognised at the outset that failures of identifiers are as much to do with poor management and governance as about failures in technology.

The project formally closed in December 2007 but received further funding to bridge a gap between the project end and the start of developing the PILIN work into an Australia-wide provision later in 2008 as part of the Australian National Data Service (ANDS).

Part of the RIDIR Project's focus has been on dealing with digital resources that become 'lost'. As noted elsewhere, there can be several reasons for this but it seems to us undeniable that a service such as that being developed in Australia could potentially bring the same benefits to the UK repository and wider communities; central services, governance and policy should reduce the number of 'lost' objects through commitments to and maintenance of explicit curation boundaries for both identifiers and resources. Such a shared infrastructure could ensure the persistence of working unique identifiers and identifier services over archival periods of time thus aiding the discovery process and contributing to the long-term preservation of the resources in question.

The RIDIR team recommend that the JISC should commission a scoping study to analyse the work of the PILIN project and to establish whether their solution could be transferred to the UK and potentially be transformed into a JISC national service within the e-Framework.

Appendix E: 'RIDIR as part of a national service' examines this recommendation in more detail and considers the potential for 'RIDIR functionality' within such a service.

RIDIR, then, has concentrated its efforts on developing services that would be complementary to those offered by the core of a shared persistent identifier management infrastructure. These services are embodied in a self-contained demonstrator that shows a potential approach to each of two different problems using a common underlying software architecture.

The first service deals with the situation that the URL for a resource no longer resolves and the content effectively becomes 'lost'. This should be an uncommon occurrence within a centrally-managed system but is potentially still possible where users or curators fail to follow the guidelines and systems available to them or where the object lies outside such a system. The service utilises identifiers that may have been associated with the original resource as a basis for a

search. Users of the system are able to record any candidate matches and these are presented as possible targets to future users of the broken URL.

The second service deals with the potential to use identifiers in order to locate resources related to one another and provides a system for recording those relationships as a network of assertions which retains contextual information. Future users of the RIDIR service finding one of the resources in a network are made aware of, and can easily navigate to, the related materials elsewhere in the network; they also have the ability to add further related items. The basis of such discovery is the recording of appropriate identifiers in the metadata associated with digital objects.

The RIDIR team urges the creators of digital objects to include in their metadata any and all associated identifiers in appropriate fields to aid the process of discovery and the creation of wide-ranging networks of related materials.

We hope that the demonstrator and its user documentation will be made available to interested parties in UK HE.

We urge users to feed back to the JISC their views on whether such value-added services would be a useful adjunct to a UK shared persistent identifier management infrastructure.

In collecting the identifiers associated with a digital object, it is beneficial not just to throw these into the metadata bucket, although this is considerably more useful than not doing so, but to have an understanding of why the identifiers are being used and how. There are two main reasons why a digital object may have multiple identifiers:

- The object may have identifiers from multiple schemes for specific purposes. For example, a Dublin Core record for a book may contain the identifier of the metadata record, the identifier(s) of associated metadata record(s) for the book itself, and identifiers for representations of that book. Any and all of these are valuable in assisting users with discovering related items and asserting a relationship between them.

- The object may be a compound object, with multiple parts, each of which can have identifiers of its own. The first example above could be modelled as a compound object, though has not always been so when described in a repository.

The recent and ongoing development of the OAI-ORE[39] (Open Archives Initiative – Object Re-use and Exchange) protocol enables the relationships between items to be described and modelled, demonstrating the 'aggregations' that are created through the establishment of such relationships.

The RIDIR project was aware of the OAI-ORE work, though has concluded prior to the full release of this protocol, and it was not possible to concretely incorporate the OAI-ORE results within the demonstrator. This standard gives the abilities for repositories to publish:

- machine-readable semantics on 'what' is being identified

- machine-readable semantics on the relationships between things being identified

[39] OAI-ORE, http://www.openarchives.org/ore

Any subsequent activity to take the work of RIDIR forward should consider the role of the OAI-ORE protocol in structuring and presenting the relationships established. This would facilitate the interoperability of the aggregations created between RIDIR and other repositories, and offer a degree of persistence for others to benefit from over time.

The RIDIR technology stack and architecture was developed in the context of existing UK services. Should there be any move to consider providing a national service based on RIDIR's work this would have to be considered alongside the results of a study into a national persistent identifier service such as that recommended above. We would therefore recommend that JISC first proceeds with the suggested study on PILIN.

The outputs of the PILIN project include an abstract informational model which could be applied to various persistent identifier services, and an implementation of a service based on Handle.

Handle has particular value in that it is not just a simple identifier resolution service, but allows other informational items to be stored alongside the persistent identifier.

PILIN makes use of this in their FRBR tool to identify both the type of resource and relationships to other resources, using the FRBR vocabulary, although it is worthwhile pointing out that the PILIN model is not limited to using FRBR.

Aspects of PILIN which are particularly relevant to the RIDIR demonstrator are:

- 'Lost Resource Finder': PILIN has a Persistent Citation Resolver Service[40] which provides similar functionality. It allows a (non-persistent) URL to be registered as a Handle which then resolves to the persistent Handle for the resource, so in the case that the original identifier becomes non-resolvable, the service can locate the Handle, and therefore the new location, for the resource. However it does not include any description of the relationship between non-persistent identifiers and their Handles. The RIDIR demonstrator goes further and allows classification of the relationships as 'authoritative' and 'candidate' (user-suggested), along with recording information about who created the relationship and when, or how many people 'agree' with proposed relationships. The outcomes of the study into PILIN will have an impact on the best architectural choices for implementation of such a service.

- 'Locate Related': PILIN does have the ability to link resources using the FRBR vocabulary, but the implementation doesn't provide the flexibility we have demonstrated in terms of usage of other vocabularies and semantics (including emergent, essentially uncontrolled vocabularies), and particularly capturing information about assertions (who, when, authority etc).

We suggest that we have demonstrated the value in using semantic descriptions of resources and the relationships between the resources, and that the outcomes from the demonstrator should be used as requirements in guiding the evaluation of PILIN (or indeed any other identifier service).

[40] See: https://www.pilin.net.au/PILIN_Implementations/Reverse_Lookup_Service.htm

That study should then be able to determine how much of the functionality we have demonstrated should be present in the architecture of a national identifier service, and how much should be provided in supporting services. It should also take note of the outcomes of the investigation of the foundational model as a basis for development of an ontology and related services as appropriate as part of future development. It would be prudent for ontology development to be undertaken with respect to the information model defined for the PILIN project and the data model defined for the OAI-ORE activity.

As well as providing a Handle-based implementation of the PILIN informational model, the PILIN project also addressed the creation of tools to enable adoption, particularly:

- JADHL – A Java API for Handle, making it easy for repository software developers and implementers to integrate persistent identifier services into repositories
- The PILIN web Handle management tools for administration of Handles
- The PILIN FRBR tool for managing types and relationships between resources using the FRBR vocabulary
- The Persistent Citation Resolver Service for managing non-persistent resolvable identifiers which then cease to resolve
- Appropriate copy and multiple resolution – an OpenURL implementation that uses Handle for storing the multiple locations of a resource against a single persistent identifier (as part of FRED – Federated Repositories for Education[41])

We recommend that any further work in determining an appropriate persistent identifier infrastructure should also use these or similar methods to address the enablement of adoption.

There are some discrete areas of RIDIR Demonstrator functionality that we believe it would be useful to integrate into existing JISC services (or develop as separate services).

Particularly, a 'lost resource finder' service could be provided in conjunction with IRS (as IRS captures the necessary metadata to drive such a service – other search services could be developed to make use of RIDIR's discovery API to broaden coverage). In essence this would involve building an enhanced OAI-PMH harvester and search service that also offers broken link resolution, with some degree of automatic suggestions for replacement links based on metadata matching. Recognising that even with a persistent identifier service there will still be corner cases, we believe there is value in proceeding with this. The actual architecture chosen would have to be determined in conjunction with the existing architecture of IRS, although the team would anticipate the outputs of the RIDIR project to be useful in terms of development of the architecture.

We recommend that the JISC and the IRS consider whether the provision of a 'lost resource finder' service could usefully be provided for the repository community.

Whatever the future for a persistent identifier management infrastructure in the UK and the contribution of the RIDIR work to it, the RIDIR team have developed a set of best practice recommendations concerning resources, their identifiers and metadata, and the relationships between resources. We would urge that those with responsibility for setting up repositories, and specifically repository policies, should take note of them. The recommendations are dealt with in

[41] See: http://fred.usq.edu.au

detail at Appendix A, but are summarised here; they are applicable to a small repository as much as to a UK-wide system:

Best practice recommendations

A. Minting identifiers for resources

1. ***Mint resolvable persistent identifiers***
2. ***Identifier structural semantics should have the same lifetime as the resource***
3. ***Use semantically opaque identifiers***
4. ***Mint universally unique semantically opaque identifiers***
5. ***Use universally unique identifiers within resolvable identifier schemes***
6. ***Consider human communication factors***
7. ***Generate identifiers early in the origination process***
8. ***Provide semantically-precise descriptions of what is being identified***
9. ***Combine preservation and resolution responsibilities***
10. ***Maintain a registry of identifier syntax***

B. Publishing and citing resources

11. ***Include descriptive metadata in resolution services***
12. ***Include descriptive metadata when citing resources***
13. ***Carry old identifiers in metadata when moving objects***
14. ***Use disambiguation services***
15. ***Provide capabilities for user-generated metadata***
16. ***Use metadata standards and provide clarification and best practices for usage of standards***

C. Resource discovery

17. ***Implement automated resource rediscovery mechanisms***
18. ***Don't rely on identifiers being persistent***

D. Linking of resources

19. ***Provide resource linking capabilities with semantics, publish relationships***

Appendices

Appendix A: Best practices and recommendations from the RIDIR demonstrator development - further detail

Note: *This Appendix forms an expanded explanation of recommendations summarised at section 4.2.5.3 ('Identifiers and persistence') and repeated in Section 8 ('Implications and recommendations'). It should not be read in isolation as it does not deal with the totality of RIDIR's recommendations, nor does it place them fully in context; for this, the reader is directed to Section 8 of this report.*

The RIDIR demonstrator project scope explicitly did not cover the provision of a persistent identifier and resolution service, and instead focussed on the usage and interoperability of identifiers in practical situations.

Building the RIDIR demonstrator covered many aspects of the identifier lifecycle, including

- the **minting** of identifiers
- making resources available through the **publishing** of resource identifiers and metadata, and the usage of identifiers and metadata when **citing** resources
- **discovery** of resources through search services
- **linking** of identified resources

The practical experience of using identifiers and their metadata and associated services during the course of developing the RIDIR demonstrator led to the formulation of a number of best practices and recommendations concerning resources, their identifiers and metadata and relationships between resources.

These best practices and recommendations are not meant to be an exhaustive list, but represent factors that would or did have an impact on the actual implementation of the Demonstrator applications; and therefore represent factors that the team feel should be taken into account when considering the implementation of identifier and associated services.

A. Minting identifiers for resources

1. Mint resolvable persistent identifiers

If the intent is for resources to be interoperable, to be consumed by other systems and to be reused by citing (through identifiers) in other publications and resources, provide persistent identifiers.

Avoid providing identifiers which, although they may resolve at a point in time, are not intended by design to have a lifetime equivalent to the resource lifetime.
Examples of this are providing search URLs which have a local identifier as a parameter, or software- and application-specific URLs which may change over the lifetime of the resources.

2. Identifier structural semantics should have the same lifetime as the resource

Any structurally-embedded semantics should have a lifetime at least as long as the resource.

- An identifier designed to refer to the latest version of a resource should not embed the date, time or version number of a particular version of the resource
- Identifiers should not encode the name of a particular software package
- Identifiers should not encode the location of a resource if that is expected to change

3. Use semantically opaque identifiers

Following on from (2), usage of semantically-opaque identifiers ensures that the identifiers have no embedded semantics that may over time become inaccurate.

4. Mint universally unique semantically opaque identifiers

Universally unique identifiers are generally not well-suited to be used as resolvable identifiers, as they require one single global resolution service, leading to potential performance and scalability issues.

However the generation of universally-unique identifiers in addition to persistent resolvable identifiers can facilitate rediscovery of a resource should the resolvable identifier cease to resolve (providing that the universally unique non-resolvable identifiers are carried with the resource and its metadata).
Categories of universally unique identifiers

- Central authority-based. Examples of these are ISBN, ISSN. Points of failure are dependency on centralised services for minting and resolution; and failure in practice (eg the minting of identifiers which claim to be ISBNs but in fact have been generated outside the central service). Central authority-based identifiers will be truly universally-unique.
- Algorithm-based. Examples are UUIDs. There will always be a statistical risk of a collision, (though very low with UUIDs). They have the advantage of having no dependence on a central service. Potential failure points include the risk of poor algorithm implementation (eg poor generation of random numbers leads to collisions)
- Content-based. Examples are MD5, CRC32. These end to be computationally intensive to generate, but have the advantage that they can be regenerated on demand from the resource itself. They are useful only for identifying resources at the representation level (one cannot, for instance, generate an identifier that can be used to identify all different versions of a resource).

5. Use universally unique identifiers within resolvable identifier schemes

One of the disadvantages of universally unique identifiers is their dependence on a single service for resolution, or the absence of any resolution service.

However if the universally unique identifier is used as part of a resolvable identifier, this disadvantage is removed.

Furthermore, should the resolution service fail for whatever reason, the universally unique component of the identifier can then be used to re-establish identity in the future.
One examples of this are using a UUID in the path of a DNS-resolvable URL. PILIN uses this technique by minting a 'local identifier' based on date/time down to millisecond and using this as the suffix part of the Handle identifier.

6. Consider human communication factors

Identifiers are often communicated by non-machine means, eg communicated by telephone or scribbled on a bit of paper. Consider these factors when deciding on an identifier scheme. Avoid identifier schemes that result in identifiers that are unwieldy in length. Take into account characters that may be easily mistaken for each other when written down (letter O and number 0) or when spoken (P and B).

7. Generate identifiers early in the origination process

Resources are persisted and transferred informally before they are ingested into institutional repositories.

Consider providing services for minting identifiers early in the origination process, before the resource is formally persisted in a repository.
The identifier could be embedded in the resource itself during the authoring process.

8. Provide semantically-precise descriptions of what is being identified

The background to the project makes it clear that metadata interoperability is a critical component of ensuring identifier interoperability. Interoperability is enhanced when the types of resources that identifiers refer to are unambiguously described.

In general, it is useful to know when an identifier refers to a representation or to a splash page, when it refers to an abstraction (such as a FRBR work) or an individual version of a resource.

Ambiguities can be eliminated by ensuring agreement between those parties or systems that must interpret the meanings of terms and any metadata expressions containing them. The RIDIR team has conducted research into (i) a foundational ontology such that differing viewpoints may be mapped onto a common underlying model of identity, reference and entities, and (ii) a means of incorporating differing vocabularies, both controlled and user-specific or ‘emergent’.

Resource metadata should be provided to achieve semantic precision, where possible using a scheme to facilitate agreement and whose meanings can be interpreted accurately in software, for instance using the DCMI type vocabulary or other standards. JISC-funded projects dealing with vocabularies and Terminology Services should be consulted as appropriate.

OAI-ORE provides a standardised mechanism for including the type of a resource, in RDF terms, in an ORE Resource Map.

9. Combine preservation and resolution responsibilities

If these responsibilities are not linked, there is a decreased motivation to continue to provide resolution services for resources when there is no responsibility over the preservation of the resources. For instance if resources are migrated to another institution, over a period of time it will become more impractical for the original institution to continue to resolve the original identifiers for these resources.

10. Maintain a registry of identifier syntax

Where there are common URL syntaxes for identifiers (usually due to common software implementations) provide a registry of the URL syntaxes so that there is a standard reference for interpreting the identifiers (for instance determining whether a resource or a splash page is being identified).

B. Publishing and citing resources

11. Include descriptive metadata in resolution services

Metadata should be carried with identifiers to add trust to what the identifier is identifying. Inclusion of descriptive metadata within resolution services also provides the ability to rediscover resources when identifiers are 'broken' or cease to resolve. Provision of descriptive metadata helps the user to 'trust' that the resource they have located is the correct one. Descriptive metadata should include information to identify the version of a resource.

12. Include descriptive metadata when citing resources

Acknowledge that even persistent identifiers may at some point cease to resolve, and therefore cite additional descriptive metadata about the resource so that it may be rediscovered through this metadata should the original identifier no longer resolve.

13. Carry old identifiers in metadata when moving objects

If an object has an identifier that will no longer resolve once the object is moved (a local repository URL for example), keep it within the resource metadata to aid rediscovery of the resource.

Provide services to resolve identifiers that no longer resolve by themselves, such as the PILIN Persistent Citation Resolver service.

14. Use disambiguation services

Try to avoid using free-text identifiers in metadata (for example people's names). Instead provide persistent identifiers where this is possible (for example, those derived from 'authority' files)

The JISC Terminology Services report[42] has a recommendation to do further work in the area of 'Named entity authority and disambiguation services'. The MIMAS Names[43] project is ongoing in this area.

[42] The JISC Terminology Services report See http://www.ukoln.ac.uk/terminology/JISC-review2006.html
[43] The Names Project. See: http://names.mimas.ac.uk

15. Provide capabilities for user-generated metadata

Interoperability enabled through metadata is only as good as the metadata itself, both in terms of quality and in terms of coverage. Services such as user annotation and tagging of resources can enhance the metadata of resources thereby enhancing interoperability.

16. Use metadata standards and provide clarification and best practices for usage of standards

Metadata Standards can be (sometimes intentionally) imprecise, and can be subject to interpretation.

Provide clear guidelines for standards usage. For instance the project found that usage of dc:identifier and dc:related was not consistent across institutional repositories whose metadata was harvested by Intute Repository Search, and it was therefore not possible to unambiguously determine the relationship between the metadata record and the identifiers provided within it.

C. Resource discovery

17. Implement automated resource rediscovery mechanisms

Acknowledge that 'persistent' identifiers may not be truly persistent, and may change over time. Discovery services that harvest metadata records (for instance using OAI-PMH) should anticipate this situation and provide services for matching old and new records, maintaining a history of previous identifiers when they have changed.

18. Don't rely on identifiers being persistent

Although covered by some of the other best practices and recommendations listed here, as a general principle when designing systems that consume resolvable identifiers assume that at some point these identifiers may cease to resolve. Build in additional functionality to deal with this situation should it arise.

D. Linking of resources

19. Provide resource linking capabilities with semantics, publish relationships

Provide the ability to link resources together with defined semantics. This aids discovery of resources by exploring relationships with related resources. Repositories should publish using standards such as OAI-ORE relationships between their resources, eg different versions, splash pages and what they describe.

Appendix B: Research and draft foundational model for issues within RIDIR scope

Overview

The abstract architecture defined roles for both a 'foundational ontology' (covering global aspects) and 'domain ontologies' (covering user-, function- or community-specific aspects). An ontology (at least partially) can defines a *terminology*, with additional *semantics* in the form of rules which constrain the circumstances under which terms within the terminology may be correctly used. The semantics ascribe a meaning to the terms defined in the ontology. If these semantics are expressed in the form of a logic which can be evaluated by computation, then the ontology in question can be said to be given a machine interpretation. Then any instance of a thing, such as a certain 'Resource', where 'Resource' is defined in the ontology, can be ascribed certain properties and relations with other instances defined in the ontology.

Together, foundational and domain ontologies represent the rules governing the behaviour and functionality of the services exposed by the RIDIR API. RIDIR's use of ontologies satisfies the requirement that metadata be explicit, machine readable and interpretable.

Foundation Ontology

Research was conducted in order to source a foundational ontology for RIDIR which met the following basic requirements:

- Has a fully-axiomatised *OWL expression* or mapping; *RIDIR requirement*: ease of compatibility with system components that work with RDF data (such as the Fedora repository), and readily-available inference facilities through OWL.

- Expression of *composite relations*, eg part/whole relationships, as well as super-class and sub-classes; *RIDIR requirement*: basic requirement from workshops that identifiers may not only refer to 'whole' resources, but their component or constituent parts

- Is *modular*, so that the ontology has a well-defined mechanism for extension of one module, without affecting the integrity of other parts. *RIDIR requirement*: a premise of RIDIR is that there may be many different 'views' as to the type of the referent of an identifier. Whilst one solution to interoperability is to mandate a single, unified ontology to classify every future eventuality, the consensus on the project was that, despite the growing influence of FRBR in the institutional repository domain, no one view would be likely to prevail if the RIDIR approach were to achieve the widespread adoption necessary for a scalable, robust offering. This consensus is recognised within the abstract architecture, which defines a 'foundational ontology' as separate to various other 'domain ontologies', any of which could overlap or even contradict each other. The RIDIR approach is intended to be resilient to any such inconsistencies at a global level.

- Ex*pressio*n of *context*, to express who asserted what, where and when. *RIDIR requirement*: Important to retain this contextual information, su*ch that RIDIR use*rs are able to adopt, extend and adapt the work of the RIDIR user community in general. This provides for *mechanisms of trust*; at one end of the spectrum, if RIDIR were to incorporate a major

standard for a controlled vocabulary such as FRBR, and as such the IFLA become a member of the RIDIR community, the likelihood of RIDIR users within the bibliographic domain adopting this vocabulary would be high. At the other, a term minted by a casual individual user for her own purposes in the style of a tag label on the popular del.icio.us bookmarking service is less likely to achieve the same level of reuse across the RIDIR user base. The overall intended effect is that cohesion is retained amongst the terminologies held by the system, whilst retaining an organic flexibility to enable a RIDIR service to adapt over time and remain robust to evolving, often unforeseen future requirements. The overall functionality is to support interoperability, and the specific RIDIR requirement is to investigate and demonstrate how interoperability is supported through through use of persistent identifiers. If those identifiers retain contextual information to support the formation of networks of trust, including those identifiers derived from vocabularies which are bottom-up or emergent, then there is an argument to say that interoperability can be achieved through the reuse of identifiers enabled by those trust networks.

- Sufficient *conceptual abstraction of roles*, such that there is a distinction between the identities of those components in software representing some object as playing *roles* and the entities themselves, rather than conflating the two. In other words, the foundational model is able to model the world as having resources that play a role in more than one process without losing its identity. *RIDIR requirement:* modelling identifier referents as entities playing roles without losing their identities in differing contexts is a fundamental requirement in terms of accurately modelling the way in which resources interact within the context of events and processes[44].

Four main candidates were (briefly) investigated:

- *BFO (Basic Formal Ontology)*[45] – extensible formal model of events and processes

- *MPEG-21 RDD (Rights Data Dictionary)*[46] – the primary reason for investigation is the RDD's implementation of the definitions arising from the <indecs> project. <indecs> also functions as the foundation of the work of the ISO TC SC9, and influential on the metadata model associated with DOIs, so is therefore significant in the context of the original proposal. Although <indecs> did not produce a readily-available published ontology, a mapping of the MPEG-21 RDD to an OWL ontology is available.[47]]

- *ABC (Abstract Base Classes)*[48] – very similarly to the MPEG-21 RDD, <indecs>, the primary focus of ABC was to enable metadata interoperability through a core expression in terms of events, such that relating resources to each other is only expressed with reference to the context (in terms of events) relating to those resources, eg creation, adaptation, etc. However, a lack of axiomatisation in the available ontology expression seems to limit the practical usability of the work.

---

[44] A discussion of various ontological approaches See http://wonderweb.semanticweb.org/deliverables/documents/D18.pdf
[45] Basic Formal Ontology (BFO) See http://www.ifomis.org/bfo
[46] Introduction to ISO/IEC 21000-6 Rights Data Dictionary See http://www.chiariglione.org/MPEG/technologies/mp21-rdd/index.htm
[47] http://rhizomik.net/semdrms/
[48] ABC Model See: http://metadata.net/harmony/Results.htm

- *IRE (Identity of Resources and Entities)*[49] – a 'reusable ontology pattern' built upon the richly axiomatised, mature and modular DOLCE foundational ontology.

Results of Evaluation

It was quickly determined that the IRE represented the only ontology that (a) explicitly met all the requirements criteria, and (b) held some promise of practical results within the very limited time frame available to this exercise. Undoubtedly, adoption of each of the other options would yield fruitful results with additional effort, and it may be useful as an exercise to evaluate these further at some point, along with more comprehensive research into any other activities that may be relevant. In particular, approaches such as 'referent tracking'[50] and the 'catalogue of entities' approach such as that taken by the Okkam project[51] are of possible interest.

Overview of IRE Ontology

Note: The descriptions of IRE given below and the theoretical assumptions lying behind it have been adapted from existing IRE literature.[52]

To quote the abstract of an article: Identity of Resources and Entities on the web by Presutti and Gangemi:[53]

> *One of the main strengths of the web is that it allows any party of its global community to share information with any other party. This goal has been achieved by making use of a unique and uniform mechanism of identification, the URI (Universal Resource Identifiers). Although URIs succeed when used for retrieving resources on the web, their suitability as a way for identifying any kind of things, for example resources that are not on the web, is not guaranteed. In this article we investigate the meaning of identity of a web resource, and how the current situation as well as existing and possible future improvements can be modeled and implemented on the web. In particular, we propose an ontology, IRE, which provides a formal way to model both the problem and the solution spaces. IRE describes the concept of resource from the viewpoint of the web, by reusing an ontology of Information Objects, built on top of DOLCE+ and its extensions. In particular, we formalize the concept of web resource, as distinguished from the concept of a generic entity, and how those and other concepts are related e.g. by different proxy for relations. Based on the analysis formalized in IRE, we propose a formal pattern for modeling and comparing different solutions to the identity problem.*

Implementation issues for RIDIR

The primary purpose of the IRE is to model resources on the web, rather than resources held within an institutional repository. However, given that institutional repositories and their associated services such as search facilities are 'on the web' (whether in the globally-resolvable extent or limited to within the institution), then the approach taken by IRE was thought to retain

---

[49] IRE homepage See http://wiki.loa-cnr.it/index.php/LoaWiki:IRE
[50] See: http://org.buffalo.edu/RTU/papers.html
[51] See: http://www2007.org/workshops/paper_150.pdf
[52] See: http://www.neon-project.org/web-content/images/Publications/towards%20an%20owl%20ontology%20for%20identity%20on%20the%20web.pdf
[53] See: http://www.igi-global.com/articles/details.asp?ID=8115

commonality with that required for RIDIR at a fundamental level: modelling the distinction between an *institutional resource* and an *entity*.

Drawing close analogies with the IRE work, the issues which emerged from an analysis between persistent identifiers created by institutional repositories, resources they curate and/or reference, and entities, the following issues were considered:

- Institutional Repository and Web semantics: How should the semantics of institutional repositories on the web be clarified and formalised, at least in terms of the basic notions involved?

- Referencing: What does it mean to reference something?

- Multiplicity of referencing: How can one clarify whether (or when) a reference to something is unique or non-unique? How can issues of 'uniqueness' be applied in principle and practice, and whether only allowing one identifier for the reference should be admitted for such a reference is desirable or feasible?

- Coupling between web and real world: How should the relationship between those things held within an institutional repository, and surfaced on the (or 'a') web, and those things in the real world (such as individual authors or books) be made explicit?

- Resolvability of references. The clarification of when and how a reference is resolvable?

The time was not available on the RIDIR project to investigate these options in detail, but only to focus on those key areas which could lead to useful insights to elaborate in developments beyond demonstrator level.

Identifiers for real-world entities and resources

Figure B1: Real-world entities and Identifiers

The relation depicted by the arrow in Figure B1 in the IRE analysis is analogous to a general assumption made in computer science, and usually in web science too: there is a 'virtual world' comprised of 'symbols', whilst there is a distinct 'real world' made up of 'things'. This provides a means for machines to recognise (or 'resolve', or 'refer to') entities 'as such', unless they are symbols as well.

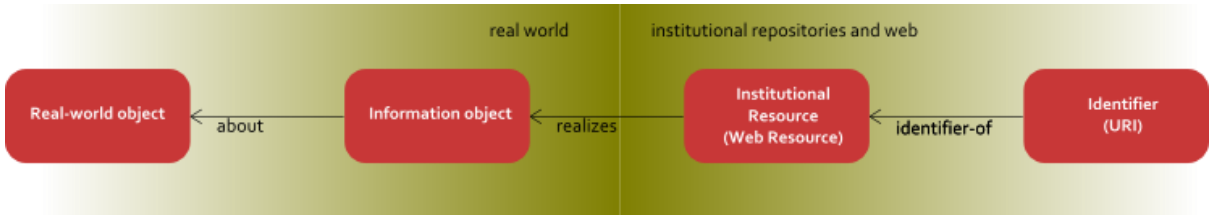

Figure B2 - four layers of referencing

Figure B2 shows the four layers of referencing that are assumed in the analysis of the institutional resource (and web) referencing problem, each of which distinguishes certain types of thing. Resources and Identifiers are taken to be 'virtual' things, and Information Objects and Real-world Objects 'real-world' things.

IRE Definitions applied to RIDIR

#1 A persistent identifier identifies an *institutional resource* (via 'abstract locations', assuming identifiers are resolvable). By analogy, in web terms, a *URI* identifies a *web resource*. The URI mechanism creates a combinatorial space made of what IRE terms *abstract web locations*. Each *abstract web location*, e.g., the one localised by 'http://deposit.edina.ac.uk', can 'contain' a *computational object*, e.g. a digital file stored within on the Depot repository's file system. A *URI* is a string that satisfies syntactical rules defined in IETF RFC 3986 (Berners-Lee et. al, 2005).

#2 If an *institutional resource* is accessible through web-based mechanisms, i.e., if the URI is resolvable, then the computational object is a *web resource*. To extend IRE, this definition is clearly applicable to other identifiers schemes such as DOI; if the DOI's Handle Identifier is resolvable, then the computational object would be a '*handle resource*'. Both web resources and handle resources are modelled as subclasses of computational objects.

#3 A *web resource* realises some *information object*. In FRBR terms, an *information object* corresponds closely with a 'Work'. So an example of an *information object* is a poem, whose '*information realization*' (American English used for precise consistency with IRE ontology terminology) could be a book or book chapter (corresponding to FRBR 'expression'). A *web resource* is a *computational object* made available on the web, hence accessible through a web protocol (e.g., a document, a web service). Given *computational objects* can be expressed as subclasses of *information realization*, *information object*s are always considered 'real-world', but *information realization*s can be *computational object*s, always considered 'virtual'. In other words, the book chapter FRBR expression may have an embodiment at the FRBR 'manifestation' level, as a resolvable *web resource* in IRE terms.

#4 An *information object* is 'about' some real world *entity*, because if we admit that at least some URIs are unique in terms of addressing web resources located in the *abstract web location* combinatorial space, the problem space is then reduced to analysing the nature of the relations between *information object*s and real world *entities*. An *information object* is some unit of information having its own identity that has been created by some agent at some time for some reason. *Information object*s range from texts to pictures, from poems to logical formulas, from diagrams to sounds, and are independent from their physical realization.

#5 The '*being-about*' relation requires that information objects are interpreted by someone that is able to conceive a 'reference' from information objects (either those contained in a resource, or others that can be associated with them), to a set of circumstances, in which real world entities are '*situated*'. An '*entity*' is anything in the real world (material, social, cognitive, etc.), and is called a 'particular' in DOLCE.

#6 URIs identify Abstract Web Locations. A URI is the identifier of an Abstract Web Location. This expresses and 'operationalises' the being-about relation, within the web setting. Again, this is extensible to e.g. handles and abstract handle locations.

#7 Abstract Web Locations are locations of Web Resources, and each Abstract Web Location can be the web location of at most one Web Resource. A Web Resource can be placed in one or more Abstract Web Location(s), which in simple terms means that the identity of a web resource is something that goes beyond its location. An abstract web location is a point in the combinatorial space that is created by the URI addressing mechanism (that is, each URI identifies one and only one abstract web location).

According to the IRE authors these definitions serve to ensure the resource identification, access and location aspects are sufficiently factored out in order to satisfy three requirements of identifiers:

- **Immutability**: an object's identifier should be the same at any point in time and everywhere (globally recognizable, or 'resolvable')
- **Uniqueness**: two objects cannot be represented by the same identifier
- **Singularity**: two different identifiers cannot represent the same object.

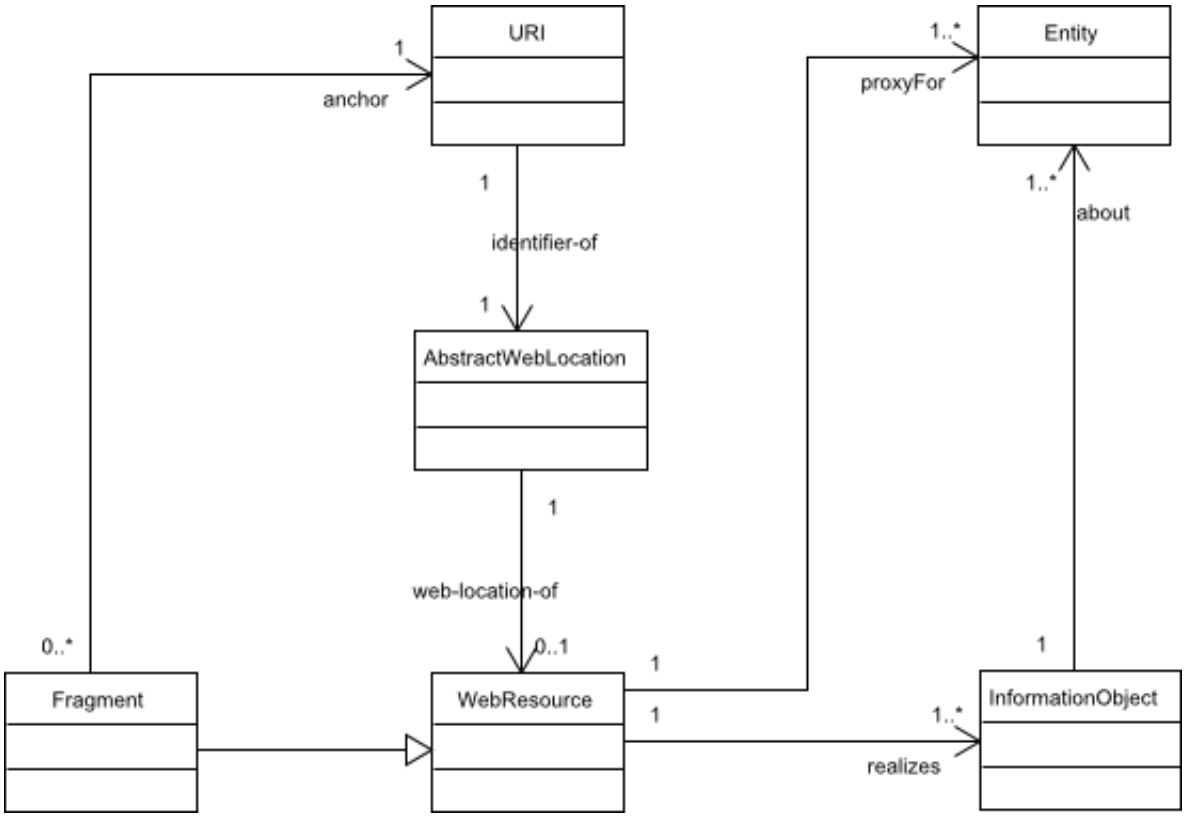

Figure B3: Key Concepts Provided by IRE

The relations may be used on the web by navigation of the basic relations depicted above. The information object could identify a specific thesis, be 'about' Ronald Reagan's US Presidency, and be 'realized by' a particular web resource. The abstract web location represents the situation whereby this web resource is located by the specific abstract web location identified by an *abstract web location at a given time* (see below).

Institutional Repositories and the Web Architecture: Resources and their Representations
Institutional repositories studied during the RIDIR project (DSpace, EPrints and Fedora) consider the 'type' of format of a web resource (or FRBR manifestation) as highly significant, so that each formatted version warrants its own persistent identifier. For example, a PDF form (IRE *computational object*, or FRBR[54] *manifestation*) of some information object (FRBR *expression*) would typically retain a distinct identifier from that of the Word form of the same object. The principal aim is to ensure the exact 'representation' of a resource is supplied over archival time spans, across all situations.

This facility is explicitly provided for in repository software offerings such as Fedora, which has the notion of a 'digital object' having multiple 'representations', 'manifestations' in FRBR terms, or 'datastreams' in Fedora. These datastreams, which may have a content type (mime type) defined and can be associated with digital object accessors called 'disseminators' to allow web access via a URL. Datastreams fairly clearly correspond to *web resources* (as computational objects) *but only in conjunction with an exact identification of form*. That is, a datastream cannot be used to refer to both a PDF and Word form of a thesis; some digital object would need to be defined as being common to both forms. A 'digital object' would typically correspond to a type of computational object, and is often more generically termed a 'resource'. In fact a 'digital object' may conceivably correspond to an information object such as a book; the purpose of such an object would be purely to store metadata, and be part of an information network overlay[55] or the repository.

The key point here is that the form is required to distinguish different 'representations' for a given 'digital object'. This contrasts with the web architecture, where different 'representations' (at 'datastream' level) of a 'resource' identified by a certain URI may be returned through the HTTP content negotiation when accessing the resource.

[54] FRBR See: http://www.ifla.org/VII/s13/wgfrbr/index.htm
[55] See: http://www.dlib.org/dlib/november05/lagoze/11lagoze.html

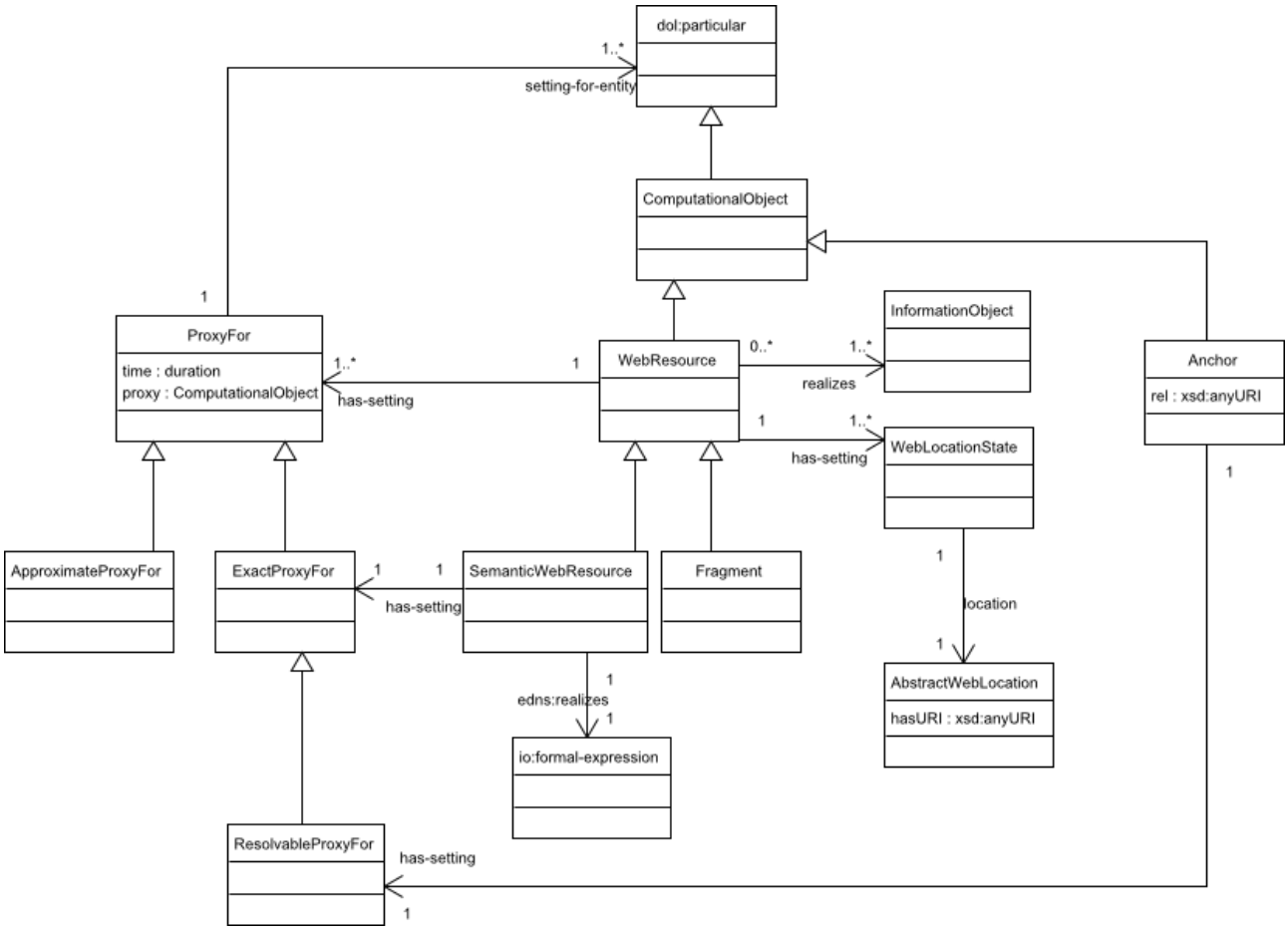

Figure B4: IRE with n-ary relations

The diagram shown above restates the same IRE ontology, using classes for the binary relations that explicitly reifies them with a time component using the n-ary relation pattern,[56] and which was the version of IRE constructed and evaluated for RIDIR since none was made available to the project by the IRE authors. It also depicts a taxonomy for the ProxyFor reified relation, to allow the representation of the triangle of relations that implement the four layer references over the web between information object, WebResource and DOLCE particular.

The taxonomy is defined as follows:

ProxyFor In general we say that a web resource *functions as a proxy for* an entity, at a given time. This association between a web resource and an entity means that the web resource realizes an information object, which is about some entity or entities at a given time.

ApproximateProxyFor is a relationship between a web resource and more than one entity at a given time, where the web resource realizes some information objects, which are about those entities. In this case the web resource approximately represents the one or more entities.

ExactProxyFor is a relationship between a semantic resource and one entity at a given time, where the semantic resource realizes an information object, which is about only that entity, and describes it through a semantic structure. For example, an individual of an OWL ontology can be an ExactProxyFor an entity.

[56] See: http://www.w3.org/TR/swbp-n-aryRelations/

ResolvableProxyFor is a relationship between an anchor[57] and a web resource at a given time; the intention is that the anchor allows access to the web resource it is *proxy for*.
For example, W3C' in a HTML document is a *resolvable proxy for* the W3C home page. The anchor specifies access to it by clicking the corresponding URI as a link. It may also have semantics by means of a semantic relation, in an OWL ontology on the web for example.

Mapping digital objects and their representations

The IRE authors note that "each specialization of *proxy for* can correspond to a different computational approach, or more specifically to a different operational semantics associated with the resolution of the web resource's URI."

Therefore, the issue arises with the use of IRE in the institutional repository context, should an institutional resource maintain an identifier for (by means of its URI) a specific 'manifestation' of some related repository 'digital object', or should it identify the digital object itself? The IRE literature does not explicitly discuss content negotiation and the multiplicity of bitstream-level representations[58].

A 'representation' could be modelled as a ResolvableProxyFor a Web Resource. ResolvableProxyFor is a situation whereby the WebResource realizes exactly one information object which is only about one entity (derived from its parent situation ExactProxyFor), and one which is also the setting for a computational object. For a digital object representing a thesis with two forms, a Word and PDF document, there would be two representations which are ResolvableProxyFor-s the thesis. The thesis could therefore have only one URI and return the distinct form based on the circumstances of web content negotiation. But taking the case of an institutional repository identifying a single abstract web location through a format-specific URI, rather than the generic case on the web, then the representation would by definition be a web resource itself: in other words, all representations would also necessarily be resources. Alternatively, a 'representation' could be modelled as being one of many instances of 'RepresentationProxyFor' a single WebResource ('digital object'), which itself is the single ProxyFor some entity. Here though, the RepresentationProxyFor is also a WebResource.

A 'representation' could be thought of as referring to exactly one time interval holding for all accesses of a 'resource' over the web. This ties the WebResource to the notion of access: the WebResource is by definition a computational object, a physical artefact, which participates in the physical computational process of access to a notional digital object using web machinery. Given an institutional repository 'surfaces' its digital objects using web machinery, then it must also be producing addressable Web Resources. A 'web resource' therefore equates to a 'representation' located at a certain abstract web location during time *t*, realizing exactly one 'digital object'. The objective of a 'persistent' identifier in URI form is therefore to ensure resolution services are provided to maximise time *t*, ensuring that all accesses using the URI are consistent to a level one can consider 'persistent'. This essentially means that a WebResource is exactly the bitstream returned by the HTTP communication event over a certain time interval; for an institutional

[57] See: http://www.w3.org/Terms
[58] although see http://lists.w3.org/Archives/Public/public-swbp-wg/2006May/0009.html for the IRE author's opinion on the relationship between IRE and the definitions given by the W3C TAG

repository minting *persistent* identifiers in URI form for their resources, multiple accesses must return that exact bitstream *over all time*.

In DOLCE, a representation language 'orders' an information object; in an institutional repository therefore, the 'format' of a digital object can be said to identify the representation language. So in the case of an institutional repository, format would also govern the process that causes (or once caused) the web resource to be the realization of the information object in question; if format is identified by content type (MIME type), then the institutional repository would govern the exact WebResources ('representations') realised at a URI (over time *t*). So to use the previous example, the thesis content is the *information object* (the repository 'digital object') that is *ordered by* the *content type* 'application/ms-word', is *about* the thesis *entity* (separately identifying the thesis), and *realized by* some process governed by the repository to produce a *ebResource* that is its Word document form. Since it also has the content type 'application/pdf' it can similarly be *realised by* some process into its PDF document form. **Each form has a separate URI which defines the combinatorial space (location) for access to that resource over HTTP**.

In a similar way, using Handle's resolution machinery, each form also has a separate handle which defines the combinatorial space (location) for access to that resource using handle machinery. The IRE model supports preservation needs through the use of multiple resolution mechanisms, affording levels resilience to changes in repository implementation over time. Were a national shared persistent identifier infrastructure to be established, changes in agreements with providers of resolution services (such as CNRI), would also be supported.

Note in the model described above, the Word form is the ResolvableProxyFor the thesis, not for the information object representing the thesis content. This allows assertions (in RDF) to be made about the thesis separately from the information objects holding the exact content about the thesis, thus enabling different versions of the content to retain a coherent semantics within the IRE context. For example, the thesis may itself have metadata representing the 'being about' relation to something, eg some individual concept identifier representing Reagan's US presidency, or it may have a dc:creator relation to the individual concept identifier representing the person concerned.

Relationship with DOLCE Upper Ontology

The 'DOLCE-Lite' module implements core ontological choices sufficient to provide the building blocks sufficient for IRE-based analysis and capable of extension to cover various domain ontologies:

- Universals, Particulars and Individual Properties
- Abstract and Concrete ('real-world') Entities
- Endurants and Perdurants
- Qualities and quality regions (spatial and temporal)
- Mereology (Parthood hierarchies)
- Temporality

The 'Descriptions and Situations' modules make basic distinctions between 'descriptive' and 'ground' entities, where the descriptive entities include social objects, like the 'student'

or ’professor’ roles, the ’being active’ task, methods, and also information objects like the text of a thesis. The module explicitly considers descriptive entities to have a lifecycle that differs from that of 'pure' information, which is an abstract entity. The module contains axioms which force a separation between ground and descriptive entities. A definition for 'context', identified as a requirement for RIDIR, can therefore be built upon concepts which define a situation that satisfies a certain description. Events and states are unified by situations and are considered therefore ground or 'real-world' in the sense that they occur in software, databases, etc (eg an http communication event); whereas the descriptive element is, for example, an HTTP Access Requesting situation ('context') satisfying the HTTP Resolution method which is a description, a 'social object' (with a completely different lifecycle to the real-world events, objects etc).
The 'information objects' module extends the descriptions and situations module to express the 'realisation' relation between physical 'information realizations' (computational objects in RIDIR), abstract 'information objects' and particulars (entities). It also covers expression and encoding, eg grammars and schema for information objects, within a descriptions and situations setting.

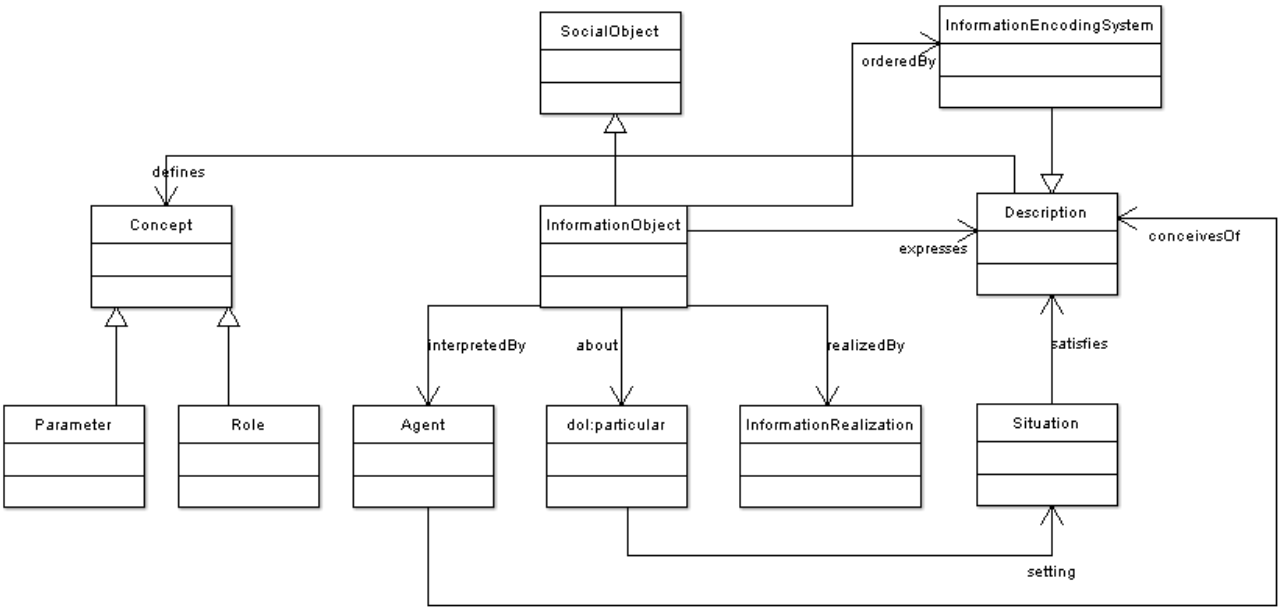

Figure B5: Ontology of Information Objects

The 'plans' module provides categories for plans and their execution, used extensively within RIDIR ontology prototyping as basic categories for descriptions and situations, as per the pattern depicted below.

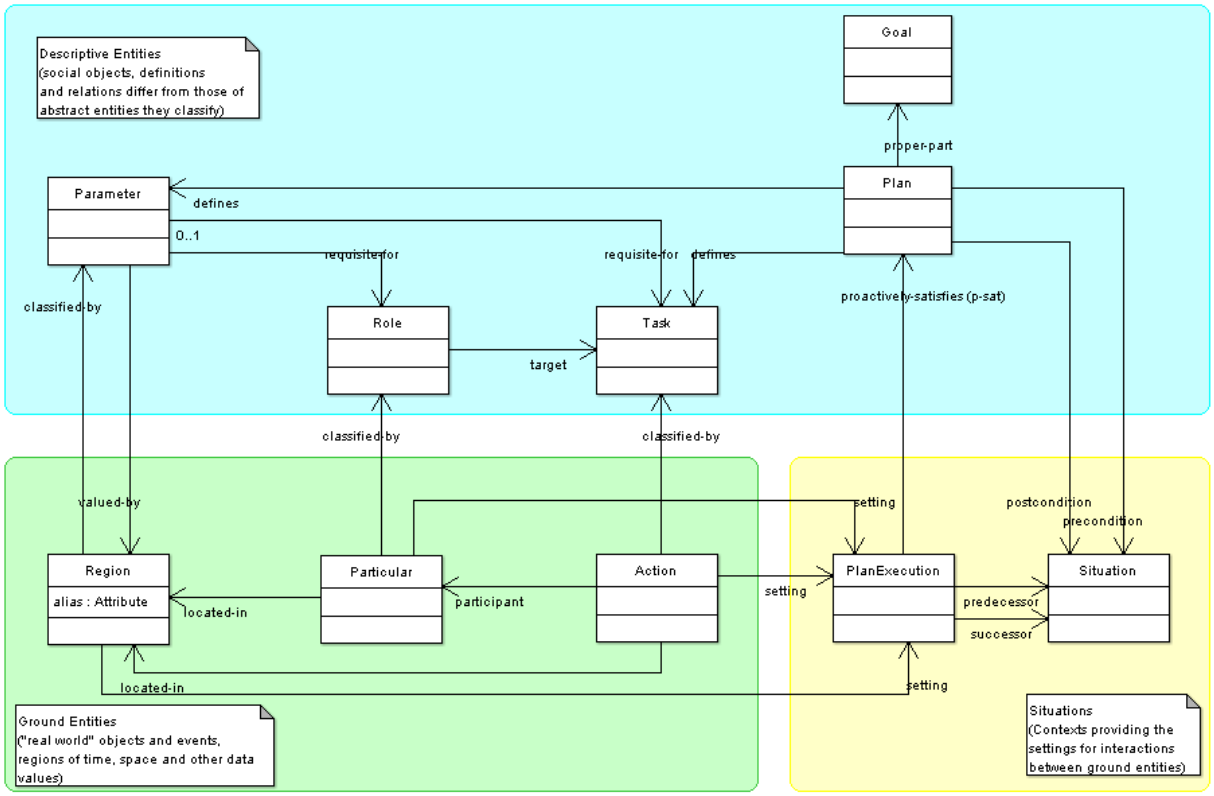

Figure B6: Plans Pattern using DOLCE

The 'temporal relations' module provides some categories for temporal relations.

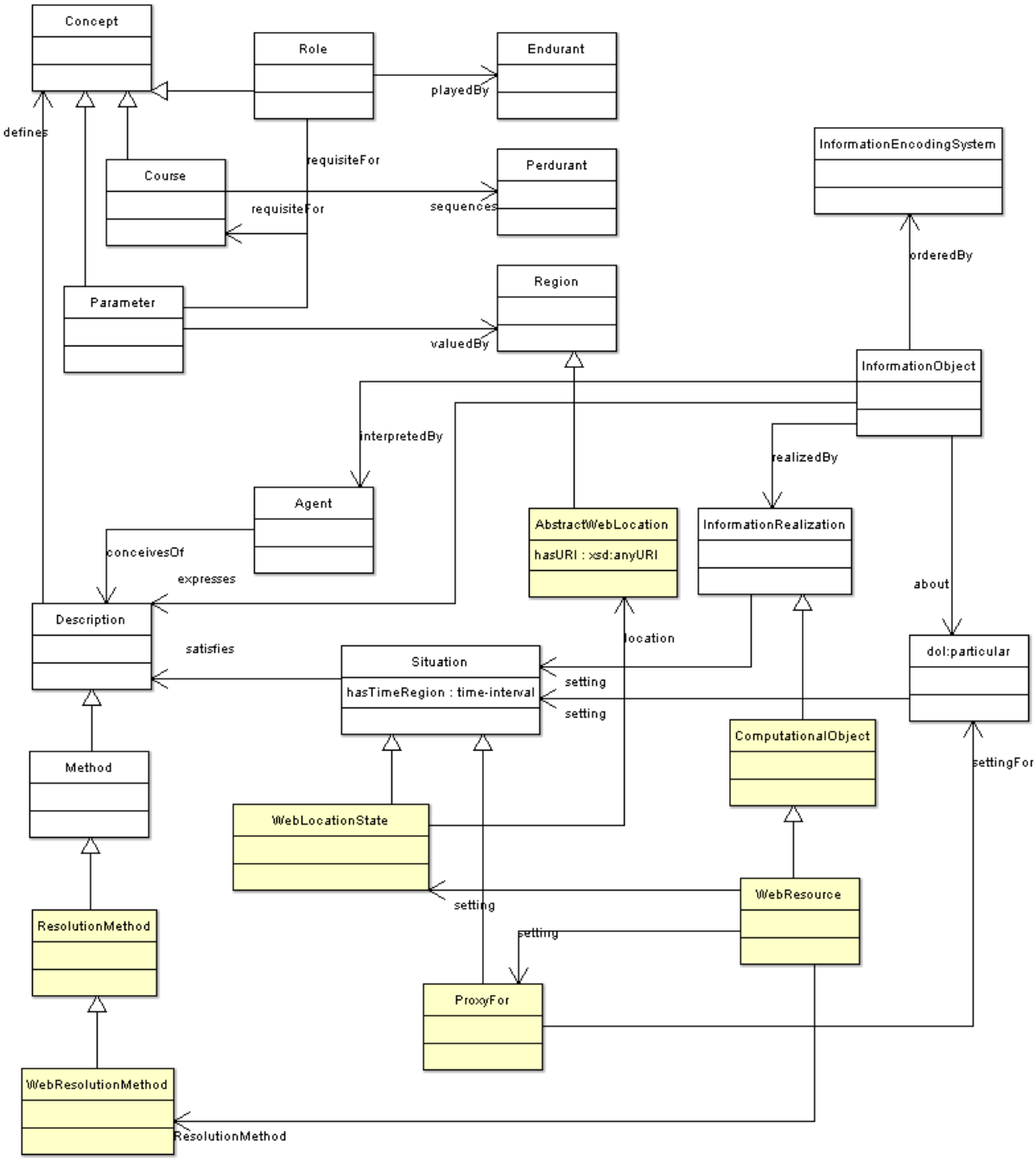

Figure B7: Relationship of IRE with the DOLCE Upper Ontology

Figure B7 depicts the basic IRE-based definitions given above mapped to classes from the DOLCE ontology and its modular extensions. Reified IRE classes are shown in yellow, whereas classes left white are from existing DOLCE modules.

Implementation of Prototype Foundational Ontology

A reified version of IRE was constructed (in the ontology language OWL-DL) based on the documentation found on the web, and various mappings to DOLCE modules introduced. This ontology was then evaluated by constructing further ontology modules based on the DOLCE framework to model application and services behaviour (see subsections below).

Overall, having RIDIR's foundational ontology based upon IRE was thought to satisfy three major requirements of an identifier derived from the application of IRE:

- **Immutability**: an object's identifier should be the same at any point in time and everywhere (globally recognizable, resolvable)
- **Uniqueness**: two objects cannot be represented by the same identifier
- **Singularity**: two different identifiers cannot represent the same object.

Adaptation of IRE and DOLCE to RIDIR

Although prototyping of a RIDIR ontology-based framework based upon the IRE and DOLCE modules was carried out to investigate the 'how' or 'cost' approach as thoroughly as possible, time constraints meant that embedding the ontology within live demonstrator software was not feasible, and experimentation was conducted within stand-alone ontology tools alone (Jena[59], Pellet,[60] SWOOP,[61] Protégé,[62] TopBraid Composer[63]]. The results of the prototyping work were promising in term of future work and are presented below for reference purposes.

Model for Retrieval of WebResources over HTTP

[59] See: http://jena.sourceforge.net
[60] See: http://pellet.owldl.com
[61] See: http://code.google.com/p/swoop/
[62] See: http://protege.stanford.edu/
[63] See: http://www.topbraidcomposer.com/

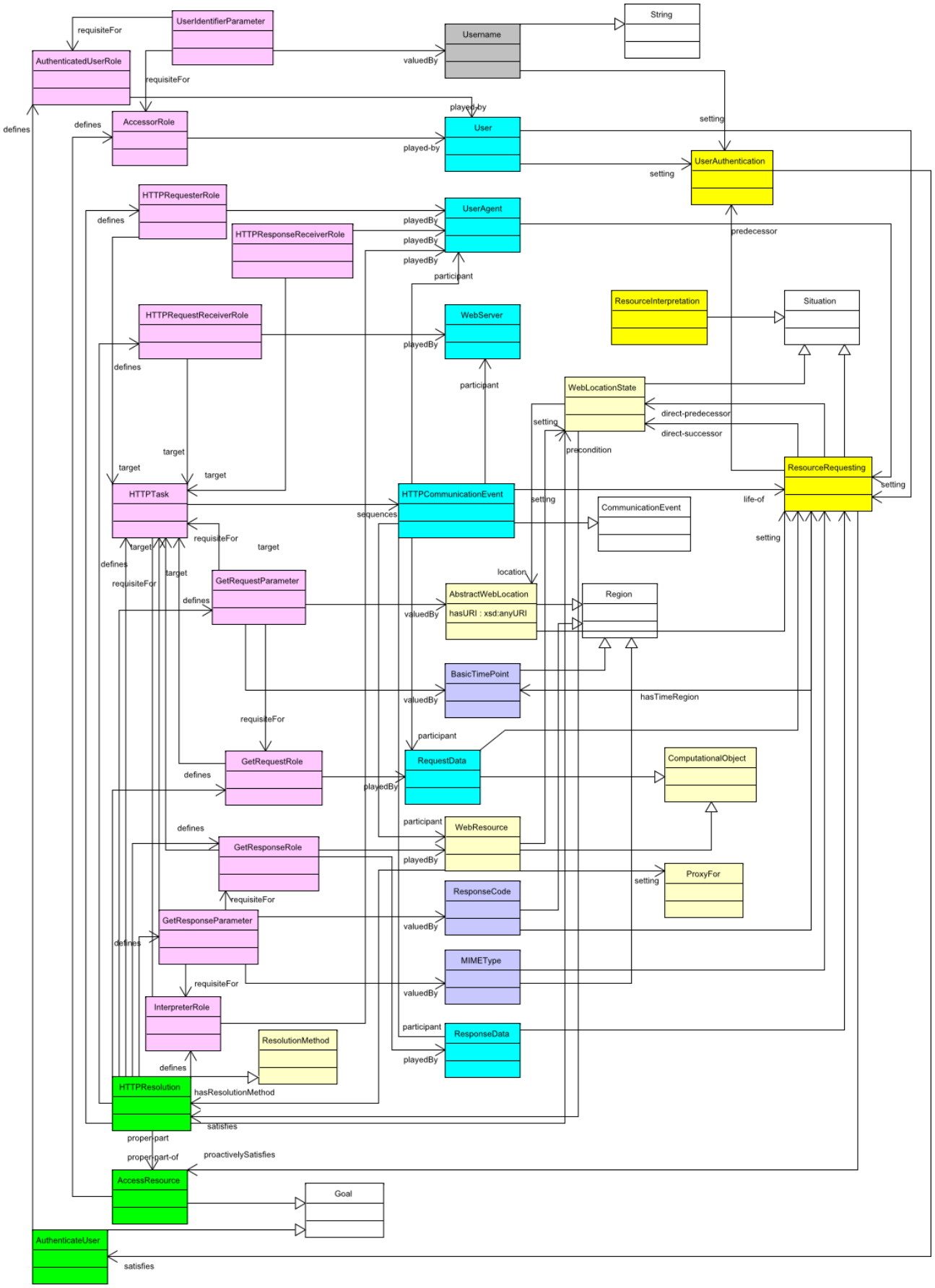

Figure B9: Foundational Pattern for Web Resource retrieval

The diagram above depicts WebResources as playing the role of a Response within the overall context of a ResourceRequesting situation that satisfies the AccessResource description, which itself is a part of an HTTP Resolution description. The real-world computational events and objects for RIDIR are shown in light blue, the 'regions' (concrete values) in grey, descriptive roles and concepts in pink, descriptions themselves in green, and RIDIR-specific situations (contexts) in darker yellow.

The pattern depicted is the core pattern within RIDIR's prototype foundational ontology. It helps the explicit expression of curation boundaries by providing:

- hooks for mapping arbitrary domain ontology via the description patterns – this includes ontology for elements implemented in application domain, such as an explicit model for claims and assertions (also implemented within the demonstrator)
- hooks for ground software components and computational machinery
- a way of representing data values (necessary for metadata and value-type (eg string) representations of identifiers)
- temporality for explicitly reifying, bounding and measuring the degree of persistence covered by the scope of an identifier
- basic ontological facilities for representing compound objects

Note that in DOLCE, and therefore IRE, Regions specify data values. During RIDIR prototyping, the MIME type was included as a Region, one participating in an HTTPCommunicationEvent and values the GetResponseParameter role. All datatype values were implemented in a separately-created datatype model, and mapped to DOLCE as subclasses of DOLCE's AbstractRegion class, following a pattern used for the Core Ontology for Multimedia (COMM)[64], which is also based on DOLCE.

Application of the IRE to the Demonstrator Application Use Cases

This section outlines a design process used to model the applications satisfying the use cases using the foundation ontology described above. The project anticipates that generic learning has been gained through this process in terms of potential future development of applications which use the RIDIR approach going forward. The analysis helped:

- define which elements were key to support within the RIDIR API
- eastablish the feasibility of using the model developed in the role of RDIR's foundation ontology, in the light of the finalised use cases
- the feasibility of using ontology tools within development process
- the feasibility of integrating an inference engine to process RDF based upon the foundational model described within the context of repository software (Fedora).
- determine a realistic scope for the iterations during the software development phase of the demonstrator.

[64] See: http://comm.semanticweb.org/

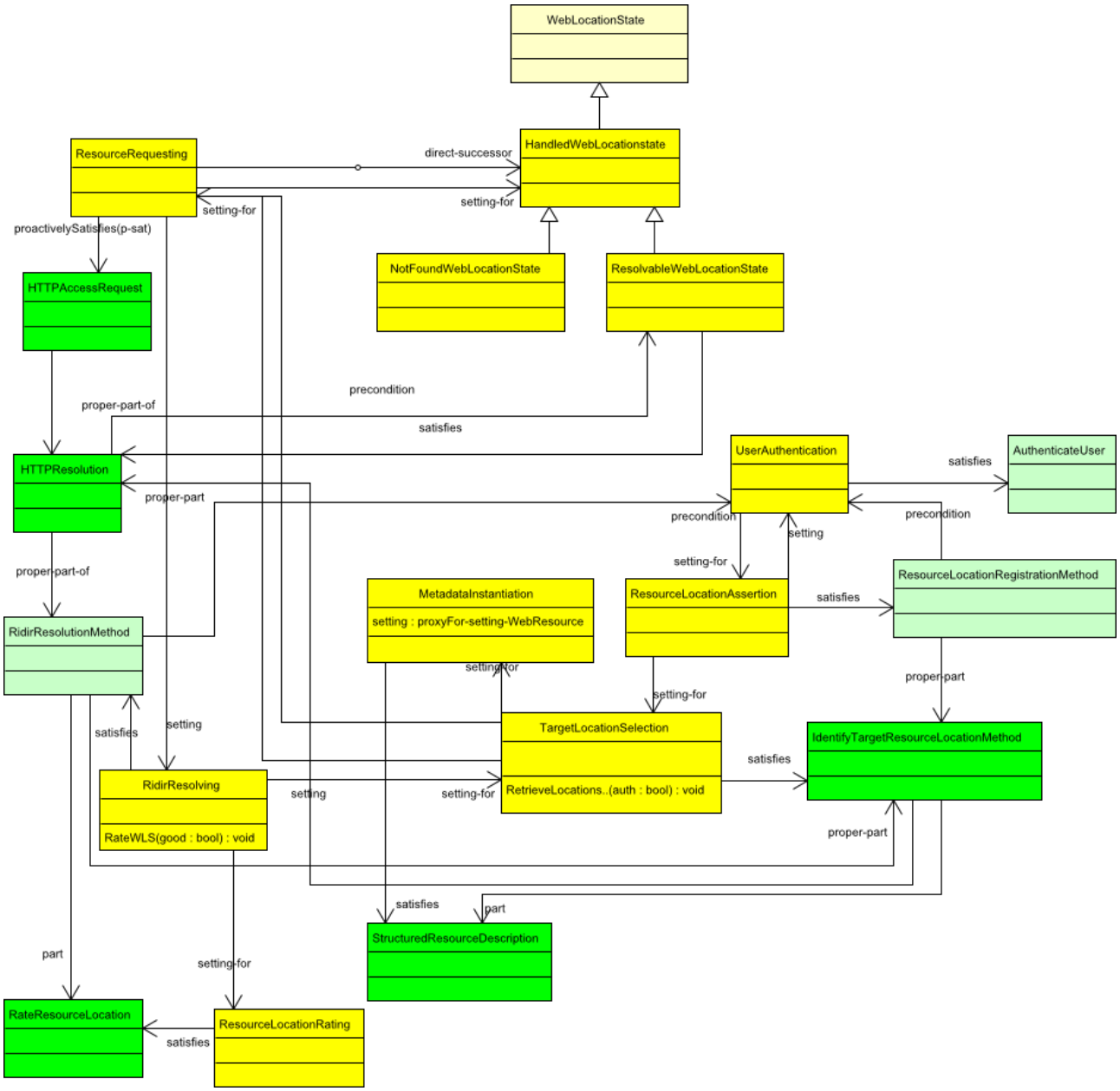

Figure B10: Use Case Descriptions and Situations

Figure B10 depicts some analysis carried out for the demonstrator use cases in terms of the foundational model. Descriptions (in green) essentially correspond to system-level use cases, and the Situations (in yellow) refer to the key contexts that satisfy them. This aspect of modelling is important since composition of the descriptions (which use part-of relationships) is significant: the light green descriptions are 'roots', initial starting points, and the dark-green ones are dependent. This analysis allows a system designer to determine *when* graphs of RDF data require committing to RIDIR (via the RIDIR API). Many Situations have setting-for relations between each other, which essentially means one Situation is only applicable within the context of another; this analysis allows the designer to determine *what* graphs of RDF to commit, since each Situation unifies (bounds or defines), a certain graph of RDF.

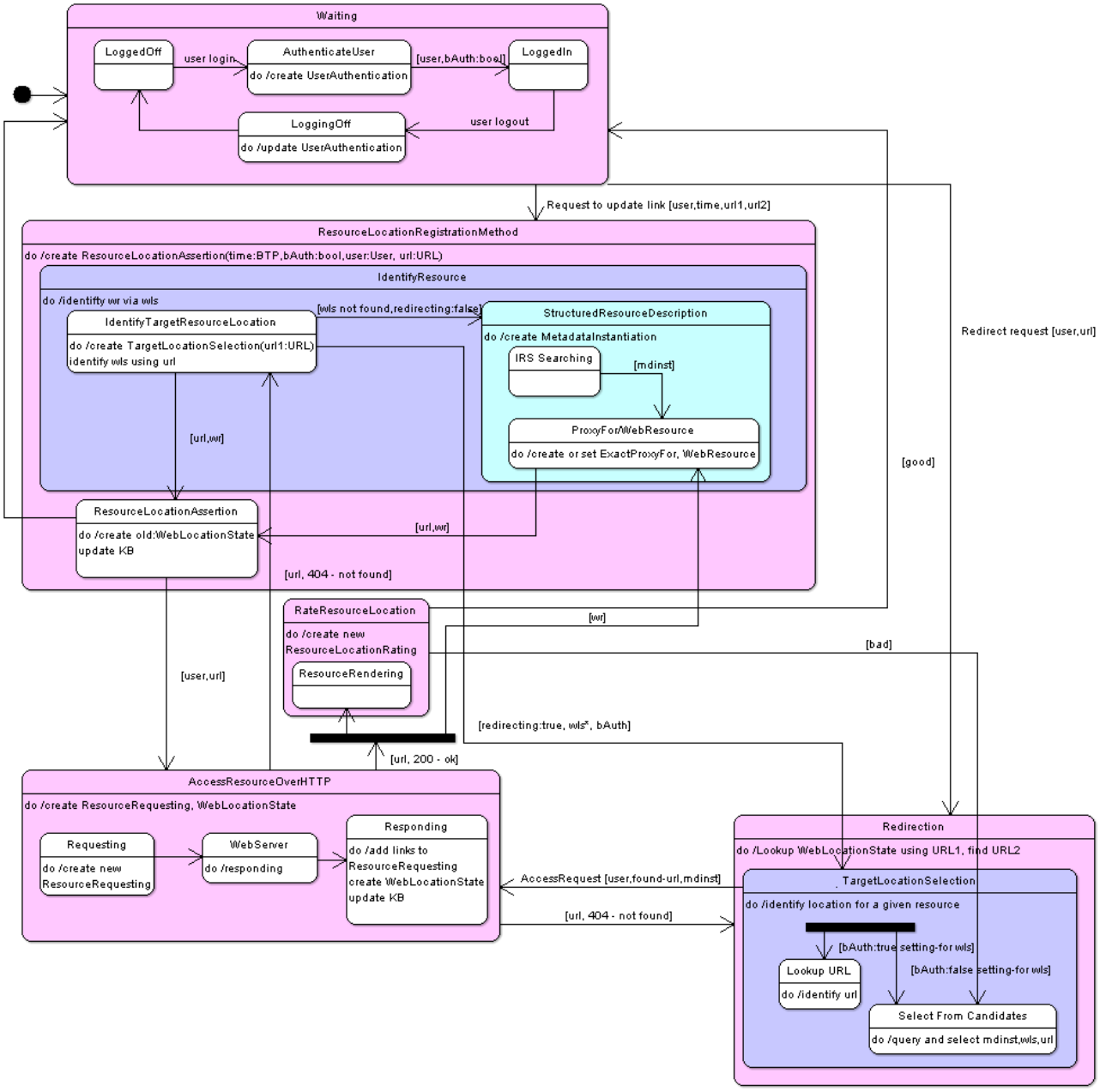

Figure B11: Preliminary Draft for Demonstrator Application State Machine Mapped to Descriptions and Situations

A state machine representation of the Descriptions and Situations involved, shown in Figure B11 was also produced during analysis to help define the key interactions within the application services with respect to the underlying repository system (which stores the RDF)

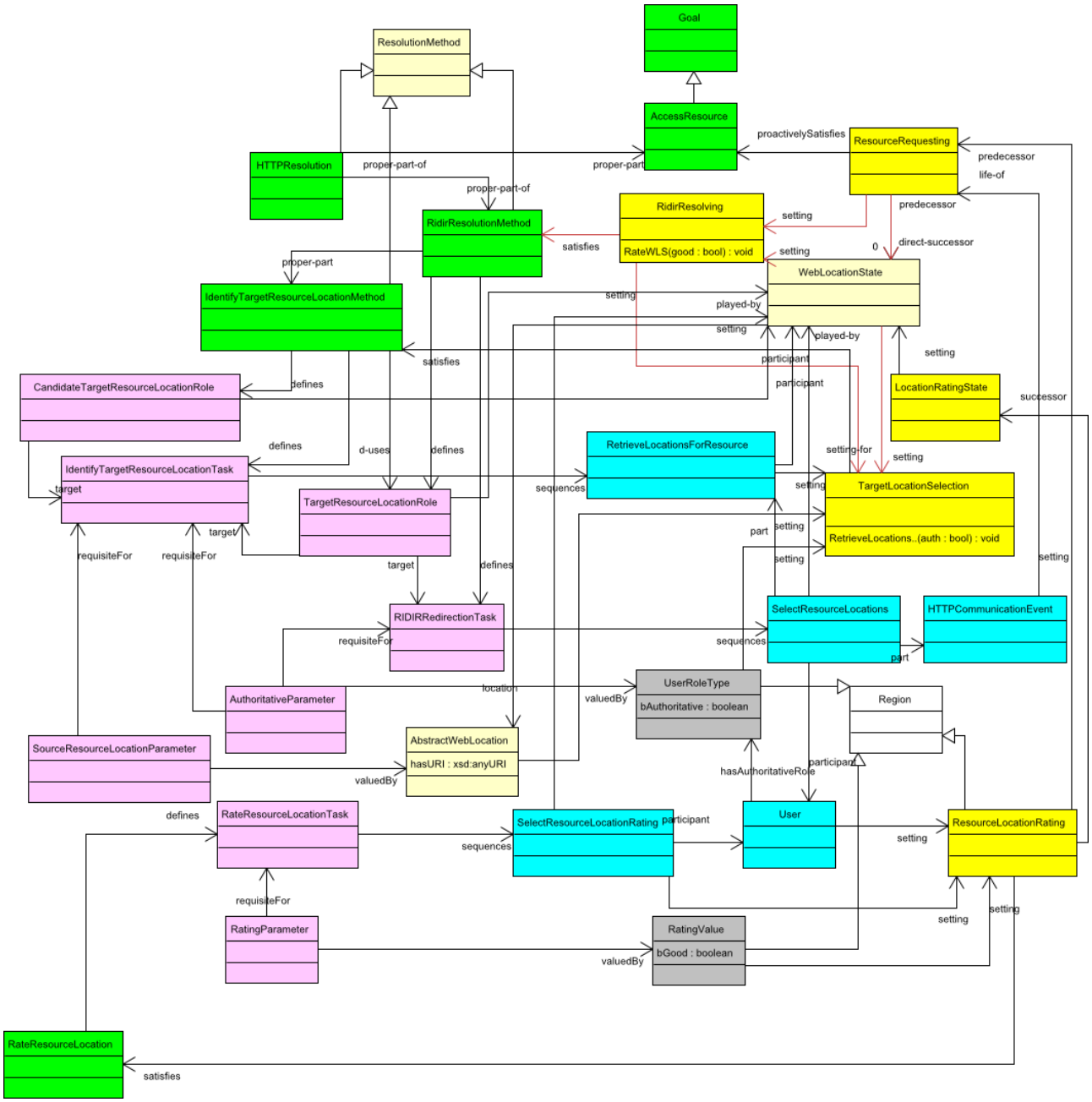

Figure B12: Redirection and Resolution Use Case

The redirection and resolution model specifies the functionality for this part of RIDIR's application services, in terms of the reifications given in the DOLCE-based ontology. A prototype version was developed in OWL-DL corresponding to the model.

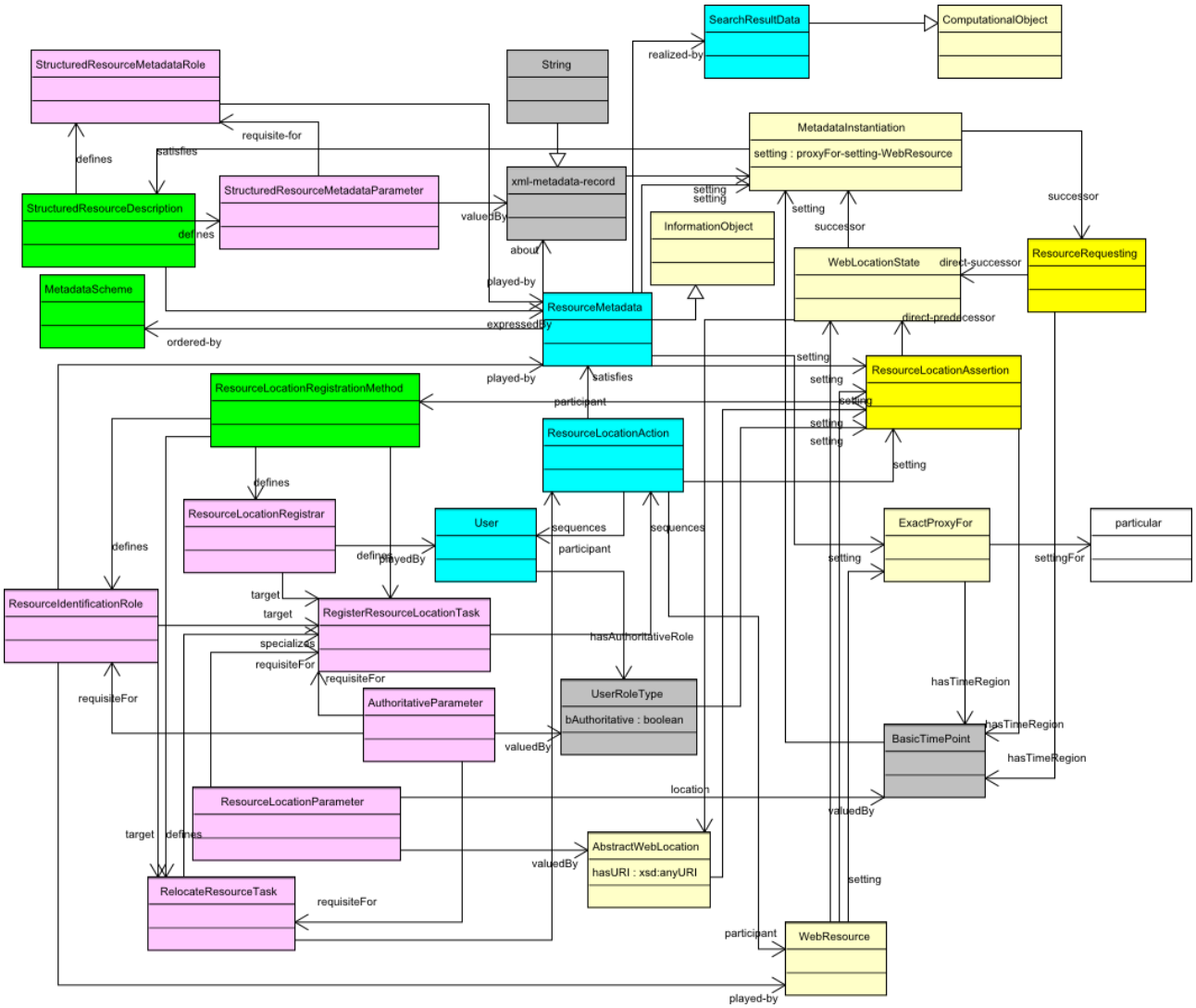

Figure B13: Redirection: Registering new resource locations

The Redirection model specifies the functionality for this part of RIDIR's application services, in semantic terms. A demonstrator-level version was implemented corresponding to the model.

The significant aspect of the overall model depicted in Figure B13 is an analysis of the concept of "metadata", its content, scheme and expression, with respect its referent, the "particular" (entity). This relation is modelled via an ExactProxyFor n-ary relation, which is also the relation between the WebResource and the entity.

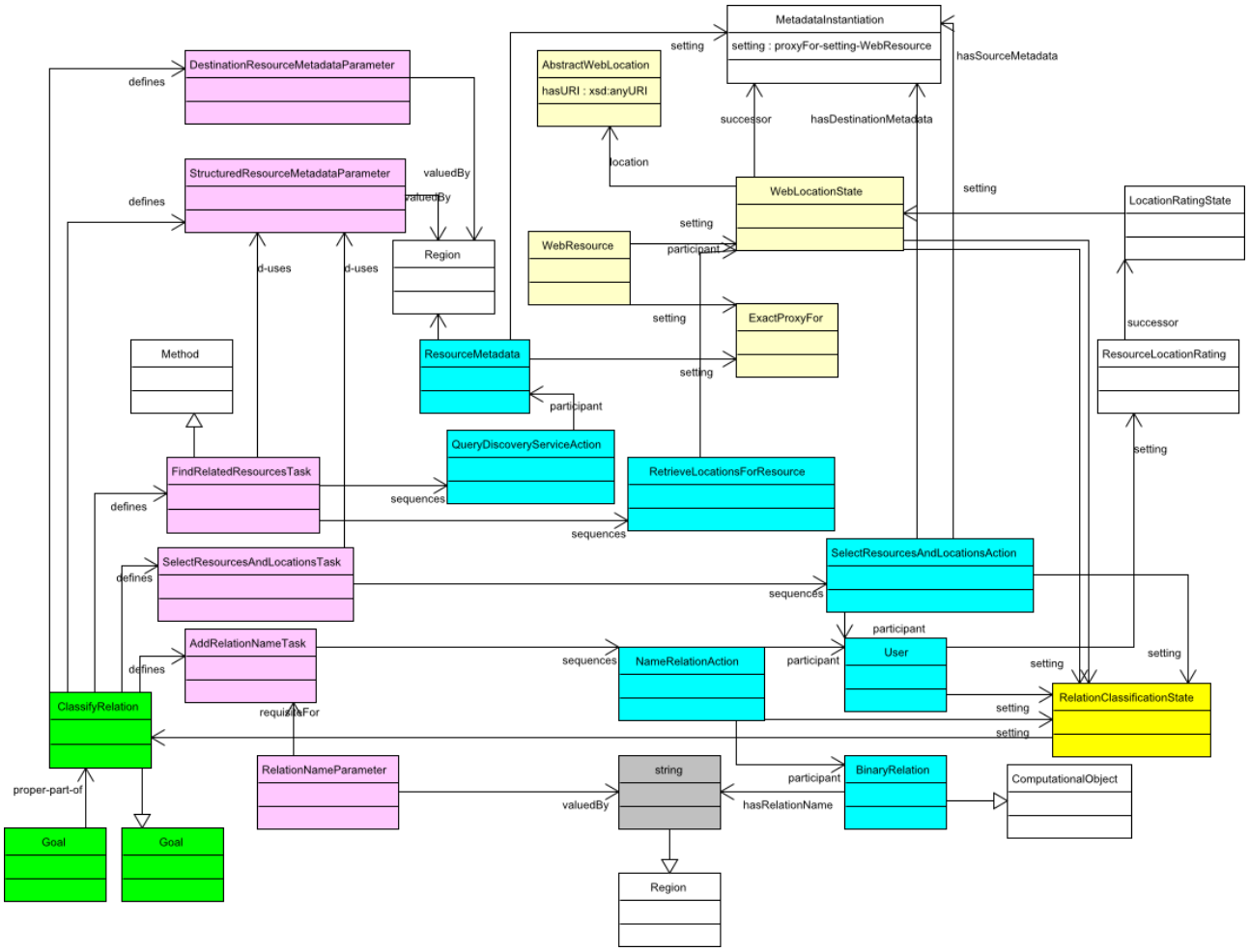

Figure B14: Locate Related Version - Classifying New Relationships

The Locate Related Versions model specifies the functionality for this part of RIDIR's application services, in terms of the reifications given in the DOLCE-based ontology. A prototype version was developed in OWL-DL corresponding to the model.

Summary of Functionality supported

Identifiers change (resource moved): resource cannot be located

The identifiers used are not true persistent identifiers. They may be URLs that are used to indicate the current location of a resource. When objects are moved from one repository to another, the URLs change, as the URL syntax identifies the location of the resource, the system used to serve the resource.

> **Solution**: the abstract web location is identified by the URI (URL), not the (web) resource (or the digital object it is a proxy for, which retains a separate identity within RIDIR.

Resource deleted; identifier refers to non-existent resource

Resources are deleted after identifiers have been published for the resource.

Solution: the web location state explicitly maintains the state of the relationship between the URI and the (web) resource itself. When is notified (via the API) that a digital object has been deleted from a repository, its WebLocationState links are updated to reflect that.

More than one copy of resource, cannot determine appropriate copy

Unable to resolve to the most appropriate copy of the resource for the user accessing the resource. User may not be able to access the resource as a license only allows access to the institution's local copy, which cannot be resolved to.

Solution: the web location state relation allows a uniquely-identified web resource to be accessed by a resolver via more than one abstract web location, each identified by a URI.

Not clear what identifier referent is (resource, splash page, metadata)

Identifiers created that refer to resources, to splash pages and to metadata for resources; no consistent usage of these different identifiers so that it is clear what is being identified in a particular context. Particularly important for machine-machine interactions (eg metadata crawling and discovery, discovery of related versions).

Solution: the facility for a web resource to be a proxyFor something allows for various domain-, application- or even single user- specific views of a referent to be attached as 'types' of digital object or 'resource'. The four layer model of referencing defined by IRE factors out the concepts such that types may be introduced without affecting reusability and interoperability, especially in the case when the 'type' is dependent on potentially differing points of view, or is simply unknown. RIDIR's approach to ensuring that each definition of such a 'type' is 'localised' only to the context of the assertion explicitly allows for multiple differing viewpoints, and giving each user the ability to 'agree' or 'reject' them, allowing circles of trust to be constructed, based upon the metadata of the context (primarily who the asserter was, and when). A 'user' may also be defined at institutional level, so that local institutional standards may be adopted; such delegation is not at present explicitly defined in the model, but is not precluded.

Free-text metadata difficult to disambiguate

Expression of people's names (for example) in free text makes it difficult to identify when one has found the 'right' John Smith.

Solution: the 'semantic' expression of metadata (in RDF) within RIDIR allows for unambiguous identification of individuals; RIDIR supports this through user-created semantic maps (assertions) currently. A future development would be to 'lift' such data using text processing techniques into RDF to provide some degree of automated support for disambiguation requirements.

Relationships between objects not persisted (objects, metadata enrichment)

No mechanisms for persisting relationships between objects once they have been discovered leads to duplication of effort in rediscovering these relationships.

> **Solution**: the use of semantic maps (user assertions) persisted in RDF within the Fedora repository software meets this requirement. Identities are created within RIDIR's semantic maps as necessary through use of the API. RIDIR is able to retrieve information through querying via the API from previously discovered relations, along with the context of each assertion.

Mapping between metadata schemes (Mediation)

Requirements to map between metadata schemes. This could also include usage of metadata (what gets indexed, what gets presented), and the syntax/packaging of the metadata.

> **Solution**: the RIDIR semantic maps approach allows users to draw relationships between objects (within the context of the four layer referencing model). RIDIR retains the contextual information about those objects. Metadata is categorised within the foundational model such that it is linked to objects and acts as a support in terms of presenting relevant human-readable metadata to support the assertion that an identifier identifies a certain object (referent). Metadata is also linked to a schema (description), so that metadata of various schema are linked via the foundational ontology (this is an enhancement over schema cross-walking in the traditional sense since a semantics is preserved across the different schemes).

Mapping/translation of taxonomies, thesauri, controlled vocabularies

Different repositories may use different semantics and different mechanisms for controlled vocabularies, taxonomies and thesauri that need mapping

> **Solution**: the use of semantic maps (user assertions) persisted in RDF within the Fedora repository software meets this requirement, when taken together with the ability to extend the four layer model with the proxyFor relation to different types. This provides the hooks for domain-specific vocabularies to be introduced. The model can be extended in principle via the Descriptions and Situations mechanism to allow authoritative vocabularies to be governed independently (each set of concepts having different lifecycles, but remaining connected via the overall framework).

Mapping between identifier schemes

Mapping between different identifier schemes, including dealing with syntactic restrictions in different schemes, dealing with semantics implicit in the identifier syntax.

> **Solution**: identifier schemes may be mapped together, since the abstract web location is identified by an identifier and the digital object (resource) retains an independent identity. Therefore, many identifiers, with their distinct 'regions' dependent on scheme, and which are further disambiguated through the relationship with the explicit identity of the

scheme's resolution method, may map to the same resource within RIDIR. Syntactic differences are handled by code in 'access modules' in the abstract architecture.

Mapping between different object/content models

Mapping between both the content models implemented in different repositories and the content models implicit from the repository software and different way the repository software chooses to model digital objects.

Solution: the foundation model covers within its scope only the fundamental issues of identification and reference. Repository content models are significant in terms of the types specified as ranges of the proxyFor relation. The model therefore allows mappings to be made and persisted between such types without affecting the underlying model.

Mapping/translation between object packaging and ingest schemes

Mapping to and from schemes for packaging and describing objects ready for ingest.

Solution: the handling of differences in digital object packaging is supported by the metadata and the networks of relationships between objects defined by the model held within RIDIR. For instance, a package may be produced linking various objects in a hierarchy and attaching various metadata based on information held within RIDIR. A future development could be to provide an OAI-PMH harvesting facility and/or OAI-ORE resource maps, to provide automated support.

Mapping/translation between different ownership and security models

Mapping between different repositories' models for handling object ownership and between repository-specific security model implementations.

Solution: the retention of user assertion metadata helps support this requirement. Mechanisms for resolution of user identity are outside the scope of RIDIR, so integration with such schemes as specified by the JISC (such as Shibboleth) would be necessary to fulfil such a requirement.

Need to handle complex objects and collections

Ability is needed to deal with part/whole relationships and collections.

Solution: Axioms held within DOLCE-Lite supports this requirement; domain- or application-specific definitions may be introduced with minimal impact through the separation of concepts of 'social agreement' within the Descriptions-and-Situations module.

Location of appropriate repositories, where to search

A list or registry of repositories with information on what resources are contained in each and details of how to access the repositories is required

Solution: out of scope of the foundational ontology model.

Information on assertion of relationships is required

It is necessary to know who claimed that a particular relationship between objects or metdata items is present to make an assessment of the authority and/or veracity of the relationship for other users.

Solution: Supported as part of the foundational ontology

Mapping/translation between different versioning schemes

Mapping between different schemes of representing versions is required.

Solution: Versioning can be handled through the separation of Entity, InformationRealization and InformationObject within the foundational ontology. A development of the proxyFor relation specific to different kinds of 'version' suggested by the VIF project would be a useful way of further refining and supporting versioning schemes compatible with the RIDIR approach.

Reintegration issues

Joining up with other services, eg integration with persistent identifier infrastructure, integration with harvesting services

Solution: the foundation model would require further investigation against specific infrastructures which are not those considered already (web/HTTP and handle). However, it is hoped that the four layer reference separation should allow for adequate hooks. Harvesting would be achieved via introduction of accessor services (modules in abstract architecture terms) using the RIDIR API.

Implicit metadata that needs making explicit

There is implicit information about objects in a repository that is not explicitly stated in metadata; for instance migrating a repository known to contain MPEG-2 clips to a general multimedia repository; the MPEG-2 repository does not explicitly state that its contents are MPEG-2; all of the users of the repository are aware that the repository is there to hold MPEG-2 objects.

Solution: out of scope of the foundational ontology model; lifting can be performed using techniques out of scope of RIDIR to present non-RDF 'semantic' data sources such as MPEG-2 as RDF.

Appendix C: RIDIR as part of a national service

We believe that the RIDIR project has identified that there is a requirement for:

- a persistent identifier resolution service: allowing resources to be identified by a single, universally-unique identifier that would resolve through a resolution service for the lifetime of the resource. We believe such a service would aid in the EThOSnet, Depot (Lost Resource Finder) and Migrate Repository use cases (and to a lesser degree in the other use cases)

- an identified resource linking service: allowing relationships between resources to be asserted and persisted. We believe such a service would aid in the Depot, Locate Related and Spoken Word use cases

The PILIN project addresses these same needs.

PILIN focuses on the building of an identifier management infrastructure based on the Handle technology. There are two particular areas of interest to the RIDIR project and its recommendations.

Not including aspects of the PILIN ontology[65] (understood to be under review at the time of writing), these are:

- The PILIN FRBR Tool, which allows the precise semantic description of what is being identified, and precise semantic description of relationships between referents, using the FRBR model[66]

- The Persistent Citation Resolver Service, which provides a service whereby a URL identifier that no longer resolves may be related back to the Handle for the resource, which can then be used to provide the new location of the resource[67]

A key aspect of the RIDIR demonstrator implementation is the **promotion of relationships between referents of identifiers to first-class identified things in their own right**. Within the RIDIR demonstrator, identifiers (PIDs) are allocated to relationships, and this allows semantic descriptions to be 'attached' directly to these relationships, for instance being able to state who asserted the relationship, when they asserted it, how many people agree with the assertion, the authority under which the assertion was made and so forth.

The benefits of this are, we believe, clear in the demonstrator:

- In the Lost Resource Finder application the user is able to differentiate between 'authoritative' new locations for resources (asserted, for instance, by a repository manager) and 'candidate' new locations for resources that have been suggested by other users of the system. When the user is presented with 'candidate' new locations for resources they are also presented with a 'confidence' rating based on how many other

[65] See: https://www.pilin.net.au/Project_Documents/PILIN_Ontology/Ontology.htm
[66] See: https://www.pilin.net.au/PILIN_Implementations/About_FRBR.htm
[67] See: https://www.pilin.net.au/PILIN_Implementations/Reverse_Lookup_Service.htm

people agreed or disagreed that the new location is in fact the correct location, the objective being that this information will help them in determining which is most appropriate.

- In the 'Locate Related' application, the user is able to see who proposed the relationship between two resources and when they proposed it, again the objective was to provide additional useful information in guiding the user through a chain of relationships. (This principle could be further enhanced, for instance allowing people to say why they created a relationship and so forth).

The RIDIR project believes that this promotion of relationships between resources to first class entities should form a key requirement when determining what should be implemented as a national identification framework service.

In conclusion, we believe that Handle and the outputs of the PILIN project should be evaluated as the basis of a national identifier framework in conjunction with the recommendations of the RIDIR project.

Particular attention should be paid to

- precise semantic identification of what is being identified
- promotion of relationships between resources to first-class identified entities
- the adoption of an appropriate ontology for the classification of identified things and their relationships

The Handle service allows the registration of Handle types, which is the mechanism through which PILIN uses the FRBR model in identifying the types of and relationships between resources in its FRBR tool. This would seem like a fruitful area for further investigation in conjunction with the RIDIR project outputs.

We have presented a candidate ontology for the RIDIR project which could in practice be implemented by attaching (for instance) fragments of RDF to identifiers. Further work is needed on the evaluation of requirements for this ontology, but we would anticipate that the registration of appropriate new Handle types could provide a mechanism for the implementation of some version of this ontology.