THE UNIVERSITY OF HULL

and

rightscom

RIDIR Project

RIDIR Focus Groups report

Richard Green and Hugh Look

September 2007

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@rightscom.com)

The RIDIR Project is being 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.

Index

Introduction

The RIDIR Project was developed in response to a successful tender to the Joint Information Systems Committee (JISC) to build a 'Persistent Identifier Interoperability Demonstrator' (JISC Circular 04/06):

> *G52 A project is sought to build a persistent identifier interoperability demonstrator across the variety of identifier schemes.*
>
> *G53 Considerable amounts of work have been done on standardising identification schemes for a particular community or sector. Equally, much work has been done on creating standard or reference metadata sets that can be used to associate key metadata descriptors with content. Much less work has been done on the impact of cross-sector (by sector we mean a particular media sector's) working. Relatively little is understood about the effect of using one industry's identifiers in another, or on attempting to import metadata from one identification scheme into a system based on another. In the long term it is clear that interoperability of all these media identifiers and metadata schemes will be required. A variety of identifier schemes are adopted across digital repositories in order for a network of repositories to flourish, interoperability across schemes will be required. One project, perhaps working in partnership across a consortium to build a demonstrator, will be funded. The project will need to create use cases and to build an interoperability demonstrator in order to test interoperability across different identifier systems, eg Handle, DOI etc. Bidders should take note of the Digital Object Identifier for Publishing and the e-Leaning Community Report.[1] This area of work is likely to feed into the application profiles that will be developed outside of this call for funding and work in collaboration with the proposed Intute Repositories Search Project[2] as persistent identifiers may be an aspect of the search criteria.*

In planning the work of the project, it was decided that RIDIR development should be informed by holding two focus group meetings with practitioners in the field. The first meeting would consult with repository managers to discover their needs in terms of identifiers and interoperability; these needs would then form the basis of use cases to be addressed by the project. The second meeting would be held with invited experts in the field to ratify and possibly expand on the findings of the first.

For the most part, this document will not ascribe views or comments to individuals; rather it will look at the generalities of the discussions that took place.

The first focus group meeting

In due course the RIDIR Project team arranged to hold the first meeting in London on 12th June, 2007 and the second in Manchester on 28th June 2007. A professional recruiter was engaged to find willing participants for the first meeting and together the team developed a list of invitees for the second. It was intended that members of the first gro up should be relatively new to repositories in the hope that they might express their own fresh views rather than perhaps simply echo more widely held 'establishment' ones.

In the event it proved almost impossible to recruit for this first meeting. Feedback from potential participants suggested that the project might be exploring areas that were not yet perceived as problems by many repository managers. Some of the invitees noted that they generated identifiers internally and had not yet considered interoperability. In itself, this was a preliminary finding.

---

[1] See www.jisc.ac.uk/index.cfm?name=project_tso

[2] Intute Repository Search Project (Interim URL) http://irs.ukoln.ac.uk

Therefore, shortly before the date of the first meeting the team decided to cancel the event as planned and instead invite a group of established repository practitioners. By a coincidence of the invitation process, it turned out that all those who attended used Fedora repository software; this fact should be noted although it is not thought unduly to have influenced the outcomes of the meeting.

Thus the group who attended the first meeting were:

Julie Allinson (UKOLN: SWORD Project and JISC-CETIS Repositories Research Team)
David Flanders (Birkbeck College, London and Bloomsbury Colleges Consortium, SOURCE Project)
Mark Hedges (Arts and Humanities Data Service)
Matthias Razum (FIZ Karlsruhe, eSciDoc Project)
Susan Thomas (University of Oxford Archives, Paradigm & Cairo Projects)

Richard Green (RIDIR Project Manager for University of Hull)
Hugh Look (RIDIR co-ordinator for Rightscom, Facilitator)

The early part of the meeting was guided by a PowerPoint presentation developed by the RIDIR team (this forms Appendix 1). The facilitator introduced the meeting by noting that the RIDIR team hoped the meeting might flush out some use cases to work with, particularly ones that were not necessarily the most obvious (the RIDIR team could probably come up with the obvious ones); and also the more challenging ones that participants might see as becoming issues in the future and that the project ought to be demonstrating if possible. The group then talked through an overview of the scope and aims of the RIDIR Project. At that stage this it was anticipated that half the meeting might be spent discussing general issues and that the second half might be used to look at possible use cases. Slide #3 was used to point up the objectives for the workshop.

As part of the process of introducing themselves, the participants were invited to give a set of initial thoughts or questions. The resulting list was:

- authors need to refer to a text and know that the identifier will still find it possibly centuries down the road
- need to be able to map objects via identifiers to variations created - perhaps for preservation
- need to deal with versioning of objects; there must be a clarity in which version an identifier refers to
- need to be able to identify component parts of compound or complex object (it was noted that these terms do not have generally accepted definitions in the repository community; here, and in the rest of this document, it was intended to refer to the atomistic approach). Also need to know what complex object or objects a component has been used in.
- need to be able to discover how parts of a compound object may have been re-used
- provenance: who owns the identifier, who maintains it, what level of trust does it (?) imply?
- do identifiers just identify or should they support some level of semantics? Should the structure of an identifier tell us something about the object?
- harvesting - what does an identifier identify? Splash page, bitstream, metadata?

- is an identifier resolvable? Does it just identify an object or should it resolve, for instance, to a page?
- how do you identify (for instance) people - does the identifier identify a warm body, a photograph, a biography, …?
- how do you cope with identifiers in the context of legitimate (simultaneous?) deposit (say an institutional repository, a Virtual Learning Environment (VLE), Jorum)?
- how do you identify software tools that may be needed to render content?
- PREMIS data dictionary[3] may be relevant to RIDIR's work
- citation - how does this work with multiple copies and/or manifestations? Provenance?
- do Handles schemes[4] offer enough granularity? (Files, folders, collections, events, agents)
- what happens when objects come with pre-existing identifiers? How are they preserved/maintained? How do you manage/administer multiple identifiers?
- 'persistent' (identifier) implies continuity and management
- there are multiple identifier schemes some of which perhaps offer unnecessary functionality which may be better dealt with in other ways
- an object can have simultaneous, multiple identifiers using different schemes
- need identifiers for aggregation objects (containers, bundles etc)
- need identifiers for different versions of an object's binary content. We need easily to locate latest version, or a particular version in time that may have been cited
- we need identifiers for dynamic collections; not all content is static
- should identifiers allow semantics or not; these could, for instance, allow users to identify a specific version of an object?
- how do you cope with a repository which depends on its content being rendered at the time of retrieval: how do you identify different renderings of the same content? Should each software rendering have its own identifier as well as an identifier for the un-rendered content?
- the notion of 'asset actions' may be relevant. Participating repositories would all support a common list of 'asset actions' verbs which, when appended to an identifier, would return something appropriate. For instance: ident:23456?getDCMetadata
- what if an object depends on a web service which is not local? What if this remote web service is down? Discontinued? Do we need to identify a version of the service in order to ensure always the same response when it is used.
- do we need a hierarchy of identifiers: (Digital Library ->) Repository -> Collection -> object -> datastream (-> TEI (eg) level -> bitstream level) This is a granularity of identification. How far do you go? Do we also need identifiers for web services? How stable would these identifiers be (repository maybe not. What happens when you

[3] See http://www.oclc.org/research/projects/pmwg/premis-final.pdf
[4] See http://www.handle.net/

migrate repository software? New repository ID? New identifiers? …?) There is a paper by Tansley about moving objects between repositories without changing identifiers.[5]

- what are the arguments for 'structural' identifiers as opposed to using identifiers and semantics. Semantics on identifiers could become too rich (ie complex) to be useful? Should identifiers have any meaning at all - even if not, should they be human-friendly so that they are easily quoted?

- how many identifiers does the world need? How complex will they need to be? Identifiers need to be able to grow exponentially so that (eg) individual characters in a document could be cited; (parts of?) annotations to the document could be cited. Machines won't care about the complexity of an identifier and there are ways to represent a sub-set of them more simply for a human (tiny URLs, for example).

At this stage the facilitator outlined the first thoughts of the RIDIR team for realising a demonstrator (Appendix 1, slide #5), noting that these had been early thoughts at the time of the proposal and may need to be changed.

It was noted that the phrase on the slide "each repository will be realised in a different technology" should probably better have read "each repository could be realised..."

There was discussion of what the Persistent Identifier Mediator Service (PIMS) might do and it was noted that whilst the RIDIR team had ideas in this regard, there was no fixed technological specification yet.

It was suggested that 'persistent identifiers' needed to be 'persistent' and a user should not need to know that an object may have been migrated from one place to another - the persistent identifier should still work. There was then a potentially conflicting argument that if an object had been migrated it wasn't really the same object anyway... The work of the Repository Bridge Project was cited as being of potential interest in this area.[6]

There was discussion about the use of a single identifier being used for several instances of the same content, perhaps in different locations. The example of an ISBN was cited. However, it was noted that an ISBN identified a class of books with a particular content, not individual copies. Three books with the same ISBN might be different in terms of condition, a dedication on the title page, marginal notes... A particular text might exist in a number of forms each with different ISBNs; behind the scenes these ISBNs might lead to different formats, different licensing arrangements etc. The ISBN does not provide this information, it is not encoded into it, but acts as a key to other back-office documents or systems. In principle it could be used within a set of relationships to bring all this information together.

The discussion then moved into the area of stewardship. It was suggested that, in some way, an identifier ought to expose the stewardship of an object. If an object were placed into a 'Hull' repository, it would gain a Hull identifier. Migrating it later to another Hull repository instance would not necessarily demand a new identifier because it was still in the same stewardship, however copying it to (eg) Jorum would require a new identifier, to indicate the stewardship of the copy, and ideally some sort of record of the objects' inter-relationship should be created.

It was suggested that care should be taken to distinguish between a system identifier and something potentially over-arching - DOI, Handle etc. This led to the expression of a concern about the use of commercial systems in the context of identifiers because the continuance of these systems is subject to the whims of their shareholders.

There was further discussion of 'what an identifier identifies'. Does one need an identifier for the slightly abstract notion of a particular paper in which "Professor X says", through which all

[5] Tansley, Robert (2006) Building a Distributed, Standards-based Repository Federation: The China Digital Museum Project Available at: http://dlib.org/dlib/july06/tansley/07tansley.html
[6] http://www.inf.aber.ac.uk/bridge/

formats and versions might be exposed, as well as specific identifiers for (say) the pdf and Word versions of it (which may well be part of the same digital object)? The nature of Humanities research suggests that it must be possible to refer to the detail of a particular representation. What is the 'promise' of an identifier to a user? There was further discussion of whether an identifier should potentially be used to navigate to a word in a text, a part of an image, a set of coordinates on a map - or whether the identifier should work at a higher level and some other system be used 'within' the material (page/line numbers, measurements, coordinates...). This led to a brief discussion of how the position of an annotation to a non-text object might be handled and whether the system needed to be machine processable.

It was noted that much of this discussion was well beyond the scope of what RIDIR might hope to tackle, but that it did point up the need for the project to be very clear about the limitations and constraints that it decided to work within, so as not to shut down linkages to wider work. These constraints will also need to be made clear in the project documentation.

Following a short break, the facilitator then went on to offer suggestions for the rest of the afternoon's discussion (slides #6, #7 and #8). He identified a number of areas that the project team would like to cover in arriving at use cases:

- Creation or selection of identifiers; how does an identifier actually get there?
- Validation of identifiers
- Linking to a resource in another repository
- Copying a resource from one repository to another
- Preservation
- Compound/complex objects

Initial discussion centred on the use case of someone discovering in one repository a reference to material in another.

Would the reference be a citation or a hyperlink direct to (say) a pdf file. One might imagine that a pdf file would render reasonably uniformly in different software agents, is the same true if the target file is (say) xml? If the link is effectively a citation, how does one handle that? Have identifiers a particular role to play in this process?

If the hyperlink is to a 'closed' publishers pdf, can the identifier help you locate an 'open' author's final copy? Likely the identifiers would be quite different; can they somehow be tied together? Google Scholar may return a number of links in response to a query about a particular paper: do we need an identifier for the abstract notion of a particular work which may then lead a user to a number of instances of it? This is straying into FRBR territory.[7] It was suggested that RIDIR may only have the resources to deal with instances where objects had identical bitstreams whilst recognising that relationships needed to be established between objects with 'essentially the same bitstream'. Some in the group felt that this would be to ignore what was the major issue in this area.

The question was raised: what does 'interoperability of persistent identifiers' actually mean? Is there a sensible use case? It was noted that interoperability would require unique persistent identifiers. (There are apparently cases of users installing EPrints software and not changing the root of the identifier so that the world has multiple, different instances of objects labelled "InsertRepositoryHere/2345" or something equally unhelpful.)

It was suggested that a real interoperability problem may be along the lines that an identical paper (at bitstream level) is simultaneously deposited in more than one repository. A particular user may only have rights to access to one of them and needs to know that the identical content resides in a number of places. It was pointed out that whilst the bitstream content may be the same, the different repositories may add value to the object as a whole in different ways so that there is a sense in which the different instances are not identical. Should the metadata in each of these objects identify the 'siblings' in other repositories, or at least point to the authoritative 'parent' object? How much is this actually a metadata issue rather

[7] Functional Requirements for Bibliographic Records See: www.ifla.org/VII/s13/frbr/frbr.pdf

than an identifier interoperability issue? Identifiers should at the least allow users to distinguish reliably between the different instances and perhaps there needs to be some form of mapping system that can assert potential relationships between objects in disparate repositories. There would be an issue that a relationship "is a copy of" would only hold true at the point the copy was made; the copy might then be changed significantly - how would the relationship system know? Ideally, metadata would record previous identifiers associated with an object.

The process of forging relationships as objects are copied may be complex where there are third parties involved. The simplest instance of it is probably migrating repository content from one system to another. Relationships need to be made, and perhaps exposed, relating 'old' identifiers to 'new'; there may also be intermediate, internal system identifiers that should be recorded. However, if by 'persistent' identifier we mean an identifier associated with a resolution service, for instance a Handle, it can be argued that the world will see no change because the persistent identifier will simply be pointed at the migrated resource. In computer terms one could argue that this is a 'move' rather than a 'copy' and that in the latter case a new persistent identifier would be needed.

Should there be a service which can be called during a migration which 'examines' an object for its current identifier and any past identifier history that may be present in its metadata and which then passes back appropriate metadata for the 'new' object containing an updated history? Could this service also update relationships that may have been declared around the old object (either locally or more widely)? Arguably a well managed repository system would do this anyway for local relationships.

There followed discussion of what is a persistent identifier, to which the answer seemed to be an identifier that is persistent! It is up to individual repositories to determine how to do this. Some may consider that (eg) a 'Handles' route is appropriate. Some may consider than an identifier based on their domain name is a better alternative (no-one could really see the University of Oxford being taken over or going out of business!). In terms of interoperability, processes perhaps do need to exist to aid recovery when a domain name (for example) thought to be stable *does* get changed. The impact of this on metadata held elsewhere could be considerable. The idea was floated of an identifier registry which held in some form of metadata the history of an identifier and its relationships to other identifiers, but it was felt that this is really part of provenance metadata and should be held with an object.

The discussion was then brought back to the issue of interoperability and the need to know what an identifier identifies: is it a splash page for the object, a metadata stream for the object, a bitstream for the object, ...? And how would a second repository know? Clearly, one answer is to give each a persistent identifier. Should those identifiers be predictable derivatives of each other? Perhaps the object is 1234, its metadata might be 1234m, its pdf datastream 1234p. If there were an agreed mechanism between 'subscribing' repositories, might this aid interoperability? Clearly to try and encourage any such system has an element of trying to 'bolt the stable door after the horse has gone', but if there is a serious need and if a number of respected organisations backed it (including the JISC) might it be possible. This is close to the idea of 'asset actions' from the DLF Aquifer Project.[8]

The asset actions approach proposed the addition of verbs to an object identifier - thus for an image object with an identifier 'institution:1234', institution:1234/getDCRecord would reliably get the DC information, institution:1234/getScreenSize might retrieve a 800x600 pixel representation of the image etc. How the verbs were implemented within a particular repository would be a local decision but the syntax and the form of response would be common. There would be a core of actions which all repositories would implement, one of these calls might be (say) 'getActions' which would return a list of all possible calls which could be successfully made on the object.

Such an approach might, for instance, allow copying of a repository object without needing to know, ahead of the event, the internal structure of that object. A query could be made on the

[8] See http://www.diglib.org/aquifer/

object to determine the range of calls which could be made to it, which taken together might provide a reasonable representation of the object at the second site. (There is, of course, scope for implementing a 'getCopy' routine.) The approach does not, in the abstract, require a particular identifier scheme or a particular form of repository software, but in its discussions the group was unable to answer the question of whether persistent identifier systems such as Handles would pass a suffix (the action part of the call) through to the resolved address.

It was noted that even if it were possible to get widespread adoption of such an approach within the education sector, there was no guarantee that other repository sectors could be persuaded to adopt it. Whilst the group could see interesting possibilities in an 'asset actions' approach it could also see a lot of implementation problems in the detail. Should the RIDIR project explore this approach, the demonstrator would very much be about showing the possibilities, not a full implementation.

The meeting wound up at this point. It was noted that the meeting had not explicitly proposed a set of use cases but that a number were sketched out during its discussions which the RIDIR project team could abstract.

Preparation for the second focus group

As noted elsewhere, it had been the Project Team's intention to take a set of use cases developed at the first focus group meeting to the meeting of 'experts' in Manchester. In view of the somewhat indeterminate outcomes of the first event it was decided to collect a range of use cases from outside the RIDIR Project which were deemed to have some appropriate elements and/or which had been outlined during part of the first meeting.

In the event, use cases were selected from an article by Norman Paskin, *Identifier Interoperability: A Report on Two Recent ISO Activities* [9] published in 2006 and cited in the JISC supporting materials for the funding call. This document drew, with acknowledgement, on work by Mark Bide who is a member of the RIDIR Project team. In addition some use cases were selected from those amassed by the PILIN Project in Australia[10] and one of the members of the first focus group provided another. These use cases, and a consideration of the life cycle of a digital object, were circulated to members of the second focus group prior to the meeting. The document is reproduced at Appendix 2.

The second focus group meeting

The second focus group was held in Manchester on 28th June. Present were

Julie Allinson (UKOLN: SWORD Project and JISC-CETIS Repositories Research Team)
Monica Duke (Software Developer - Repositories and aggregation, UKOLN)
Gordon Dunsire (Deputy Director, Centre for Digital Library Research)
Richard Jones (Repository Developer, Imperial College, London)
Jim Rutherford (Researcher, Hewlett Packard Laboratories, DSpace committer)
Rob Sanderson (University of Liverpool)
Frances Shipsey (Repository Manager, London School of Economics and Political Science)

Chris Awre (Domain specialist, University of Hull)
Mark Bide (Project Identifier Consultant, Rightscom)
Martin Dow (Project Software Architect, Rightscom)
Richard Green (RIDIR Project Manager for University of Hull)
Hugh Look (RIDIR Co-ordinator for Rightscom, Facilitator)

[9] http://www.dlib.org/dlib/april06/paskin/04paskin.html
[10] http://www.arrow.edu.au/PILIN

The meeting was facilitated by Hugh Look using a PowerPoint presentation (reproduced as Appendix 3) for focus.

Whilst introducing themselves and their interests, the group identified the following initial thoughts and questions:

- need to understand how identifiers can be most effectively used in the context of repositories
- definitions, vocabularies and terms are a minefield of ambiguity in current papers
- what is an object? How far do you take granularity, especially in the context of data sets?
- what bits of a 'complex' object need an identifier?
- should an identifier be simply that, or should it be able to carry semantics?
- what does an identifier identify? Just objects, or also concepts, people, ...?
- what is identifier interoperability?
- how do identifiers work when applied to 'complex' objects?
- what does an identifier identify? If a person, can it be resolved? What to?
- should an identifier be 'opaque' or can it be allowed a visible structure?

Following introductions, the meeting turned to a scoping discussion. The first question posed was "what might we mean by identifier interoperability, and why might it be a good thing?"

As a starting point, it was suggested that it would be useful to know when a pair of identifiers identified the same thing - perhaps through some supporting structure: this could be termed interoperability. Or we might be talking about identifiers for components of a 'thing' which has a higher order identifier: interoperation between the two orders of identifier. These might both be acceptable definitions but they are not the same thing at all and need to be discussed separately.

It was suggested that trying to hold any sort of 'tree' of identifier relationships was unrealistic, and that such relationships were better dealt with in a graph. Producing and/or maintaining a graph would clearly use identifiers but the services would lie outside the identifier systems.

Identifier interoperability is maybe therefore about processes through which systems communicate with each other about identifiers (in metadata) that they know about. We need to clearly distinguish an identifier, which merely identifies something else, an identifier as a metadata element, metadata records which can have all sorts of structural information in them, and the things that the metadata itself identifies.

If an identifier is a semantic-free label, it may be necessary for systems to exchange it alongside some contextual information about where it came from.

If you have a system that can assert identifier 'A' and identifier 'B' identify the "same thing", then that relationship probably needs an identifier, 'C'. Once you have 'C' you don't need 'A' and 'B'? This begs the question of what "same thing" might mean in this context.

Interoperability requires a common understanding of the *basis* for the interoperability. For instance OAI relies on knowing what it is you will get. Such a system needs a framework within which institutions, repositories, systems, can work. Anything that RIDIR will do,

therefore, needs to be clearly defined so that the context and associated set of constraints are clear. You need to be very clear about what you want to *do* by making the identifiers interoperable.

It was asserted that there are three 'parts' to this forming a triangle. An identifier, metadata and data. The identifier can point to the metadata or data, the metadata is about the data. In general interoperability is achieved with the metadata.

It was suggested that the problem needed to be approached from the object 'end' rather than at the identifier end. If you have an object and give it an identifier you know if you see the identifier in another context what it refers to. If you are given simply an identifier you have to seek out someone (?) who can tell you what it relates to and how then actually to locate that object.

In considering where RIDIR might look to find additional expertise in this area, the following were identified:

- the ORE Project[11]
- Life Science Identifiers[12]
- InChis[13]
- semantic web registries - identifiers for metadata rather than object identifiers

The discussion then turned to the slides on 'life stages' (Appendix 3, slides #7-#12).

It was suggested that a unique identifier almost necessarily carried some notion of 'authority' by indicating the repository in which an object originates. This might be achieved by explicit semantics or by implicit; in the second case some form of lookup might be required.

It was also suggested that when an object is created two identifiers should be generated: an identifier which is an 'electronic location' for the object, and an identifier for the metadata record which describes the object. Further, it was suggested that the only way into an object should be through the metadata so that only the metadata needs to understand any semantic of the object identifier. This approach avoids the 'public' use of a repository-issued identifier for an object which potentially could give false provenance information (say if an object has been transferred from one repository to another but somehow retaining an original identifier). We should move away from the idea that an object identifier itself contains any useful semantics. The metadata would provide the context for the object, not the object identifier. There was further discussion of this model in the context of search engines, for instance Google.

It was agreed that 'destruction' of objects was a big issue. A persistent identifier should clearly point to its object; if that object is destroyed the identifier should then point at some form of record that explains, for instance, when, why and by whom. This discussion moved into what might happen if the relationship between a metadata record and its associated object became broken or irresolvable.

It was suggested that if we try to enforce a necessary separation between the metadata and the object then we create problems that don't perhaps exist. Instead what we need to do is to understand that there is such a thing as metadata that is actually practical and useful that is in some sense part of a larger object that has a content package as well. This comes back to the ORE attempt to structure an object - where some of the representations of an object are metadata representations. The identification framework has to be able to deal with all of the contexts in which you might expose the metadata.

[11] http://www.openarchives.org/ore/
[12] http://lsids.sourceforge.net/
[13] http://www.iupac.org/inchi/

There was discussion of whether the project should work exclusively with unique identifiers or whether something like a title might function as an identifier in a particular context. Whilst it was acknowledged that a title might serve as an identifier if the context was very tightly scoped it would otherwise be ambiguous.

There was a (deliberately) brief aside about the use of a part of one object's content in another (for instance a few paragraphs quoted from a larger text). Should the subset have its own identifier? This was considered to be a functional question: if the subset needs to be identified separately then it needs its own identifier.

There followed a long discussion about what identifiers should be exposed for 'human' (as opposed to possible machine-machine) use. Should the only way to a bitstream for a human be via some form of metadata page (perhaps a splash page)? If so then it is the identifier for the metadata record that should be exposed, the identifier for the bitstream should be encapsulated in the metadata. In the case of the video posited in the first part of slide #8, it is possible that "the titles mention other" videos which are identified not as bitstreams but as classes (my video contains extracts from 'Gone with the Wind') which may have instance identifiers elsewhere in the system but not explicit here (but this instance identifier may itself be for a metadata page in the first instance). If you use a bitstream identifier are you really asserting that the extract came from that particular instance?

As the discussion progressed it became apparent to the group that FRBR terms could usefully be used:

Suppose the title screen of the new video (slide #8) says "inspired by 'Gone with the Wind' by Margaret Mitchell". This is referencing a 'work'.

If the title screen says "this video contains sequences extracted from the 1939 film of 'Gone with the Wind' directed by Victor Fleming" it is referring to an expression.

If it says "this video contains an extract from the Warner Home Video 2006 DVD release of "Gone with the Wind"'. This is a manifestation.

It could say "this video contains an extract from the Warner DVD of 'Gone with the Wind' held in this library" - an item.

Each of these four levels potentially needs an identifier and they are inter-related in various ways.

At this stage it was decided that the ideas expressed thus far should be tested against the potential use cases that had been proposed in the papers for the meeting.

Use cases

In view of the wide range of use cases provided in the briefing papers, the first part of this discussion was devoted to narrowing down the range to the three that seemed, to the group, to be most relevant. In the end, four rather than three were quickly identified:

- Related versions (slide #15)
- Locate originals of derived components (slide #16)
- Identifier chains (slide #19)
- Migrate repository (slide #28)

It was agreed that there was overlap between 'related versions' and 'identifier chains'.

Related versions and event chains

A specific instance was related of a repository that gathers items together from a variety of sources and pulls them together into an administrative system. This is managed by library staff. When they are happy with the content they have acquired it is published to the repository proper. There is therefore a long chain of events connected by an extremely flexible, non-linear workflow. The final object exists on two systems and parts of the object reside elsewhere. All these components need clear identification at each stage of the process, and at particular points in time, and the final object evolves through a number of versions as it is assembled. It would be useful to be able to step backwards and forwards through the workflow to understand the development process and audit it. [This concept was referred to as "bidirectionality" during the rest of the workshop.] The 'final' object might then be copied to another repository external to the organisation.

RIDIR might provide tools to help. There are risks connected with the loss of relationships between objects/versions. It was suggested that given the efforts put into the Scholarly Works Application Profile and the development of FRBR these could contribute to a background architecture (for identifier relationships). RIDIR's work might contribute to the on-going discussion about the use of FRBR-based architectures and the specific discussion about the nature of FRBR 'expressions'. There are apparently schemes similar to FRBR more applicable to, for example, multimedia.

Locate sources [previously 'originals'] *of derived components*

It was felt that 'source' was a better word than 'original'.

This example has a level of complexity because of the various part/whole relationships involved. It is equally as applicable to datasets as to, say, learning objects. There may not be identifiers associated with some of the sources. Even if there are identifiers, these may not be at a level of granularity appropriate to the new object (uses extracts from 'Gone with the Wind'). This, again, has bidirectional aspects. Getting identification of sources wrong could have legal implications for the teacher identified in the case study.

RIDIR might demonstrate how one deals with a lack of formal identifiers, what might be achieved by automation given various levels of identifier availability, what might be done to make such work more effective.

Migrate repository

This is a real problem that repository managers are going to have to face in a number of guises.

At one level this is a workflow issue - where there is a global identifier mapped to a local locator only the mapping needs changing; this assumes that the identity model in the new repository is the same as in the old. Relationships between objects (that may not all be in the repository in question) need to be maintained. Users need to be able to get back what they 'expect' even after the move. There are issues of provenance depending on whether migrate means 'copy' or 'move': copy implies extension of the relationship chain.

This use case was conceived as a 'whole repository' issue, but there are related use cases when, say, an object's owner wants to take it with him to a new institution's repository. The transfer might be done by the owner or technicians on his behalf. How do you cope with the situation where he seizes the opportunity to 'clean it up' in some way. The migrated object is not the same as the one that has been removed at the old institution.

RIDIR might investigate how much these workflows could reasonably be automated. Particularly it might investigate how platform migration might be achieved where the internal structures in the 'old' repository are not immediately compatible with those in the 'new'. Does one try, for instance, to transfer 'lowest common denominator' metadata? What if a change of

manifestation is mandated: perhaps the 'old' repository accepted Word documents but the 'new' one only accepts pdf.

Wish list

Finally, participants were asked, if they could determine the one or two things that the RIDIR demonstrator would demonstrate, what would they be. In other words they were asked to express personal priorities. These ideas were not necessarily intended to demonstrate 'solvable' problems but perhaps to show how much of a problem something is at the moment.

The list was:

- de-duplication/grouping
- machine-to-machine which will take an identifier and a relationship and return all matches (find me all equivalents to ISBN 123456789X which might return 'doi:12.3456/789' etc, find all children of...) This is effectively working with an 'identifier cloud'.
- a tool which crawls such identifiers and builds/maintains a big semantic map (it would also be useful to be able to 'push' new relationships into the map rather than rely wholly on crawling)
- better exposure of identifier issues to systems developers
- bidirectional functionality in identifier chains

It was emphasised that we need to stay aware of other projects and contexts, for example, ORE. RIDIR's work needs to complement such projects.

Outcome and follow-up

The RIDIR Project Plan makes clear that, following the focus group meetings, the RIDIR team would consult with the JISC to agree the best way forward. Thus, subsequent to these meetings, the RIDIR team met to develop a business proposal for submission to the JISC. The final version of the proposal forms Appendix 4 to this document. It identified five basic use cases with which to ground the project in real issues:

1. the work of EThOSnet
2. the work of Spoken Word Services
3. Migrate repository as an explicit, demonstrated example - both simple and compound objects.
4. The Depot and its eventual need for migration
5. Locate original

Taken together, these illustrate some of the different aspects of the main use cases proposed through our workshops:

- Versions with a long chain of connecting events
- Locate originals of derived components also locate 'unknown' children (bi-direction)
- Migrate repositories

These use cases are described in detail at Section 3 of Appendix 4.

It seemed that two approaches were valid, one showing the value of interoperability and one showing its cost:

The value of interoperability

The use cases we have identified show examples of issues where the ability to work with identifiers yields value. In this case, the role of the demonstrator would be to show some of these examples in action. We 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 know to be very limited. We would focus on cases where there is a unique, *accessible and usable* 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 advantage of this approach is that it would serve as an important demonstration to the repository community of the importance of the role of identifiers in achieving interoperability and therefore encourage them to plan for it, and also to participate in any subsequent policy development as well as adopt working practices that will help the situation in future.
It might also be used to develop a business case for the development of (for example) any necessary Shared Infrastructure Services or other products or services that might facilitate interoperability.

The cost of interoperability

In this approach, using our use cases 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. There is a range of possible ways in which these tasks might be undertaken.

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.

Common ground

Whichever approach is taken, 'value' or 'cost', there will be common elements in the construction of the demonstrator. The choice will determine which aspects of the demonstrator need to be emphasised and thus take more of the project's resources. Amongst these common elements will be some of the functionality of the Persistent Identifier Mediator Service (PIMS), proposed in the RIDIR Project Plan

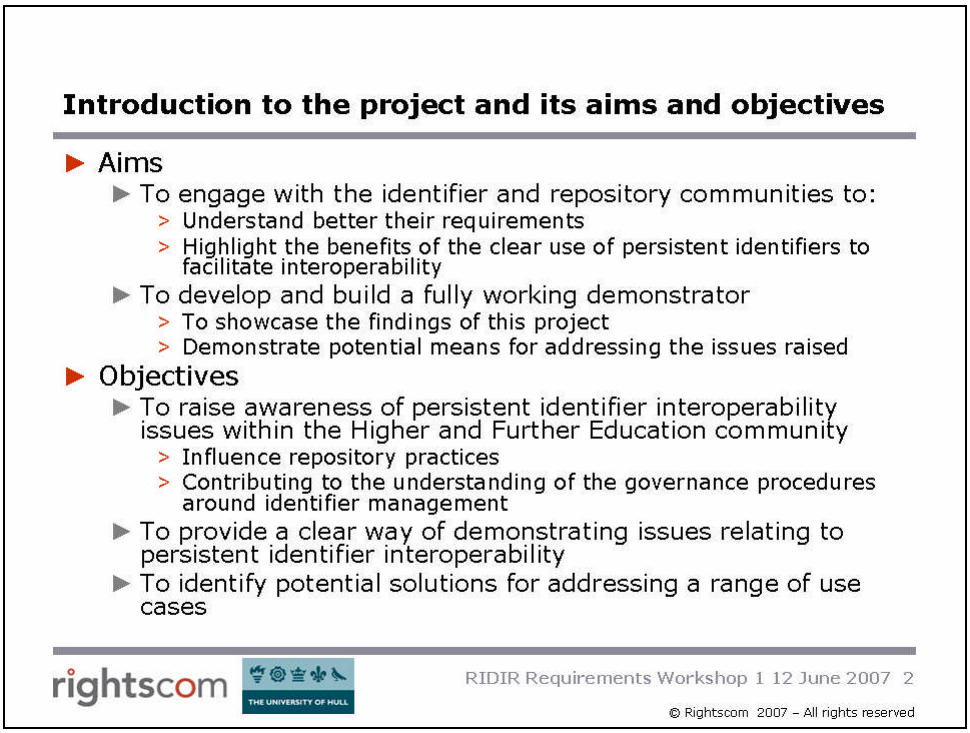
There will also be some important common outputs. These might include recommendations on policies, the relevance of standards, and working practices that would benefit content-sharing and interoperability.

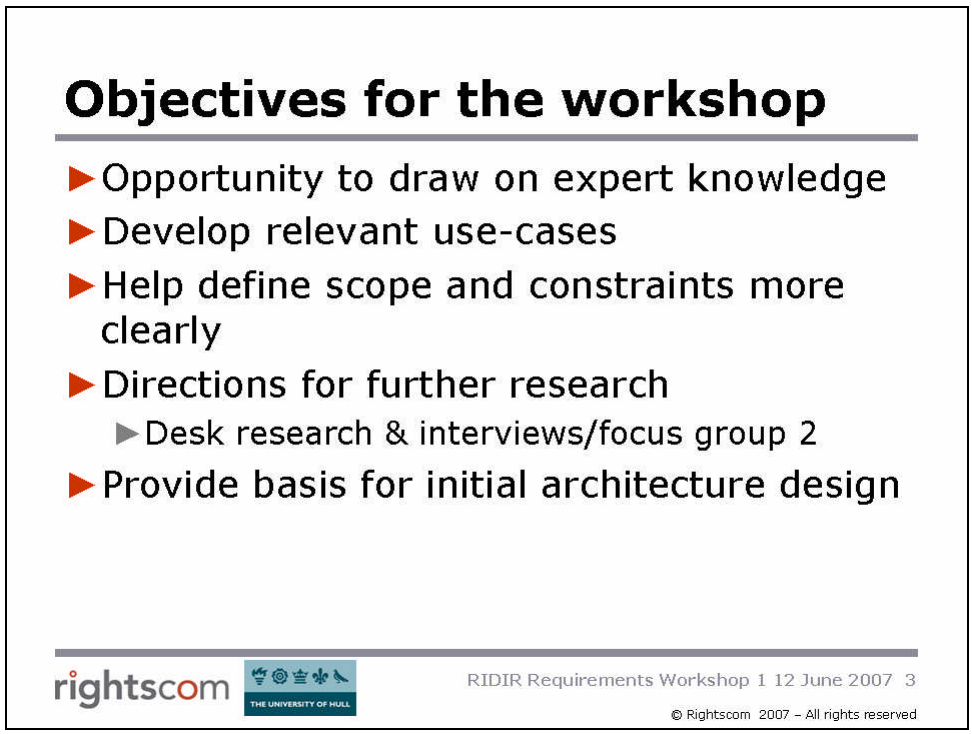
The development of policies and working practices in particular is important: without some level of compliance and shared practice, it is unlikely that interoperability would be feasible in the long-term.

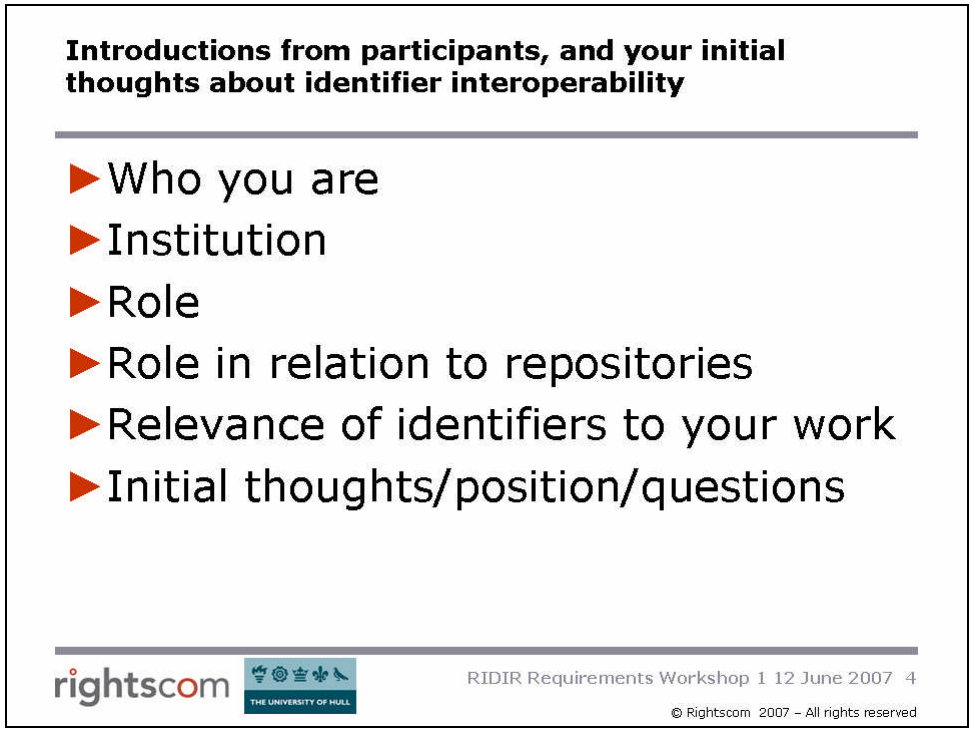
Final decision

The business proposal was presented to the JISC on 8th August 2007. On 21st August the JISC requested that we follow an approach which demonstrated our second option, the cost of interoperability.

Appendix 1
PowerPoint presentation to first focus group




Objectives for the workshop
Opportunity to draw on expert knowledge
Develop relevant use-cases
Help define scope and constraints more clearly
Directions for further research
Desk research & interviews/focus group 2
Provide basis for initial architecture design
rightscom
THE UNIVERSITY OF HULL
RIDIR Requirements Workshop 1 12 June 2007 3
© Rightscom 2007 – All rights reserved


Introductions from participants, and your initial thoughts about identifier interoperability
Who you are
Institution
Role
Role in relation to repositories
Relevance of identifiers to your work
Initial thoughts/position/questions
rightscom
THE UNIVERSITY OF HULL
RIDIR Requirements Workshop 1 12 June 2007 4
© Rightscom 2007 – All rights reserved

Our first thoughts: two building blocks

- Persistent Identifier Mediator Service
  - Function is to sit as a shared service between 'client' Repositories, and ensure that resources which have persistent repository identifiers are available to any other client repository
  - In a production context mightt interoperate as a shared infrastructural service within the JISC IE
- Client Repositories
  - To verify and demonstrate the satisfaction of a key identifier interoperability use case
    - We will transfer sets of digital objects from one repository to another
    - Preserving the identity of the relations between them
    - Preserving their resource identifiers
  - Each repository will be realised in a different technology
  - Each will be constructed with a different content model
    - Will show how the interoperability solution is able to resolve between identifiers minted for various digital objects within each client repository.
  - 'Edge cases' where preservation of a digital object is
    - Difficult
    - Requires manual intervention
    - Is completely incompatible

rightscom THE UNIVERSITY OF HULL

RIDIR Requirements Workshop 1 12 June 2007 5

Scoping discussion

- Different views of:
  - What identifier interoperability would mean
  - How to achieve it
- Is there any "state of the art" thinking that we could benefit from?

rightscom THE UNIVERSITY OF HULL

RIDIR Requirements Workshop 1 12 June 2007 6

Priorities

- What are the priority problems that the demonstrator should be aimed at?
- Who are those problems important to?

rightscom THE UNIVERSITY OF HULL

RIDIR Requirements Workshop 1 12 June 2007 7

Development of use-cases based on the discussion

- Typical use-cases
  - Creation or selection of identifiers
  - Validation, QA etc.
  - Linking to a resource in another repository
  - Copying a resource from one repository to another
- Edge-cases
  - Use of identifiers to support preservation
  - Complex objects that include many resources

rightscom THE UNIVERSITY OF HULL

RIDIR Requirements Workshop 1 12 June 2007 8

Checklist

- Have the use cases addressed the main issues raised in the rest of the discussion?

Constraints that the demonstrator will have to take into account

- Technical
- Policy
- Standards
- Other?


Thank you
Contact: Richard Green
RIDIR@jiscmail.ac.uk
http://www.hull.ac.uk/ridir/index.html
rightscom
THE UNIVERSITY OF HULL
RIDIR Requirements Workshop 1 12 June 2007 11
© Rightscom 2007 – All rights reserved

Appendix 2
Life stages and use cases available to second focus group

Life stages

Key events in the life of a digital object that could lead to identifying use cases in a less abstract, more scenario-driven, way. The focus is on concrete events that happen to the object, and may lead to information that would help with the use-cases.

We are happy to hear if all or any part of this example is in any way misleading or irrelevant.

The objective of a life-stages narrative is to flush out general points that could feed use-cases - don't take the example given as trying to be a use-case. It is intended to show the level of information that we hope to analyse.

- Object is created. In this case, a digital video recoding - this is an entirely original work to the academic who created it. The titles mention other objects that relate to this video that are also available online from the institution.
- The object is placed in the institution's repository, and some metadata is created to help manage the object
- As it is the output from a funded project, the object also has to be placed in a subject repository
- Shortly after that, someone (Author B) asks the original author A if he can use a long extract in a multimedia work he is producing for his PhD. The author grants permission as long as credit is given.
- The extract used in the new work makes the original popular, and author A improves it by digitally cleaning his original - no editing or any other changes are made. He replaces the older version with this cleaned on.
- Author A would rather that this version was seen, so he contacts Author B and asks him to replace the original footage with the cleaned footage. Author B happily complies but doesn't make any other changes.
- Author B then moves institution. He puts a copy of the work into the repository of his new institution. Other people start to reference that.

Life-stages can divide one or many times:

- In parallel, Author A begins a re-edit of the original, which he does not make immediately available. No content is added or removed. This version is intended for use on handheld devices.
- The re-edit changes a number of things besides the sequence of the content, including the file format, the compression used and the screen resolution so that navigation and display on a handheld become acceptable.
- This version is downloaded by Student D, who adds some audio commentary of her own and posts it to YouTube……..

- The original is licensed for inclusion in a package of recordings as part of an online course. This course is only licensed in its entirety and the technical means to extract an individual sequence are not supplied.
- The institutions that buy this collection typically makes it available to students through a VLE
- The entire collection is licensed internationally to different publishers, and is overdubbed into a variety of languages…..

These branches are now growing in parallel, and will result in several objects owing their original to the first recording but potentially very different in content and located in different resources or collections of resources. We are interested to find out your views on the role identifiers will play in this scenario, and the areas in which they will need to interoperate.

RIDIR Draft Use Cases

We offer these two sets of uses cases for initial consideration. We feel that some of these are irrelevant or misleading, but hope that the reason you might identify them as such may be different from ours. The two sets are presented differently mainly to avoid distorting the original intent: during the workshop itself we hope to draw out all the common themes despite this difference.

We are happy to hear if you think or any parts of these use cases are in any way misleading or irrelevant.

This first section of possible Use Cases for RIDIR to consider is based on a section of an article by Norman Paskin in the April 2006 issue of DLIb. This part of the article, in turn, draws heavily, with permission, on the report of the Dec 2005 ISO TC46/SC9 identifier interoperability workshop prepared by Mark Bide of Rightscom Ltd., which itself was based on input from the registration agencies responsible for ISAN, ISWC, ISRC, ISAN and DOI, and invited experts.

The use cases as proposed in the Paskin article have here been reinterpreted to have application to the world of academia rather than that of publishing.

Use Cases Part 1

USE CASE 1:
DISCOVERY OF "RELATED CONTENT" ITEMS

Who	• A researcher who has discovered a paper relevant to his/her work
What	• Wishes to discover and explore content "related" in some way to the first paper, including (for example): • Other papers by the same author • Other papers on the same subject • Other versions of "the same" paper (e.g., French language version)
Why	• To explore related content
Where & When	• Online • Any time
How	• Requires a discovery mechanism through which content with arbitrary shared attributes can be discovered • Implies that the attribute sets used with respect to different content types *either* • Use common semantics, *or* • Have a mechanism through which disparate semantics can be mapped
Issues	• Which specific attributes might be used as discovery keys? Any?

USE CASE 2:
DISCOVERY OF DIFFERENT VERSIONS OF THE SAME WORK

Who	• A researcher who has discovered a paper relevant to his/her work but knows or suspects that this is not the 'original' authoritative version
What	• Locate the original, authoritative version of the paper in its home repository
Why	• Authority and provenance
Where & When	• On line • Any time
How	• Mechanism for linking copied, cloned or derived objects back to their original source object
Issues	

USE CASE 3:
LOCATE ORIGINALS OF DERIVED COMPONENTS

Who	• A teacher who has located an interesting 'compond' learning object and wishes to re-purpose parts of it which may themselves have been repurposed from elsewhere
What	• Identify content used in the compound object
Why	• Tracking back to original material for the purpose of copyright/DRM
Where & When	• Online • Any time
How	• Mechanism for linking copied, cloned or derived material back to their original source object
Issues	

USE CASE 4:
TRACK OBJECTS WITH RELATED IDENTIFIER

Who	• Central service?
What	• Where multiple persistent identifiers under (Handles, PURL …) are allocated to the same object access parallel identifier
Why	• Broken identifier under one system
Where & When	• Online • Any time
How	• Automated registration of linked identifiers
Issues	Automated identifier issue?

USE CASE 5:
ESTABLISHMENT OF COMPREHENSIVE IDENTITY AND METADATA RECORD TO FACILITATE SUBSEQUENT DISCOVERY OF RELATIONSHIPS

Who	• Creator of a learning object • Creator of a research paper/thesis with particularly complex copyright issues
What	• Using multiple copyright elements in a ‘derived’ object • Licensed to use the elements subject to the creation of a comprehensive metadata record for the work, which includes the identity of the original materials and their source
Why	• To meet contractual obligation under terms of licence • To aid subsequent rights administration
Where & When	• Online • Any time
How	•
Issues	•

USE CASE 6:
IDENTIFIER CHAINS

Who	• Academic with a particular research paper
What	• Discover and download parallel publications of the same article
Why	• Check for additional, related materials
Where & When	• Online • Any time
How	
Issues	

USE CASE 7:
COMPILING MULTIMEDIA OBJECTS

Who	• e-Learning Course Designer
What	• Discover identifiers for a range of different content types to be included in a multimedia e-learning course, which might include one or more of: • Journal article • Chapter from a book • Audio-visual clip • Sound recording • Photograph • Graphic image • Musical score • Software application • Any of this content might be self created or have rights owned by third party
Why	• Rights ownership discovery • Could be contextual: where can I clear rights for this specific use in this particular territory for this time?

	• Rights clearance • Rights usage reporting • Providing comprehensive metadata for users (see Use Case 1)
Where & When	• Online • Any time
How	• Online identifier discovery and use • Online metadata discovery and use
Issues	• Substantial extension of metadata availability to include mechanisms for discovering rights manager identities for particular objects

USE CASE 8:
IDENTIFIER CHAINS

Who	• Researcher, student, teacher…
What	Discover and download different expressions (FRBR - see below) of the same work, (this may imply following a chain to manifestations and then items)
Why	• To compare and contrast different expressions
Where & When	• Online • Any time
How	• Availability of some identifier system for nodes in the FRBR hierarchy?
Issues	• Access to some sort of authority files?

USE CASE 9:
UNAMBIGUOUS LINKING OF RIGHTS TO PEOPLE AND/OR ORGANISATIONS

Who	• User wanting to re-use material in some way
What	• Unambiguously link materials to rightsholders
Why	• To ensure correct copyright clearance
Where & When	• On line • All the time
How	• Requires a link from the object to an unambiguous party identifier for the author (International Standard Interested Party Identifier etc)
Issues	• Lack of any widely deployed party identification systems

USE CASE 10:
COLLOCATION IN LIBRARY CATALOGUE

As things have been altered, this is effectively a repetition of Use Case #6 but it has been left for completeness. For 'library catalogue' read repository?

Who	• Librarian
What	• Linking an article published in several different serial publications
Why	• Collocation
Where & When	• In library catalogue system • At any time
How	• Link each article to an ISTC for the article • Link each article to the ISSN of the relevant serial publication
Issues	• Article identification (DOI?) • Discovering and linking using the ISTC

USE CASE 11:
LINKING REPERTOIRE TO "USAGE TERM SETS" IN A NATIONAL LIBRARY

There are many potential examples of the requirement to link a list of repertoire covered by a specific set of usage terms with that set of usage terms. This particular example may be slightly unfamiliar, but is included for precisely that reason.

Who	• A national library
What	• Linking a set of resource identifiers with a set of usage terms
Why	• Unless all resources in a national library archive are managed under "lowest common denominator" usage rights, it is necessary to identify the particular set of usage terms that applies to a particular resource in a particular context • Repertoires are likely to overlap (in other words, more than one set of usage terms may relate to the same resource in different contexts -- for example, some uses of a resource may be governed by legislation and others by licence) • A set of usage terms may relate to a single resource or to a complete collection or anything in between
Where & When	• In library archive system • In perpetuity...
How	• Likely to be at least semi-automated when resources are ingested
Issues	• Implies an appropriate identifier is available for the set of usage terms as well as the resource -- a "licence identifier" perhaps • Only known standard licence identifier known within the group is the Musical Works Licence identifier (MWLI) which was developed as part of the MI3P initiative and which is managed by CISAC

Use cases part 2

The second set of use cases is abstracted from the work of the PILIN Project in Australia. This is a relatively small subset of their use case work but represents the ones that seem possibly applicable to RIDIR's brief. There is some obvious overlap with the use cases above.

The presentation of these is rather different to avoid distorting the originals to fit the same pattern as the first set. However, we anticipate that the basic issues identified in the workshop are likely to have much in common with those identified in the first set.

RESOLUTION TO MULTIPLE EXPRESSIONS.

A digital work can have several datastreams delivering its content in different formats—according to user platform, user preference, disability access, etc. These are different FRBR expressions of the same work. The work identifier serves to yoke the expressions together. Individual expressions can be accessed through parameterised services, or through individual identifiers. If an actionable identifier provides multiple resolution, the appropriate expression can be selected through parameterised resolution, or resolution driven by user preferences. (TSO DOI report use case #5)

- A work—today's lecture on hippos—is published in three expressions: a 200x400 pixel version, a 400x800 pixel version, and a text transcript. There is an identifier for the work, as an aggregation.
- The identifier for the work is capable of multiple resolution.
- The resolution service is tunable. Given user preferences informing of visual impairment, a non-visual resolution (the transcription) is selected as the default resolution.
- A parameterised invocation of the resolution can select a particular resolution, based on discriminatory metadata (in an agreed scheme) rather than a distinct identifier; e.g. http://www.example.com/resolve/1.2.3/32893?res=lowres
- Otherwise, the resolution of the work ID can be to a screen allowing the appropriate expression locator to be selected.
- The individual expressions can have identifiers of their own, although that only makes business sense if the expressions will plausibly be disaggregated/accessed separately, rather than through a selector service as expressions of the work.

MANAGED VIRTUAL COLLECTION

A virtual collection of digital objects can be assembled by a collection managed out of existing items on repositories, independently of repository managers. The result is a managed collection that spans different repositories.

- Collection manager identifies digital object to include in their virtual collection. Identification includes procuring global identifier.
- Collection manager aggregates objects through identifiers as a collection.
- Collection manager assembles enough metadata and infrastructure that the collection can be queried or at least browsed. (NOT managed or stored; that's what repositories do.)
- Request to resolve identifier discovered through virtual collection will redirect to source repository.

HARVESTING SET

OAI PMH allows any subset of objects to be harvested have a set identifier (http://www.openarchives.org/OAI/openarchivesprotocol.html#Set), and that membership of the set be queryable by the harvester (3)

- A harvesting set is defined on a repository through some membership criteria.
- The set is assigned an identifier.
- The identifier is operated on by membership services ("is item a part of set b?"), rather than resolving to a first order digital object: it actually defines an aggregation
- Membership services are required by OAI PMH to support services. The target repository must be able to respond with membership information when a set identifier is parameterised to the service requests ListIdentifiers?, ListRecords? and GetRecord? .
- The identifier is used in the OAI PMH context, through these services.
- This is an example of a service operating on an identifier that is NOT a resolution service. Moreover, the identifier does not NEED a resolution service to be useful.

(3)Much much more generally: arbitrary vocabulary items in metadata can be indexed through identifiers, and those identifiers are defined through non-resolving services. Resolution is optional but recommended for accountability, possibly to definitions rather than digital objects—cf. Handle's treatment of types.

MIGRATE REPOSITORY

Yes it's a trivial use case, but what matters is the actor. A repository will migrate its repository from ePrints to Fez. In the process all the item locators will need to change. The identifiers do not. This is a predictable and managed process of redirection.

- An item is deposited in a repository. A global identifier is mapped to its local locator.
- The item will be transferred to a new repository.
- Item is ingested from old to new repository.
- Global identifier is redirected to new locator.
- Old repository is decommissioned.
- End user clicks on identifier and gets same resolution as before.
- Persistence was guaranteed only because the item was managed—the repository manager was an actor who could initiate the identifier redirection in time. With an unmanaged item, this is implausible.

CONTACT INFORMATION

A digital object identifies its manager for authority purposes. The identifier needs to be fluid (the person filling the role may change), actionable (I should be able to contact the manager to escalate a query), and extensible (there are different ways to contact a person, and different attributes to do so with, such as contact hours). The contact information is captured in an aggregate digital object, which is indexed by a global identifier.

- An aggregate digital object is created for the contact information of the PILIN Business Analyst. It contains a name, a landline phone number, a mobile phone number, an email address, a snail mail address, a room number, and a Skype address. And for contact hours.
- An identifier is created for this aggregate. Individual fields are accessible by specifying field names.
- Handle doesn't do field names, it does field numbers and types. That's an implementation issue, which probably involves a metadata schema mapping field names to field numbers. The metadata scheme would have its own identifier, natch. Field types will not necessarily be a practical solution.
- Individual fields can be updated or added with impunity; the outside world access this information through the aggregate identifier and a field name parameter.
- The service doing this accessing is fully a web service. One can have invocations like:
- http://www.arrow.edu.au/contact/1159.1/62453?email returns ninichol@lib.monash.edu.au
- http://www.arrow.edu.au/contact/1159.1/62453?skype returns opoudjis
- http://www.arrow.edu.au/contact/1159.1/62453?snailaddress returns <address> <inst>ARROW Project</inst> <bldg>Building 4</bldg> <city>Monash University</city> <state>VIC</state> <postcode>3800</postcode> </address>

- These service calls can be orchestrated; http://www.arrow.edu.au/contact/1159.1/62453?skype can be parameterised in turn into a skype frontend (given the skype API)

- The contact identifier can be used wherever a human needs to be identified, provided the context of use places no data model requirements on its resolution (or the resolution happens to match the required profile). For example, Creator Identifier in LOM?

- Once more: this is resolution, but not a one-to-one mapping of a URL to a discrete web page.

APPROPRIATE COPIES—ONE SERVICE

An object is stored in two different locations. The identifier allows either location to be resolved to. The decision on which location to resolve to is up to the server, and can be intelligent.

- An item is deposited in a repository, and has a locator.

- The same item is deposited in a different repository, with a different locator.

- The two copies are kept in sync by the repository managers.

- An identifier is created which links to both locators.

- Resolution is provided by a service, which can pick either locator.

- The choice of locator that the service makes can be informed by repository uptime and physical location; digital rights; accessibility constraints; etc.

DEDUPLICATION

A repository wishes to guarantee that it does not accidentally ingest an exact duplicate of an object already ingested elsewhere in the same repository, or another repository in the same federation. Ingestion will normally assume the object is new to the repository and assign it a unique identifier; if this is interpreted as identifying a unique new work, unacceptable confusion will result as the two instances each have their own metadata built up around them. To forestall that, once a digital object is ingested, it needs to be branded with the identifier; and ingesting within the federation needs to check that the candidate object (or datastream) has not already been ingested. The branding cannot be restricted to the metadata record, since the same content item can be submitted for ingestion using two distinct metadata records. The branding cannot be specific to an instance of the digital object: a copy of the digital object may have been made prior to submission, and then submitted independently for ingestion.

- A digital object is submitted for ingestion in a repository.

- An identifier is associated with the digital object.

- The discriminant attributes coded for the association are not restricted to the locator (since that does not rule out other copies), but depend on the digital object content.

- The extracted or calculated attribute coded for the identifier truly needs to be unique for the object, but common among all instances of the object. (So not the location, and relying on object content which is common to all copies, but more specific than say the document title).

- Another digital object is submitted for ingestion in the repository.

PERIPATETIC ACADEMIC

. An academic changes institutions. They reuse a digital object they've originally created in their new job's LMS. This is a derivative work (localisation), preserving same author, but different authorisation (different employer).

- Academic Jane (who is peripatetically employed) creates a learning object at Duck Uni. It is deposited at the Duck U LMS with a global identifier.
- Jane retains the rights to create derivative works.
- Jane's contract is up, and she ends up at Gander Uni.
- Jane has an (inappropriate) copy of her learning object, which she customises (localises) to Gander Uni requirements.
- Jane publishes the customised object at the Gander U LMS, with a distinct identifier.
- Per CUSTOMISE, a lineage of the Gander Uni object from the Duck Uni object is acknowledged, and discoverable through a Relationship Service.
- The Relationship Service does not provide accessioning: a student at Duck U can discover that there is more up to date version of the object now at Gander U, but they are not allowed to access it.
- Either Jane or Duck U pull the Duck U course from the Duck U LMS. The lineage information is preserved in the relationship service, because the identifier is persistent: the ID has not been destroyed. If Duck U is nice, the metadata about the obsoleted course is still exposed, though the object is not.
- Jane's course goes commercial, and in a paroxysm of lawsuits, her new employer gets the exposed metadata about her old course pulled. The identifier still survives in the wild; though it's not particularly actionable, and the lineage may still be reconstructable.

DATASET LIFECYCLE MANAGEMENT

Per Lyle, APSR paper indicates that not all data sets are guaranteed long-term preservation, because of the appreciable size of resources involved. Datasets are contractually mandated by the funding authority to be made available, but only for a limited time. Once that period expires, a value decision needs to be made on what data is preserved. The bases for that decision are outside of our scope and domain specific; but they are not infallible. They need to factor in who is still using the dataset (although deletion does not allow for long-tail effects: "I saw your paper from thirty years ago the other day…"). But that determination of ongoing use needs to be robust enough to cope with: changes in location of dataset (including both changes of repository location and move to different repository); changes in location of user; off-line access to the dataset (e.g. use by downloader of downloaded copy past dataset expiration, or use of downloaded copy by another party). When the decision comes to delete the dataset, a reasonable effort needs to be made to notify potential stakeholders. An identifier-base scheme only partly addresses the problems

- A researcher accesses a dataset. As part of their initial access, they are asked to subscribe to notifications of changes in the lifecycle of the dataset. If they do not, they risk losing access to the data long-term.
- The subscription takes the form of a mapping between a persistent identifier of the dataset, and a reasonably persistent locator of the user.

- A new stage in the lifecycle of the dataset is decided. This includes deletion, access restriction, and service enhancement. It would not normally include relocation and rehoming, since the persistent identifier should cope with such changes.

- The subscribers to the dataset retain their association with the dataset despite rehoming, because of the persistent identifier; so they can still be discovered.

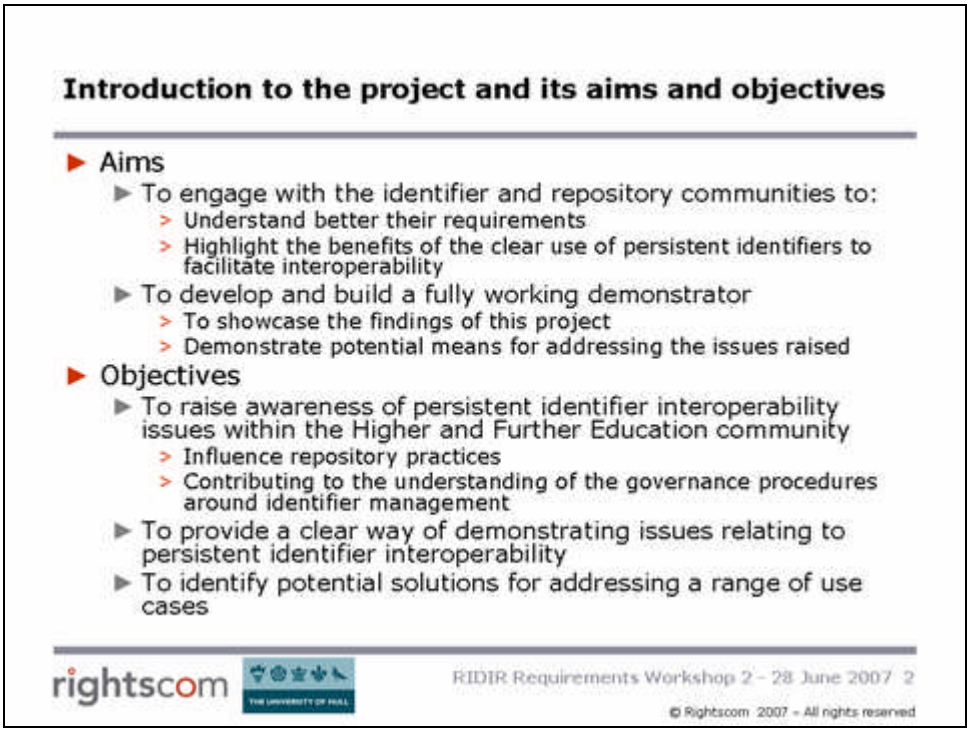
- The subscribers are alerted as to the new stage coming up through their persistent locators.

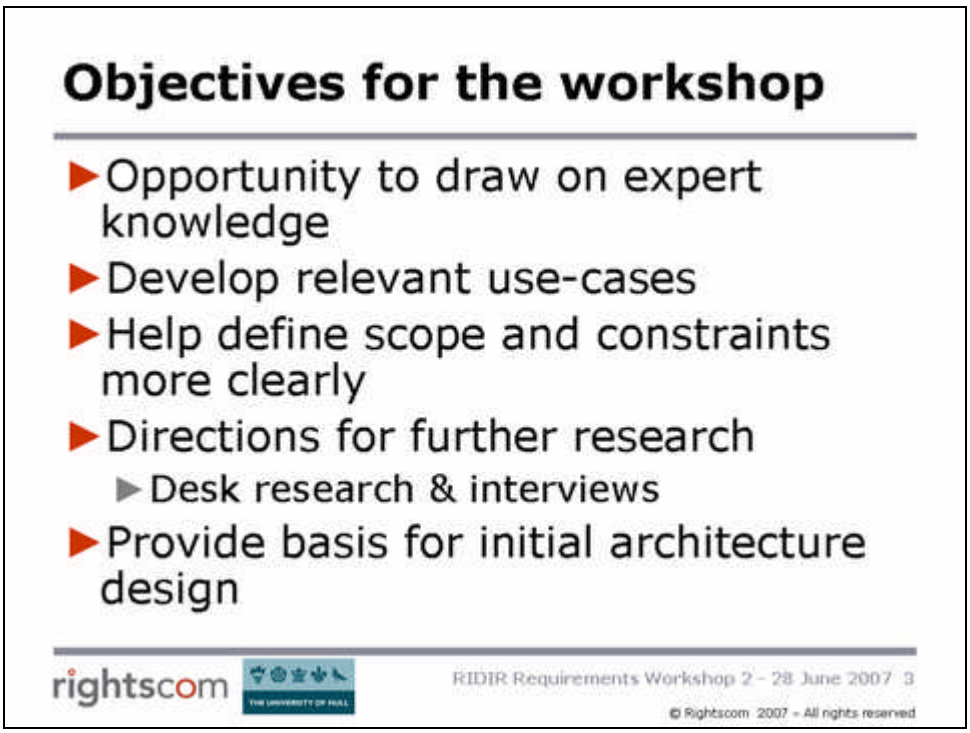
And a final use case submitted by David Flanders who was part of the first RIDIR focus group.

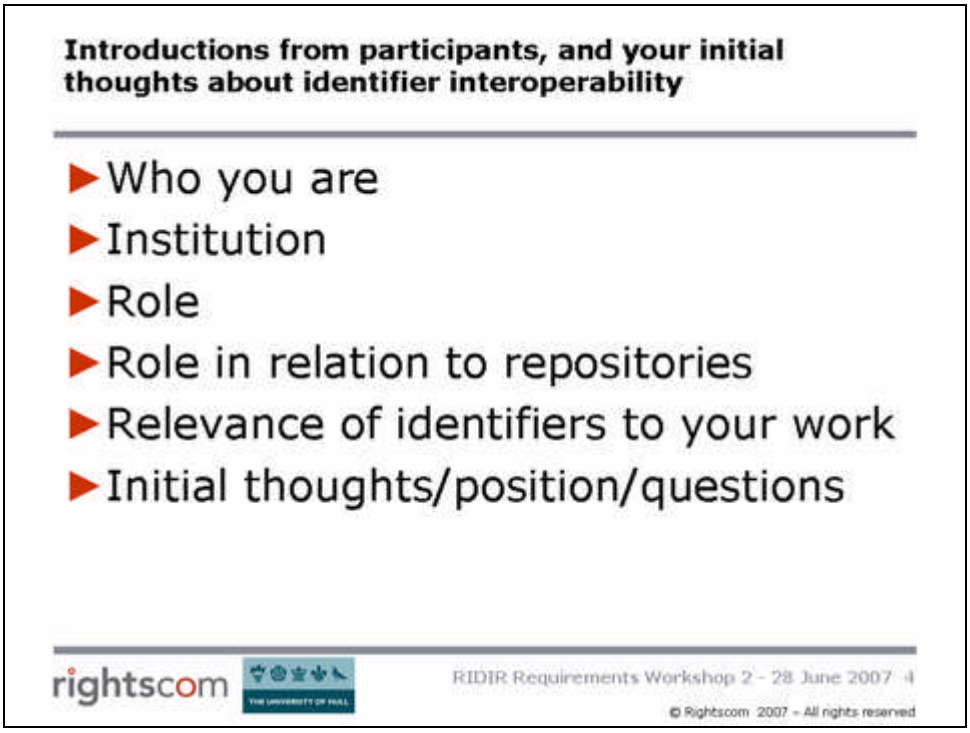
METADTA FOR OBJECTS MIGRATED ACROSS REPOSITORIES

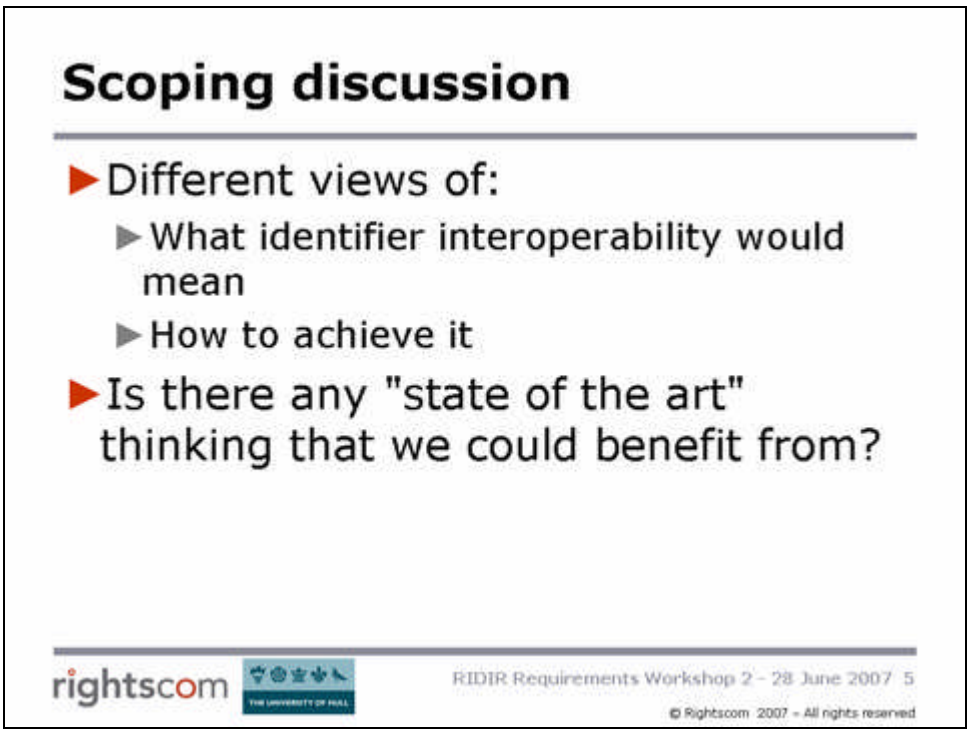
A user migrates an object (an image) to another repository. In the process of migration a metadata crosswalk service is evoked to crosswalk metadata from DC to LOM (this is for the purpose of making the metadata interoperable with the new repository which is a Learning Object repository (this repository requires LOM for its indexing and browsing functionality: it is dependent upon specific element fields to populate the learning taxonomy browse tree), where the previous existence of the object was in a research repository). All of these metadata streams are carried on with the object, but just as an additional metadata stream. This crosswalking idea could be continued, say the image was transferred to a publisher repository (Amazon) where ONIX metadata was required for sales purposes. Accordingly, there are three different metadata streams in the object all of which are "primary" metadata (and perhaps were even augmented by additional information from each of the separate repositories). A key question is which metadata to use??

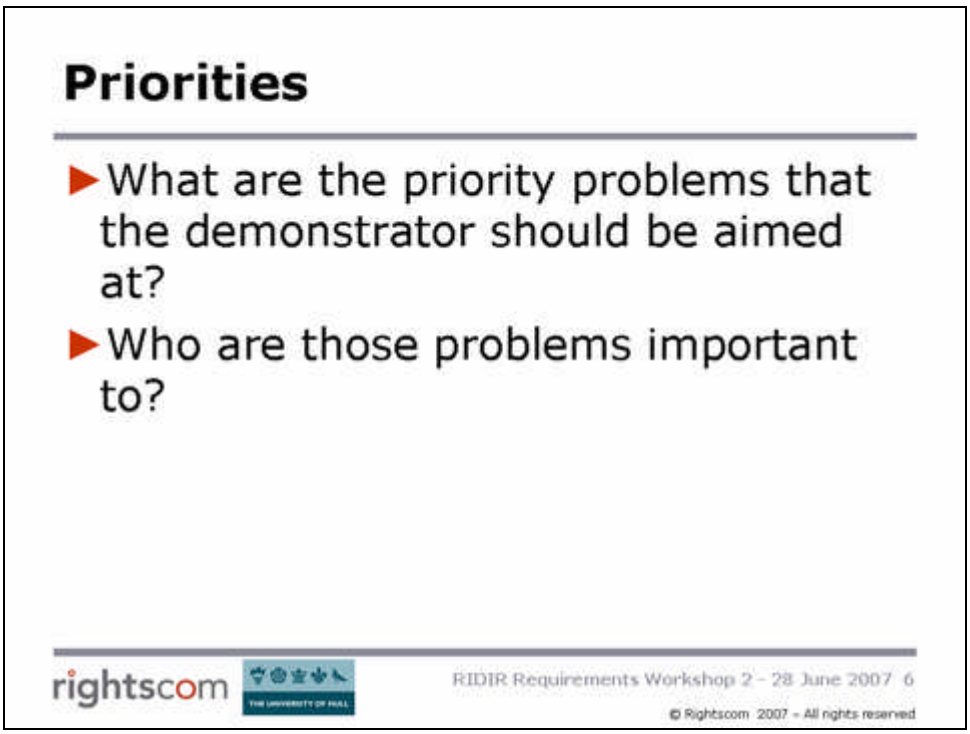
Appendix 3
PowerPoint presentation to second focus group

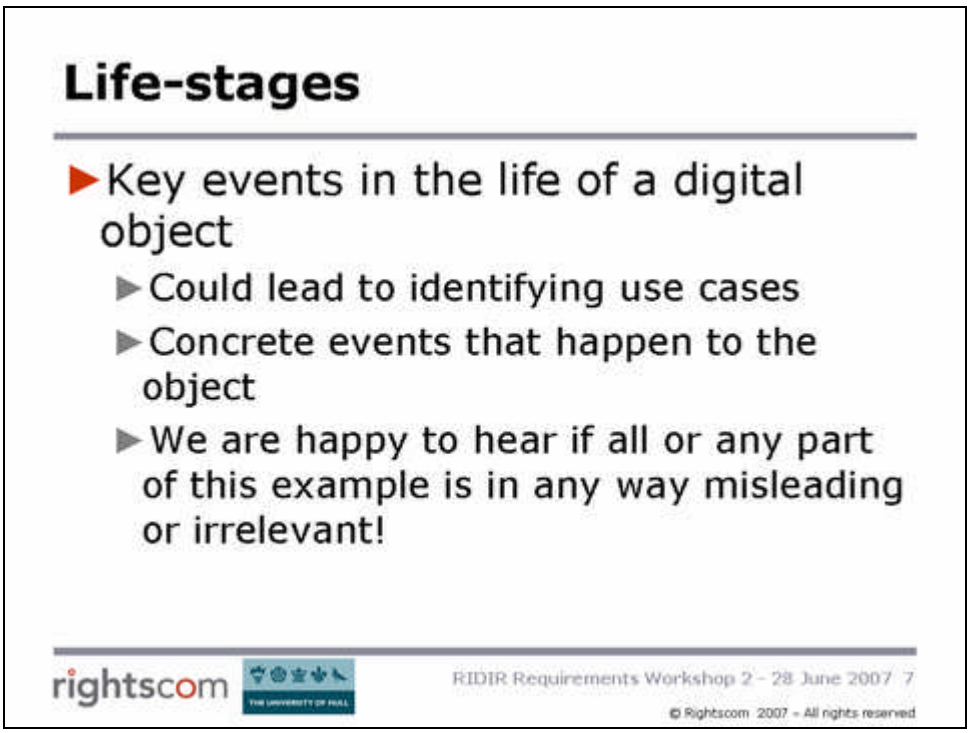


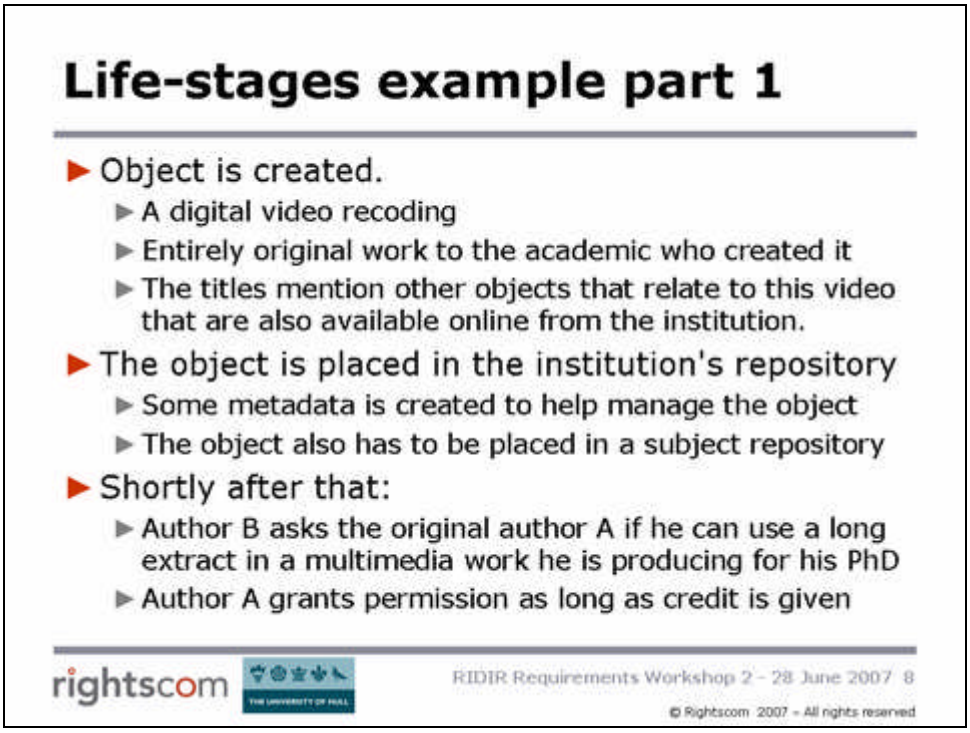

Objectives for the workshop
Opportunity to draw on expert knowledge
Develop relevant use-cases
Help define scope and constraints more clearly
Directions for further research
Desk research & interviews
Provide basis for initial architecture design
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 3
© Rightscom 2007 - All rights reserved

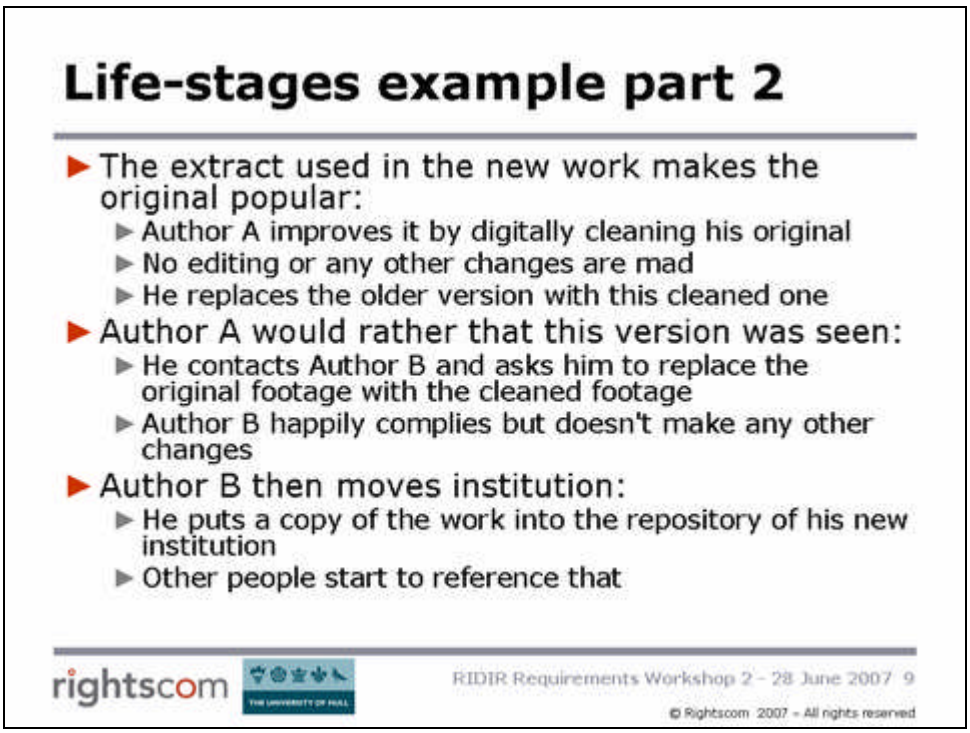

Introductions from participants, and your initial thoughts about identifier interoperability
Who you are
Institution
Role
Role in relation to repositories
Relevance of identifiers to your work
Initial thoughts/position/questions
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 4
© Rightscom 2007 - All rights reserved

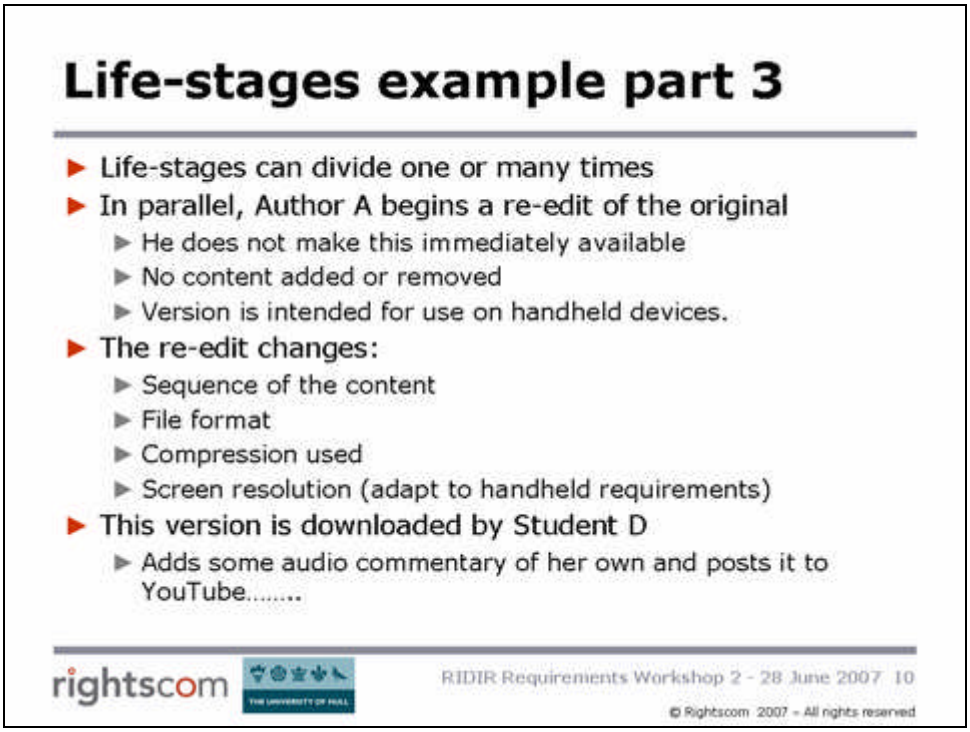

Scoping discussion
Different views of:
What identifier interoperability would mean
How to achieve it
Is there any "state of the art" thinking that we could benefit from?
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 5
© Rightscom 2007 - All rights reserved

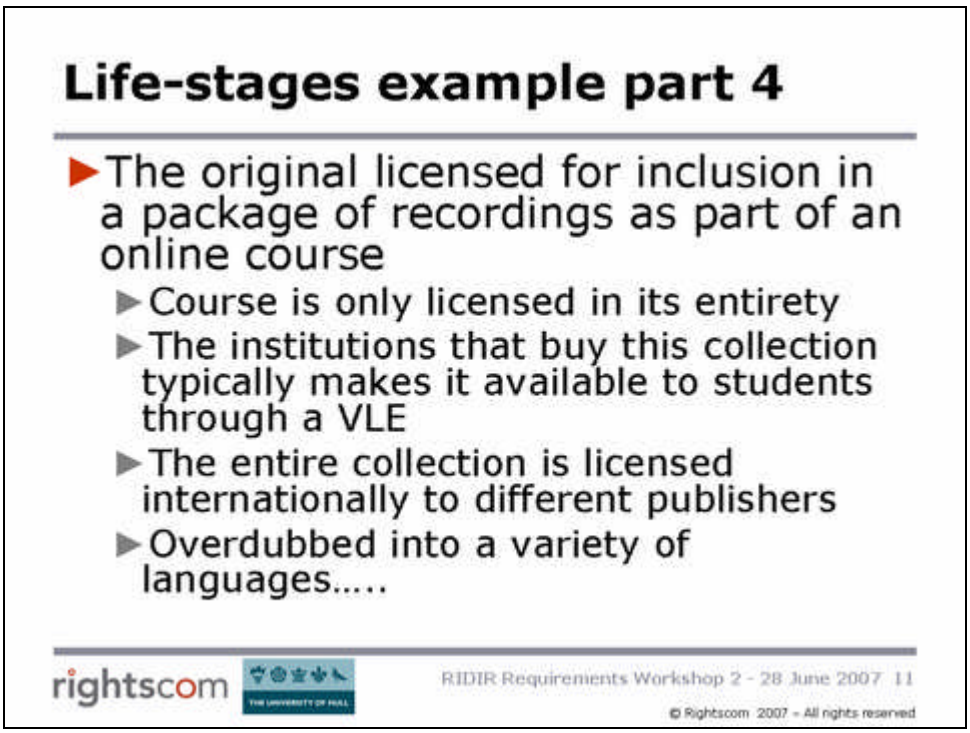

Priorities
What are the priority problems that the demonstrator should be aimed at?
Who are those problems important to?
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 6
© Rightscom 2007 - All rights reserved


Life-stages
Key events in the life of a digital object
Could lead to identifying use cases
Concrete events that happen to the object
We are happy to hear if all or any part of this example is in any way misleading or irrelevant!
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 7
© Rightscom 2007 - All rights reserved

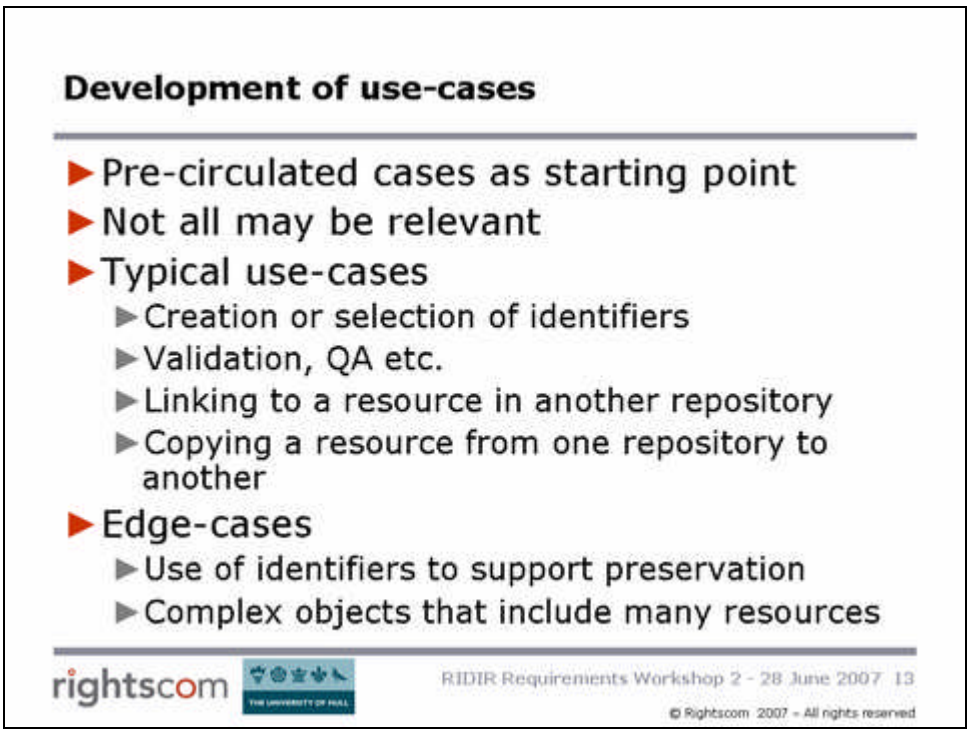

Life-stages example part 1
Object is created.
A digital video recoding
Entirely original work to the academic who created it
The titles mention other objects that relate to this video that are also available online from the institution.
The object is placed in the institution's repository
Some metadata is created to help manage the object
The object also has to be placed in a subject repository
Shortly after that:
Author B asks the original author A if he can use a long extract in a multimedia work he is producing for his PhD
Author A grants permission as long as credit is given
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 8
© Rightscom 2007 - All rights reserved

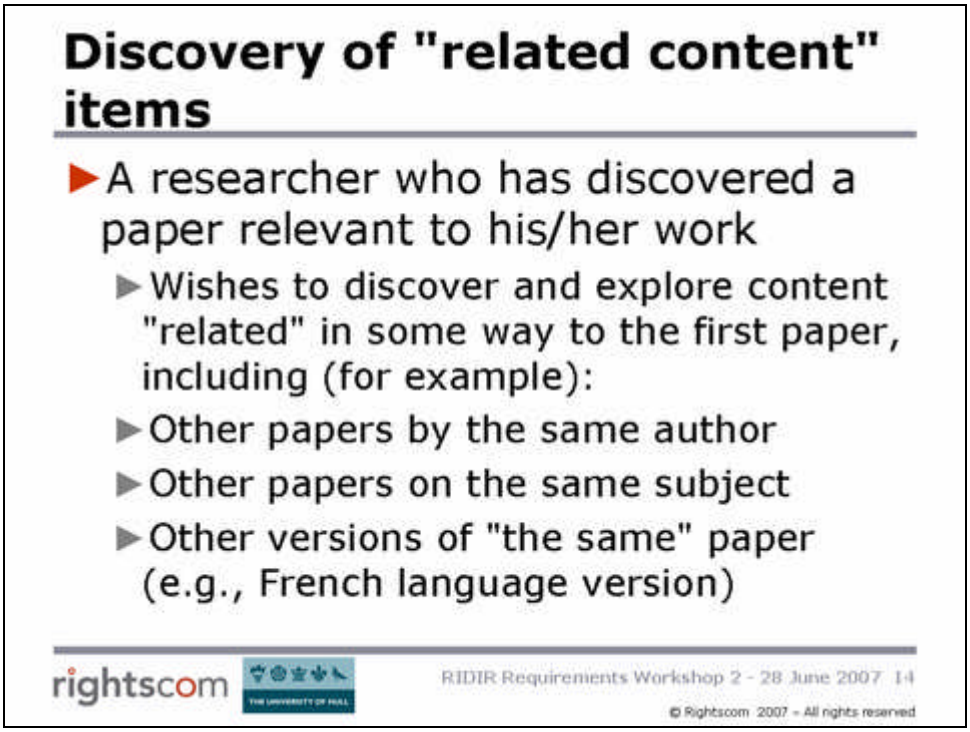

Life-stages example part 2
The extract used in the new work makes the original popular:
Author A improves it by digitally cleaning his original
No editing or any other changes are mad
He replaces the older version with this cleaned one
Author A would rather that this version was seen:
He contacts Author B and asks him to replace the original footage with the cleaned footage
Author B happily complies but doesn't make any other changes
Author B then moves institution:
He puts a copy of the work into the repository of his new institution
Other people start to reference that
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 9
© Rightscom 2007 - All rights reserved

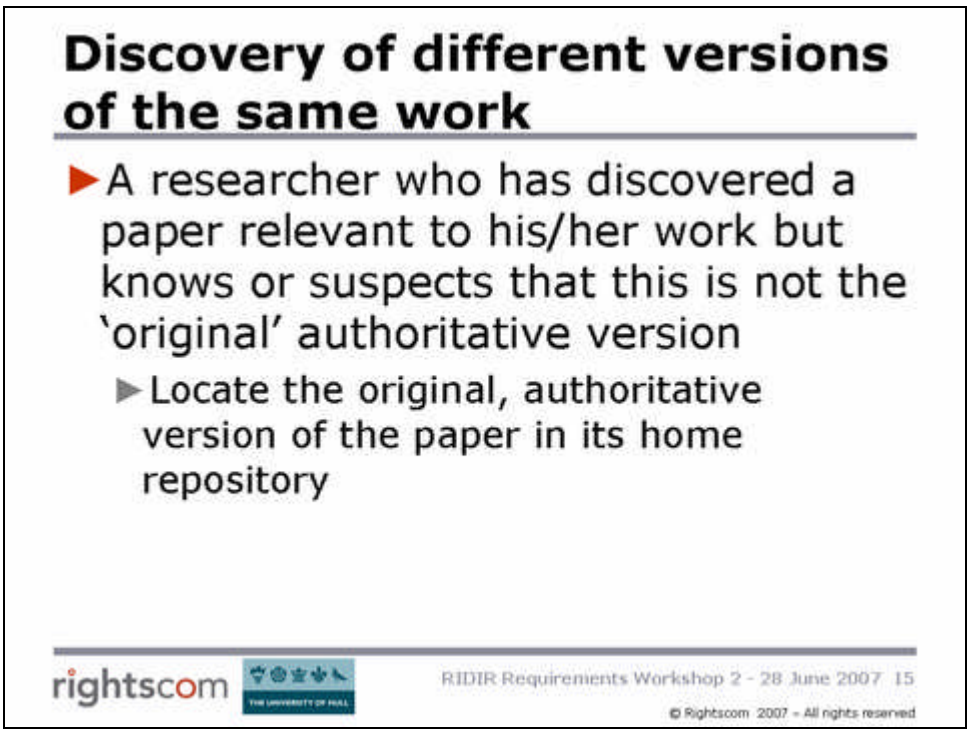

Life-stages example part 3
Life-stages can divide one or many times
In parallel, Author A begins a re-edit of the original
He does not make this immediately available
No content added or removed
Version is intended for use on handheld devices.
The re-edit changes:
Sequence of the content
File format
Compression used
Screen resolution (adapt to handheld requirements)
This version is downloaded by Student D
Adds some audio commentary of her own and posts it to YouTube.....
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 10
© Rightscom 2007 - All rights reserved

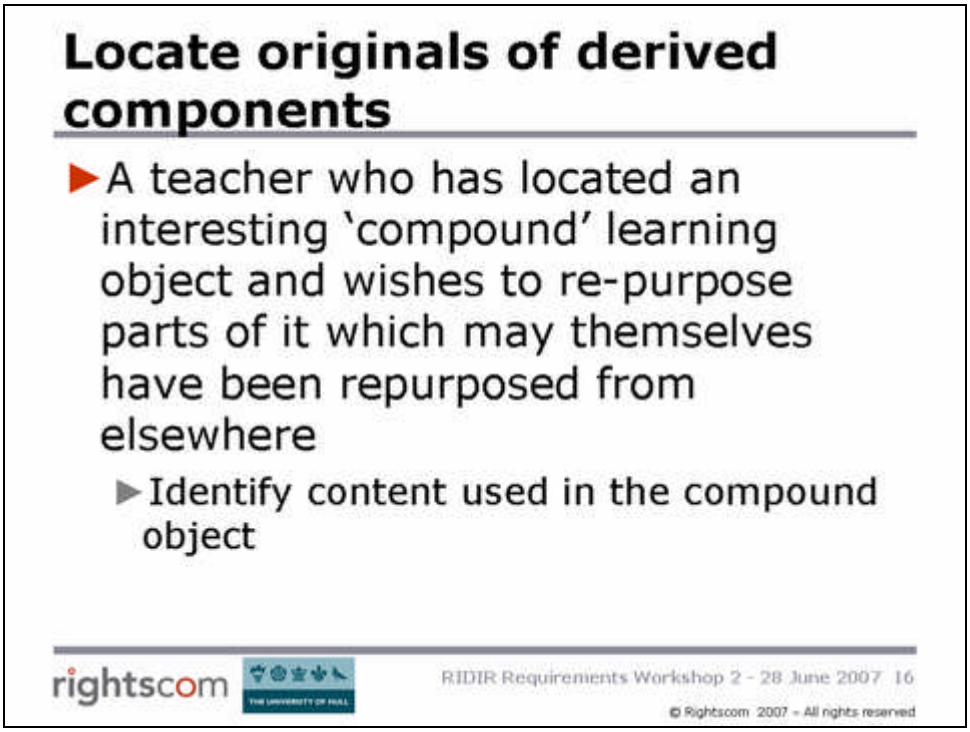

Life-stages example part 4
The original licensed for inclusion in a package of recordings as part of an online course
Course is only licensed in its entirety
The institutions that buy this collection typically makes it available to students through a VLE
The entire collection is licensed internationally to different publishers
Overdubbed into a variety of languages.....
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 11
© Rightscom 2007 - All rights reserved

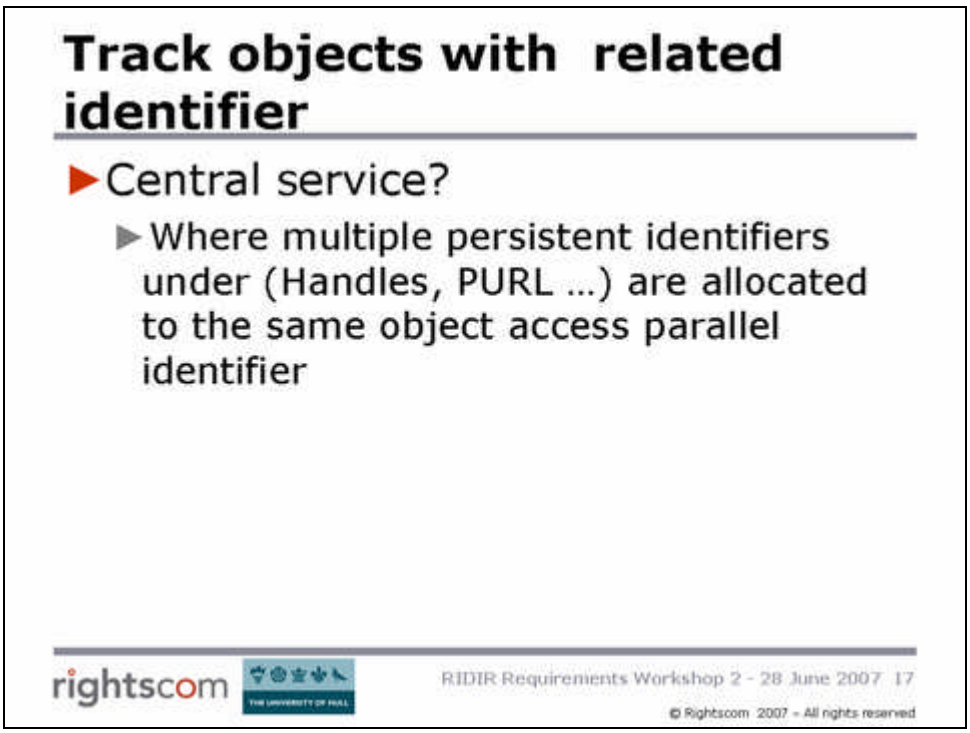

Life-stages – what next?
These branches are now growing in parallel, and will result in several objects owing their original to the first recording but potentially very different in content and located in different resources or collections of resources. We are interested to find out your views on the role identifiers will play in this scenario, and the areas in which they will need to interoperate.
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 12
© Rightscom 2007 - All rights reserved

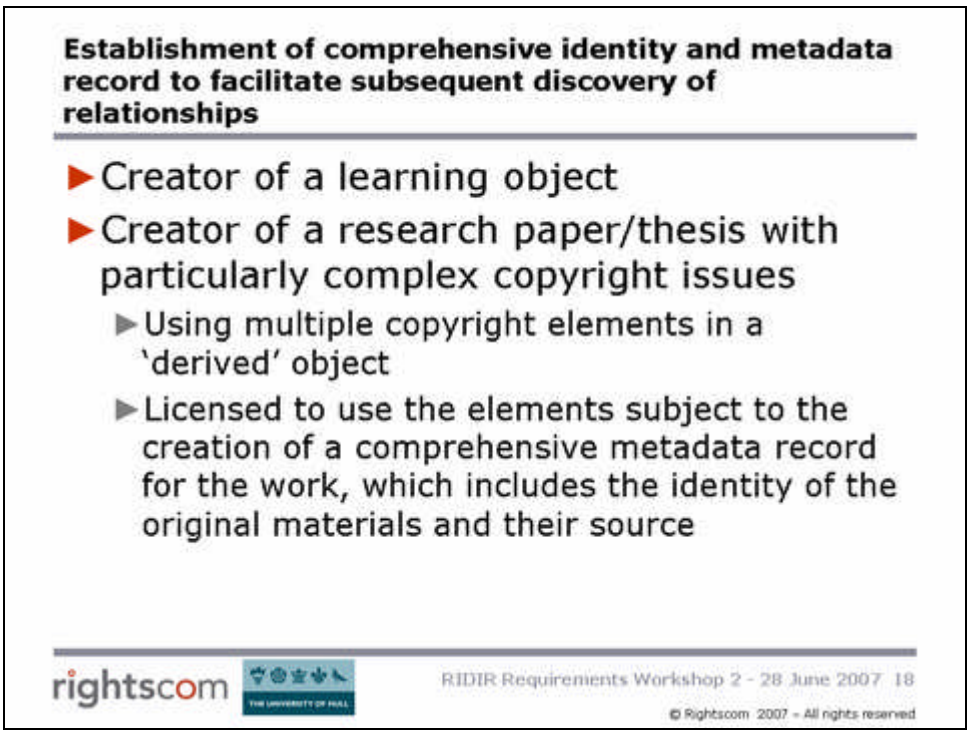

Development of use-cases
Pre-circulated cases as starting point
Not all may be relevant
Typical use-cases
Creation or selection of identifiers
Validation, QA etc.
Linking to a resource in another repository
Copying a resource from one repository to another
Edge-cases
Use of identifiers to support preservation
Complex objects that include many resources
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 13
© Rightscom 2007 - All rights reserved

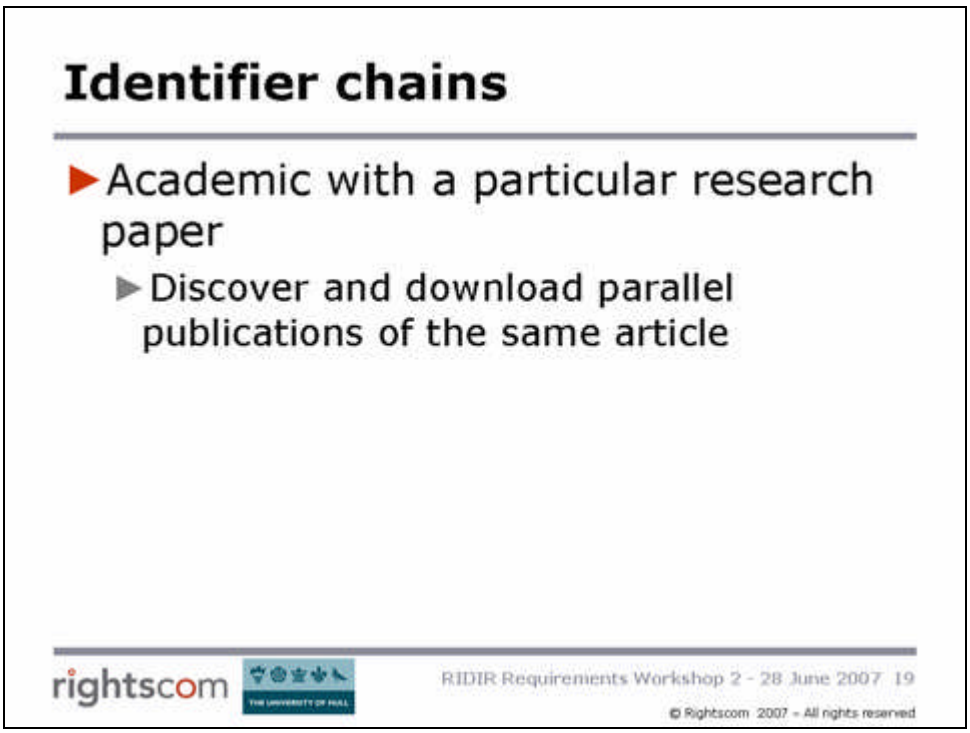

Discovery of "related content" items
A researcher who has discovered a paper relevant to his/her work
Wishes to discover and explore content "related" in some way to the first paper, including (for example):
Other papers by the same author
Other papers on the same subject
Other versions of "the same" paper (e.g., French language version)
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 14
© Rightscom 2007 - All rights reserved

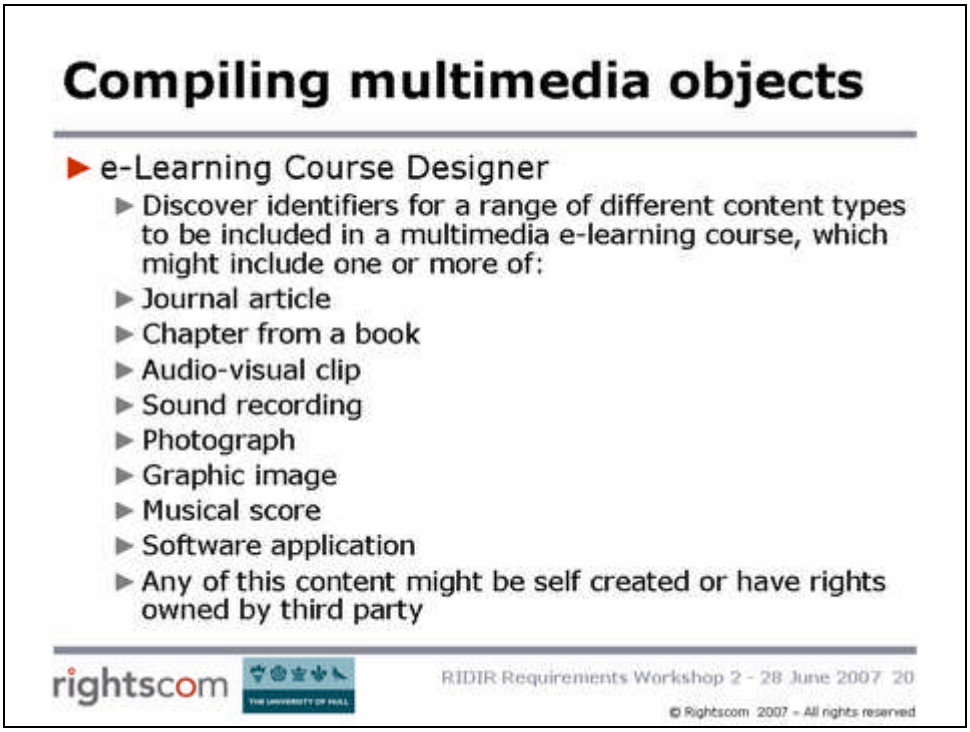

Discovery of different versions of the same work
A researcher who has discovered a paper relevant to his/her work but knows or suspects that this is not the 'original' authoritative version
Locate the original, authoritative version of the paper in its home repository
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 15
© Rightscom 2007 - All rights reserved

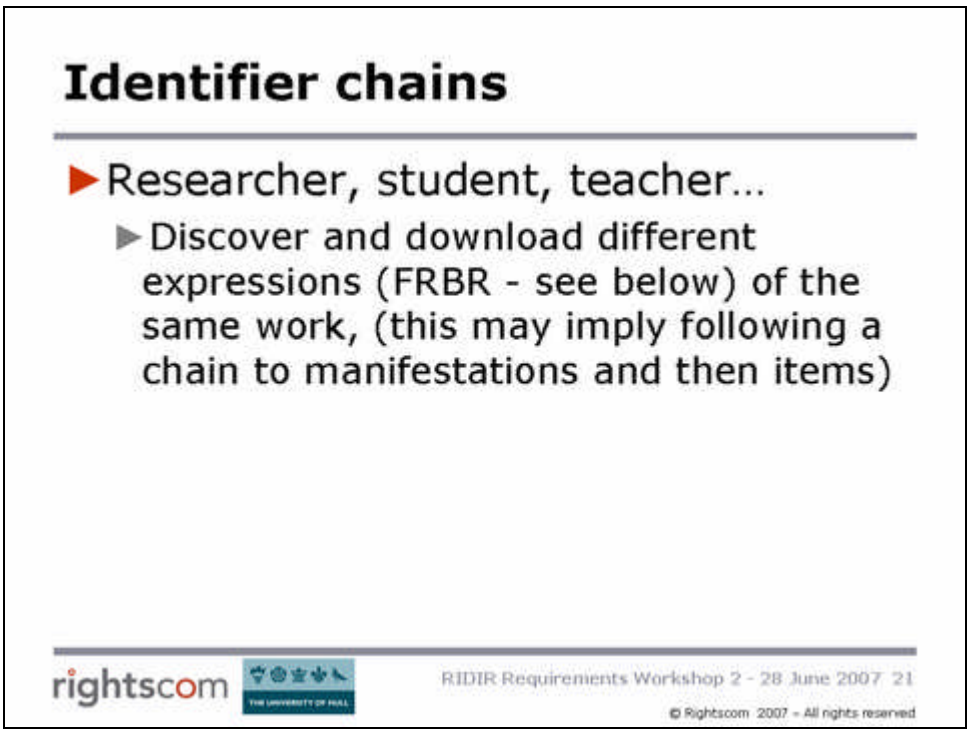

Locate originals of derived components
A teacher who has located an interesting 'compound' learning object and wishes to re-purpose parts of it which may themselves have been repurposed from elsewhere
Identify content used in the compound object
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 16
© Rightscom 2007 - All rights reserved


Track objects with related identifier
Central service?
Where multiple persistent identifiers under (Handles, PURL …) are allocated to the same object access parallel identifier
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 17
© Rightscom 2007 - All rights reserved

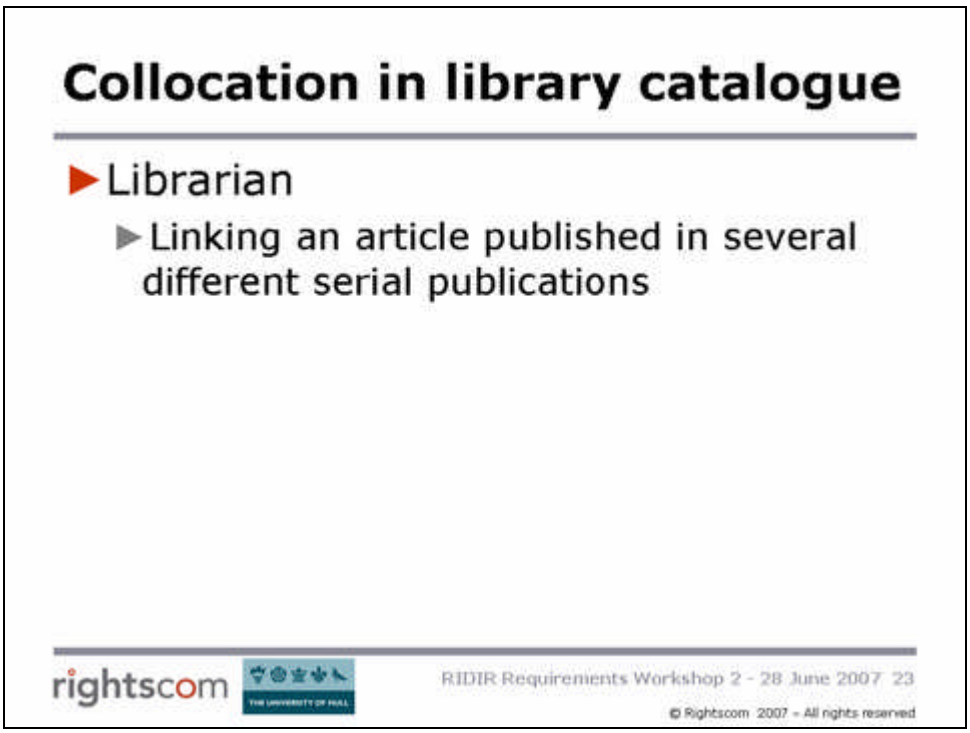

Establishment of comprehensive identity and metadata record to facilitate subsequent discovery of relationships
Creator of a learning object
Creator of a research paper/thesis with particularly complex copyright issues
Using multiple copyright elements in a 'derived' object
Licensed to use the elements subject to the creation of a comprehensive metadata record for the work, which includes the identity of the original materials and their source
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 18
© Rightscom 2007 - All rights reserved

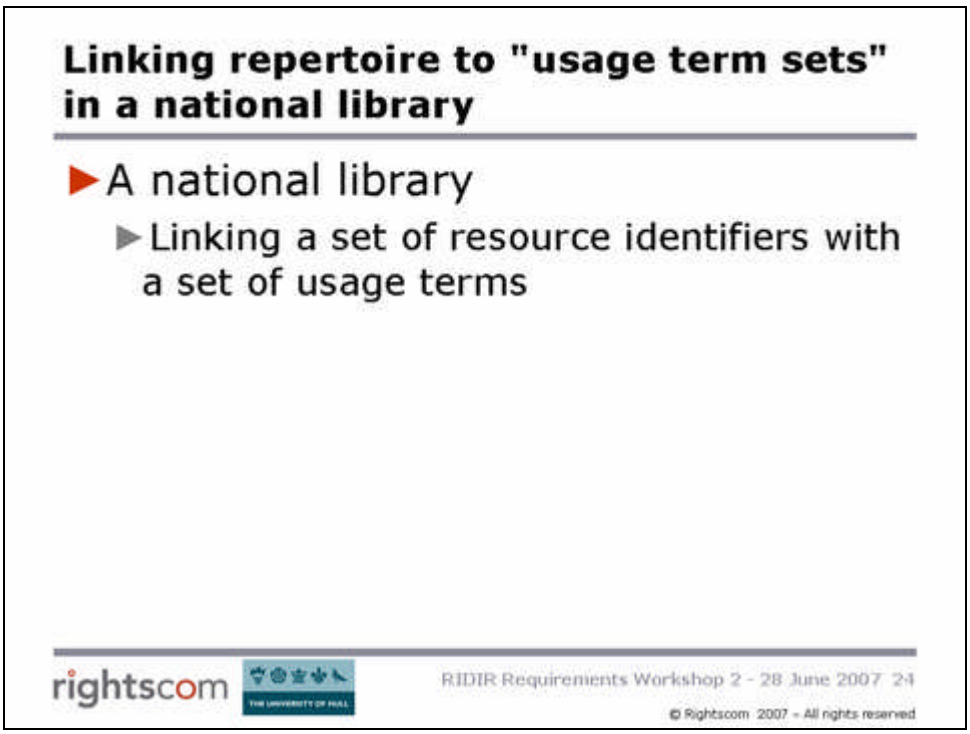

Identifier chains
Academic with a particular research paper
Discover and download parallel publications of the same article
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 19
© Rightscom 2007 - All rights reserved

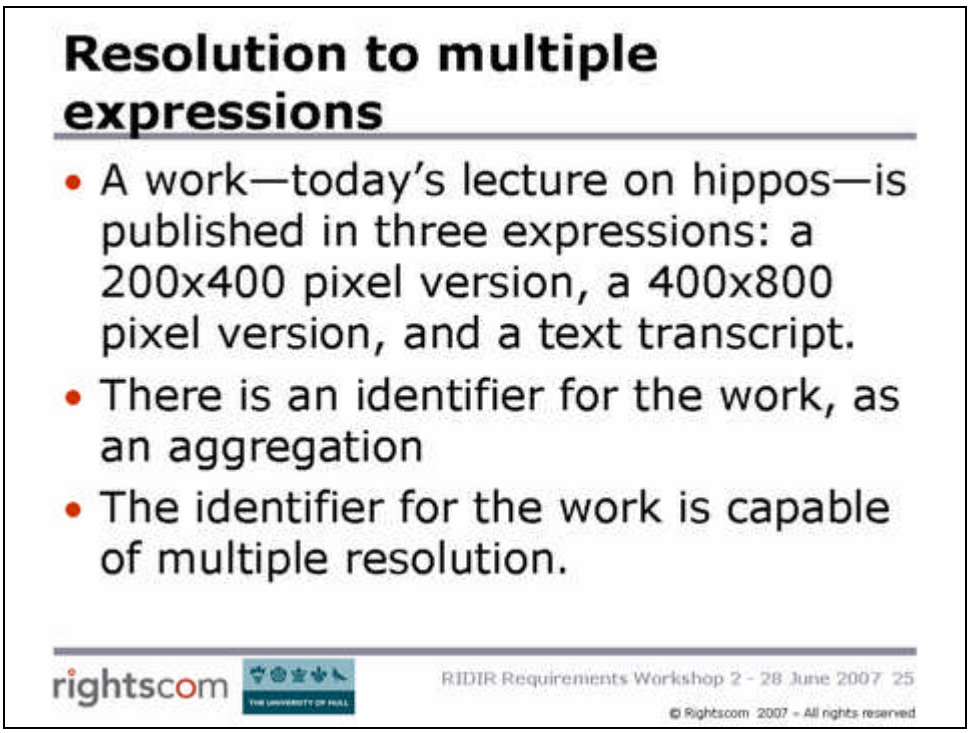

Compiling multimedia objects
e-Learning Course Designer
Discover identifiers for a range of different content types to be included in a multimedia e-learning course, which might include one or more of:
Journal article
Chapter from a book
Audio-visual clip
Sound recording
Photograph
Graphic image
Musical score
Software application
Any of this content might be self created or have rights owned by third party
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 20
© Rightscom 2007 - All rights reserved

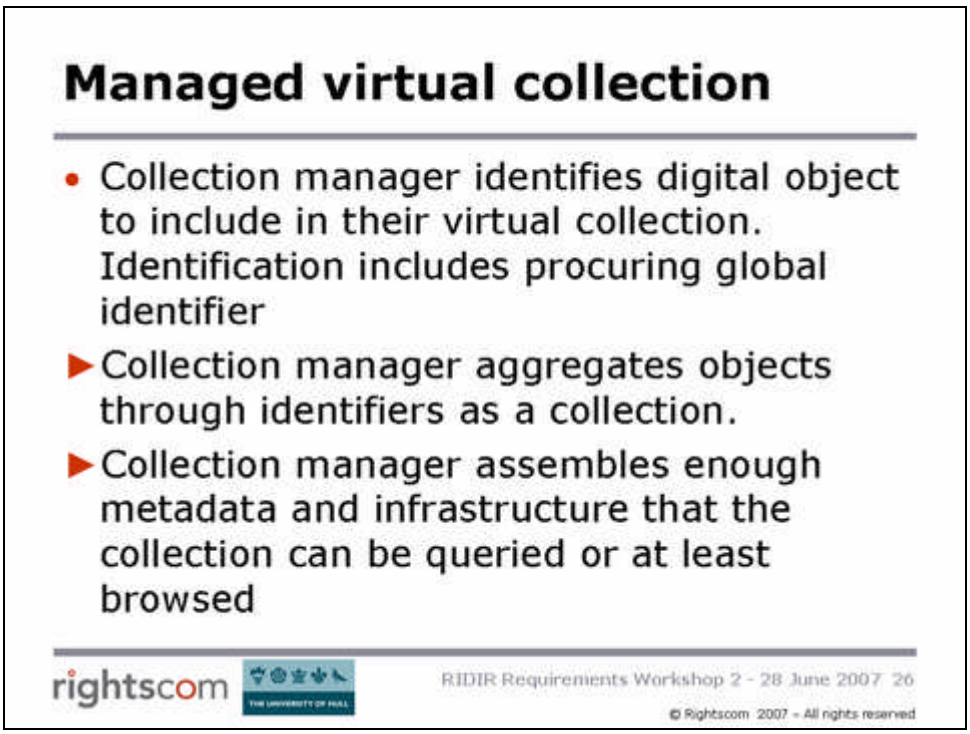

Identifier chains
Researcher, student, teacher…
Discover and download different expressions (FRBR - see below) of the same work, (this may imply following a chain to manifestations and then items)
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 21
© Rightscom 2007 - All rights reserved

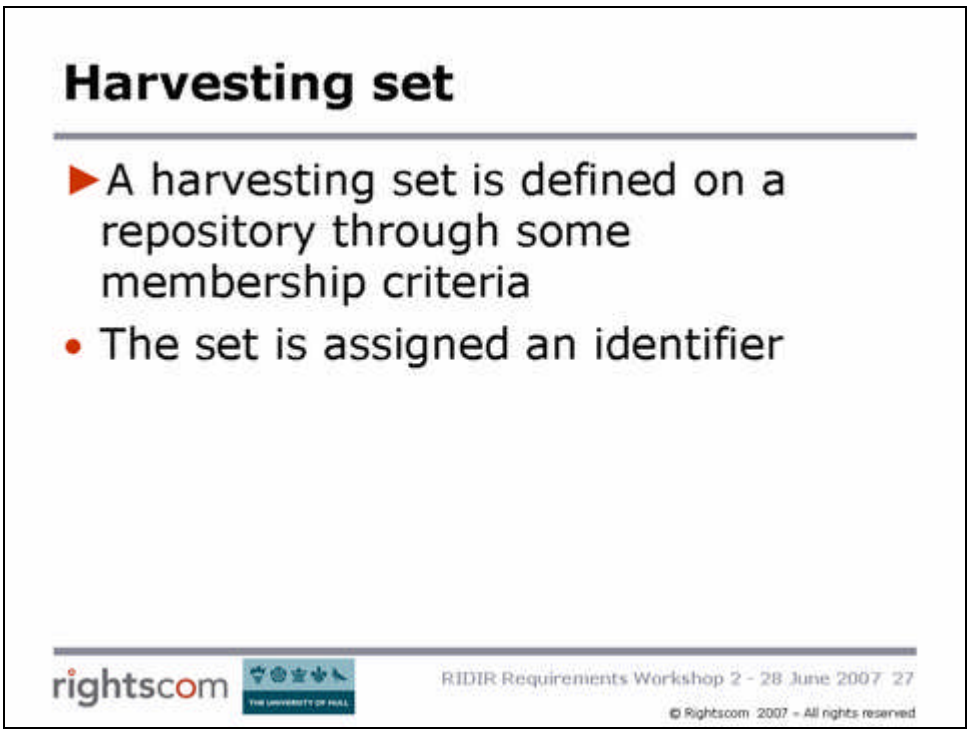

Unambiguous linking of rights to people and/or organisations
User wanting to re-use material in some way
Unambiguously link materials to rightsholders
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 22
© Rightscom 2007 - All rights reserved

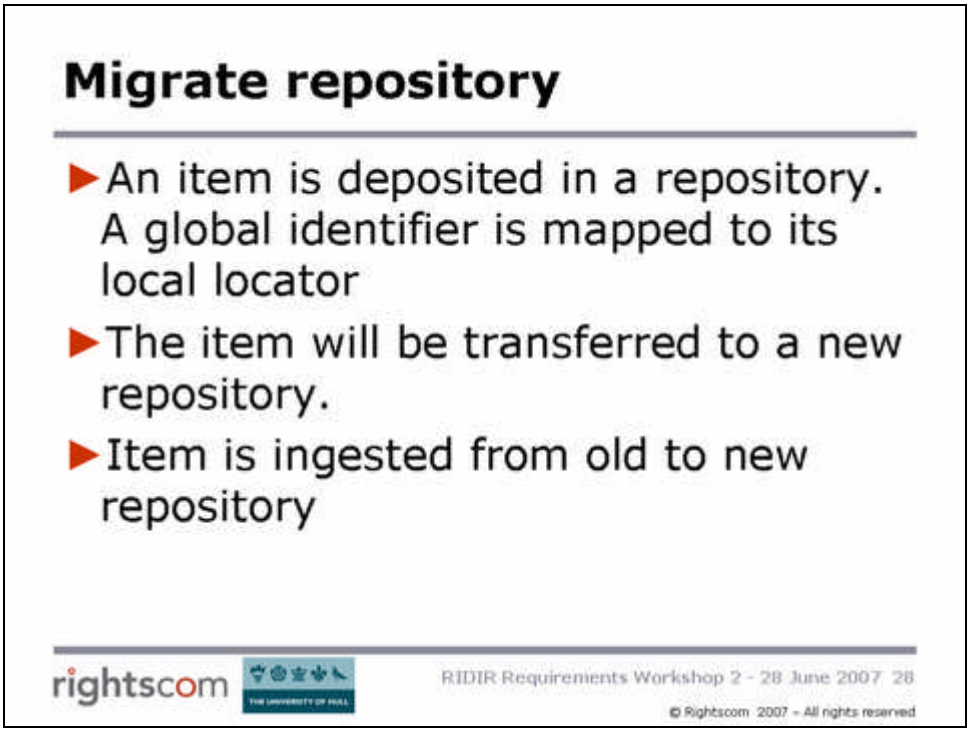

Collocation in library catalogue
Librarian
Linking an article published in several different serial publications
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 23
© Rightscom 2007 - All rights reserved

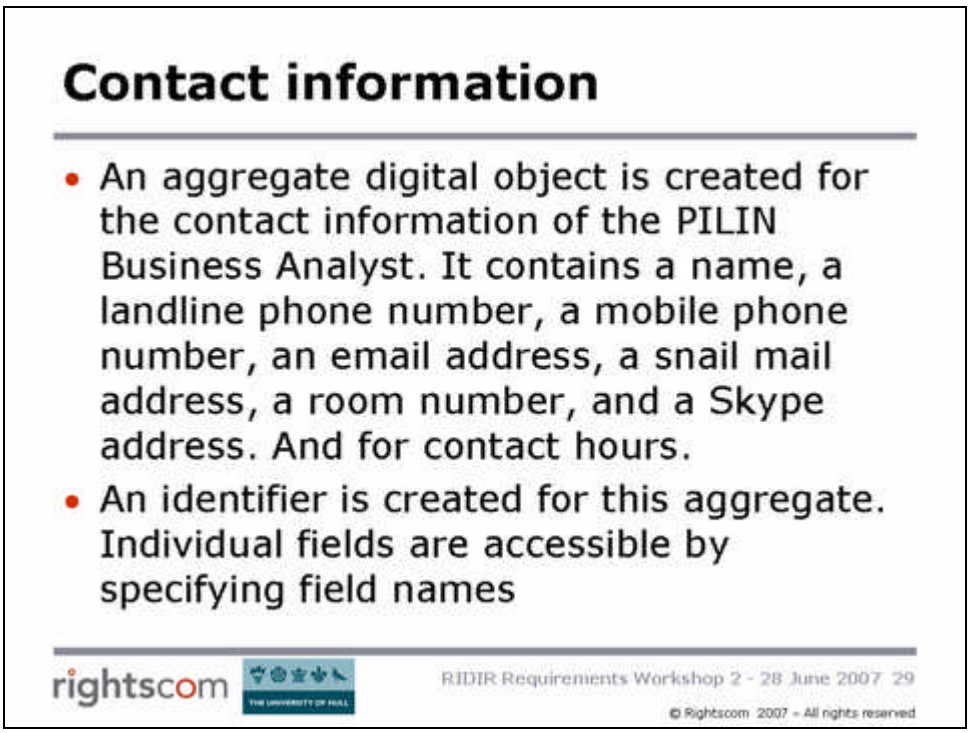

Linking repertoire to "usage term sets" in a national library
A national library
Linking a set of resource identifiers with a set of usage terms
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 24
© Rightscom 2007 - All rights reserved

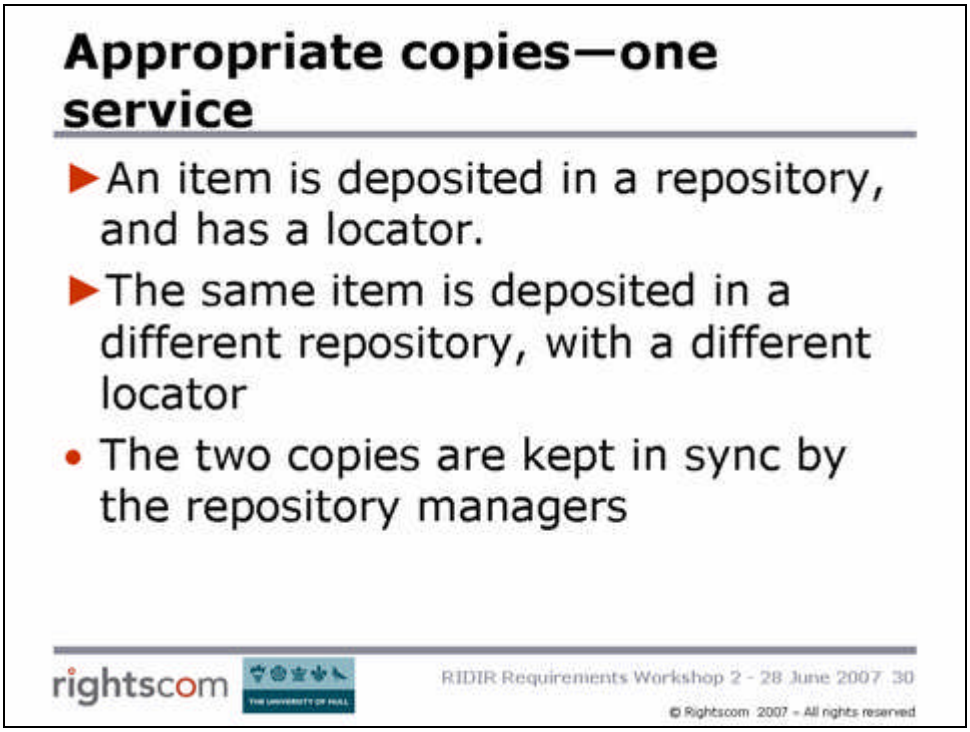

Resolution to multiple expressions
A work—today's lecture on hippos—is published in three expressions: a 200x400 pixel version, a 400x800 pixel version, and a text transcript.
There is an identifier for the work, as an aggregation
The identifier for the work is capable of multiple resolution.
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 25
© Rightscom 2007 - All rights reserved

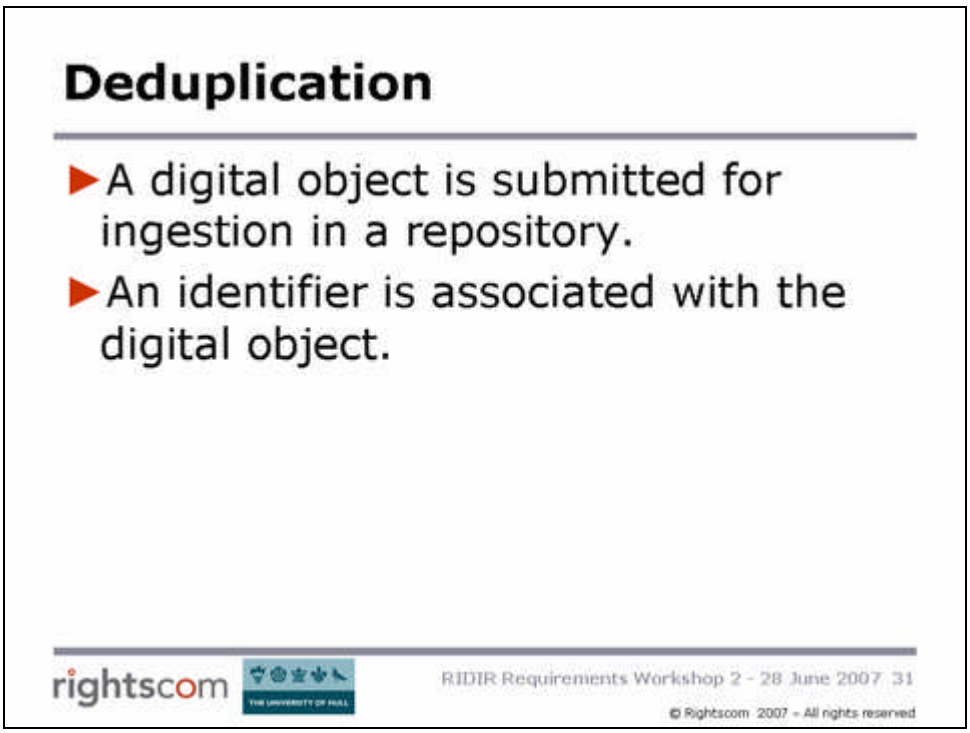

Managed virtual collection
Collection manager identifies digital object to include in their virtual collection. Identification includes procuring global identifier
Collection manager aggregates objects through identifiers as a collection.
Collection manager assembles enough metadata and infrastructure that the collection can be queried or at least browsed
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 26
© Rightscom 2007 - All rights reserved

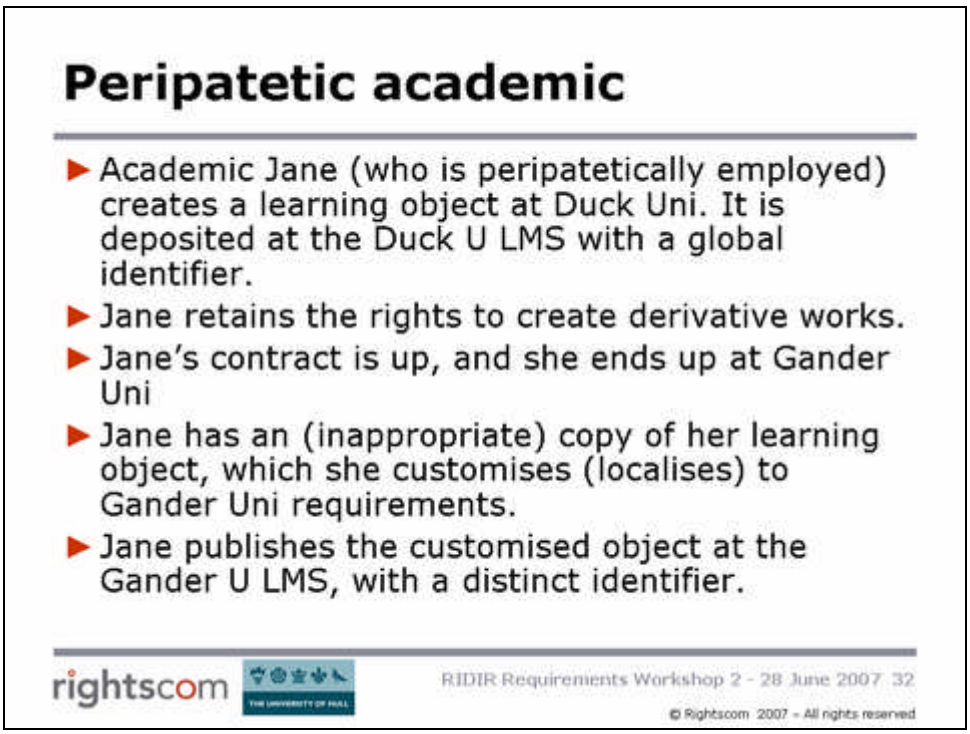

Harvesting set
A harvesting set is defined on a repository through some membership criteria
The set is assigned an identifier
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 27
© Rightscom 2007 - All rights reserved

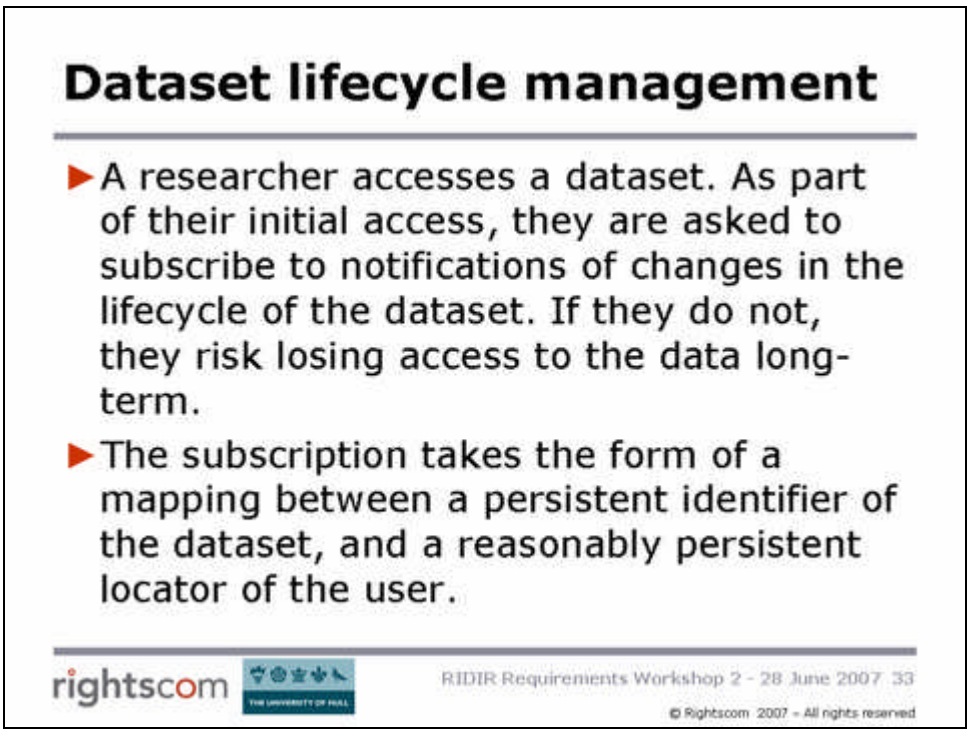

Migrate repository
An item is deposited in a repository. A global identifier is mapped to its local locator
The item will be transferred to a new repository.
Item is ingested from old to new repository
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 28
© Rightscom 2007 - All rights reserved

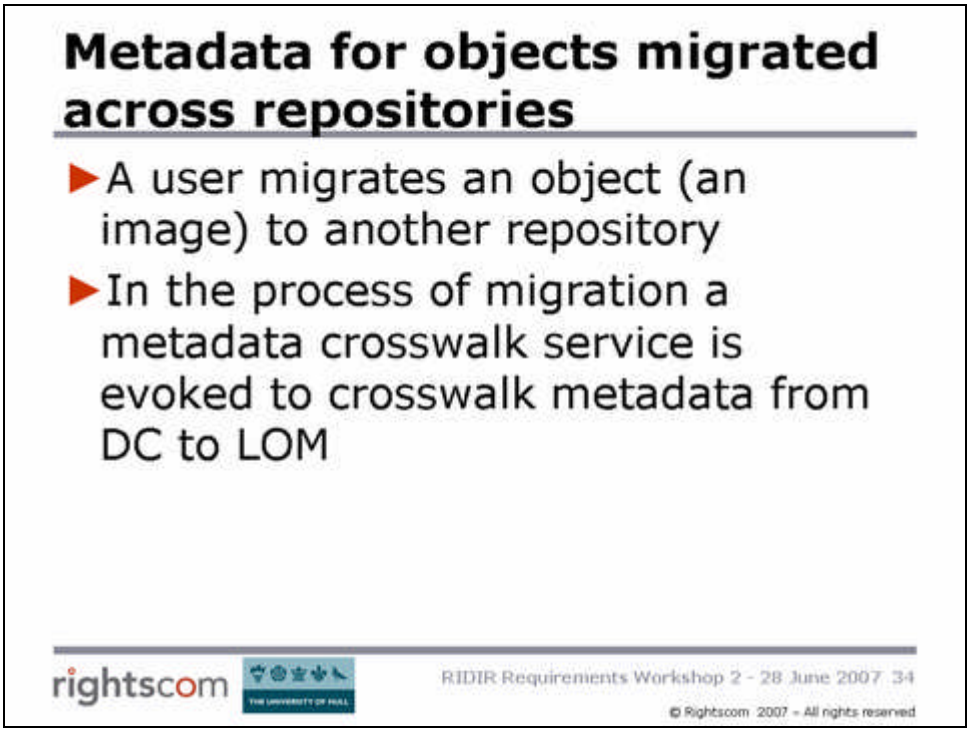

Contact information
• An aggregate digital object is created for the contact information of the PILIN Business Analyst. It contains a name, a landline phone number, a mobile phone number, an email address, a snail mail address, a room number, and a Skype address. And for contact hours.
• An identifier is created for this aggregate. Individual fields are accessible by specifying field names
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 29
© Rightscom 2007 - All rights reserved

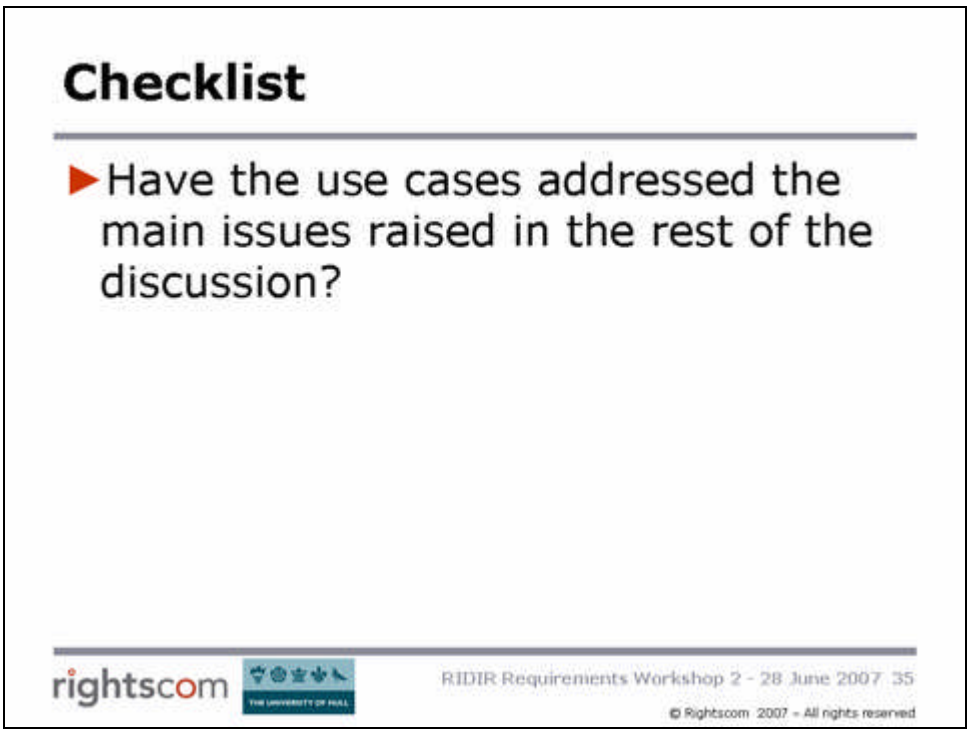

Appropriate copies—one service
▶ An item is deposited in a repository, and has a locator.
▶ The same item is deposited in a different repository, with a different locator
• The two copies are kept in sync by the repository managers
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 30
© Rightscom 2007 - All rights reserved

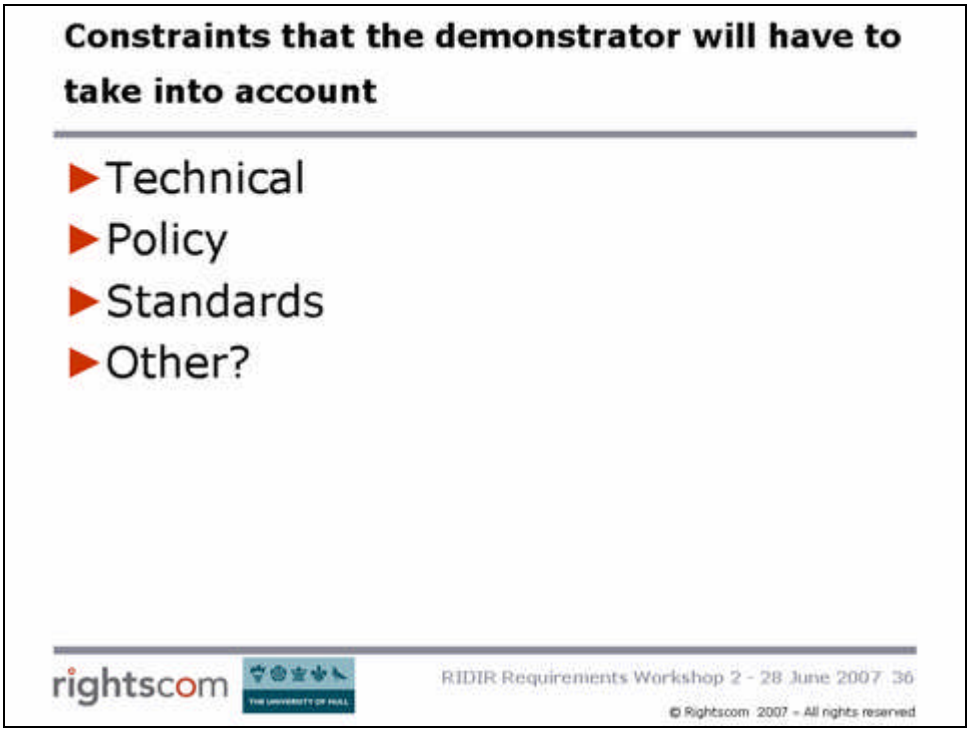

Deduplication
A digital object is submitted for ingestion in a repository.
An identifier is associated with the digital object.
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 31
© Rightscom 2007 - All rights reserved


Peripatetic academic
Academic Jane (who is peripatetically employed) creates a learning object at Duck Uni. It is deposited at the Duck U LMS with a global identifier.
Jane retains the rights to create derivative works.
Jane's contract is up, and she ends up at Gander Uni
Jane has an (inappropriate) copy of her learning object, which she customises (localises) to Gander Uni requirements.
Jane publishes the customised object at the Gander U LMS, with a distinct identifier.
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 32
© Rightscom 2007 - All rights reserved


Dataset lifecycle management
A researcher accesses a dataset. As part of their initial access, they are asked to subscribe to notifications of changes in the lifecycle of the dataset. If they do not, they risk losing access to the data long-term.
The subscription takes the form of a mapping between a persistent identifier of the dataset, and a reasonably persistent locator of the user.
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 33
© Rightscom 2007 - All rights reserved


Metadata for objects migrated across repositories
A user migrates an object (an image) to another repository
In the process of migration a metadata crosswalk service is evoked to crosswalk metadata from DC to LOM
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 34
© Rightscom 2007 - All rights reserved


Checklist
Have the use cases addressed the main issues raised in the rest of the discussion?
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 35
© Rightscom 2007 - All rights reserved


Constraints that the demonstrator will have to take into account
Technical
Policy
Standards
Other?
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 36
© Rightscom 2007 - All rights reserved


Thank you
Contact: Richard Green
RIDIR@jiscmail.ac.uk
http://www.hull.ac.uk/ridir/index.html
rightscom
RIDIR Requirements Workshop 2 - 28 June 2007 37
© Rightscom 2007 - All rights reserved

Appendix 4
RIDIR Business Proposal to the JISC

1 Starting-points

1.1 Current understanding of scope

Our background research and the two workshops we have run have suggested that there is within the repository community a small group of repository managers and technologists deeply engaged in understanding "interoperability" and how to enable the objects in their repositories to be identified in a way that would allow relationships between objects also to be identified; how to make the objects and their relationships sharable with other repositories; and how to benefit from objects and relationships that other repositories are making sharable. There is a much larger group within the repository community that has other pressing concerns that must be dealt with before interoperability and content-sharing becomes an issue. Our experience suggests that this group knows little about the issue and it does not seem to attach a high priority to it.
As is almost always the case where a group of technical specialists is solving a problem in advance of "market need" there is little sign of a middle group: the position is highly polarised. This is both normal and almost inevitable in the case of most forms of innovation. When the first group has done its work, its output will be adopted by the second group as soon as their needs are pressing enough and as soon as any solutions can be implemented by non-specialists in normal operating environments.
The work we have done with those involved with interoperability suggests that they are concerned about two things in particular:

- How to actually make it work
- What the value of making it work would be (including the consequences if it does not work)

A key decision for the project at this stage will be how to balance these in the focus of the demonstrator. The technological approaches required by each may be different.
In either case, we would outline issues and solutions for the approaches *not* taken in the demonstrator.
We feel that the choice between the two approaches should be made by considering the audience for which the JISC intend the demonstrator.
If the audience is intended to be those involved in repository development and management at a 'grass roots' level then we believe that the "value" approach will be best. If we are going to suggest that repositories will need to take a much more systematic approach to the use of identifiers in order to promote interoperability, this group will ask "Why?" and the demonstrator will need to be focussed to answer this question.
If the audience is intended to be those working at the cutting edge of repository development, if you like the "movers and shakers" of the repositories world, then we believe that the "how" approach will be better. With this audience it can be taken as axiomatic that interoperability is a good thing and the demonstrator can be focused on showing how identifiers might best be created and managed to achieve it.
This issue is one that should be resolved in discussion with JISC, rather than by the project team alone. We are currently developing abstract architectures that might support either approach, but need to focus our resources before we finalise those and begin the development work itself.

1.1.1 Value of interoperability

The use cases we have identified show examples of issues where the ability to work with identifiers yields value. In this case, the role of the demonstrator would be to show some of these examples in action. We 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 know to be very limited. We would focus on cases where there is a unique, accessible and usable 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 advantage of this approach is that it would serve as an important demonstration to the repository community of the importance of the role of identifiers in achieving interoperability and therefore encourage them to plan for it, and also to participate in any subsequent policy development as well as adopt working practices that will help the situation in future.
It might also be used to develop a business case for the development of (for example) any necessary Shared Infrastructure Services or other products or services that might facilitate interoperability.

1.1.2 Cost of interoperability

In this approach, using our use cases 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. There is a range of possible ways in which these tasks might be undertaken.
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.

1.1.3 Common components

Whichever approach is taken, "value" or "cost," there will be common elements in the construction of the demonstrator. The choice will determine which aspects of the demonstrator need to be emphasised and thus take more of the project's resources. Amongst these common elements will be some of the functionality of the Persistent Identifier Mediator Service (PIMS), but see also Section 2.
There will also be some important common outputs. These might include recommendations on policies, the relevance of standards, and working practices that would benefit content-sharing and interoperability.
The development of policies and working practices in particular is important: without some level of compliance and shared practice, it is unlikely that interoperability would be feasible in the long-term.

1.1.4 Development directions

It is considered that the approach outlined at 1.1.1, showing the value of interoperability, is the closer fit with the outline expressed in the JISC call to which we responded and we consider this eminently achievable. The second approach outlined at 1.1.2 would address a number of potentially important practical issues in relation to identifiers more generally though potentially it has greater risks in terms of certainty of outcomes from demonstrator development. We have discussed a concern that there would potentially be overlap with the work of the DEST-funded PILIN Project in Australia but have come to the conclusion that either of our approaches would, in fact, be complementary to their work (see Appendix). Having identified (sic) the scope of work that we consider can be carried out, we wish now to engage JISC in the discussion about the preferred route forward to ensure the most appropriate benefit for the JISC community. Our reasons for believing that JISC should be closely involved in this decision are that there may be strategic factors that we are unaware of, there is a trade-off between outcomes to be made, and it is not possible to establish a reliable risk/benefit model for a research and development project.

1.2 Use cases and scenarios

We are developing a range of scenarios to ground our proposal and identify the issues that are likely to arise: these are described in more detail in Section 0. These scenarios are equally applicable whichever audience this consultation identifies as the focus of the demonstrator.

1. EThOSnet

2. Spoken Word Services
3. Migrate repository as an explicit, demonstrated example - both simple and compound objects.
4. Depot
5. Locate original

Taken together, these illustrate some of the different aspects of the main use cases proposed through our workshops:

- Versions with a long chain of connecting events
- Locate originals of derived components also locate 'unknown' children (bi-direction)
- Migrate repositories

We would like to address all these use cases if we can, but it is too early to be sure if resources will allow this.

2 Clarification of proposal

It is appropriate to consider if the objectives of our original proposal, given the additional knowledge that has been uncovered by the desk research and workshops, remain relevant and feasible.

Two key parts of the proposal are:

> ***6.3 Persistent Identifier Mediator Service***
> *The persistent identifier mediator service is the integral element of the identifier interoperability solution whose function is to sit as a shared service between "client" repositories, and ensure that various resources – and their constituent metadata elements – which have repository identifiers "persistent" within the context of an individual repository, are available to any other client repository. If the persistent identifier mediator service were to be moved into a production context it will be required to interoperate as a shared infrastructural service within the context of other shared services within the JISC IE, in particular those dealing with user identity and rights of access for particular identified contexts of use.*

We are not committed to building this as part of the demonstrator, but we remain aware that it may be a very important part of solving the problem in the long term. If we investigate the "value" approach we shall need to understand what such a service might provide and how this might be shown in a demonstrator. If we investigate the 'cost' approach there will be a more obvious role for the PIMS, but one that requires dedicated effort to bring about. We are aware that, in a short, moderately resourced project, there could be a tension between actually building a Mediator module as a discrete entity and simply incorporating Mediator functionality within the demonstrator. As noted above, the focus of our development effort will be considerably influenced by the decision on the intended audience for the demonstrator.

We also said:

> ***6.4 Client Repositories***
> *In order to verify and demonstrate the satisfaction of a key identifier interoperability use case, we propose to transfer sets of digital objects from one repository to another, preserving the identity of the relations between them and the resource identifiers themselves. Each repository will not only be realised in a different technology but also be constructed with a different content model that will map into the interoperable metamodel. This will show how the interoperability solution is able to resolve between identifiers minted for various digital objects within each client repository. We shall also investigate the "edge cases", where preservation of a digital object is difficult, requires manual intervention, or is completely incompatible.*
> *Sample content will be provided by the University of Hull or specifically generated for the project.*

We are still engaged in scoping this. It is clearly very relevant to the 'value' or 'why' approach, but might need some development in detail to ensure that the key points that illustrate value are emphasised in the demonstrator.

3 Scenarios

3.1 EThOSnet

Notes

This poses potential challenges in managing identifier chains between the copies of an e-thesis and how you manage copying and migration of e-theses between repositories. Versioning is not, or should not, be an issue, as e-theses are unlikely to be altered once created.

The EThOSnet project is the second phase in setting up a national e-theses service based at the British Library. The first project, EThOS, concluded in Autumn 2006: it delivered a prototype service based on a heavily adapted version of EPrints 2 software. This is now being scaled up for the live service, which is also called EThOS.

The EThOS service will involve establishing relationships between the central hub repository and institutional repositories on a number of different levels, as described below:

- An institution will hold its own e-theses content in its own repository (with its own identifiers). Metadata about the e-theses, including the identifier and possibly a separate identifier for the metadata itself, will be made available for harvesting by EThOS and become searchable through EThOS. When a user discovers such a thesis in EThOS they will be offered a link back to the institutional repository to access the full content, using an appropriate identifier to maintain the link. Such a link should be persistent if the institutional repository changes.
- An institution may follow the above, but also deposit a copy of the e-theses at the BL (through a deposit tool) for preservation purposes. The BL will place this copy in its preservation repository and associate preservation metadata with this (including its own identifier, though the preservation metadata has yet to be formally agreed completely). The institution will decide which copy the EThOS service should point to (and hence which link is created).
- An institution may submit its paper theses for digitisation by the BL, creating an e-thesis with associated metadata and identifier. All such generated e-theses will be deposited in the BL's preservation repository, and they can also be returned to the institution for inclusion in its institutional repository, dependent on the choice of the institution. In both cases, metadata from the e-theses will be available through the EThOS service, and the necessary link to the full content appended to this.

The BL is deliberately being flexible in order to encourage participation in the EThOS service.

This poses potential challenges in managing identifier chains between the copies of an e-thesis and how you manage copying and migration of e-theses between repositories. Versioning is not, or should not, be an issue, as e-theses are unlikely to be altered once created.

Within the community of e-theses it may be feasible to establish a commonly agreed resource to assist in tracking where e-theses sit. This would also assist in establishing and having confidence in the authenticity of a thesis.

3.2 Spoken Word

This illustrates the issues surrounding parent/child object relationships in particular: how children can be linked back to the parent.

A national broadcasting organisation has a large library of media clips which it makes available to other such organisations. Each of the clips has an identifier conforming to the organisation's adopted standard for such things. The organisation keeps a set of metadata for each object which facilitates search and discovery by registered users of its library, but which also records information relating to rights management and royalties such as each transmission of the clip. In order to keep this metadata as complete as possible they want to be able to 'crawl' or otherwise interrogate the libraries of other national broadcasting organisations worldwide which may have copied the clip. The purpose of the crawl would be to pull back any relevant metadata in the remote library (including transmission information) to enhance their own record. This process would be more straightforward if the second broadcaster includes the

original identifier in their metadata; if they do not, then it may be possible to identify candidate objects in a remote repository by matching other characteristics.

3.3 Repository migration

It is likely that with the development of technology and with institutional reorganisation, migration of repositories from one platform to another will be a regular, if not frequent, activity. The use case will consider the identifier issues involved in such a migration, especially those where an object may be related to other objects and those relationships need to be preserved. This is essentially the use case for Section 6.5 of the original proposal.
In this scenario, we have envisaged some technical issues representative of ones that could typically arise. The institution may have an established repository which uses an 'out-of-the-box' solution such as EPrints or DSpace; and the repository contains both simple and compound objects. They now wish to migrate its contents to a more flexible and extensible solution based on, say, Fedora; in doing so, they wish to maintain the relationships expressed in the existing compound objects. They wish to ensure that, following the migration, any use of an identifier associated with the 'old' repository results in an appropriate action on the object in the 'new' repository.
Such migration will not always happen en-masse: it is very likely that there will be a need for frequent or regular migration of individual objects or small groups.

3.4 Depot

The JISC-funded Depot service provides a potentially interesting scenario. The Depot is a holding place for academics whose institution does not have a repository in which to deposit their own articles. In due course, it is anticipated that the objects in the Depot will be transferred to an institutional repository as these come on-stream. There will thus be a need to ensure that identifier information and relationships are not lost during the process. When an object is transferred from the Depot to an institutional repository it is currently envisaged that the handle for the object will be changed to reflect the new location. It is possible, though, that the institutional repository will create its own, visible, identifier for the same object. It would be appropriate to expose a formal relationship between the two identifiers.

3.5 Location of original material where a robust chain of identity is not present

A member of staff at a University discovers in a remote institutional repository a particularly interesting piece of text/digitised old film. It seems to illustrate well a point in some material they are preparing, but the staff member wants to establish the context from which this extract originally came. If (s)he does then want to use this, or the original, it will be necessary to establish the arrangements made (if any) for derivative rights. The metadata associated with the object (s)he has discovered is inadequate to this purpose and does not provide an explicit identification of the original repository or object from which this derivative came. The original needs to be located and that information made available to others who may later have a use for it.

4 Key points

To summarise the position that we have reached to date, and the key decision that we have to make in consultation with the JISC Programme Manager:

1. Our desk research, workshops and discussions have revealed that very few repository managers are aware of or concerned about this issue at the moment;
2. It has therefore been hard for us to identify many real use-cases and requirements;
3. Those that we have been able to identify are "real-world" but are complex and specialist and at the leading edge of practice;
4. Everyone seems to agree that more general requirements will follow, but it's hard to predict when there will be a significant number of repositories with a reasonable volume of content for which interoperability will be an issue;
5. The generic use-cases and scenarios that we have identified will illustrate the need for coherent policies for the creation, mediation and management of identifiers and their relationships within repositories and also what might happen if policies are not in place or are not followed;
6. They are also likely to illustrate the scope of benefits and capabilities that could be attained from such changes;
7. They may also illustrate the kind of decision- and practical-support that might be needed to improve the creation of identifiers and relationships;
8. We also know that there will be many repositories that do not follow policies or practice because they were created before the value of good policy and practice was recognised;
9. We also know that it is unrealistic to expect every repository to follow policy closely, so there will always be many objects that are not identified in a consistent fashion, and in particular whose relationships with one another may not be identified at all;
10. We are therefore proposing to focus on a demonstrator that either shows the value of identifiers and the relationships between them that can support interoperability (and the loss of value that might be entailed in not doing so), or that directly addresses some of the issues in a way that suggests how the creation, mediation and therefore management of identifiers might be handled;
11. If resources permit, we will also aim to scope the policies in outline; investigate what the risks are if they are not developed and used; and identify what might be done in cases where there has been no policy applied.
12. The two possible approaches we have outlined each have benefits and risks attached. The "value" or "why" approach perhaps best fits the outline in the JISC call but we are aware that it would preclude detailed work on the "cost" or "how" approach; it may be that the JISC would prefer the "cost" or "how" approach for strategic reasons or in the light of the information produced during our requirements-gathering process. It is appropriate that we seek the input of the JISC Programme Manager at this stage to help us decide the appropriate course and to clarify 'the JISC view' on elements within it.

The implication of this is that while Ridir cannot present a complete solution to the challenge of identifiers and their relationships because of its restricted scope and resourcing, it can provide the foundation for other projects to build on; without Ridir or something very like it, further progress is very unlikely.

Appendix to the business proposal

At the time the RIDIR Project was established, the JISC sought reassurance that its work would not duplicate that of the DEST-funded PILIN Project in Australia. This document also refers to the PILIN work.

Having met with PILIN, and being given access to the PILIN project wiki, our current understanding is that the two projects would be complementary, and any overlap would assist with future re-use of both project outcomes.

Being focussed on persistent identifiers, PILIN would be expected to share certain sorts of use case, and indeed several of these were used as material for the RIDIR workshops. However, it is our current understanding that the focus of PILIN is on infrastructure and services, and identifier management within the context of that infrastructure, regardless of what the semantics are that users of identifiers in PILIN ascribe. Complementary to this, RIDIR is focussed on investigating the means by which identifier interoperability can be achieved, regardless of which infrastructure is used at the "back-end," based on metadata definitions and identification of precisely what those identifiers refer to, and mediating between these. Seen in this context, the outcomes of each project clearly share mutual territory – an analogy might be with a rail transport system: PILIN would build the track, others, for example CNRI and DOI, the carriages/engines and trains respectively, and RIDIR guides how the carriages and trains fit and work together with respect to the overall function of transporting people and goods. For example, the RIDIR architecture is likely to contain a repository for the identifier-related artefacts it needs to store in order to function: the role of this repository could in principle be fulfilled by PILIN infrastructure services in due course.

We quote below from the PILIN Project wiki in order that readers can understand these points:

The PILIN Project

The emphasis in the PILIN Project will be on building identifier management infrastructure based on a technology (Handle) that is now under development through the auspices of CNRI to underpin sustainable global identifier infrastructure. PILIN aims to meet a specific need common to e-Research communities, the proposed work to be undertaken will be transferable to other communities, such as the VTE sector, the Le@rning Federation and the TILIS Project. The project aims to take advantage of existing governance and consultative mechanisms within the ARROW environment to ensure relevant and sustainable outcomes and optimal return on investment. The project will be run in partnership between ARROW and the University of Southern Queensland (USQ), specifically through the RUBRIC Project.

PILIN Aims and Objectives

- Support adoption and use of persistent identifiers and shared persistent identifier management services by the project stakeholders.

- Plan for a sustainable, shared identifier management infrastructure that enables persistence of identifiers and associated services over archival lengths of time.

Project Outputs

- Best practice and policy guides for the use of persistent identifiers in Australian e-learning, e-research, and e-science communities.

- Use cases describing community requirements for identifiers and business process analysis relating to these use cases.

- E-Framework representations of persistent identifier management services that support the business requirements for identifiers.

- A “pilot” shared persistent identifier management infrastructure usable by the project stakeholders over the lifetime of the project. The pilot infrastructure will include services for creating, accessing and managing persistent digital identifiers over their lifetime. The pilot infrastructure will interoperate with other DEST funded systemic infrastructure.

- The development phase of the pilot will use an agile development methodology that will allow the inclusion of “value-added” services for managing resources using persistent identifiers to be included in the development program if resources permit.

- Software tools to help applications use the shared persistent identifier infrastructure more easily.

- Report on options and proposals for sustaining, supporting (including outreach) and governing shared persistent identifier management infrastructure.