Optimising Metadata Workflows in a Distributed Information Environment

R. John Robertson and Jane Barton

Centre for Digital Library Research, Department of Computer & Information Sciences, University of Strathclyde, Livingstone Tower, 26 Richmond Street, Glasgow G1 1XH, UK
robert.robertson@cis.strath.ac.uk, jane.barton@strath.ac.uk

Abstract. The different purposes present within a distributed information environment create the potential for repositories to enhance their metadata by capitalising on the diversity of metadata available for any given object. This paper presents three conceptual reference models required to achieve this optimisation of metadata workflow: the ecology of repositories, the object lifecycle model, and the metadata lifecycle model. It suggests a methodology for developing the metadata lifecycle model, and illustrates how it might be used to enhance metadata within a network of repositories and services.

1. Introduction

In a distributed information environment (dIE), comprised of repositories and repository-based services (hereafter repositories), a clear understanding, in the form of a model or framework, of the movement of metadata is required.[1] This need arises as repositories move from pilot to operational status and have to address two key problems: how to ensure the quality of their metadata can support the functionality they offer and how to operate within a sustainable budget. The community has reached a point where getting a repository to work is not the problem – the critical factor is finding resources to keep it working. Part of the solution is a reference model, describing the movement of metadata within the dIE and enabling the academic community to harness their collective efforts and implement more efficient metadata creation, augmentation and enhancement processes throughout the metadata lifecycle. This paper will begin to develop the conceptual components of a metadata lifecycle model and to illustrate how such a model can be used to optimise metadata workflows on a local and wider scale.

A holistic approach to metadata workflow and its optimisation allows the workflow used to create or transform metadata in an individual repository to take into account, and where possible exploit, metadata creation and transformation processes taking place elsewhere in the dIE. These processes generate the potential for metadata

[1] Weibel's understanding of metadata interoperability (and by extension metadata itself) as having structural, semantic, and syntactical aspects is assumed throughout [1].

enhancement as, within the dIE, diverse purposes exist due to the different user communities creating and managing repositories and the different expectations and requirements of those communities. Just as the services of museums and libraries differ, so too will the functions of their digital embodiments and the metadata required to support them.[2] This may involve using different metadata standards, different controlled vocabularies within the same standard, or different approaches to recording names. Even within repositories with similar purposes, differences in metadata requirements may occur due to the organisational culture of those administering the repository, or the scale of the repository, amongst other variables. Clearly, no repository would wish to record every possible piece of metadata about the items it chooses to manage. Rather, each repository will have different metadata requirements for the same object and will adopt different methods of generating that metadata. This diversity works against the notion that there can be a one size fits all approach to repository metadata requirements.[3] As such diversity is often seen a challenge to interoperability and efforts are made to develop standards and controlled vocabularies which are as generic as possible in an effort to overcome it. These efforts are necessary for many reasons, yet it is these challenges for interoperability which allow metadata workflow optimisation. When different repositories record diverse metadata about the same object interesting things can happen.

2. Towards Optimising Workflow

The widespread realisation of the potential benefits of these differences for metadata workflow requires a model of the dIE (including detailed profiles of repositories and their relationships), which could facilitate strategic partnerships, inform divisions of labour and funding, and foster a holistic approach to the creation, augmentation and enhancement of metadata. To achieve this two conditions must be met: firstly, the local workflow must be articulated; and secondly this local workflow must be placed in the context of the wider environment, so that its relationship with other relevant workflows can be understood and acted upon. As repositories have to compete for funding and strive to create more efficient and sustainable working practices, the first condition, the articulation of local workflow, is being met. Those repositories using workflows for object management are now also articulating the metadata creation process at the point of object ingest (e.g. the digital preservation community [4], [5]). The second of these conditions requires an understanding of the dIE at various levels of granularity as follows:

- an ecology of repositories
- an object lifecycle model
- a metadata lifecycle model

---

[2] These tensions are well illustrated by a repository serving both purposes [2]

[3] At best there can be regions of such agreement on some requirements within a dIE (very tightly-controlled distributed repositories, such as learning objects in a military environment as discussed by Collis and Strijker [3], are effectively a single entity within the dIE as they don't so much interoperate as integrate).

2.1 The Model of an Ecology of Repositories

An ecology is both "the relationship between organisms and their environment" [6] and the study of those relationships. Within a dIE repositories inevitably relate to each other, these relations whether driven by common interest or external pressure (e.g. funding body) create an ecology where the actions of one repository or service can have a significant effect on others. Most of the effort in developing such a model for repositories is attempting to classify the different types of repositories, the levels of relationship between them, or their domains. This is the most developed of the three reference models, even though as yet those working on it are not modelling the relationships between repositories.

2.2 The Model of an Object Lifecycle

An object lifecycle model attempts to profile the movement, transformation, and adaptation of digital objects within the dIE. This is the hardest model to develop as there is currently little agreement about or experience of what happens, or even what should happen, to objects within the dIE. This is because issues arising over object use, movement, re-use, re-purposing, granularity and content-packaging are obstacles for objects within large sections of the dIE, where constraints due to ongoing discussions over ownership and identifiers, among others are hindering the intentions of many repositories. Fortunately these obstacles are recognised and solutions are under development.[4] These attempts to address object management questions contribute to the object profile part of this lifecycle model, but developing the model to show how objects evolve within the dIE will require resolution of these issues.

2.3 The Metadata Lifecycle Model

The metadata lifecycle model attempts to profile the metadata held in a repository and its movements and transformations within the dIE. The development of the model is not dependent on the unresolved issues in the object lifecycle model as they can be contained within that layer. This paper will suggest the nature of a metadata lifecycle model, a way to develop it, and its relevance to the optimisation of metadata workflow.

2.4 Relating these Reference Models to Existing Models

There are existing models which relate to parts of the three conceptual models suggested above. They are: the E-Learning Framework (ELF), the cosmic wheel of McLean and Blinco, the JISC Information Environment, the model being developed by CORDRA, and the formal model developed by Gonçalves et al. The ELF is attempting to develop a "common approach to Service Oriented Architectures for

[4] For example, the efforts to produce an agreed rights expression languages [6]

education". It is creating a definitional model of service components and developing standards and tools to support their interoperability. Its model is addressing a particular domain of the dIE and it can provide a typology of functions in that part of the ecology of repositories [7]. McLean and Blinco have further developed a service domain typology of repositories; their 'cosmic' view of repositories space develops a more comprehensive if necessarily less-detailed typology of repository service domains for the ecology [8]. The JISC Information Environment can be viewed as a specific implementation of the ELF with the specific focus of providing "convenient access to a comprehensive collection of scholarly and educational materials". It models a superstructure to co-ordinate technical infrastructure development and support integration and interoperability of e-resources for further and higher education in the UK [9]. Like the ELF it focuses on technical solutions to support structural and syntactical interoperability. It is also taking a lead in developing solutions to some of the thorny issues in the object lifecycle model. The approach of the new CORDRA initiative is to enable access to a wide range of learning object repositories through federated searching [10]. As such it is attempting, through negotiation and standardisation, to create a high common denominator for participating distributed repositories – this particular approach will create a community of repositories (and an interoperability boundary) but has difficulties in a larger context as it assumes federation as the ideal method of repository interaction, as such it cannot easily take advantage of metadata workflow optimisation. Within its bounds it is attempting to produce metadata that can be integrated (rather than interoperate), this will raise the cost of participation and limit it to the education community.[5] Gonçalves et al are developing a complex formal taxonomy of repositories [12]. It attempts to comprehensively catalogue repositories through providing five views of a repository. Although it engages with concepts present in all three of this paper's conceptual models, it is perhaps of limited use in that it is not attempting to address repository relationships, object lifecycles, or metadata lifecycle and only offers a static view of a repository.

Overall, existing models address the structural or syntactical aspects of repository interactions but have thus far underdeveloped the modelling of semantic interactions. Existing models are about the voices, vocabulary, and grammar of repositories. Metadata workflow optimisation requires this modelling to be extended to also profile what repositories do and are interested in discussing -- personal ads for repositories if you will.

3. A Methodology for Developing a Metadata Lifecycle Model

For a metadata lifecycle model to support the intelligent harvesting or exchange of metadata records, elements and other information (e.g. controlled vocabularies), a metadata profile within it would ideally describe a repository's metadata at element

[5] This is of course its intention; the analysis is solely from the point of view of a wider dIE. There are also other attempts to develop architectures for repositories but they are similarly focused on issues of technical interoperability e.g. [11].

level. It would comprise of a profile of the metadata produced to meet local demands, plus any requirements to meet external commitments. These building blocks would then be joined together by processes which move or transform the metadata as it travels between repositories. The effort required to produce such a model at element level however, is not initially practicable.[6] A more feasible approach is to create a metadata profile that describes in more general terms a repository's metadata requirements (structural, semantic, and syntactic). Such a profile would allow the most relevant external workflows to be pinpointed. Individual repositories would then create an element-level profile of selected metadata relationships, and integrated metadata workflows incorporating intelligent harvesting or exchange could be facilitated.

As part of the development of integrated workflows, this internal metadata profile would then be supplemented in the model by other components. These map the processes that repositories or services carry out when importing or exporting data (such as normalisation, adding collection metadata etc.), but also the processes they carry out to make data suitable for agreed exchange (e.g. with a partner site or large-scale service).[7] Within the model there will be regions which have similar requirements for agreed exchange, these correspond to the notion of interoperability boundaries - the requirements a repository must meet to participate in a service or other repository community. Although the metadata implemented by a repository should reflect the purpose of the repository, the metadata model is agnostic about what a given repository does. The relationships within the metadata lifecycle model could be represented along the following lines:

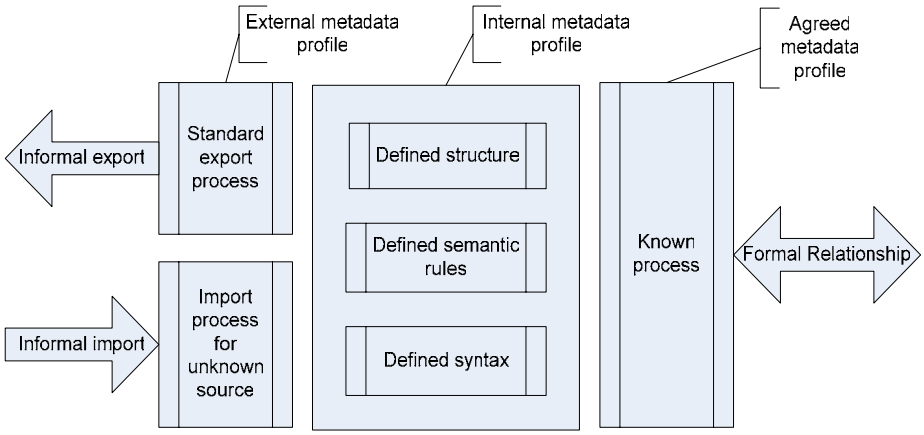

Fig. 1. A building block for the metadata lifecycle model

[6] Similarly attempts at any of the three reference models will initially occur for sections of the dIE only, as some communities will be able to develop and exploit partial models more quickly than others.

[7] This distinction between metadata for internal requirements and the 'published' metadata is useful as this transformation process may be more negotiable than the internal workflow.

4. Using the Metadata Lifecycle Model for Workflow Optimisation

The value of the model to repositories lies in its potential to exploit, through formal (i.e. not random) relationships, known metadata sources elsewhere in the dIE and so optimise local repository workflow. Where persistent identifiers are used, there is some benefit to be gleaned from the informal harvesting of multiple metadata records for a given object (parallel to the approach in a union catalogue); however, significant and ongoing effort is required to manage such imports. In a formal relationship known records (or elements) are the harvested and imported much more efficiently as they have known properties and quality. Consequently importing the data can be more automated and there is the possibility negotiated changes in what the harvested site exports. Similarly an awareness of interoperability boundaries (i.e. points within the dIE where a minimal level of metadata quality is required to participate) provides the basis for the establishment of minimal relationships with groups of repositories. The model also provides a basis for discussing the cost and benefits of participating in consortia or implementing metadata elements in particular ways. The value of the model for workflow optimisation is further illustrated by the following scenarios:

1. The NSDL operates what is effectively a mini dIE. They break harvested metadata records into their component elements and assign provenance metadata to them (in effect creating a metadata profile of sorts for each harvested repository). This approach allows them to create an optimum record by combining metadata elements from different sources, with human input at collection-level only. As much of the content is created for them, they have known relationships with those repositories. They can then create customised profiles and import processes for the truly external repositories and effectively turn them into known sources. Their approach, though one-way rather than interactive, is proof of the benefit of known relationships in optimising metadata and demonstrates that intelligent harvesting already works under some conditions [13].

2. A repository of learning objects using the LOM wants to harvest metadata records about zoology. They have crosswalks and mappings to convert metadata structure and syntax from a selection of other metadata formats. To decrease the amount of work they have to do in adjusting importing records to fit their implementation they want to find repositories using similar guidelines for the creation of semantic information (e.g. using the same guidelines for writing descriptions or a particular controlled vocabulary).

3. A federated search service wants to be able to dynamically select search targets that can support a user's choice of subject scheme (e.g. MESH).

4. A departmental repository wants to enhance its records but can't demand more from its metadata creators. It discovers that its records are harvested by a relevant subject repository and its own university library. It re-harvests the domain specific subject terms from the subject repository and the general

subject terms from the university library and incorporates these into its records, and is now able to support subject access to its materials.

5. A central repository harvests metadata and using automatic processes improves it by adding missing file type, file size, and language metadata. It then re-exposes this metadata for harvest. The original source re-harvests the improved record.

6. A virtual museum is trying to add a view of its digital collections for teachers. It discovers that a college repository has harvested their metadata and is adding educational descriptions to it to use the images in their courses. The museum harvests the educational metadata created by the college and negotiates with them to produce metadata for other parts of their collection.

5. Developing the model and the optimisation of workflow

This model will enable the whole community to develop an understanding of how best to create each piece of a metadata record. In theory, any given repository could create every piece of metadata required to support every possible desired access point for an object, but in practice resources are limited. Optimizing metadata workflow ensures that local workflows capitalise on all the resources available, both locally and elsewhere in the dIE. In this way repositories can expand their metadata element set without compromising on quality and so expand what functions they can offer. Higher level repositories can speed up their ingest processes through known relationships and support more automatic transformations and enhancements of metadata.

To move the development and use of these models forward there needs to be an agreed way to describe repositories on these three levels. This would allow exemplar optimised workflows to be developed and promoted. If these models are also to develop as a technical or machine readable solution, methods of developing and integrating existing registry projects which record repositories, standards, application profiles, and controlled vocabularies need to be encouraged and expanded. Developing the metadata lifecycle model is an extension of utilising these existing developments. As they are used and populated essential components become available and metadata workflow optimisation comes one step closer. Even if the development of a ‘reference index’ based on these models proves impracticable in the short term, encouraging repositories to think about models in this way will advance how they describe themselves and decide workflows and so promote the development of instances of metadata workflows that take advantage of the wider dIE.

References

1 Weibel, S.L.: The Metadata Landscape: Conventions for Semantics, Syntax, and Structure in the Internet Commons. Metadiversity: Proceedings of the Conference, Natural Bridge, VA. (1998).

2 Caplan, P., Haas, S.: Metadata Rematrixed: Merging Museum and Library Boundaries. Library Hi Tech **22**, 3 (2004) 263-269

3 Collis, B., Strijker, A.: Technology and Human Issues in Reusing Learning Objects. Journal of Interactive Media in Education, Vol. 4. Special Issue on the Educational Semantic Web. (2004)

4 Rauch, C.: Workflows in Digital Preservation. ERPANET Workshop on Workflow, Budapest Hungary. Available at http://www.erpanet.org/events/2004/budapest/presentations/Workflows_in_Digital_Preservation_2004-10-13.pdf (2004)

5 DELOS Digital Preservation Cluster: DELOS Summer school 2005. Available at http://www.dpc.delos.info/registration/ (2005)

6 Dictionary.com Available at http://yourdictionary.com/ (2000)

7 Wang, X.: MPEG-21 Rights Expression Language: Enabling Interoperable Digital Rights Management. IEEE Multimedia (2004) 84-87

8 The E-Learning Framework. Available at http://www.elframework.org/ (2005)

9 McLean, N.: The Ecology of Repository Services: A Cosmic View. Keynote address given at ECDL 2004 Available at http://www.ecdl2004.org/presentations/mclean/ (2004)

10 JISC: Strategic activities: Information Environment. Available at: http://www.jisc.ac.uk/about_info_env.html (2003)

11 Krann, W., Mason, J.: Issues in Federating Repositories: A Report on the First International CORDRA™ Workshop. D-Lib Magazine, Vol. 11 **3** (2005)

12 Schek, H., Türker, C.: The Work and Vision of Work Package 1: Digital Library Architecture. Delos Newsletter **2** (2004)

13 Gonçalves, M.A., Fox, E.A., Watson, L.T., Kipp, N.A.: Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries. ACM Transactions on Information Systems, Vol. 22 **2** (2004) 270-312

14 Hillmann, D.I., Dushay, N., Phipps, J.: Improving Metadata Quality: Augmentation and Recombination. DC-2004: International Conference on Dublin Core and Metadata Applications, Shanghai, China (2004)