Building Quality Assurance into Metadata Creation: an Analysis based on the Learning Objects and e-Prints Communities of Practice

Jane Barton
Centre for Digital Library Research, University of Strathclyde, UK
jane.barton@strath.ac.uk

Sarah Currier
Centre for Academic Practice, University of Strathclyde, UK
sarah.currier@strath.ac.uk

Jessie M. N. Hey
University of Southampton Libraries and
Intelligence, Agents, Multimedia Group, University of Southampton, UK
jessie.hey@soton.ac.uk

Abstract

This paper challenges some of the assumptions underlying the metadata creation process in the context of two communities of practice, based around learning object repositories and open e-Print archives. The importance of quality assurance for metadata creation is discussed and evidence from the literature, from the practical experiences of repositories and archives, and from related research and practices within other communities is presented. Issues for debate and further investigation are identified, formulated as a series of key research questions. Although there is much work to be done in the area of quality assurance for metadata creation, this paper represents an important first step towards a fuller understanding of the subject.

Keywords: *Metadata creation, quality assurance, learning object repositories, open e-Print archives, resource discovery.*

1. Introduction

Communities of practice are recognised to be increasingly important for creating, sharing and applying organisational knowledge. A community of practice is a relatively loose, distributed group of people connected by a shared interest in a task, problem, job or practice [1]. Here we take the opportunity to explore an issue concerning two parallel communities of practice which have emerged within the academic environment in recent years. One is based around principles of sharing and reusing learning objects in e-learning delivery, facilitated by the use of digital learning object repositories, which may be institutional or shared across communities or subject areas. The other is developing open archives of e-Prints, typically comprising published papers and pre-prints, although they may include other research outputs such as reports and theses; initially these were subject-based but more recently fledgling institutional archives have been appearing. Both of these areas are underpinned by the concept of interoperability for educational resources and systems, and by a growing awareness of the need to optimise the value of resources created within educational institutions [2],[3],[4]. The two communities also share a number of goals, including ensuring the long-term sustainability of digital resources and systems in education, minimising the cost of creating and providing access to resources for individuals and institutions, and improving access to a wider variety of learning materials for teachers and learners on the one hand and to the latest research for academics and researchers on the other.

Standardised metadata is central to interoperability; at its best it is a powerful tool that enables the user to discover and select relevant materials quickly and easily. At worst, poor quality metadata can mean that a resource is essentially invisible within a repository or archive and remains unused. Clearly metadata quality has an important role to play in realising the goals of learning object repositories and e-Print archives, and much effort has already gone into developing standardised approaches to metadata structure, but as yet the issues surrounding the creation of good quality metadata within that structure have received surprisingly little attention.

In this paper, we seek to challenge four of the assumptions which underlie both the absence of inquiry into how metadata should best be created, and the trend for authors of learning objects and e-Prints to create the metadata for their own resources. These four assumptions are:

- that, in the context of the culture of the Internet, mediation by controlling authorities is detrimental and undesirable.

- that rigorous metadata creation is too time-consuming and costly a barrier in an arena where the supposed benefits include savings in time, effort and cost.
- that only authors and/or users of resources have the necessary knowledge or expertise to create metadata that will be meaningful to their colleagues.
- that given a standard metadata structure, metadata content can be generated or resolved by machine.

Repositories and archives are now being more widely implemented and practical problems resulting from a poor understanding of the metadata creation process are beginning to emerge. It is therefore timely to scope the issue of metadata creation with a view to quality assurance for repositories and archives, by drawing together the few studies so far published, the practical experiences of learning object repositories and e-Print archives, and related and potentially useful research and practices within other communities, including the library community. From these, we will identify issues for debate and further investigation, formulating them as a series of research questions. We will conclude by revisiting the assumptions put forward above to see whether they hold true for our communities of practice.

2. The development of learning object repositories and e-Prints archives

Much discussion, research and exploratory work has been applied in the area of learning objects and interoperability, moving towards a future "learning object economy" [5], where teachers, course developers and learners involved in online education will be able to share and re-purpose digital learning materials. In recent years, various projects have been developing repositories of reusable learning objects [6], supported by international standardization work, most notably the suite of specifications produced by the IMS Global Learning Consortium. Downes suggests that the next stage of development in this "economy of education" should be the development of a distributed learning object repository network [6].

The metadata work in this area has mainly centred on the development of the world's first formal e-learning standard, the IEEE Learning Object Metadata (known as the LOM), which was ratified by IEEE in 2002 [7]. The IEEE worked closely with interoperability specification bodies, including IMS, in creating the LOM; it is integral to other IMS specifications such as IMS Digital Repositories Interoperability [9]. In the UK and elsewhere, key work is being done in developing good practice and common usage of these specifications [10],[11]. However, there has been little formal investigation of the processes involved in actually creating metadata that describes learning objects. In fact, from the start, the issue has been elided. In his 2001 paper on the necessity for a learning object economy, Stephen Downes, considered by many to be one of the seminal thinkers in e-learning, had only this to say on the issue:

> "The authoring of metadata itself will be straightforward for most course designers. Because metadata files are machine-writable, authors will simply access a form into which they enter the appropriate metadata information." [5].

However, there is a growing number of repository development projects in the UK whose early experiences suggest that there is more to the creation of good metadata than simply filling in an online form, as will be shown in Section 4.

Meanwhile, the e-Prints community has adopted a standards-based interoperable framework within which metadata can be harvested from individual data providers and delivered to end users via centralised or federated services. The initial emphasis has been on producing a low barrier mechanism for achieving this by creating the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [12] to harvest XML-formatted metadata and mandating Dublin Core as the common metadata format [13]. Version 2 of the OAI-PMH has achieved the relative stability of being a production release. Early implementations and prototypes were predominantly subject-based, but more recently a number of institutional archives have begun to appear.

Having identified a practical technical solution, some of the emphasis has shifted to examining and changing the culture within academic institutions so as to encourage deposit, with the wider goal of changing the increasingly unsustainable economics of scholarly communication:

> "The development of institutional repositories emerged as a new strategy that allows universities to apply serious, systematic leverage to accelerate changes taking place in scholarship and scholarly communication." [14].

One example of activity in this area is the current lobbying for mandatory deposit of the full text of publications by academics in their institutional repositories for the UK's Research Assessment Exercise [15]. Another local example is the University of Glasgow's Create Change initiative [16].

Although there is a greater recognition of the need for quality assurance for metadata creation within the e-Prints community, the current focus on participation means that anything that is perceived as a barrier between academics and their parent institutions tends to be played down. However, metadata quality has a profound bearing on the quality of service that can be offered to end users, particularly in a federated system, as the examples presented in Section 5 will demonstrate, and this in turn may have a detrimental effect on long term participation.

3. The need for quality assurance in metadata creation

So, given the existence of all this work, why is there a need for further quality assurance? The key to answering this question is to separate out the concepts of structure and content. The developments mentioned above deal primarily with the structure of the metadata, whilst this paper is concerned primarily with the content of the metadata fields. Once a metadata standard has been implemented within a system, the specified fields must be filled out with real data about real resources, and this process brings its own problems. For end users, these problems manifest themselves in various ways, including poor recall, poor precision, inconsistency of search results, ambiguities and so on. They arise due to errors, omissions and ambiguities in the metadata, many of which are known and understood in other communities of practice, often having tried and tested solutions.

Some, but by no means all, of the areas where problems commonly arise, and where quality assurance is needed to achieve a corresponding quality of service for users, are outlined here:

- Spelling, abbreviations and other such errors and ambiguities which occur at the data entry stage.
  This issue is nicely illustrated by Doctorow:
  "Even when there's a positive benefit to creating good metadata, people steadfastly refuse to exercise care and diligence in their metadata creation. Take eBay: every seller there has a damned good reason for double-checking their listings for typos and misspellings. Try searching for "plam" on eBay. Right now, that turns up nine typoed listings for "Plam Pilots". Misspelled listings don't show up in correctly spelled searches and hence garner fewer bids and lower sale-prices. You can almost always get a bargain on a Plam Pilot at eBay." [17].
- Author and other contributor fields.
  If the same person's name is entered differently each time, if they get married, for instance, or if initials are used inconsistently, you won't retrieve all of their works when you search using only one representation of their name. Conversely, if there is more than one person with the same name, the search results will be ambiguous. Similar problems arise around corporate names, used in fields such as author affiliation or publisher. These problems are fairly readily addressed by applying rules and conventions, and through the use of authority files.
- Title.
  This is a surprisingly difficult area. Determining the title of a published paper or PhD thesis may be relatively straightforward, but many resources have more than one possible title, for example composite learning packages, while others, particularly non-textual resources, may have no title at all. In these cases, we have to determine who decides what the title of a given resource is and according to what criteria. The library community has an extensive set of rules to deal with this issue.
- Subject, in the form of keywords and classifications.
  This is one of the most difficult, and most controversial, areas of metadata creation, and a detailed discussion is beyond the scope of this paper. The basic problem is this: who is best placed to add subject-related metadata for maximum resource discovery, the author, who may know the subject area and its terminology well, or a metadata specialist, who may be better placed to step back and think about all the potential users of a resource, or about consistency of subject terms and classifications across a repository, archive or network, so that like really is classified with like. The use of taxonomies and subject classification schemes is part of the solution, but in turn creates other problems, as demonstrated in the SeSDL case study at Section 4.3.
- Date.
  Two sets of problems arise here. The format of the date - whether to use 1 October 2003 or 01/10/03 - is fairly trivial and lends itself to machine solutions. However, the issue of semantics - what the date actually means, for example whether it refers to the date of creation or the date of publication - is more complex and requires an understanding of the context in which the metadata is being created and the uses to which it will be put.

At a local level, the context in which the metadata is being created can have a bearing on the importance of quality assurance, particularly as it relates to specific fields. In some cases, the larger the dataset, the greater the likelihood that a problem will manifest itself. For example, in a large population of authors, name authority files may be needed to disambiguate one John Smith from another. In other cases, the degree of diversity can determine whether the quality of the metadata becomes an issue. For example, in an archive of papers and reports originating from a single research group, author affiliation can be set as a default value, whilst a subject-based archive may need to use corporate name authority files to ensure that papers can be retrieved effectively by organisation. However, in an environment where each repository or archive is part of a wider system predicated on interoperability, the importance of quality assurance for metadata creation goes far beyond that which the local context might suggest. The possible population of authors is that of the whole world, the diversity limitless. Metadata that supports successful resource discovery perfectly

adequately in the local context may not be as effective in an aggregated system.

There will always be some aspects of the metadata that are inaccurate, inconsistent or out of date, even in systems which have extensive quality assurance procedures in place and have invested heavily in the creation of good quality metadata. For example, when a published subject classification scheme is updated, new resources may be classified using new subject terms but existing resources may not be reclassified, giving rise to inconsistent subject-based searches. Furthermore, in established systems, there may be a drift over time between policy and practice; a study into cataloguing practices in Scottish libraries as part of the CAIRNS Project [18] found this issue to be widespread. Nevertheless, it is essential that quality assurance is built into the metadata creation process at the outset, that its scope extends beyond the local context and that the resulting metadata is as 'good' as it can be within the inevitable limitations of time and cost.

4. Evidence from the learning objects community

In line with the development of e-learning standards and specifications, a growing number of learning object repositories are now being implemented. Some of these repositories are beginning to encounter problems with the metadata creation process and to report that the quality of their metadata is having adverse effects on resource discovery. The three case studies described below give an indication of the range of issues that have already emerged and demonstrate that although pragmatic solutions can generally be found on a case-by-case basis, there is clearly some way to go before cost-effective and scalable approaches to quality assurance become established within this community.

4.1. The Higher Level Skills for Industry Repository (HLSI)

This project is currently developing a repository for digital learning objects that aims to support the delivery of learning programmes in the broad subject areas of engineering and manufacturing at a level ranging from high school to higher education [19]. Based at the University of Huddersfield, UK and funded by local development agency Yorkshire Forward, it is implementing the IEEE LOM v.1.0 and has collected approximately 6,500 objects in a variety of sizes and file formats, together with metadata records. Resources are generally uploaded by their authors, who add the metadata themselves:

> "The people who submitted resources also provide the metadata, which gives them some ownership over the records. The drawback is that the quality of metadata varies." [19].

The assumption that those who submit resources want "ownership" of the metadata records is interesting and points to an underlying cultural issue within the community that warrants further investigation.

The project's problems with metadata quality are detailed further as:

- the same metadata records were applied to many or all components of a package of educational content
- the terminology used by the metadata authors was not consistent
- when searching the repository the terminology used by the metadata authors was interpreted in different ways
- some metadata authors described the facets and characteristics of the educational object and not the educational content of the object
- the software allowed default values and these were over-utilised [20].

This has been found to have an adverse effect on the performance of the repository and as a result three steps are being taken to improve the metadata creation process:

- explaining why metadata is important to resource authors
- providing more documentation to guide authors through the process of entering metadata
- employing cataloguers to validate resources and improve the metadata.

The results of this last step are being recorded and analysed and will be written up as a research paper later in 2003. Ryan notes that, as of June 2003, 2,500 metadata records have been re-edited, taking about 550 hours and costing around £6,500, or about £2.60 per record [20].

In conclusion:

> "The HLSI project team considers obtaining consistent metadata content to be a major difficulty. The technical obstacles involved with metadata were considered less difficult to solve." [19].

4.2. The Bolton Woods Local History Project

This project was a community-based initiative in which members of the community created digital resources, mainly family and local history materials, which in turn were used as informal learning resources by their peers. A small study, funded by BECTa, was carried out to investigate whether the creators of the resources could also create metadata for their resources, and to assess how well they could do this in comparison with the information specialists involved in the project [21].

The key findings of the study are as follows:

- In general terms, resource creators did not have a good understanding of the purpose of metadata or an appreciation of its value;
- Resource creators did understand and appreciate the context of their resources and focused on these elements within the metadata;
- Information specialists had a better understanding of the purpose of metadata and included a wider range of metadata elements;
- Information specialists "struggled" with contextual aspects of the metadata;
- Neither the resource creators nor the information specialists handled pedagogic aspects of the resources well.

The study concludes that a collaborative approach to metadata creation is needed to optimise the quality of the metadata in this context [22].

4.3. The Scottish electronic Staff Development Library (SeSDL) Taxonomy Evaluation

SeSDL was an early, seminal project investigating the creation of a learning object repository based on IMS specifications, including the IMS Learning Resource Meta-data specification (v1.1). The project brought in a librarian to create a subject-specific classification scheme, the SeSDL Taxonomy [23]. A small-scale peer evaluation of the Taxonomy was carried out, in which six consultants, drawn from the project's user community, were provided with eight learning objects, or granules, to be classified using the Taxonomy. While this evaluation was not designed to assess the proficiency of users in creating metadata, it did provide some interesting results of relevance. Even with guidance notes and explanations provided for the purposes of the evaluation, the ability of the consultants to understand and carry out the task varied considerably. One consultant commented in the post-evaluation focus group:

"The whole exercise has given me more admiration and respect for librarians." [23].

To summarise, the results of the evaluation seem to indicate that users of SeSDL will assign a wide variety of classifications to their granules, and will do so fairly inconsistently in comparison with each other. This means that learning objects listed under a particular branch of the SeSDL browse tree may appear to browsers to be randomly or inconsistently classified. This in turn could have an impact on the users' perception of the quality of the repository as a whole, and on their willingness to keep searching. The Evaluation Report concluded with a number of recommendations, the most pertinent of which relate to user support, as follows:

> "One of the main areas highlighted by this evaluation was the necessity for adequate user support in classifying granules whilst uploading them. Without this support, the classification of granules is likely to be so inconsistent as to make the browse tree unusable." [23].

5. Evidence from the e-Prints community

With the development of global networks, traditional scholarly communication practices have been transferring to electronic form, speeding up access by involving authors in the publication deposit process; the scientific publication archive of e-Prints, developed by Paul Ginsparg, now known as arXiv being a pioneering example. As the number of archives from different disciplines has increased, so the need for cross search services has arisen. Technical solutions have developed around the concept of harvesting metadata from 'open' archives which are compliant with the same protocol into a federated search service. With this technical challenge overcome but with the diversity of authors and users increasing, the issue of metadata quality is becoming more visible, as the following examples demonstrate. The issue must be addressed by current projects, such as those within the two European programmes described below, which aim to develop institutional e-Print archives alongside archives already established for specific disciplines.

5.1. The experience of prototype federated search services

The Universal Preprint Service (UPS) Prototype was developed as a proof-of-concept ahead of the first meeting of what was to become known as the Open Archives Initiative [24]. The UPS Prototype used the NCSTRL+ protocol [25] to harvest about 200,000 records from a number of existing archives of scholarly material and made them available to the end user through a single service interface [26]. The project encountered significant metadata-related problems:

> "The lack of quality of the metadata available in the UPS Prototype project has an important, baleful influence on the creation of cross-archive services as well as on the quality of services that can be created." [26].

The following year, the Arc service became the first federated search service based on the OAI protocol [12]. It grew out of the UPS Prototype but was able to take advantage of the greater capabilities and wider uptake of the new protocol to move beyond the prototype stage and offer a fully fledged service. However, the quality of the metadata on which the service relied continued to represent a significant problem:

> "Construction of this prototype demonstrated several issues that are likely to recur in any attempt to build an OAI service provider. The effort of maintaining a quality federation service is highly dependent on the quality of the data providers. Some are meticulous in maintaining exacting

metadata records that need no corrective actions. Other data providers have problems maintaining even a minimum set of metadata and the records harvested are useless." [27].

The OAI community has focused on machine-based solutions to problems of metadata quality and this was certainly the case with both the UPS prototype and the Arc service. Techniques such as automatic generation of authority files were implemented at the harvester end of the system with some degree of success. However, the limitations of this approach are acknowledged:

> "Even extensive interventions during the metadata conversion phase could not prevent the negative impact that poor metadata quality has on the search and linking facilities developed in the course of the project." [26].

Arc researchers seem to be referring implicitly to a need for quality assurance in the metadata creation process when they comment that:

> "There is a limit to the quality of services that can be offered on metadata from archives that allow free text entries from contributors for fields such as 'subject', 'type' and 'language'." [27].

The UPS team are more explicit, saying:

> "In order to solve this problem, data enhancement procedures need to be run to improve the quality of existing metadata in archives. In parallel with that, an exploration of submission techniques is required, in order to identify ways in which the data quality can be improved at the source, without demotivating authors by requiring them to submit material with lengthy and complex submission mechanisms." [26].

The TARDIS project [28], described in more detail below, is specifically addressing the submission process for e-Prints with the aid of librarians and a human computer interaction expert.

5.2. The next step: improving access to institutional resources

Cross searching different disciplines introduces various new issues which may fundamentally impact on quality and consistency. This problem is compounded when the whole spectrum of disciplines is encountered, as within an institutional e-Print archive, and is particularly important in the context of interdisciplinary research and inter-institutional collaboration. Countries such as the Netherlands and the UK have been putting new national programmes in place to encourage the disclosure of institutional resources and to research the issues involved. While not identical in scope, their approaches complement each other.

In the Netherlands, the SURF programme, Digital Academic Repositories (DARE) [29], is a significant joint initiative of the Dutch universities, announced in 2002, which aims to make all their research results digitally accessible. A common approach has been adopted, which should encourage consistency in the metadata. The standards used are being chosen to be robust in relation to future advances and are closely allied with international developments, enabling information to be exchanged nationally and internationally in a highly efficient way. In 2003 the focus will be on two main goals:

- implementing the basic infrastructure by setting up and linking repositories within participating institutions;
- starting and promoting the submission of scientific content to these repositories.

The second of these goals is inextricably linked to the quality of the metadata, as quality of service will be an important factor in enlisting and encouraging champions for the submission process. However, it is not yet clear how this issue will be addressed within the SURF programme.

In the UK, the JISC-funded Focus on Access to Institutional Resources (FAIR) programme [30] commenced in August 2002 and will run for three years. Inspired by the success of the Open Archives Initiative, the FAIR programme aims to evaluate and explore different mechanisms for disclosure and sharing of content to fulfil the vision of a web of resources built by groups with a long term stake in the future of those resources, but made available to the whole community of learning. Within the programme, the e-Prints and e-Theses cluster of projects are investigating a variety of issues which complement each other. Some projects, such as Project DAEDALUS, are developing e-Prints and e-Theses archives within a single institution [31]. Some are focused on a single issue, for example Project RoMEO, which will investigate the addition of rights metadata fields [32]. Others are more broadly based, such as Project SHERPA [33] which aims to create a substantial corpus of research papers from several of the leading research institutions in the UK by establishing e-Print archives which comply with the OAI-PMH using the free GNU EPrints software [34]. Advocacy, or the fostering of a culture of participation within institutions, is a key element of many of the projects. From the outset it is recognised that:

> "Advocates need to ensure that their attempts to persuade colleagues of the advantages of open archives should be accompanied by new services to enable those colleagues to self-archive more easily. Examples of such enabling services might be assisting researchers with copyright issues, and self-archiving by proxy." [35].

Metadata assurance issues are starting to be addressed by the TARDIS project [28], which is exploring the most effective options for e-Print archiving using both self-archiving and mediated deposit. The aim is to build a sustainable multidisciplinary archive with which to leverage the

research output of the institution, and in this context it is trialing simpler interfaces to the GNU EPrints software to encourage quality metadata entry for academics from different cultural backgrounds. Learning from some of the inconsistencies produced in early local databases, it is now testing the value of targeted help, more logical field order and citation examples created by information specialists to steer the author. However, where authors are daunted by either the quantity or quality of their own efforts at input, then a mediated service is also being offered and evaluated.

As yet, most FAIR projects do not have strategies in place to deal with the issue of quality assurance for metadata creation within their fledgling archives. This is in part due to the fact that as yet many institutional archives have very little content, such that metadata creation is not happening in a 'real' context. For example, much of the initial content in Project DAEDALUS's e-Print archive was authored within the university library itself, while metadata for new content is being created by project staff, also within the library. Based on the current level of activity, it may be some time before the problems associated with metadata creation by e-Print authors begin to manifest themselves. Project staff are aware that the issue must be addressed at some point, but for now the need to encourage participation among its academic staff outweighs the need to create metadata of an acceptable quality in a sustainable and scalable way [36].

However, a key feature of the FAIR programme is that similar projects are clustered together, with mailing lists and joint meetings. This presents an easy mechanism for sharing experiences, discussing common problems, such as metadata quality, and evaluating possible solutions. This should ensure that the findings of individual projects are disseminated across the programme in a timely manner and that effective strategies can be put in place as institutional archives move beyond the pilot phase and begin to amass significant amounts of content. The result should be a range of effective e-Prints search services covering both subject based archives and institutional archives, and deposit processes that work well for the communities they are serving.

6. Relevant research from other communities of practice

One key study which has taken place outside the core academic community provides food for thought on the subject of author-generated metadata. The study [37], at the National Institute of Environmental Health Sciences in the US, investigated the hypothesis that resource authors can create metadata of sufficient quality to support effective resource discovery on an organisational web site. A second strand to the study investigated whether a simple web form, with textual guidance and selective use of features such as drop-down menus, could assist authors in this process. The Dublin Core schema was adopted and a controlled metadata creation experiment was carried out, also a survey of authors' views on metadata creation.

The results of the study indicate that, with the assistance of a simple web form, resource authors can indeed create good quality metadata and in some circumstances may be better placed to do so than metadata specialists. Authors recognised not only the value of metadata but also the value of their own contribution to the metadata creation process, although Greenberg notes that authors may be more reluctant to participate when metadata creation is seen as "a bureaucratic order or extra chore as opposed to an option that has rewarding benefits" [37].

This study builds on previous work within the same community [38] and elsewhere to develop "metadata metrics", that is, a set of criteria on which the evaluation of metadata can be based. Greenberg has also developed a metadata generation framework [39]and notes:

> "Decisions about the processes, persons and tools to employ for metadata generation depend on a project's architecture, complexity of desired metadata schema, time allotment and project deliverables and the availability of human, financial and time resources. Clearly, different combinations of these metadata generation components will be more effective in different environments. Research efforts testing various combinations of processes, people and tools will help establish useful models to guide metadata generation activities." [39].

Greenberg's team are currently developing a model to facilitate efficient and effective metadata generation for web-based resources within scientific research centres by integrating human and automatic processes [40]. A number of other research projects, based in the Centre for Natural Language Processing at the University of Syracuse, have also been investigating the automatic generation of metadata for text-based resources, and the implications that this has for the development of the Semantic Web [41]. As with the TARDIS project, some of this work brings together librarians and human computer interaction specialists, in this case to evaluate the effectiveness of automatically generated metadata.

7. Key research questions

The evidence suggests that good quality metadata is a key component in the successful implementation of learning object repositories and open institutional archives, yet the issues surrounding the creation of good quality metadata are not well understood and continue

to receive little attention from researchers and practitioners alike. However, an analysis of the evidence does enable us to identify a number of research questions, which could produce useful information on which developers and managers of repositories and archives could base their decisions.

The list that follows is not exhaustive. The intention is to stimulate debate in the area of quality assurance for metadata creation across a range of communities of practice, and to raise awareness of the need for further research into this area and of the potential significance of the results of such research.

The research questions can be grouped as follows:

- How do cultural factors influence a community's approach to metadata creation?
  For example, why is ownership of metadata perceived to be important within e-learning?
- What constitutes good quality metadata, both within individual repositories and archives, and within the global networked environment?
  For example, to what extent does metadata which is 'good enough' for local purposes also support effective retrieval by remote users operating in a different contextual setting? And can a set of 'metadata metrics' be agreed within communities and beyond?
- Who is best placed to create the metadata in any given context?
  For example, to what extent does the type of metadata (subject metadata, educational metadata, etc) have a bearing? Is a collaborative approach to metadata creation the best way forward, and if so, how can this be managed effectively? How effective is automatically generated metadata?
- What kinds of tools can be used to facilitate the metadata creation process and how effective are they?
  For example, does the use of online forms encourage the creation of good quality metadata among resource authors? To what extent can metadata cleaning be automated?
- To what extent can the provision of guidelines, training and support improve metadata creation?
  For example, can information specialists provide adequate guidelines to enable non-specialists to use a taxonomy effectively? And can librarians be trained to create educational metadata?
- What are the costs and benefits associated with the various approaches to metadata creation?
  For example, to what extent are savings at the initial metadata creation stage eroded by subsequent costs such as data cleaning? And does reducing metadata costs within the repository or archive simply increase the cost, in terms of time and effort, to the end user?

Clearly there is much work to be done before the e-learning and e-Prints communities have a good understanding of the issues surrounding metadata creation, such that effective policies and practices can be put in place to assure the quality of their metadata and hence the quality of the services they offer.

8. Concluding remarks

Returning to the four assumptions outlined in the opening section of this paper, the evidence from the literature and from practical experiences within the e-learning and e-Prints communities and beyond is sufficient to at least challenge, if not completely refute, all of them.

The intense activity and substantial resources now being directed at the development of a more organised approach to open archives and repositories, in which content can be discovered more effectively, is an acknowledgement that the uncontrolled nature of the Internet has its limitations, and that in some contexts a degree of mediation and control is beneficial. Following on from this development, there is a growing awareness that poor quality metadata has a detrimental effect on the services that can be offered by these archives and repositories and that some investment in metadata creation is necessary if the potential benefits are to be realised. The increasing number of repositories, archives and other collections of digital resources which are adopting collaborative approaches to metadata creation indicate that both authors and metadata specialists have an important role to play in the process, whilst the experiences of large scale prototype federated services have shown that not all problems of metadata quality can be addressed effectively by machine solutions.

Those of us who find the metadata creation process to be a fascinating area of study may never convince the majority of practitioners that it is anything other than a tedious but necessary evil. However, as the implementation of learning object repositories and open institutional archives continues apace, the e-learning and e-Prints communities must turn their attention to a more thorough investigation of the issues surrounding metadata quality, and ultimately to the development of policy and guidelines on the creation of metadata, so as to ensure that metadata quality is not unduly compromised, and effective discovery and reuse of resources is not adversely affected.

Acknowledgements

The authors would like to thank the following: Phil Barker, Lorna M. Campbell, Jackie Carter, Charles Duncan, Gordon Dunsire, Jane Greenberg, Neil McLean, David Nicol, William Nixon, Ronan

O'Beirne, Andy Powell, Ben Ryan, Pauline Simpson, Steve Walmsley.

References

[1] Wenger, E., McDermott, R.L. & Snyder, W. (2002). *Cultivating Communities of Practice*. Harvard Business School Press, Cambridge, Mass., 2002.

[2] CATRIONA II Final Report. [currently offline]

[3] Crow, R., (2002). *The Case for Institutional Repositories: A SPARC Position Paper*. SPARC, Washington, DC, 2002. Retrieved 16 May 2003 from http://www.arl.org/sparc/IR/ir.html.

[4] van Bentum, M., Brandsma, R., Place, T., & Roes, H. (2001). Reclaiming academic output through university archive servers. *New Review of Information Networking,* 7: 251-264. Retrieved 16 May 2003 from http://drcwww.kub.nl/~roes/articles/arno_art.htm.

[5] Downes, S. (2001). Learning objects: resources for distance education worldwide. *International Review of Research in Open and Distance Learning*, July 2001. Retrieved 16 May 2003 from http://www.irrodl.org/content/v2.1/downes.html.

[6] Downes, S. (2003). Design and reusability of learning objects in an academic context: a new economy of education? *USDLA Journal*, 17(1). Retrieved 16 May 2003 from http://www.usdla.org/html/journal/JAN03_Issue/article01.html.

[7] IEEE Learning Object Metadata Standard (2000). Retrieved 16 May 2003 from http://ltsc.ieee.org/doc/wg12/LOM_WD4.htm.

[8] IMS Global Learning Consortium, Inc. (2001). *IMS Learning Resource Meta-data Best Practice and Implementation Guide. Version 1.2.1 Final Specification*. Retrieved 16 May 2003 from http://imsglobal.org/metadata/imsmdv1p2p1/imsmd_bestv1p2p1.html.

[9] IMS Global Learning Consortium, Inc. (2003). *IMS Digital Repositories Interoperability – Core Functions Best Practice Guide. Version 1.0 Final Specification*. Retrieved 16 May 2003 from http://imsglobal.org/digitalrepositories/driv1p0/imsdri_bestv1p0.html.

[10] Graham, G. & Campbell, L. (2003). UK Common Metadata Framework. Draft 0.1, May 2003. Retrieved 16 May 2003 from http://www.cetis.ac.uk/profiles/ukcmf/ukcmf_v0p1.doc.

[11] Canadian Core Learning Object Metadata Application Profile. Retrieved 16 May 2003 from http://www.cancore.ca/.

[12] Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Retrieved 16 May 2003 from http://www.openarchives.org/OAI/openarchivesprotocol.html.

[13] Dublin Core Metadata Initiative. Retrieved 16 May 2003 from http://dublincore.org/.

[14] Lynch, C.A. (2003). Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. *ARL Bimonthly Report* 226, February 2003. Retrieved 16 May 2003 from http://www.arl.org/newsltr/226/ir.html.

[15] Harnad, S., Carr, L., Brody, T. & Oppenheim, C. (2003). Mandated online RAE CVs Linked to University Eprint Archives. *Ariadne* 35, 30 April 2003. Retrieved 16 May 2003 from http://www.ariadne.ac.uk/issue35/harnad/intro.htm.

[16] University of Glasgow, UK. (2003). *Create Change: Challenging the Crisis in Scholarly Communication*. Retrieved 16 May 2003 from http://www.gla.ac.uk/createchange/.

[17] Doctorow, C. (2002). Metacrap: putting the torch to seven straw men of the meta-utopia. *E-Learning Guru Newsletter*. Retrieved 16 May 2003 from http://www.e-learningguru.com/articles/metacrap.htm.

[18] Nicholson, D., Dunsire, G., Denham, M. & Gillis, H. (2001). CAIRNS Final Report, University of Glasgow, 2001. Retrieved 16 May 2003 from http://cairns.lib.gla.ac.uk/cairnsfinal.pdf.

[19] Barker, E. & Ryan B. (2003). The Higher Level Skills for Industry Repository. *Case Studies in Implementing Metadata Standards*, CETIS, 2003. Retrieved 16 May 2003 from http://metadata.cetis.ac.uk/guides/usage_survey/cs_hlsi.pdf.

[20] Ryan, B. & Walmsley, S. (2003). Implementing metadata collection: a project's problems and solutions. *Learning Technology*, 5(1), January 2003. Retrieved 16 May 2003 from http://lttf.ieee.org/learn_tech/issues/january2003/index.html#3.

[21] O'Beirne, R. (2002). *Learner created resources: Bolton Woods*. PowerPoint presentation. Retrieved 16 May 2003 from http://www.ukoln.ac.uk/metadata/education/meetings/agendas/2002-04-18/ronan.ppt.

[22] O'Beirne, R. (2003). Personal communication.

[23] Currier, S. (2001). *SeSDL Taxonomy Evaluation Report.* University of Strathclyde, 2001. Retrieved 16 May 2003 from http://www.sesdl.scotcit.ac.uk:8082/taxon_eval/SeSDL TaxFinRep.doc.

[24] Ginsparg, P., Luce, R. & Van de Sompel, H. (1999). *First meeting of the Open Archives initiative*. October 1999. Retrieved 16 May 2003 from http://www.openarchives.org/ups1-press.htm.

[25] Nelson, M.L. et al. (1998). *NCSTRL+: Adding Multi-Discipline and Multi-Genre Support to the Dienst Protocol Using Clusters and Buckets*. Proceedings of the IEEE Forum on Research & Technology: Advances in Digital Libraries, Santa Barbara, CA., 22 to 24 April, 1998, pp. 128-136. Retrieved 16 July 2003 from http://techreports.larc.nasa.gov/ltrs/PDF/1998/mtg/NASA-98-ieeedl-mln.pdf.

[26] Van de Sompel, H. et al. (2000). The UPS Prototype: An experimental end-user service across e-print archives. *D-Lib Magazine*, 6(2), February 2000. Retrieved 16 May 2003 from

http://www.dlib.org/dlib/february00/vandesompel-ups/02vandesompel-ups.html.
[27] Liu, X. et al. (2001) Arc - an OAI service provider for Digital Library Federation. *D-Lib Magazine*, 7(4), April 2001. Retrieved 16 May 2003 from http://www.dlib.org/dlib/april01/liu/04liu.html.
[28] Simpson, P. (2003). e-Prints at Southampton. *IAMSLIC Newsletter*, 86, February 2003.
[29] SURF's Digital Academic Repositories (DARE) programme. Retrieved 16 May 2003 from http://www.surf.nl/en/download/DARE-summary.pdf.
[30] JISC's Focus on Access to Institutional Resources (FAIR) programme. Retrieved 16 May 2003 from http://www.jisc.ac.uk/index.cfm?name=programme_fair.
[31] Project DAEDALUS, University of Glasgow, UK. Retrieved 16 May 2003 from http://www.lib.gla.ac.uk/daedalus/.
[32] Gadd, E., Oppenheim, C. & Probets, S. (2003). RoMEO studies 1: the impact of copyright ownership on academic author self-archiving. *Journal of Documentation*, 59(3): 243-277.
[33] Project SHERPA, University of Nottingham, UK. Retrieved 16 May 2003 from http://www.sherpa.ac.uk/index.html.
[34] GNU EPrints software. Retrieved 16 May 2003 from http://software.eprints.org/.
[35] Pinfield, S. (2003). Open Archives and UK Institutions: An Overview. *D-Lib Magazine*, 9(3), March 2003. Retrieved 16 May 2003 from http://www.dlib.org/dlib/march03/pinfield/03pinfield.html.
[36] Nixon, W. (2003). Personal communication.
[37] Greenberg, J., Pattuelli, M. C., Parsia, B., & Robertson, W. D. (2001). Author-generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization. *Journal of Digital Information*, 2(2). Retrieved 16 May 2003 from http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Greenberg/.
[38] Moen, W. E., Stewart, E. L. & McClure, C. R. (1997). The Role of Content Analysis in Evaluating Metadata for the US Government Information Locator Service (GILS): results from an exploratory study. Retrieved 16 May 2003 from http://www.unt.edu/wmoen/publications/GILSMDContentAnalysis.htm.
[39] Greenberg, J. (2003). *Metadata Generation: Processes, People and Tools*. Bulletin of the American Society for Information Science and Technology, 29(2), December/January 2003, pp. 18-21. Retrieved 16 July 2003 from http://www.asis.org/Bulletin/Dec-02/ASISTDecJan.pdf.
[40] Metadata Generation Research Project, School of Information & Library Science, University of North Carolina. Retrieved 16 July 2003 from http://ils.unc.edu/~janeg/mgr/.
[41] Liddy, E.D. (2003). *Automating and evaluating metadata generation.* Presented at: Libraries in the Digital Age, 26 to 30 May 2003, Dubrovnik & Mljet, Croatia.