



| Project Information | | | |
|--|---|-----------------|-------------------|
| Project Acronym | - | | |
| Project Title | Scoping Study: Aggregations of Metadata about Images and Time-based Media | | |
| Start Date | June 2010 | End Date | 17 September 2010 |
| Lead Institution | EDINA, University of Edinburgh | | |
| Project Director | Christine Rees | | |
| Project Manager & contact details | Sheila Fraser sheila.fraser@ed.ac.uk 0131 651 7715 | | |
| Partner Institutions | - | | |
| Project Web URL | http://edina.ac.uk/projects/Aggregations_Scoping_Summary.html | | |
| Programme Name (and number) | Resource Discovery Programme | | |
| Programme Manager | Andy McGregor | | |

| Document Name | | | |
|-------------------------------------|--|---|---|
| Document Title | Metadata Aggregations Scoping Study Final Report | | |
| Reporting Period | - | | |
| Author(s) & project role | Sheila Fraser, Leah Halliday, James Stewart, Caroline Ingram and Tim Stickland | | |
| Date | 29/10/2010 | Filename | Metadata_Aggregations_Scoping_Study_Final_Report_v1.1.pdf |
| URL | http://edina.ac.uk/projects/docs/metadata_aggregations_final_v1.1 (Previously http://mass.blogs.edina.ac.uk/) | | |
| Access | <input type="checkbox"/> Project and JISC internal | <input checked="" type="checkbox"/> General dissemination | |

| Document History | | |
|------------------|-----------|--|
| Version | Date | Comments |
| 1.0 | 17 Nov 10 | Including final comments for release |
| 1.1 | 21 Apr 11 | Final version including comments and updates from blog. Additional appendix considering use of blog software for gathering input and comments. |

Scoping Study: Aggregations of Metadata about Images and Time-based Media

EDINA (JISC National Data Centre)

Sheila Fraser, Leah Halliday, James Stewart, Caroline Ingram and Tim Stickland

Contact:
The Administrator
EDINA
Causewayside House
160 Causewayside
Edinburgh
EH9 1PR

Tel: 0131 650 3302
Fax: 0131 650 3308



This work is licensed under a Creative Commons Attribution 2.5 UK: Scotland License. Any reuse should attribute EDINA, University of Edinburgh. To view a copy of this licence, please visit <http://creativecommons.org/licenses/by/2.5/scotland/>

Table of Contents

| | | |
|-----|--|----|
| 1 | Acknowledgements | 2 |
| 2 | Glossary of Selected Terms & Acronyms | 2 |
| 3 | Executive Summary | 3 |
| 4 | Background | 6 |
| 4.1 | Aims of the Study | 6 |
| 4.2 | Setting the Scene | 6 |
| 5 | Methodology and Sources..... | 7 |
| 5.1 | Desk Research..... | 8 |
| 5.2 | The Consultation Exercise | 9 |
| 6 | Main Findings | 12 |
| 6.1 | What can be learnt from other metadata aggregators? | 12 |
| 6.2 | Would an aggregation of metadata for images, films and sounds be used? | 13 |
| 6.3 | What collections are sought by learners, teachers and researchers? | 15 |
| 6.4 | What are the metadata considerations? | 17 |
| 6.5 | What would it take to participate in an aggregation of metadata about images, films and sounds?..... | 21 |
| 7 | What challenges and barriers must be resolved? | 24 |
| 7.1 | Legal: IPR, Contractual | 24 |
| 7.2 | Metadata Quality | 25 |
| 7.3 | Organisational | 25 |
| 7.4 | Financial | 25 |
| 8 | Conclusions | 25 |
| 8.1 | Metadata | 26 |
| 8.2 | Metadata Contributors..... | 26 |
| 8.3 | Metadata Aggregation Model | 27 |
| 8.4 | Aggregated Metadata Consumers | 27 |
| 9 | Recommendations | 27 |
| 9.1 | Aggregator Recommendations | 27 |
| 9.2 | RDTF Management Framework Recommendation | 29 |
| 10 | Bibliography..... | 29 |
| 11 | End Notes..... | 31 |

Appendix A – Desk Research

Appendix B – Online Survey

Appendix C – Models for Aggregating Metadata

Appendix D – Online Survey Analysis

Appendix E – Use of Blog and Comment Software for Feedback

1 Acknowledgements


This project was funded by JISC and was undertaken as part of the Resource Discovery Programme. The team would like to thank all participants in the consultation process for their time and expertise.

2 Glossary of Selected Terms & Acronyms

Time-based media: items such as films, videos and sounds which play over time.

Item: describing a unique thing in a collection, a record or a resource, equivalent to a book or a journal.

Metadata: information about an item, but not the item itself; e.g. for the image below:

| Metadata | Describing this item |
|---|--|
| <p>Author: H R Bowers Date: 1910-1913 : British Antarctic Expedition (Terra Nova) Title: Scott's party at the South Pole with Amundsen's tent in the background. Description: Scott's party at the South Pole. From left to right, Scott, Oates, Wilson, Evans Copyright: Royal Geographical Society Location: South Pole (-90N, 0E) Tags: Scott, Oates, Wilson, Evans, South Pole, Antarctic, Terra Nova Format: JPEG image Size: 37kB Acknowledgement: Scott's party at the South Pole with Amundsen's tent in the background. © Royal Geographical Society. URL: http://edina.ac.uk/purl/eig2/rgs.S0000236.jpg</p> |  |

Thumbnail: a small version of the original image¹.

Aggregation: collection of different sources into a (possibly unorganised) whole, e.g. a library catalogue collects all bibliographic items in a library or group of libraries.

API: Application Programming Interface, that facilitates interaction between different software programs, such as to request and receive metadata.

HE: Higher Education.

FE: Further Education.

URI: Uniform Resource Identifier, a string of characters used to identify a resource.

URL: Uniform Resource Locator, a URI that specifies where an identified resource is available and the mechanism for retrieving it.

RDF: Resource Description Framework, provides a lightweight ontology system to support the exchange of knowledge on the Web.

RDF triple: a statement about a resource in the form of subject-predicate-object, where the subject is a resource, the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and object. For example for the image above, the image (subject) has the location (predicate) "South Pole, Antarctica" (object).

Semantic Web: provides a common framework that allows data (including metadata) to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by the World Wide Web Consortium (W3C) and is based on RDFⁱⁱ.

Linked Data: the collection of inter-related data available on the Web in a standard, reachable and manageable format, whose relationships are also made available.

SPARQL: an RDF query language (SPARQL Protocol and RDF Query Language).

OAI-PMH: Open Archives Initiative – Protocol for Metadata Harvesting, a low-barrier mechanism for repository interoperability.

OAI-ORE: Open Archives Initiative – Object Reuse and Exchange, standards for the description and exchange of aggregations of web resources.

XML: Extensible Markup Language, a flexible way to create common information formats and share both the format and the data using web protocols.

IPR: Intellectual Property Rights.

3 Executive Summary

“Having to search sooo many sources - users spend longer finding the sources than they do finding the actual information they are looking for.”

The aim of this short (3-month) scoping study was to determine the feasibility, viability and value of creating an aggregation of metadata about images and time-based media (films and sounds). The research was conducted by EDINA, and is intended to contribute to the implementation of the Resource Discovery Taskforce vision of having a *collaborative, aggregated and integrated resource discovery and delivery framework*ⁱⁱⁱ for UK Higher and Further Education.

The scoping study sought to elicit views from a wide range of stakeholder groups: which has revealed a wide range of views, sometimes opposing. Integrating the views of such a range of stakeholders with such diverse experience has been a challenging task.

A total of 80 respondents took part through interviews and online survey: 47 completing the online survey and 40 being interviewed (of which 7 were follow-ups to the online survey). Following the interviews and online survey period, a further 8 people who attended the UK Metadata Forum meeting at the Repository Fringe seminar 2010 participated in a break-out group which further explored the issues addressed during this consultation exercises.

The range of stakeholder groups and range of views was very diverse. In order to extend the consultation exercise of this 3 month study and verify the findings, feedback on this report was encouraged via the open blog at <http://mass.blogs.edina.ac.uk/about/> until April 2011.

The main conclusions drawn were:

- There is little that is distinctive about aggregating metadata about images and time-based media, although the variety of implementations of standard metadata profiles leads to complexity in any harmonisation attempt. The conclusions and recommendations arising from the analysis are often not specific to the media format being aggregated, but are related more generally to aggregating metadata.
- An aggregation of metadata about images and time-based media is useful, only if the purpose and use is clear. This study describes a number of possible uses that indicate the purpose of such an aggregation thus suggesting it is desirable that these metadata be aggregated. It is also inherently valuable to have digital metadata to enable discoverability of related physical resources.
- To generate potential service provider interest in the aggregation, it may be practical, at least initially, for such an aggregation to focus on aggregating metadata from smaller or lesser-known collections with clear licence terms that are not readily visible to search engines. This would need to be balanced with achieving a sufficient amount of metadata to encourage use.
- Metadata:
 - Metadata are particularly important when searching for images and time-based media because search based on full-text-indexing cannot be applied to the resources themselves. Content-based image recognition has not yet developed to the point where textual metadata are not needed. However, metadata for images and time-based media are often more sparse than for journals and other publications. Among other things, a thumbnail or clip is considered critical to service users but, unfortunately, is often lacking in the metadata. In order to meet user need, collection owners would be required to make thumbnails or video or sound clips available freely and harvestable from their collections.
 - Enrichment of metadata is valuable, especially if this can be automated. However, if the metadata record is already sparse, it can be more difficult to enrich it by automated means. Enabling users to enrich metadata with tags (via crowdsourcing) can be beneficial; the new ‘value-added’ metadata should be treated as a commodity that can also be aggregated, which also can be shared.
 - Direct links to the resources described by the metadata should be included in the metadata wherever possible.
- Metadata Contributors:

- The metadata gathered in the aggregation will likely be more useful for applications and hence users if it is ingested in the highest standard and most appropriate format that a collection owner can provide. This will also make management of an aggregation project easier. Ideally collection owners should describe their own content, and self-deposit as much as possible.
- The level of support required from the aggregator by different metadata contributors may vary according to their level of digital readiness.
- Sufficient numbers of collections of different types of content, probably around a theme, would be required to gain significant benefit. To support gathering of a useful body of content, the aggregator should evaluate the collections described in collection-level aggregations, such as in IESR, to determine whether the content is suitable for an aggregation of metadata about images and time-based media about individual digital resources.
- Metadata Aggregation Model:
 - There is little agreement about whether metadata for images and time-based media should be standardised into a common schema and if so who should do this. The experiences of other aggregators suggest that agreeing a schema would be time consuming. Deploying a common schema based on simple Dublin Core may actually lose information and so may not provide sufficiently useful metadata for images and time-based media.
 - An aggregator should consider the mixed model to centralise the metadata and provide it to others whenever there is a use case to do so.
 - The aggregator should make the aggregated metadata accessible by providing APIs and supporting standard protocols to support developers, and adding to access mechanisms when needed by keeping a watching brief of technology and emerging developer requirements.
- Aggregated Metadata Consumers:
 - Aggregations of metadata must provide added value and adhere to open access principles so as to maximise their exposure and encourage service developers and others, such as researchers, to make use of them.
 - It is important to work with service developers to provide the capabilities they need, and thus the metadata to drive the services that their end-users need. The aggregator should also be prepared to provide aggregated metadata to other aggregators such as Europeana in the format of the defined schema.

The report recommends the following:

- Clearly communicate the benefits of an aggregation with descriptions tailored to each of the different stakeholder types (considering collection owners with different levels of digital readiness separately) in language they understand to encourage participation.
- Make it as easy as possible for collection owners to contribute metadata: support multiple different ingest formats and protocols due to the varying levels of digital readiness of collection owners. Gather metadata and updates through both harvest and submission, and provide guidance to those who would like to digitise their metadata but have not yet done so. Work with lawyers to develop a process for legal agreement that minimises effort for metadata contributors.
- Adopt different approaches for collections with different levels of digital readiness, all of which should focus on getting collection owners to describe their own content, and deposit metadata themselves as far as possible, whilst providing guidance when needed. For those who have some technical experience, tools should be provided to enable them to deposit metadata easily, in the best form they can provide.
- Determine the aggregation model that would be most appropriate for an aggregation of metadata for images and time-based media. Initial indications are that the mixed model may be most appropriate; although further work is required to arrive at a conclusive recommendation.
- Develop a collections policy in line with end-user needs, and prioritise inclusion of such collections in an aggregation of metadata.
- Include in the collection policy or contributor agreement a requirement that contributors grant permission to the aggregator to provide publicly accessible or

harvestable thumbnail images for all visual resources (images and moving images), and clips for film and sound resources.

- Engage with service providers early on in development of any aggregation.

4 Background

Recent work has been undertaken by the Resource Discovery Taskforce, a joint JISC and Research Libraries UK (RLUK) initiative, to develop a vision of having a *collaborative, aggregated and integrated resource discovery and delivery framework*^{iv}. This will establish the framework for a shared UK resource discovery infrastructure to support research and learning, to which libraries, archives, museums and other resource providers can contribute open metadata for access and reuse.

The benefits of aggregation, with regard to bibliographic metadata are understood. Among others, it supports one-stop searching, provides alternative, sometimes tailored, routes into the content, and increases the exposure of collections. Increased exposure of multi-media collections can also result in valuable, often small, collections held in diverse organisations being targeted for preservation (see Appendix A – Desk Research). This scoping study seeks to build on previous work to scope the issues and barriers to creating and making available an aggregation of open and shareable metadata of images and time-based media, and to identify potential opportunities and benefits that such an aggregation would provide.

One route to providing better access to digital collections is by including the collections in aggregations that are promoted and exposed through commonly used channels such as commercial search services. Google is one type of aggregator. Many organisations are using the standards of the Open Archives Initiative (OAI) to both publish and to harvest metadata.

4.1 Aims of the Study

A metadata aggregator is an entity which collects, proofs and publishes metadata from an alternate location than one associated with any individual collection. The scoping study aimed to:

- Explore what it means to have aggregations of metadata about images and time-based media (films and sounds), through a number of questions, which included:
 - What collections should be considered?
 - What are the metadata considerations?
 - What, or who would such a metadata aggregation serve?
 - How can the metadata be gathered and shared?
 - What are the issues that must be resolved?
- Provide insight into the benefits of, and opportunities for, such aggregations.
- Better understand the challenges and barriers to making collections of open metadata available.
- Describe some scenarios in which aggregations of open metadata about images and time-based media could be useful or required.

Where feasible within the project time constraints, consideration was also given to:

- Whether and where normalisation and enrichment of metadata takes place.
- Working with commonly used web interfaces (e.g. RESTful APIs, RSS^v and ATOM^{vi}, which may be used to access the aggregation) and those standards used within the academic, museum, and archive sectors (e.g. OAI-PMH and SRU (Search/Retrieval via URL), which enable metadata harvesting and access).
- The focus being the aggregation of metadata rather than a user interface(s) onto the aggregation (for a particular use case such as search).
- Whether aggregations could be exploited more effectively by being made available in linked data format.

4.2 Setting the Scene

There are many collections of images, films and sounds that could be useful in education. Learners, researchers and teachers can be frustrated by technical, licensing, organisational and other barriers to discovering, exploring and using these collections, and many collections may remain little used, e.g. through lack of exposure to online search facilities.

One approach is to aggregate the item information from multiple collections, and support the development of services that bridge individual collections. The following illustration describes how an aggregation of metadata might work and the main stakeholders involved.

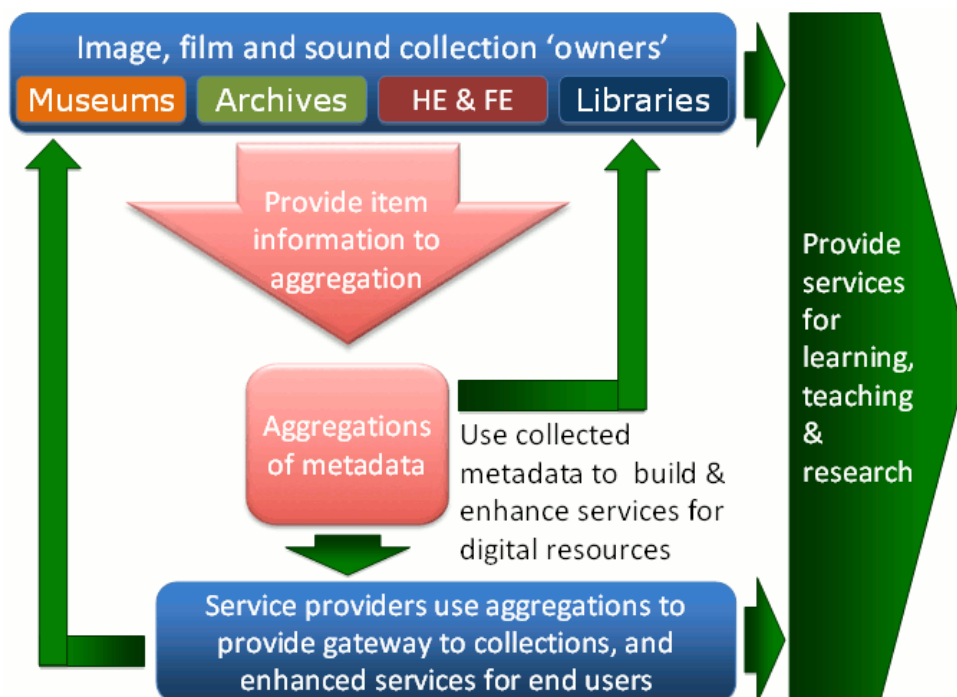


Figure 1: Metadata Aggregations and Stakeholders

5 Methodology and Sources

The RDTF implementation plan^{vii} identifies the key stakeholders in achieving the vision. As there is potential for most to be involved in an aggregation of metadata about images and time-based media, a corresponding wide range of views was sought. The approach taken was desk research followed by a consultation exercise and is summarised as follows:

- The desk research informed background on current metadata activity, identification of initial stakeholders with whom to consult, issues around aggregations of metadata and identification of a preliminary set of questions for the consultation exercise.
- The questions were refined in initial interviews with 9 participants.
- Semi-structured interviews^{viii} were the primary method used to determine:
 - What metadata are collected.
 - How they are used.
 - Attitudes to sharing (and willingness to participate).
 - Potential uses.
 - Benefits and risks as well as barriers to sharing.
 - The attitudes of vendors to such sharing.
- To supplement the interviews the team posted an online survey (see Appendix B – Online Survey) and publicised it through relevant JISC email lists and the JISC blog. A project website was set up with further information about the study and a link to the study (at http://edina.ac.uk/projects/Aggregations_Scoping_summary.html).
- Some stakeholders who had completed the online survey were then also interviewed as a way of further exploring some particularly interesting responses from the online survey. Some also answered follow-up clarification queries via email.
- During the course of the project an opportunity arose to gain additional feedback at the UK Metadata Forum meeting during Repository Fringe 2010 in September 2010. This was gathered through a discussion with a group of 8 participants.

5.1 Desk Research

A significant amount of work has been done in relation to aggregations of metadata and some organisations have already aggregated metadata for different types of media. A review of the literature is provided in Appendix A – Desk Research and key points are highlighted here.

5.1.1 Importance of Metadata

Metadata about images and time-based media may contain descriptive information that helps users to discover content as well as technical and administrative data (e.g. related to creation, quality control, rights and preservation) that help rights holders to better 'manage and exploit their assets' (The Technology Strategy Board^{ix}). Technical and administrative data may facilitate management, tracking, migration and re-use of digital assets. Clearly, metadata is valuable only if it is persistently linked to the content it describes (Nicolas et al, 2009) – whether this means that it is stored separately or as part of the digital object described.

5.1.2 Why aggregate metadata?

Aggregating metadata provides a discovery mechanism for content that is distributed. This saves the user time and gives more exposure to collections many of which may be small and difficult, otherwise, to find. Aggregation may also provide opportunities for the aggregator to add value through e.g. *inter alia*: enhancement of metadata; provision of authentication and authorisation; facilitation of preservation; provision of personalisation and alert about a variety of sources; links to related materials which provide subject-entry points; provision of a single point of information for statistics about access and downloads (Swann and Awre (2006). In learning and teaching, aggregation presents an opportunity to provide new ways to access and present information e.g. multiple routes that are tailored to different learning styles or access to resources relevant to a topic by type of resource (Pitts and Sharp, 2003). With regard to images and time-based media, a single entry point is important because content is distributed across a large range of different types and sizes of organisation. Increased exposure not only makes it more accessible, it also increases the likelihood that it will be targeted for preservation (small organisations often cannot undertake preservation off their own bats).

5.1.3 Uses and Users of Aggregation

While aggregation is clearly beneficial, it is also challenging. Arms et al. (2002) and Shreeves et al. (2003) noted that variation in metadata authoring practices and consistency of metadata records challenge service providers' abilities to build consistently searchable systems.

5.1.4 What Are The Barriers To Sharing, Re-using And Aggregating Metadata?

There is little written about the barriers to sharing, reusing and aggregating metadata for images and time-based media but barriers related to metadata about other types of content have been documented. Legal and cultural issues are highlighted as key barriers (e.g. McGill et al, 2008, Romer and MacMahon 2007) as is lack of time and a concern among collection owners that they lack the technical understanding or expertise to share are also important (Rogers and Barker 2007, Romer and MacMahon 2007).

Other barriers to sharing include:

- Organisational issues – i.e. difficulty securing cooperation within the organisation to participate; a desire to prioritise exposure of metadata through the organisation's own services before participating in an aggregation.
- Trust, attribution, incentive, reputation and approbation may also important – lineage is often an important guide to quality and thus should be clear when content is accessed through an aggregation (Whitelaw 2007) but if the aggregator has a trusted reputation, this may be enough.
- Legal issues may also be a concern – collection owners may be less concerned about rights in metadata than in content but the metadata increases the exposure of the content so raises concerns about the legal status of the content.
- Financial – many collections have been created with project funding and while project funders often require that the content be exposed as widely as possible there is often

little or no funding available beyond the project end date for sharing metadata (Rogers and Barker, 2007). In those instances, projects may be willing to participate if the effort required is minimal and if they are provided with clear guidelines on how to participate.

- Language – aggregation is often undertaken by digital library initiatives whereas much of content in this field is generated by broadcasters and cultural heritage organisations where technical staff know nothing of the language and technology used by digital libraries.
- Standards – requirement to comply with a specific metadata standard often deters participation either because it presents a technical or resource hurdle or because collection owners consider that compliance requires them to ‘dumb down’ their metadata.
- Quality – poor quality metadata may not meet the needs of aggregator and users but collection owners may lack resource or incentive to improve their metadata.

5.1.5 What can be learned from other initiatives?

A broad range of initiatives throughout the world provide examples of how aggregations are created and managed. These provide models of possible approaches and some provide specific learning. Europeana^x, for example, collects multimedia library, museum and archives into one digital website combined with Web 2.0 features. This project has engaged extensively with end-users and has achieved loyalty and frequency of use among them. Its strengths are considered to be the quality and authenticity of the content, guaranteed by the cultural organisations behind the service, and its openness to participation by cultural institutions. Because it delivers a variety of different types of materials, it can bring together relevant work in different formats e.g. the works of a painter along with an archive of documents related to her life. Europeana prefers to work with aggregators rather than individual collection owners of which there are a huge number. It has automated much of the process of content ingestion, in particular by developing a tool which allows aggregators to check that their metadata functions correctly and to view their metadata displayed in a dummy Europeana interface. This reveals any problems and the onus is on the aggregator to correct these before final submission. Europeana will shortly require ‘rights labelling’ of all content contributed.

Universeum is a European network concerned with academic heritage in its broad sense. It aims at the preservation, study, access and promotion of university collections, museums, Enrichment of metadata. One of the questions being addressed in this study is whether metadata contributed to an aggregation should be enriched (enhanced) and, if so, how should that be done and by whom. Enrichment may be achieved through, for example, allowing users to enhance metadata, or incorporating into the metadata information that may be contextual within the native service (e.g. a service about a specific historical figure does not need that person’s name in the metadata - the context makes this clear). Metadata may also be enhanced indirectly by translating records into different forms or providing related information from authority files or other records describing the same item but much of this would depend on accurate matching of records which is, itself, a formidable challenge.

5.1.6 Standards

The desk research outlines the source and purpose of a broad range of standards currently in use to create and distribute metadata about images and time-based media. These are related to items rather than collections. As yet, there is no dominant metadata standard for describing collections, although in the last few years there has been substantial progress towards this goal^{xi}. NISO has also recently released a set of guidelines for building good digital collections^{xii} and UKOLN’s Collection Description Focus offers advice in this area, including a tutorial.

5.2 The Consultation Exercise

A total of 80 respondents took part in the interviews and online surveys. Some interviewees, obviously, have experience extending beyond their current formal role. For example, the Head of Information Management at a JISC Band D University who is responsible for, among other things, the institutional repository, also has experience, over many years, of working at

a national level on development of national aggregations of different types. This interviewee brings a broad understanding of the issues, and many others contributed valuable lessons and insights from current and previous work.

5.2.1 Initial ‘pilot’ 1:1 interviews

During the desk research initial stakeholders were identified for interview and a preliminary set of questions was drawn up. The preliminary questions were discussed with 9 stakeholders, who answered the questions where appropriate, suggested improvements to the questions for a wider audience, and, in some cases, identified others for interview.

The stakeholders who answered the preliminary questions were multimedia and bibliographic collection owners, existing image and time-based media aggregators, existing aggregators of geospatial, repository and learning material metadata and an advisor in digital media for HE and FE institutions. Their input and advice was used to formulate a more comprehensive set of questions. These questions were then used as the basis for the subsequent consultations.

5.2.2 Main 1:1 interviews

During the main interview phase a total of 31 interviews were conducted. Twenty four of the interviewees were identified through desk research or during initial interviews. The remaining 7 were interviewed following their completion of the online survey.

The interviewees were from museums, libraries, archives, HE and FE institutions, or were existing metadata aggregators, existing service providers, and academics with experiences in metadata, aggregations of metadata and related technologies such as the semantic web.

A range of collection owners participated from museums, libraries, archives and HE institutions. These ranged from large organisations with well over a million resources to smaller organisations with less than 100 resources.

5.2.3 Online survey

The questions developed for the main interview phase were also posted online (see Appendix B – Online Survey) to supplement the interviews. The online survey was publicised through relevant email discussion lists, the JISC blog and the EDINA website.

A total of 79 people responded to the online survey and of these 47 responses form the basis of the online survey analysis in this report; 32 responses were excluded from the analysis having completed only questions relating to role, organisation and stakeholder type.

Respondents classified themselves in one or more of the categories of potential stakeholders in an aggregation of metadata. It was possible for each respondent to select more than one role (Q2.3); the following chart therefore shows a total of 92 responses received from 47 respondents.

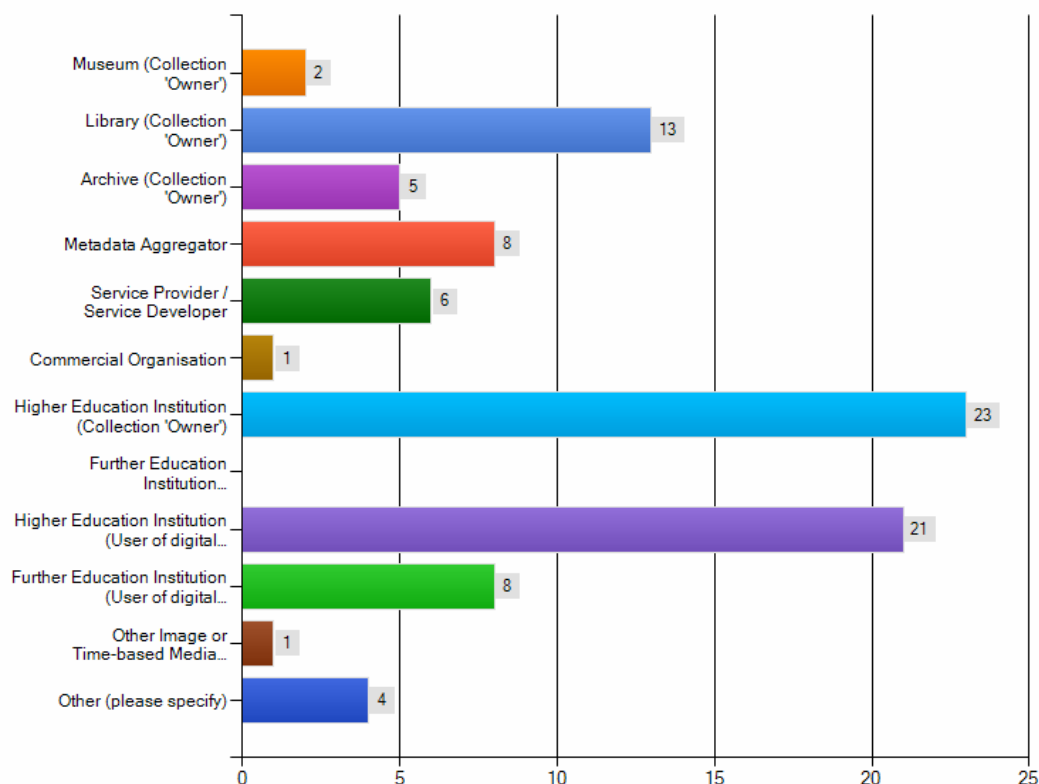


Figure 2: Number of online survey responses to “Which of these describes your organisation?” (Q2.3)

A high proportion of respondents were HE collection owners or users, with a lower number of FE users, but there were no FE collection owners. Thirty respondents identified themselves as one or more types of collection owner, and 27 as users, with a high degree of overlap between these (16 selecting roles of at least both owner and user).

The others identified themselves as a freelance author and trainer, an access aggregator for museums and cultural institutions, a university inter-departmental research centre and a JISC project supporting institutional repository development.

5.2.4 Further consultation

Following the interviews and online survey period, a further 8 people participated in a breakout group which discussed the models of aggregation (see Appendix C – Metadata Aggregation Models) during a meeting of the UK Metadata Forum held at the Repository Fringe Meeting 2010.

The scoping activity has been carried out over a short period of time and there are many types of stakeholder with a diverse range of views. Feedback on this report via comments on the blog hosting this Final Report (<http://mass.blogs.edina.ac.uk/about/>) was therefore encouraged as part of the consultation exercise.

Several requests were made for a printable version of the report rather than on the blog. The comments received are included in this final version of the report that will be made available as a pdf on the project website and the blog will no longer be available for further comment from April 2011. An Appendix E is added to this report which describes the use of the blog and comment software, including the process of gathering input and the challenges and solutions of using the blog in more detail.

6 Main Findings

6.1 What can be learnt from other metadata aggregators?

The respondents worked with a variety of content and metadata aggregations covering geospatial information, bibliographic information, learning materials, as well as those with image, film and sound metadata or metadata plus content. Academics familiar with aggregating metadata also shared thoughts and practical lessons that would be applicable.

A key finding was there are multiple different models for aggregating metadata, and that these different models could be applied to metadata for images and time-based media. Five of these models are summarised as follows:

- Common (standardised) schema metadata aggregation model: a central aggregator defines a particular metadata schema and each of the collection owners provides metadata in the common schema format.
- Multiple schemas metadata aggregation model: a central aggregator, but one which does not define the metadata schema; instead each of the collection owners provides metadata in their existing format together with the schema.
- Federated search model: a central aggregator with a common schema which links in real-time to each of the source collections, which provide metadata to the requesting service (this is not strictly a model of metadata aggregation, since the metadata remains in the collection owners' databases).
- Linked data model: no central aggregator and no common metadata schema defined. Instead each of the collection owners provides their metadata in linked data format with a search interface from their own site.
- Mixed model: a mixture of the multiple schemas aggregation model and the linked data model, with a central aggregator and a common metadata schema guideline, but where each of the collection owners provides their metadata in their existing format together with their schema; the aggregator creates the links between the schemas and metadata using linked data.

See Appendix C – Metadata Aggregation Models for more information on each model, with diagrams and the main advantages and disadvantages of each model, as well as challenges that relate to all of the models.

Several of the aggregators use a common schema metadata aggregation model and a significant amount of time has been spent agreeing these schemas. However, although the standards have been agreed, in some cases based on industry standards, these still need to be checked for ingest as there are differences in interpretation of the rules.

Regardless of which model is used, each aggregator reported that some human intervention is needed during the ingest process. The aggregators endeavour to minimise this effort and noted that when using a protocol such as OAI-PMH to harvest metadata less ongoing intervention was required than using custom exports. However from the experience of those responding, use of OAI-PMH to harvest metadata can still present technical challenges including:

- Basic errors such as wrong character-set encoding.
- Servers not being properly OAI-PMH compliant.
- OAI-PMH client not being able to work with a server, e.g. tokens that should have been unique in every batch were not served uniquely.

Most, if not all, of the metadata aggregators offer something back to the collection owners as contributors to the aggregation. For example, those libraries contributing to SUNCAT may download records in a MARC format, and one image and time-based media aggregator offers their software to contributors to help them organise their own collections (with advantage to the aggregator of a known platform for harvesting).

Some funding bodies mandate deposit of research outputs into a subject or institutional repository as a condition of funding. This has increased the rate of deposit into those services, which act as aggregators for both data and metadata. It is hard to envisage how this model could apply to image and time-based content, other than JISC, as a major funder of

digitisation initiatives stipulating that metadata be provided for aggregation as part of realising the RDTF vision.

Collection owners are concerned to maintain their brands. They do not want to see the aggregator's brand, if it provides its own interface onto the aggregation, take over as the brand that users associate with their collection. Two aggregators of images and time-based media have addressed this in their search service by providing the logo or name of the collection owner as part of the search results for end users. Even if the RDTF aggregation does not have a user interface, it could encourage participation in the aggregation if collection owner logos were available with some requirement that services making use of the metadata display them.

6.2 Would an aggregation of metadata for images, films and sounds be used?

6.2.1 Who do you think the principal users of an aggregation of metadata would be?

Most understood a 'user' to mean someone using an interface onto the metadata aggregation. Most respondents considered Higher and Further Educational institutions to be the principal users of an aggregation, followed by collection owners and non-commercial service developers (multiple responses were allowed). There does not appear to be a correlation between the stakeholder type and their view of the principal users of an aggregation of metadata, i.e. all stakeholder types appear to follow the same pattern.

Some of the questionnaire respondents said that the general public would be users. The fact that the general public might be users was also highlighted in a number of interviews, including with one large collection owner. Two interviewees responded that researchers would be users, including private research and historians. Further users of an aggregation of metadata about images and time-based media were: charities, independent artists and filmmakers, arts institutions, professional academic organisations and users searching for items in collections.

Those describing users as other services or service providers included several suggesting that aggregators may be users i.e. if one aggregation provides a feed to another such as Europeana.

One interviewee suggested that non-commercial service providers would be principal users of such an aggregation because JISC would probably fund development of such services thus stimulating this interest. S/he commented also that commercial service developers should have access because they are better placed to resource developments i.e. they will take risks to develop services that ultimately benefit the HE and FE community as well as others. S/he acknowledged that some stakeholders within HE are loathe to share with commercial organisations but noted that the open access community has learned, over the last 10 years, that exposing metadata and content is not enough to ensure that services are developed onto that content. A viable business model is also required. Any commercial organisation that believes it can create a viable service onto open metadata occupies a role that is valuable to all. A second interviewee represented the view that resources created within HE with HE funding should not be used for commercial benefit by other organisations. S/he said that s/he would not participate in an initiative that would benefit commercial service providers.

One aggregator found that there is a real demand for research and development metadata, and this can provide a saving to JISC. This occurs because when JISC funds small projects for rapid innovation those projects often spend 2 months creating an aggregation that they need in order to do the research and development. Thus, there is a strong case for aggregating the metadata in advance then providing tailored feeds to the projects as needed.

6.2.2 Benefits and Opportunities

It is clear from the extent of the responses that a wide range of benefits is believed to be associated with an aggregation of metadata. There were several recurring themes:

- Simple, easy access from a single point to many resources that could be reused.
- Making aggregations available to be indexed by commercial search engines, so users could find resources through their normal search techniques.

- Like Google, but having a better understanding of what could be done with the resources found.

Many of the benefits described by respondents are descriptions of the facilities that they would expect to be available in the services making use of the aggregations. This section describes the main benefits only, further benefits may be found in Appendix D – Online Survey Analysis.

The benefits for learners relate to quick, easy access to a wide range of additional content without having to search multiple databases. This is simple, saves time, enhances knowledge and facilitates discovery of resources that, otherwise, would not be found. There is also a benefit in the potential reuse of resources, including those that have been annotated by others; and 'safer' results (described as those with clear Creative Commons licenses).

For teachers the benefits relate to the ability to query multiple collections in one place to find rights-cleared resources for use in lectures or assignments. These include: time saving, making lesson development easier, enhancing the teaching experience, and the ability to search from a single place.

The benefits for researchers are largely the same as those identified for learners and teachers, as well as to make it easier to obtain research resource, the potential for making unlikely or surprising connections (through a deeper, more intuitive browse experience), and developing the understanding that image, film and sound resources are available and may be used and cited in a scholarly way – as with journals and books.

For aggregators the benefits include opportunities to exploit links between collection metadata, improving metadata globally, more efficient metadata gathering and the potential to create a 'pyramid' of aggregators at national, regional and local level which may help encourage and maintain links with collection owners.

The main benefits for service developers are development opportunities and marketability, with better resource availability to enable quicker development of services as well as provision of richer services and tools for end-users.

For collection owners the main benefits include raising the profile of their collections, gathering information about use of their collection(s), to enable cross-linking to other collections and brand enhance brand recognition for smaller collection owners who can have their material listed on a service next to higher profile collections.

An aggregation of metadata would also provide benefits for other beneficiaries, these, with the benefits identified include:

- End-users in general –
 - Potentially improved experience of web based search and retrieval.
 - To search for images for presentations.
 - Searching across aggregated records to discover new uses, research areas, editorial possibilities and programming opportunities.
- Universities – to raise the university profile through inclusion of special collections.
- Content producers – to provide additional channels to disseminate their content.
- Archivists – to facilitate persistent archiving and share the cost, with a view to reducing archive costs by holding only one copy of a work, and knowing which archive holds it.

6.2.3 Scenarios where aggregations of metadata could be useful or required

6.2.3.1 Search services

The most frequent suggestion given for a search of aggregated metadata being useful was to find images, videos, audio for use in research, teaching and presentations. Examples given were usually discipline specific: search for images related to colorectal cancer; track individual dancers or other contributors over time; trace non-theatre space performance work; discover images that can be reused to develop an art history course; find image, film and sound material for creating an online language course for use in business situations.

One participant visualised what was needed as ‘a Google-like intelligent solution that crawls databases like Google crawls the web’.

6.2.3.2 Tools

Respondents suggested that tools to remix content would be useful, as follows, although these relate to the digital resources themselves rather than the metadata:

- To search, play, annotate and re-use multiple video streams:
 - To produce montages.
 - For live remixing.
- To build subject-specific content.

Two scenarios for mashup tools using image and time-based media metadata were suggested:

- Combining mapping for geo-spatial with a range of moving image indexed metadata media sources (e.g. YouTube, Vimeo, etc.).
- Integrating metadata into learning material such as virtual learning environments and learning materials themselves.

6.2.3.3 Other services and tools

A number of respondents gave specific suggestions for other services and tools (again some related to the digital resources rather than the metadata) that would be useful to them including those that would facilitate:

- Repackaging and repurposing of metadata on demand for purposes such as for research analysis and provision to other aggregators.
- Creation of user-generated tags that would allow lecturers or classes to tag resources in a way that would facilitate sharing with others in the class.
- Annotation of resources held online, i.e. without first downloading them (e.g. Caboto or the output of the Crew Project).
- Determination of whether two audiovisual files have *roughly* the same content (e.g. video A is video B with a splash screen).
- Extraction of words from a video/audio file for categorisation.
- Easy comparison of multiple versions of same dance work.
- Segmentation of an image, such as to identify particular structures or boundaries within an image.

6.3 What collections are sought by learners, teachers and researchers?

JISC has previously supported work in relation to digitisation of still images and time-based media through the JISC Digitisation Programme^{xiii}. Many HE and FE institutions also hold their own collections or include multimedia content in their institutional repositories. Museums and Archives have collections of interest for teaching and research.

6.3.1 Should there be any restrictions on what collections are in scope?

A wide variety of images, films and sounds are sought by users of digital resources, therefore there should be no restriction on the collections in scope. However, subject based coverage for image and time-based media will be important, as many of the scenarios and examples given implied a subject-driven perspective. It may not be practical to achieve wide coverage across a broad variety of subjects in a short period of time, so focus on a limited number of disciplines would allow for critical mass to be gathered.

Examples of images sought by respondents were many and varied covering a wide range of academic and vocational subjects: arts images, medical images, positive imagery for promoting equality and well-being, design, graphics, manuscripts and books, animals, birds, plants, insects, fish, engineering, people, objects, animation, architecture, crafts, design, fashion, film, monochrome and colour photographs, earth science, literature, illustrations, paintings and landscapes - at scales from macro to microscopic.

Film and sound examples were similarly for a broad subject range and equally varied:

- Films: non-specific documentaries, demonstrations, techniques, concepts, broadcasts or recordings of talks, student or self-produced materials, language materials (including those for learning English), performing arts, primarily contemporary or modern dance, sports, natural history, environmental, agriculture, forestry, people, engineering, fisheries, art historical, research materials, photography, video, brand management, product design, earth science, animation and feature films/movies.
- Sounds: non-specific podcasts, recordings of talks or spoken word (for pronunciation, talks about architecture, arts, crafts, design, fashion, film, photography, video, brand management, product design; and court recordings), voice overs and radio broadcasts, music, student or self-produced materials, birdsong, and ambience.

One interviewee from an HE institution said that s/he would focus on content for which either the institution holds a licence or that is available with a Creative Commons licence – because the library should endorse only those resources that may, legally, be used by its end-users. At another point in the interview, however, s/he said that an aggregation should include the large commercial services like Flickr and YouTube because users look for that content and it would be useful if they could find that through an educational portal.

Tracking the Reel World^{xiv} identified 0.9 million hours of film, 9.4 million hours of audio, and 10.5 million hours of video, the majority of which is concentrated in a handful of extremely large collections (national audiovisual archives, broadcasters, deposit libraries). However, many of the collections also recorded in the survey were small or very small; their results show that about 65% of film and around 40% of audio and video collections consist of no more than 500 hours of materials.

It is difficult to reconcile views gathered by this study regarding inclusion of metadata describing resources from commercial services such as YouTube and Flickr. Those working in HE and FE know that their users consult these services and thus, if their metadata are included in an aggregation or portal service, users need consult fewer resources and thus their searching might be more efficient. On the other hand, some staff in HE and FE wish to present users only with content that has clear licence terms attached such as subscription content and that licensed under Creative Commons.

6.3.2 Is it necessary that a digital asset be available, or are metadata referring to physical objects useful?

The primary aim of this scoping study was to explore metadata for digital image, film and sound resources, but many collection owners have other related assets in their collections which they would like to make discoverable for others to access and use, such as:

- Film scripts, special collections, the analogue film and other related film items.
- An analogue sound collection and an archive of student films and performance documentation that have not yet been digitised.
- Documents and journals, PhD theses, old books about local history, maps and photographs of the local area, and a collection of sheet music.

Although an aggregation of digital metadata is most convenient when it describes digital resources (because these can be accessed remotely), digital metadata describing physical resources are also useful. These metadata enable discovery of those resources, and could establish linkage between physical and digital resources, or between digital resources and applications that use them (e.g. algorithms (software) for analysis of digital medical images).

Further, by interlinking several aggregations of metadata covering multiple media types, subjects or themes in different ways, an aggregator could facilitate development of services that are more valuable to end-users, particularly researchers. For example, in the field of genetics there are books, drawings, images, biomedical papers, films, and software algorithms, which would all be valuable to someone researching genetics or the history of genetics.

6.3.3 What formats of media types are needed?

This question “What formats of digital resources do you need for your work?” was asked to understand if there are specific types of digital resources that should be sought, and therefore which collections should potentially be prioritised for inclusion.

The primary type of image sought was jpeg, followed by tiff. For sounds the majority sought the mp3 format, followed by wav. Collections that include digital resources with these formats may therefore be more suitable for inclusion in an aggregation of metadata, however as formats change frequently these preferences may also change. With regard to preferred formats for films there was no strong majority. Thus it may be appropriate to include in an aggregation metadata describing a wider variety of film collections than would be the case if the preferred format were clear.

Although this is a small number of responses it is worth bearing in mind that those searching for content may not want to search separately for images, films and sounds but may wish to search for all formats (i.e. images and time-based media alongside text) at the same time.

6.4 What are the metadata considerations?

Most metadata-related issues identified in this study are not specific to images and time-based media, but are, nevertheless, relevant as they apply to all aggregations of metadata including those for images and time-based media.

6.4.1 Should metadata be normalised into a common standard? If so, who should do this?

The reasons for having a common standard, and the issues associated with normalising (standardising) metadata were explored. The responses were split, with many assuming an aggregator would introduce a common schema, and others cautioning against it due to practicality and other reasons.

More online respondents indicated that metadata should be normalised (8) than not (5), but the largest number of respondents selected 'other' (14) in response to this question. More of the comments of those selecting 'other' supported the view that metadata should be normalised (7) than not (2), the remainder proposed caveats to normalisation or alternatives. The alternatives were:

- *“Both the arguments are valid, surely there must exist some sort of interface which could accommodate both.”*
- *“If you go to semantics, you can keep the original metadata based on cataloguing rules and standards and enrich them with alias to the semantic model.”*

From this small sample there is little agreement about whether metadata should be normalised and who should do this. Of interest from the online survey is that the majority of respondents who were collection owners thought that metadata should not be normalised, and the majority of those who thought that the collection owner should normalise metadata were users.

Several interviewees had experience of agreeing a common schema for normalising metadata, and a common view was that it would be useful for consistency and cataloguing but would also be difficult to put into practice. One noted that smaller archives are concerned that larger ones would take over in setting the standard and that getting agreement between different partners would be challenging.

Concerns were raised by several interviewees that metadata quality would be lost through aggregation if trying to “shoehorn things into Dublin Core”, for example a large collection owner said that it uses specialised fields for video and that information would be lost if the metadata must adhere to a standard schema. However, this problem is mitigated by providing, within the aggregated metadata record, a link back to the full record on the collection owners' site.

One interviewee from a large aggregation said that metadata does not necessarily need to be normalised - that normalising is an impossible task - but as long as it is possible to translate between the different schemas the commonality can be extracted when needed. This has been found, by an existing aggregator, to be particularly true with regard to image collections and metadata sets. S/he recommends that trying to normalise too much would be difficult since different organisations in different subject areas use different schemas.

6.4.2 What is the minimal profile for metadata?

From the high number of relevant responses giving a wide variety of proposed schemas it seems that gaining agreement on a single metadata schema would be a significant challenge if the common metadata schema model were to be followed. One respondent neatly summed this up as: *“Hard to find one that would please all parties”*. Respondents also added different fields as being desirable to the various schemas suggested.

Several respondents specified a list of fields that would constitute their minimum schema rather than naming an existing schema, and no two of these were the same. A selection of those suggested as minimum fields were:

- Title, description, unique identifier, copyright.
- Title, description/abstract, keywords, file format.
- Title, creator, summary/description, file format, original production date.
- Title, related words, dates, file type, file format.
- How the item has been labelled (e.g. ‘INT.27.rtf’ for the 27th interview in a collection); duration; description; theme; and format.

The following profiles were suggested as suitable schemas for a minimal profile for metadata for images and time-based media:

- Dublin Core (DC) (although each respondent suggesting DC also gave a different enhancement as none thought simple DC would be sufficient - these included geospatial data, basic keywords, RDF triples, qualifiers not used for ones’ own collection and a number of other fields).
- PBCore (Public Broadcasting Core)^{xv}.
- EBUCore (European Broadcasting Union Core)^{xvi}.
- DICOM (Digital Imaging and Communications in Medicine) based with extension to include metadata useful to research (this was in relation to medical images).
- The schema used by the European Collected Library of Artistic Performance (ECLAP).
- People’s Network Discovery Service DC Application Profile^{xvii} but with no “complicated FRBR based stuff though!”^{xviii}.
- METS (Metadata Encoding and Transmission Standard, a Library of Congress metadata packaging format using XML^{xix}) MODS and RightsMD.
- Images Application Profile^{xx} and Time-based Media Application Profile^{xxi}.
- A simplified version of the schema used by the Siobhan Davies Replay Archive^{xxii}.
- PADS (Performance Art Data Structure)^{xxiii}.
- STARS (Semantic Tools for Screen Arts Research)^{xxiv}.
- BBC Class Clips Learning Zone^{xxv}.
- Web standards such as RSS.

Two interviewees said that consideration should be given to what people will be doing with the collection before a minimum profile is decided upon. For example, subject or discipline-specific searching may be really useful, but this must then be accommodated in the metadata structure.

One interviewee from a specialist HE museum agreed as it holds a variety of schemas for different subjects e.g. entomology, zoology, palaeontology, and has not standardised these. S/he ‘counsels strongly against trying to attempt to fit into the [single schema] and the time spent agonising [or you will] invent 10,000 fields to accommodate everyone’. Furthermore, s/he believes that standardisation would ‘degrade the information available’.

Interestingly nobody suggested the use of the W3C ‘Ontology for Media Resource 1.0^{xxvi}’, which combines multiple different media types and maps this to a number of common vocabularies such as DC, METS, OGG and YouTube. This may be because this is relatively new and therefore has not been tested and adopted.

From the high number of relevant responses and the variety of proposed schemas it seems that gaining agreement on a single metadata schema would be a significant challenge. Clearly, the specifics of any schema would relate to the envisaged use. This would explain why different survey respondents and interviewees suggested different fields – there is probably an implicit use case behind many responses. However, the idea behind the

aggregation is that it be made available to service providers who would create services that meet the user needs of which they are aware or as they become aware of them. This adds weight to the argument that a minimal profile is inappropriate as any loss of metadata would restrict the potential use and application before the specific functionality needed by users is known. It may also stymie the capacity of the aggregation to meet emergent user needs.

6.4.3 Should metadata be enriched?

All but one of the respondents to the online survey believed that metadata should be enriched but there was no conclusive outcome regarding who should be responsible for enrichment. Ten respondents indicated that the collection owner should be given technical support to enrich the metadata as they know their metadata best. Nine indicated that the aggregator should enrich the metadata as they can act consistently for all collections. Five indicated that the service provider should enrich the metadata depending on the demands of users for the service. A further four responses suggested that it would depend on a number of different factors, such as the resources of the collection owners, the use of standardised labels (e.g. top-level library of congress categories) or niche discipline labels and the target audience.

The online survey respondent who did not believe that metadata should be enriched had experience of this with an HE repository, where metadata were automatically uploaded from another database which did not work effectively. However, s/he would not limit reuse of metadata from the repository as it can be searched and re-presented online, and she would be willing for machines to enrich it whether to make it into linked data or for any other purpose.

Among the interviewees, those supporting enrichment by collection owners tend to acknowledge that in many instances collection owners will lack the resource and/or expertise to enrich metadata and thus would require support from the aggregator to do this. Those supporting aggregators and service providers to enrich metadata suggested that text mining may be used to automatically enhance the metadata with references to items such as places, names, dates and languages. However, if the common schema metadata aggregation model were used with a minimal schema, the metadata available could be sparse so automatic enhancement would be limited. If the multiple schemas metadata aggregation model were used then there may be more metadata available, which may give more opportunity for automatic enhancement.

Many interviewees thought that contributions from users should be taken into account: they may have the knowledge to enrich the metadata, but care should be taken as 'incorrect' metadata may hinder searching; therefore the weight that user generated metadata plays in a search would be important. Three examples were suggested:

- Using search behaviour of users to enhance the discovery path of subsequent users.
- Crowdsourcing: asking users to add metadata such as tags, categories or comments.
- Asking subject area experts to add metadata tags for particular subjects; for some interviewees metadata provenance was important.

One interviewee said that if the opportunity to expose enriched metadata through RDFa or Linked Data format were taken, then it is likely this would still be useful in 5-10 years time (Google, Twitter and Facebook appear to be thinking along these lines^{xxvii}). Further, the NSDL (National Science Digital Library^{xxviii}) in the US has been looking at metadata recombination^{xxix}, i.e. getting metadata from different places and combining it into a single better record.

Several interviewees commented that none of the existing technical protocols were good at giving metadata back to the collection owners, and that if metadata owners were to check the metadata that were added to their original records this would be 'a project in its own right'.

The increasing use of digital cameras to take digital images and film means that a large quantity of technical metadata can be captured automatically (e.g. date, time, focal length, F-number, exposure time and in some cases geo-location) from the digital device. It is not clear whether such metadata that is stored within the asset should be extracted by the collection owner, aggregator or service provider; this was not explored further.

Perhaps most appropriate would be that the aggregator undertakes some enrichment of metadata and that it provides support for enrichment to those collection owners who have the

resource and the will to enrich their own metadata. Enrichment by users through services could also be incorporated through the addition of user tags that are clearly differentiated from 'official' metadata.

6.4.4 Other metadata issues

A number of non-technical challenges are known to EDINA through experience of developing the Visual Sound and Materials (VSM) Portal Demonstrator. These include:

- Difficulty identifying the appropriate person for metadata within the collection owning organisation and the possibility that they have little time or budget to devote to this initiative.
- Lack of understanding, within the collection owning organisation, of what was being requested by the aggregator.
- The agenda of the portal did not match the internal agenda of the collection owning organisation so a lower priority was given to the job of contributing metadata.

One interviewee said that there was, perhaps surprisingly, a poor awareness of metadata among collection owners, and some were not aware of the potential value to the community of the collections that they held, or of their associated metadata.

A number of interviews touched briefly on the challenges associated with aggregations of aggregations such as differences in metadata in aggregations or services due to metadata enrichment or model used, duplication of entries, and potentially increased effort from collection owners if supplying multiple aggregators. It would be useful to investigate these challenges further, with the caveat that the value of this work may be transitory if linked data can be implemented effectively by collection owners.

6.4.4.1 Metadata quality

A number of the respondents were concerned about the content and quality of metadata provided for aggregation given the highly specific historical and disciplinary nature of much of the metadata in the catalogues of collections, and because metadata is often created for particular uses under specific circumstances (and thereby, is less useful for broader use).

Two interviewees stated that the "description" field, especially for images, may not contribute any useful information: it may simply repeat the title, or be blank. Another said that frequently the subject of the image is not contained in the metadata. Keywords may be more useful, and easier to keep in a common, searchable format, however they may be inconsistent.

One interviewee explained that for art images there is often no description or title as, in the arts, many things are untitled as well as anonymous. A challenge shared with scholarly works is that there may be many creators rather than just one.

One interviewee involved with normalising metadata for images said that the metadata which commercial agencies compile for images tends to be oriented around specific keywords of interest to the advertising community, which may not be helpful or relevant for educational use. This is less so for organisations such as galleries, whose metadata are more suitable for academic use. In another instance, the required information was contained in the full metadata record but was not included among the metadata fields in the specified schema.

6.4.4.2 Samples of images, films and sounds

As text-based documents are commonly 'full-text searchable' it is possible to locate the search string within the document and present it in context, i.e. to present the extract of the document (the snippet) that contains that search term with the search term highlighted. A relevant 'snippet' to aid discovery would be a thumbnail for images or films, or a short clip for films or sounds. For those searching for such resources, such a sample is essential to determine whether the resource is the right one: one interviewee said that if an aggregation did not provide the facility for users to view the resources, for example, as thumbnails, it would be a 'critical failure'; another said 'a thumbnail still has to be there'. If such a sample is not available, then those searching for resources are reliant on the metadata, which may include the metadata embedded in the object if it has been extracted.

6.4.4.3 Complexity in metadata for images and time-based media

The JISC-funded application profiles developed for images^{xxx} and time-based media^{xxxi} reflect a systematic understanding of the complexities involved in metadata for images and time-based media. Both application profiles were based on the Functional Requirements for Bibliographic Records (FRBR)^{xxxii} entity-relationship model, which, although developed for bibliographic records, identifies and considers many of these issues which are relevant for images and time-based media. The scenarios for use of aggregated metadata given by respondents were in line with those identified as use cases for these two application profiles.

Complexity in metadata about images and time-based media reflect complexity in the resources themselves – and differences between different media types. The media types are used to capture a range of different resources e.g. an image may represent a photograph of a painting or medical problem or the pages in an original manuscript. Users searching for a photograph of a painting may want information about both the original painting and about the photograph (e.g. who was the painter? Who was painted? Where is the painting currently stored or displayed? Who was the photographer? When was it painted? When was the photograph taken?). Those searching for a medical image will want information about the diagnosis but, in this instance, the anonymity of the subject is essential and access to view may be restricted to specific audiences. Effective use of page images from a manuscript may require that the metadata include the specific page order and if available that individual pages be linked to a transcript of the full text. The grouping of images to reflect page flow in a manuscript is important, and hence metadata to support this rendering.

A film may have multiple authors and may consist of various elements e.g. film segments, soundtrack, subtitles and a printed cover, each of which has various contributors e.g. the designer, illustrator and editor for the cover. Thus, a film is one resource made up of multiple other resources, with corresponding layers of metadata. The relationships between these different layers of the resource are important to both contributors and users and thus should be possible to reflect in the metadata.

6.4.4.4 'Time' and 'date' mean various different things

As indicated above the field label 'date' is not straightforward as, often, many dates are relevant to a resource and different users look for different dates. A documentary produced in 2001 may describe an event from 1600 using some footage produced in 1970. If it is made available in a repository, the date of ingest will also be recorded. Similarly, a digital image of an analogue photograph of a painting of a famous battle has several important dates: the date of the battle, the date of the painting, the date on which the analogue photograph was taken, the date on which the digital version was created, and the date of ingest into a repository. To an art-history student one date is most important and to a student of a photography another.

'Time' may also refer to different things, e.g. the time of day on which a recording was made or the duration of the work.

6.5 What would it take to participate in an aggregation of metadata about images, films and sounds?

In general collection owners appear willing to take part in an aggregation of metadata about images and time-based media, as long as they are not required to invest too much work, support is provided where required, and the barriers identified (see section 7) are addressed.

6.5.1 Three levels of digital readiness need different support

During the course of the scoping study it became apparent that there are many different collection owners who would like to improve access to their collections, and are interested in the potential for an aggregation of metadata to do this.

These can be grouped into three distinct types by their degree of digital readiness. These are collections with:

- Metadata digitised and already available for harvest.

- Metadata digitised, but either not easily harvested or not available for harvesting. An example of the former is collections whose information is published as web pages, and of the latter collections that have databases of metadata but only used internally within the organisation.
- Metadata not digitised.

In general the larger collections have metadata digitised and available for harvest, though a number of smaller collections also fall into this category.

6.5.2 Would collection owners participate?

While the vast majority of respondents to the survey (8 of 9 respondents) and all of the interviewees were generally open to sharing metadata with an aggregator – primarily as a way of increasing exposure – there were caveats and previous disappointments which would be important for an aggregator to address^{xxxiii}. There are also, in some HE institutions, organisational issues that may delay participation, and one large collection owner that already makes their metadata available to Europeana could not see a benefit in participating in another aggregation. On a positive note, there was indication that in recent years, the capacity of larger collections to contribute to aggregations has increased, although smaller collection owners still lag behind.

An interest in increased exposure echoes the findings of an earlier survey conducted by VADS which covered 89 collections across the UK and related to digital objects as well as the associated metadata^{xxxiv}. VADS found that the overwhelming majority (88%) of respondents were willing to explore participation in cross-search services or prototypes in future. The main reason for participating would be to benefit from ‘marketing and publicity’ as collection owners expected that the aggregation may draw new traffic to their websites. Anecdotal evidence suggests that since the survey was conducted, some of these image collections have contributed to Culture Grid, and some to VADS, but no firm figures are available.

6.5.2.1 Caveats to sharing

For some collection owners this relates to the question of who would be the principal users of metadata, since they wanted to have a say in how their metadata might be shared (see section 6.2.1). Other conditions on sharing metadata with an aggregator (some contradictory from different collection owners) are that:

- Permissions may need to be sought for metadata to be made open for organisations that did not clear rights for metadata and objects originally.
- The collection owner would have access to download statistics.
- An indication that consideration would be given for a regional portal for all kinds of documents.
- Limited metadata only should be provided to aggregator as the full catalogue has commercial value (and has been sold in the past).
- Any conditions imposed by the aggregator are acceptable, including conditions of use of the metadata by others.
- Metadata should be made available to the general public and not just for use in education.
- Metadata should not be made available for commercial use.
- Metadata should not be restricted to non-commercial use, as commercial service providers are in a better position to take risks if they feel that there is a business model to support development.

Although of concern, some of these may be addressed through discussion about open licensing of the metadata such as use of Creative Commons.

It would be valuable to explore further whether collection owners would consider a middle position between the last two bullet points i.e. that their metadata be made available freely on the internet in one context *and* be available in another context to commercial service developers to include in innovative new services that may serve the principal community of users of the content. It would be helpful to support this with education about the role of commercial organisations. These may fund development of services that may not otherwise

be developed or develop free services and thus contribute something useful to users in the UK HE/FE community.

6.5.3 How can the metadata be gathered and shared?

6.5.3.1 How can the metadata be gathered?

The Information Environment Service Repository (IESR) is a Mimas service funded by JISC, providing an aggregation of collection-level descriptions and facilitating access to the collections themselves. There are approximately 6,300 collections registered in IESR, and the majority of these collections are available via web pages. Approximately 1,850 state that they provide a method of metadata harvesting. The vast majority provide harvest capability via OAI-PMH, a much smaller number support Z39:50, and fewer still SRU, SRW and RSS.

Similarly, the most commonly discussed standard for gathering metadata was OAI-PMH, with several aggregators already using this to harvest metadata from other collections. OAI-PMH is considered to work very well at taking metadata from a source into a cache and is well supported by existing repository interfaces such as ePrints and Fedora. Once past the initial harvest some considered it is less useful in relation to managing the state of records over time. A further option described was support for a custom schema and metadata in HTTP and XML format (which it is noted would need effort by both to collection owner and aggregator to agree and understand the schema).

A number of the collection owners interviewed had already contributed metadata to other organisations or aggregations, such as Europeana and the VSM Portal Demonstrator, and they had contributed using different methods. Some produced metadata to OAI compliant standards, some used their own tools to produce metadata in a required format (e.g. for Europeana), and others contributed their own schema and data. One collection owner had not shared metadata with an aggregator but did issue RSS and podcast feeds.

In order to develop an aggregation to which collection owners would contribute it should be as easy as possible for them to do this. Experiences of other aggregators suggest that it is necessary to support multiple different ingest formats due to the varying levels of digital readiness and that some collection owners need more support than others. This poses a challenge for sustainability of any aggregation, not just one of images and time-based media, if collection owners have to do work (although minimal) themselves to contribute.

6.5.3.2 How can the aggregated metadata be shared?

An aggregator would develop expertise in managing and manipulating the metadata contained within the aggregation, and therefore would be best placed to provide metadata in as many different ways as possible to support discoverability. The most commonly encountered standard to facilitate reuse was OAI-PMH. This, however, is not the only way that the metadata could be shared: several of those interviewed recommended making aggregations available through RESTful APIs or in linked data format, which would enable service providers to exploit the aggregation.

There is current activity looking at the use of RSS in the learning materials area^{xxxv}, which looks at the possibility of a repository subscribing to another repository to learn about new content and offer this in turn to its own end users. This has raised a number of issues, such as how to indicate item updates and deletions; these issues would likely apply to use of RSS to provide feeds from aggregations of metadata. It would be valuable to continue to track development in this area, from the perspective of repositories 'pushing' data via RSS and also those subscribing to the feeds.

An aggregator of metadata for images and time-based media should be prepared to provide data in the required format to other aggregators (such as Europeana) to facilitate discoverability, and, if possible, to support making the items discoverable by commercial search engines.

See Appendix D – Online Survey Analysis for selected graphs of the online survey responses and some more detailed responses to selected questions.

7 What challenges and barriers must be resolved?

Survey respondents and interviewees were invited to select from a list the key barriers to participation in an aggregation and to identify any barriers that were not listed. Legal, organisational and financial issues were most important for survey respondents while interviewees considered legal issues and metadata quality to be most important.

7.1 Legal: IPR, Contractual

Legal and IPR issues are potential barriers to participation because they are considered to be complex and can take time and resource (including legal advice) to resolve.

7.1.1 Ownership and willingness to share metadata

When considering value and ownership, the content described by the metadata is often the focus, while ownership in the metadata itself may remain undefined (although, clearly, ownership in the content may also be unclear). The VADS survey^{xxxiv} found that legal and copyright issues could be a barrier for collection owners in sharing metadata, and this was a key challenge for many interviewees and survey respondents. There appears to have been some progress in this regard however as several respondents from HE institutions indicated that they have policies related to metadata that are distinct from those relating to the content described within institutional repositories. Seven collection owners who were interviewed or responded to the online survey said that the organisation whose staff had created the metadata owned it. Four survey respondents identified specific professional roles within the institution as owners of the metadata and one interviewee said that the BBC owned and had licensed the metadata to the institution for use in an online service.

In institutions that lack policies related to metadata it seems that securing permission may take some time as the issues raised must first be explored at an institutional level, and this may mean clearing rights with many individual content creators; those responsible for the collections may be willing to share the metadata but may not, independently, have the authority to do so. Other collection owners are clear that they do not wish to share metadata unless they can exert some control over its use. Ownership of metadata may be less clear where the metadata themselves are more complex. In some instances, the metadata are works in themselves e.g. sleeve notes and a musical score may be metadata for discovering a recording but are also important original works attracting IPR.

7.1.2 Tension between contractual responsibility and cost of agreeing the contract

One interviewee who had previously agreed to take part in an aggregation, subsequently withdrew after receiving 'a massive contract through the post' which, her/his legal department said was too big to handle. Clearly, if a long contract is a deterrent to participation, it would be preferable that any agreement between the aggregator and the collection owner be kept as short and simple as possible. Unfortunately, in one project, advice received from legal experts within the UK HE/FE community and from university lawyers generated a lengthy contributor agreement which sought to protect the aggregator by securing all of the permissions required to deliver the aggregation and ensuring that the collection owner was responsible for any liability in the data. This creates a tension between the need for a simple, low-cost process for collection owners and the need for the aggregator to minimise risk. (This need is often greater for a large organisation than for a small start up; the former is more likely to be a target because it is viewed as having the financial resources to pay any penalty.) Nevertheless, there has been a move towards lightweight licensing in UK HE/FE in recent years. It would be important, when developing a licensing process, that any aggregator seek to minimise the effort required by collection owners.

7.1.3 Metadata and licence terms for the content it describes

One interviewee explained that because the content described in metadata is distributed, i.e. owned by different bodies and published under different licences, it can be difficult to ensure that users are aware of the terms on which that content is licensed. This is a challenge for any aggregation, including for images and time-based media. Collection owners want to be reassured that users will be directed to the terms and conditions of use before they link to the

content and use it. Users want to access the content described in the aggregation via a service in as few clicks as possible. The combination of aggregator and service provider must find a way to: explain to the user the terms and conditions associated with the metadata aggregation; differentiate those from the terms of use of the content described by the metadata; direct the user to the latter; and make clear to the user that s/he is responsible for complying with it.

7.2 Metadata Quality

Metadata quality was considered a barrier because improving the quality of metadata requires resource and, often, the organisation sees no benefit in committing further resource to improve the quality of metadata which meet the organisational needs in their current state. One interviewee referred to the British Library (BL) EThOS initiative as an unusual example where academic institutions were willing to engage to improve and normalise data in order that the institution be accepted as a participant because non-participation in a BL national aggregation would reflect badly on the institution. Metadata quality issues may disproportionately affect small, specialist collections. An interviewee responsible for a collection housed within HE suggested that an aggregation would be valuable and that s/he would participate only if it were 'inclusive' of the small, specialist collections that are housed in a heterogeneous range of small organisations. S/he acknowledged that these are least likely to be able to provide quality metadata, adhere to metadata standards or resource participation.

7.3 Organisational

'Organisational issues' were identified as important barriers to participation by respondents to the online survey and this was elaborated by interviewees. It seems that this issue applies generally to images and time-based media rather than specifically to metadata and participation in an aggregation. It can be difficult, particularly within HE and FE institutions, to secure agreement to commit resource to initiatives related to images and time-based media because these are not valued equally alongside other resources for scholarship, teaching and learning. Many academics are still resistant to using these resources, preferring, instead, traditional, text-based resources.

7.4 Financial

Several interviewees discussed financial gain by others as a potential barrier. Some were concerned that by contributing to an aggregation, they may lose potential commercial opportunities through sale of their metadata. Others wished to secure agreement that if the aggregator generated profit, it should be shared with the contributing organisations. While financial gain is not the primary motive for many collection owners participating in this study, they do not wish to see commercial outfits profiting from their collections and if profit is to be generated, they want their organisations to benefit.

8 Conclusions^{xxxvi}

An aggregation of metadata about images and time-based media is useful, only if the purpose and use is clear. This study describes a number of possible uses that indicate the purpose of such an aggregation thus suggesting it is desirable that these metadata be aggregated. It is also inherently valuable to have digital metadata to enable discoverability of related physical resources.

There is a general willingness to contribute to an aggregation of metadata about images and time-based media, but the overheads for doing so must be kept to a minimum, with support available when required, and as long as the barriers identified in this scoping study are addressed.

To generate potential service provider interest in the aggregation, it may be practical, at least initially, for such an aggregation to focus on aggregating metadata from smaller or lesser-known collections with clear licence terms that are not readily visible to search engines. This would need to be balanced with achieving a sufficient amount of metadata to encourage use.

Use of descriptions and language that address specific audiences, avoid jargon and thus are easy to understand may be important. For example, a collection owner may not understand the benefits of the semantic web in simple terms in this context, but may more easily grasp that they could increase linkage to their website.

8.1 Metadata

Metadata are particularly important when searching for images and time-based media because search based on full-text-indexing cannot be applied to the resources themselves. Content-based image recognition has not yet developed to the point where textual metadata are not needed. However, metadata for images and time-based media are often more sparse than for journals and other publications. Among other things, a thumbnail or clip is considered critical to service users but, unfortunately, is often lacking in the metadata. In order to meet user need, collection owners would be required to make thumbnails or video or sound clips available freely and harvestable from their collections.

While aggregations of metadata about images and time based media that lack links to the resources they describe are useful in some scenarios, those that include links to the resources are more useful for research. Therefore direct links to the resources described should be included in the metadata wherever possible.

Metadata about images and time-based media can be complex: one resource may consist of multiple other resources that could be considered works in their own right. For example one film can have separate teams producing the trailer, soundtrack, computer based animation, marketing posters and film, each of which is an individual resource in its own right.

For some sound resources, containing e.g. dialects and oral history, author and title are less relevant for discovery than date and place. Thus time and space must be captured in the metadata for such sounds to be discoverable through a useful service.

Enrichment of metadata is valuable, especially if this can be automated. However, if the metadata record is already sparse, it can be more difficult to enrich it by automated means. Enabling users to enrich metadata with tags (via crowdsourcing) can be beneficial; the new 'value-added' metadata should be treated as a commodity that can also be aggregated, which also can be shared.

8.2 Metadata Contributors

The metadata gathered in the aggregation will likely be more useful for applications and hence users if it is ingested in the highest standard and most appropriate format that a collection owner can provide. This will also make management of an aggregation project easier. Ideally collection owners should describe their own content, and self-deposit as much as possible.

The level of support required from the aggregator by different metadata contributors may vary according to their level of digital readiness. Those contributors who have limited IT capability would benefit from support for example through provision of a range of standard tools to help with metadata extraction and manipulation which thus help to standardise metadata with a common schema. Contributors with some IT capability could benefit from technical guidance and support with some requirements, for example if using HTTP and XML to transfer metadata as resource is needed from both aggregator and content provider to agree and understand the schema. For those with strong IT capability it would be necessary to engage them to provide metadata that could be regularly harvested.

Sufficient numbers of collections of different types of content, probably around a theme, would be required to gain significant benefit. To support gathering of a useful body of content, the aggregator should evaluate the collections described in collection-level aggregations, such as in IESR, to determine whether the content is suitable for an aggregation of metadata about images and time-based media about individual digital resources. Clearly, this should be based on the aggregator's collection policy.

Manually created and imported metadata are not sustainable for an aggregation in the long-term, so (dependant on the aggregation model implemented) guidance should be given to collection owners to develop their systems in such a way that their metadata and updates can

be gathered automatically on a regular basis. The agreement with metadata contributors must acknowledge and agree on onward sharing of aggregated metadata to other aggregators.

Collection owners have different views on use of metadata with reference to definitions of 'commercial' and 'educational', which affects their willingness to contribute to an aggregation. Further work on this would be useful to determine whether collection owners' perceptions of commercial services that charge for their services differ from their perceptions of services that are provided free of charge but display adverts, and whether this is different for images, films and sounds than for text-based resources.

8.3 Metadata Aggregation Model

There is little agreement about whether metadata for images and time-based media should be standardised into a common schema and if so who should do this. It is interesting that a small majority of respondents to the online survey who thought that the collection owner should normalise metadata were users. The experiences of other aggregators suggest that agreeing a schema would be time consuming. Deploying a common schema based on simple Dublin Core may actually lose information and so may not provide sufficiently useful metadata for images and time-based media. Many of the scenarios for using an aggregation of these metadata may not be realisable with such a schema: the key exception being the case where an aggregation takes very basic descriptive metadata (title and keywords, plus thumbnails or clips) and acts as a signpost to the collections themselves, which continue to maintain detailed metadata.

An aggregator should consider the mixed model to centralise the metadata and provide it to others whenever there is a use case to do so. This would enable the aggregator to retain as much metadata as possible in a form that is as close as possible to the original collection owner's schema, and to develop expertise in understanding and manipulating the metadata to provide it to others in suitable formats. Use of linked data to create connections between collections could reduce maintenance issues and facilitate inter-linking with other aggregations.

The aggregator should make the aggregated metadata accessible by providing APIs and supporting standard protocols to support developers, and adding to access mechanisms when needed by keeping a watching brief of technology and emerging developer requirements.

8.4 Aggregated Metadata Consumers

Aggregations of metadata must provide added value and adhere to open access principles so as to maximise their exposure and encourage service developers and others, such as researchers, to make use of them. It is important to work with service developers to provide the capabilities they need, and thus the metadata to drive the services that their end-users need. The aggregator should also be prepared to provide aggregated metadata to other aggregators such as Europeana in the format of the defined schema. Furthermore, as many users use Google to search, aggregations of metadata should facilitate indexing by Google and other search providers to support discoverability.

9 Recommendations

Our analysis provides evidence to support the development of an aggregation of metadata for images and time-based media and generates a series of recommendations which should contribute to the sustainability and cost-effectiveness of any aggregation service. There is little that is distinctive about aggregating metadata about images and time-based media, although the variety of implementations of standard metadata profiles leads to complexity in any harmonisation attempt. The conclusions and recommendations arising from the analysis are often not specific to the media format being aggregated, but are related more generally to aggregating metadata.

9.1 Aggregator Recommendations

- Clearly communicate the benefits of an aggregation with descriptions tailored to each of the different stakeholder types (considering collection owners with different levels

of digital readiness separately) in language they understand to encourage participation.

- This support and communication should be extended throughout the life of an aggregation, so that the benefit of participating is continuously clear to stakeholders, and particularly collection owners.
- For large collection owners whose metadata are readily available for harvest and are less likely to need technical support than others, make apparent the benefits of contributing to the aggregation to encourage deposit.
- As an integral part of any project to develop aggregations of metadata educate potential contributors within HE and FE institutions regarding the role and potential value of commercial service developers to the HE and FE community in this context.
- Make the resources described within an aggregation of metadata accessible at a granular level, i.e. a direct link to the full description of the resource in the host collection owner's site should be included in the metadata.
- Make it as easy as possible for collection owners to contribute metadata: support multiple different ingest formats and protocols due to the varying levels of digital readiness of collection owners. Gather metadata and updates through both harvest and submission, and provide guidance to those who would like to digitise their metadata but have not yet done so. Work with lawyers to develop a process for legal agreement that minimises effort for metadata contributors.
- Adopt different approaches for collections with different levels of digital readiness, all of which should focus on getting collection owners to describe their own content, and deposit metadata themselves as far as possible, whilst providing guidance when needed. For those who have some technical experience, tools should be provided to enable them to deposit metadata easily, in the best form they can provide.
- Make aggregations of metadata accessible to the collection owners who contribute, and, if possible, to the general public, i.e. do not restrict an aggregation to use for educational purposes only.
- Include in the collection policy or contributor agreement a requirement that contributors grant permission to the aggregator to provide publicly accessible or harvestable thumbnail images for all visual resources (images and moving images), and clips for film and sound resources.
- Provide a simple decision-making tool for collection owners, such as an online decision tree regarding what, if anything, collection owners can do with their collections in relation to the legal rights for metadata. This would be of particular help to smaller collections, and those with mixed provenance.
- Determine the aggregation model that would be most appropriate for an aggregation of metadata for images and time-based media. Initial indications are that the mixed model may be most appropriate; although further work is required to arrive at a conclusive recommendation.
 - As it is likely to be difficult to agree a common schema, an aggregator should standardise metadata only if there is a clear need from service developers or end-users to do this, and then the aggregator should standardise into a suitable schema (agreed by relevant interested parties) only those fields that will be used. This schema should contain a link back to the full record on the collection owner's site.
- Develop a collections policy in line with end-user needs, and prioritise inclusion of such collections in an aggregation of metadata.
 - Determine end-user needs for image and time-based media discovery and create aggregations accordingly, possibly around subject areas or themes, as a precursor to creating a more comprehensive aggregation. The latter would contain metadata about a large number of resources across a broad range of subject areas and those metadata would be structured in such a way as to would facilitate filtering by, among other things, subject, licence type and resource type (i.e. these things should be included in the metadata).
 - Consider prioritising inclusion of collections that use formats favoured by end-users that can be readily used without additional software – and use other formats where they are particularly important for specific disciplines or areas (e.g. performing arts).

- Engage with service providers early on in development of any aggregation and work closely to understand their needs and those of end-users. Produce guidance for service providers to encourage recognition of collection owner brands, particularly alongside search results, and provide a link from search results that takes the user directly to the resource on the collection owner's site.

9.2 RDTF Management Framework Recommendation

- Make clear in the establishment of the RDTF management framework that an aggregation service that 'may be used by somebody occasionally' is valuable, as even infrequent access may enable service providers to create new services or tools that will in turn have a healthy end-user population.

10 Bibliography

All online reports and articles accessed 10 September 2010

Arms, W.Y., Dushay, N., Fulker, D. and Lagoze, C. (2002) A Case Study in Metadata Harvesting: the NSDL. *Library Hi Tech*, Vol. 21 (2), 228-237, <http://www.cs.cornell.edu/lagoze/papers/Arms-et-al-LibraryHiTech.pdf>

Blossom, J. (2009) *Content Nation: surviving and thriving as social media changes our work, our lives, and our future*. John Wiley & Sons, Inc.

Brand, A. Daly F and Myers B, (2003) Metadata Demystified, http://www.niso.org/standards/resources/Metadata_Demystified.pdf

Brosnan K (2005) Final report: Learning to Learn Project, <http://www.ioe.stir.ac.uk/research/projects/l2l/docs/finalreport.pdf>

Casey J (2004) Intellectual Property Rights (IPR) In Networked E-Learning – a beginners guide for content developers. JISC Legal service, http://www.jisclegal.ac.uk/publications/johncasey_1.htm

Charlesworth A (2005) Rights in digital environments, www.jisc.ac.uk/uploaded_documents/JISC%20Rights%20in%20Digital%20Environment%20Report.pdf

Charlesworth,A., Ferguson,N., Schmoller,S., Smith,N. and Tice,R. (2007) *Sharing eLearning Content: a synthesis and commentary*. Project Report

Fay, E., (2008) Metadata Tools for JISC Digitisation Projects of still images and text, http://www.jisc.ac.uk/media/documents/programmes/digitisation/jisc_metadata_tools_fay.ppt

Groat, G. (2009) Future Directions in Metadata Remediation for Metadata Aggregators, <http://www.diglib.org/aquifer/dlf110.pdf>

L'Hours, H. (2007) Content Packaging for Complex Objects: The METS Model, http://www.jisc.ac.uk/media/documents/programmes/digitisation/jisc_metadata_seminar_content_packaging_web_final.ppt

Liu, X., Maly, K., Zubair, M., Hong, Q., Nelson, M.L., Knudson, F. and Holtkamp, I. (2002) Federated searching interface techniques for heterogeneous OAI repositories. *Journal of Digital Information*, Vol. 2 (4), Article No. 106

McGill, L., Currier, S., Duncan, C., and Douglas, P. (2008) Good intentions: improving the evidence base in support of sharing learning materials, <http://ie-repository.jisc.ac.uk/265/1/goodintentionspublic.pdf>

Nicolas, N., Ward, N., and Blinco, K. (2009) A Policy Checklist for Enabling Persistence of Identifiers D-Lib Magazine Vol 15, Number 1/2, <http://www.dlib.org/dlib/january09/nicholas/01nicholas.html>

Pitts, K. and Sharp, K. (2003) The Technical development and benefits of a metadata aggregation and insertion tool. In G.Crisp, D.Thiele, I.Scholten, S.Barker and J.Baron (Eds),

Interact, Integrate, Impact: Proceedings of the 20th Annual Conference of ASCILITE.
Adelaide

Pringle M (2005) The Digital Picture: Final Report,
http://thedigitalpicture.ac.uk/documents/pdf/digital_picture_final_report.pdf

Rightscom (2009) Information Gathering Exercise for the Resource Discovery Taskforce Final Report, <http://rdtf.jiscinvolve.org/wp/files/2009/09/jisc-resource-discovery-report-final-20090908.pdf>

Rogers, L and Barker, P (2007) *Image Case Study: Community Led Engineering Image Collections*, <http://www.icbl.hw.ac.uk/images/ImagesCaseStudy.pdf>

Romer, W and MacMahon, C (2007) *Archaeology Image Bank Case Study: Final Report*. Project Report, <http://ie-repository.jisc.ac.uk/19/>

Shreeves, S.L. (2005) Barriers to Metadata Sharing via the OAI Protocol: A White Paper. IMLS Digital Collections and Content Project, University of Illinois at Urbana-Champaign, <http://imlsdcc.granger.uiuc.edu/3yearreport/docs/BarriersToInteroperability.pdf>

Shreeves, S. (2007) The Dynamics of Sharing: An Introduction to Shareable Metadata and Interoperability. *Society of American Archivists (SAA) 2007 Annual Meeting in Chicago*, <https://www.ideals.illinois.edu/handle/2142/2263>

Shreeves, S.L., Kaczmarek, J. and Cole, T.W. (2003) Harvesting Cultural Heritage Metadata Using the OAI Protocol. *Library Hi Tech*, Vol. 21, No.2, 159-169

Shreeves, S.L. and Kirkham, C.M. (2004) Experiences of Educators Using a Portal of Aggregated Metadata. *Journal of Digital Information*, Vol 5 (3)

Swann and Awre (2006), Linking UK Repositories, a scoping study a report commissioned by the Joint Information Systems Committee of the Higher Education Funding Councils, http://eprints.ecs.soton.ac.uk/14000/1/Linking_repositories_report.pdf

Tracking the Reel World (2008), http://www.tape-online.net/docs/tracking_the_reel_world.pdf

Waller, G. (2009) Issues surrounding syndicated feed deposit into institutional repositories, http://community.jorum.ac.uk/file.php/25/Issues_surrounding_syndicated_feed_into_institutional_repositories_GW.pdf

Whitelaw, L. (2007) Personal repositories online wiki environment (PROWE) project, Metadata report, <http://www.prowe.ac.uk/documents/PROWEmetadatareportv6final.doc>

11 End Notes

- ⁱ Note that a thumbnail is not typically considered as metadata as it cannot be harvested by a protocol like OAI-PMH that provides only text. However, the URL in the metadata can be harvested so the thumbnail may be displayed from its source server without the user realising it is in a different place.
- ⁱⁱ “The Semantic Web, Linked and Open Data”, page 2,
http://wiki.cetis.ac.uk/images/1/1a/The_Semantic_Web.pdf (Accessed 07/09/10)
- ⁱⁱⁱ <http://rdtf.jiscinvolve.org/wp/> and <http://ie-repository.jisc.ac.uk/475/>
- ^{iv} <http://rdtf.jiscinvolve.org/wp/>
- ^v Really Simple Syndication, an XML based document format for the syndication of web content so that it can be republished on other sites or downloaded periodically and presented to users. <http://www.rssboard.org/rss-profile>
- ^{vi} The *Atom Syndication Format* is an XML language used for web feeds, while the *Atom Publishing Protocol* is a simple HTTP-based protocol for creating and updating web resources. http://en.wikipedia.org/wiki/Atom_%28standard%29
- ^{vii} <http://rdtf.jiscinvolve.org/wp/implementation-plan/>
- ^{viii} The primary consultation method chosen was semi-structured interviews because of the exploratory nature of this project and the broad range of roles and experiences that bear on this topic. Semi-structured interviews gather richer information than online surveys. However, interviews are resource intensive. The questions posed to specific interviewees varied depending on their role in relation to the topic.
^{ix} http://www.innovateuk.org/_assets/pdf/competition-documents/briefs/tsb_metadatavalueindigcontentcomp-final.pdf
- ^x <http://www.europeana.eu/portal/>
- ^{xi} <http://imlsdcc.grainger.uiuc.edu/resources.asp>
- ^{xii} <http://framework.niso.org/node/5>
- ^{xiii} <http://www.jisc.ac.uk/whatwedo/programmes/digitisation>
- ^{xiv} http://www.tape-online.net/docs/tracking_the_reel_world.pdf
- ^{xv} <http://pbcore.org/2.0/> (Accessed 31/08/10; PB is Public Broadcasting)
- ^{xvi} <http://tech.ebu.ch/lang/en/MetadataEbuCore> (Accessed 31/08/10; EBU is the collective organisation of Europe’s 75 national broadcasters)
- ^{xvii} <http://www.ukoln.ac.uk/metadata/pns/pnstdcap/> (Accessed 31/08/10)
- ^{xviii} This comment was submitted as feedback via the blog on 18/11/2011 in relation to the People’s Network Discover Service DC Application Profile: “a discussion paper has recently been published [October 2010] outlining a number of possible options for future revision of People’s Network Discover Service DC Application Profile which was adopted by the Culture Grid, UK aggregator service. Please see <http://museum-api.pbworks.com/w/page/Culture-Grid-Profile> to access and respond to the paper.”
- ^{xix} Metadata Encoding and Transmission Standard is librarians XML standard, contrasted with for example the multimedia standard DIDL (Digital Item Declaration Language, also known as part 2 of the MPEG-21 standard) See Richard Gartner http://www.jisc.ac.uk/media/documents/techwatch/tsw_0801pdf.pdf
- ^{xx} http://www.ukoln.ac.uk/repositories/digirep/index/Images_Application_Profile (Accessed 15/09/10)
- ^{xxi} http://wiki.manchester.ac.uk/tbmap/index.php/Project_Outputs (Accessed 15/09/10)
- ^{xxii} <http://www.siobhandaviesreplay.com> (Accessed 31/08/10)
- ^{xxiii} Referenced at <http://www.jiscdigitalmedia.ac.uk/training/digital-performance-seminars/> with example at <http://www.jiscdigitalmedia.ac.uk/seminars/elements/> (Both accessed 31/08/10))
- ^{xxiv} <http://www.dshed.net/stars/preview> (Accessed 31/08/10)
- ^{xxv} <http://www.bbc.co.uk/learningzone/clips/> (Accessed 31/08/10)
- ^{xxvi} <http://www.w3.org/TR/mediaont-10/>
- ^{xxvii} <http://www.newscientist.com/article/mg20727715.400-google-twitter-and-facebook-build-the-semantic-web.html> (Article from New Scientist online 02/08/10 by Jim Giles, accessed 30/08/10)
- ^{xxviii} <http://nsdl.org/> (Accessed 06/09/10)

^{xxix} http://ecommons.cornell.edu/bitstream/1813/7897/1/Paper_21.pdf (Accessed 06/09/10)

^{xxx} http://www.ukoln.ac.uk/repositories/digirep/index/The_Images_Application_Profile

^{xxxi} http://wiki.manchester.ac.uk/tbmap/index.php/Project_Outputs

^{xxxii} <http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records>

^{xxxiii} The single respondent who would not participate in an aggregation owns approximately 2,200 digital images, films and sounds. The metadata are digitised but are available only for internal use. No follow-up was possible as the respondent did not provide contact information.

^{xxxiv} http://www.vads.ac.uk/picshare/report/picshare_final_report.pdf (Accessed 07/09/10)

^{xxxv}

http://community.jorum.ac.uk/file.php/25/Issues_surrounding_syndicated_feed_into_institutional_repositories_GW.pdf (Accessed 07/09/10)

^{xxxvi} It seems that when discussing an aggregation of metadata about images and time-based media, interviewees and respondents to the online questionnaire often confused the metadata with the content. This is perhaps an easy distinction to miss in a world characterised by services providing streamlined access to content via search results (e.g. Google). During the interviews the distinction between metadata that describe a digital resource and the resource itself could be clarified by the interviewer during the discussion. Similarly the distinction between services built on an aggregation and the metadata aggregation itself was clarified during discussions, particularly in relation to search services. In some instances, it appeared that respondents to the questionnaire lost those distinctions and, as there was no interviewer present, this was not clarified. Thus, the interview findings might be more robust than those of the online questionnaire. A number of respondents also stated that their view was not necessarily those of their organisation.