Jisc Content Projects and the Discovery Principles

Jisc Content Programme 2011-13

David Kay and Owen Stephens

September 2013

Contents

1. Scope

In January 2013, the 24 projects funded in the three strands of the JISC Content programme (2011-13)[1] were invited to participate in a self-assessment with reference to the 'practical steps' towards more effective resource discovery as identified by the Jisc and RLUK Discovery initiative[2].

The objectives of this review were

» To position project objectives and outcomes in the context of the Discovery principles[3]

» To derive intelligence that can be synthesised to provide constructive advice to future projects and to inform programme developments

» To inform the development of road maps and business cases for future digital content programmes and projects, where discoverability and reuse will be key factors

The content programme's call for proposals placed strong emphasis on discoverability, stating that 'projects should ensure they have embedded the technical means for their digitised content and metadata to be discovered and reused … employing stable URLs … processes for metadata to be harvested … exposing their metadata as linked data and via Application Programme Interfaces (APIs) … indexable by search engines such as Google.'[4]

In this context, it is therefore particularly interesting to understand the extent to which projects gravitated towards the practical approaches recommended in the Discovery principles. In particular, we would expect attention to be paid to open licensing and to the structuring of metadata and its exposure to discovery mechanisms. However, we may not necessarily expect digitisation projects to prioritise investment in significant additional technical tasks such as the development of APIs.

Set in this context, this report reflects on three sources of information

1. Responses from 20 projects to an online self-assessment survey, summarised in Section 3 and further detailed in Section 5
2. An external review of the web presence of 19 projects, based on readily repeatable 'manual' measures, reported in Section 4 and informing observations in Section 5
3. Case Studies of four projects presented Section 6

The report concludes (Section 7) with priority practical observations and recommendations on what works in practice for content focused undertakings, with constrained financial and technical resources.

[1] http://www.jisc.ac.uk/whatwedo/programmes/digitisation/content2011_2013.aspx
[2] http://discovery.ac.uk
[3] http://discovery.ac.uk/principles
[4] http://www.jisc.ac.uk/fundingopportunities/funding_calls/2011/06/econtentcapital.aspx

2. The Discovery Principles

The Discovery programme (2010-13) was funded by Jisc, working with RLUK and stakeholders in the UK libraries, archives and museums communities. With an overarching objective of making UK scholarly assets more widely, consistently and sustainably discoverable within the evolving online environment, the programme focused strongly on key challenges for the development of discoverability including

Business Cases ranging from aggregations to individual institutions and from academic researchers to the wider public

- Modes of discovery from cross-domain linkages to serendipity
- The relationship between licensing and reuse
- The role of persistent identifiers and key authorities, notably subjects, people and places

For the Jisc Content projects referenced in this report, there are particular dimensions that resonate more with archives and museums than with libraries, not least the balance between the resource itself and its metadata description in the context of discoverability.

Discovery recognised the importance of guidance in a field that is both fast-moving and technically challenging (where 'technicalities' involve legal, domain and IT expertise). In response the team at Mimas developed a set of twelve criteria, Practical Principles intended as a headline checklist for ensuring discoverability to assist institutions and individuals undertaking work involving the description and publication of scholarly assets[5]. The principles fall under the broad headings of Licensing, Metadata, Interfaces for Discoverability and Service.

Discovery also developed a micro-site to elaborate these Principles[6] and to provide practical guidance and exemplification through Case Studies, to which the four Case Studies from Section 6 of the report have been added. It is emphasised throughout that no project should expect to address all the principles, but rather that projects should prioritise carefully on the basis of relevance, timescales, funding and expertise. All the case studies are therefore cross-referenced against the Principles they chose to prioritise.

However, regardless of such constraints, it should be recognized, as proposed in section 7, that there are baseline responsibilities to be addressed in publishing clear terms of use, implementing persistent identifiers, following current Search Engine practice and developing a sustainability plan.

[5] http://guidance.discovery.ac.uk/approach
[6] http://guidance.discovery.ac.uk

3. Project Self-Assessment

3.1 – Method

Projects undertook the online self-assessment in January 2013, at which point two thirds of the projects had completed, with the remainder running to July 2013. Projects assessed their alignment with the 12 Discovery Principles, which had been previously applied in case studies to a wide range of scholarly services and initiatives. Responses were received from 20 of the 24 projects.

The practical steps (A to L) are broken in to four categories – licensing, metadata, interfaces and service as follows. Detail on each step is available at **http://guidance.discovery.ac.uk/approach**[7].

Licensing

A – Adopting open licensing of content

B – Adopting open licensing of metadata

C – Providing clear and reasonable Terms for reuse, where not freely open

Metadata

D – Establishing persistent identifiers for content assets (DOIs or URIs)

E – Using widely authoritative identifiers for such as Place and Name

F – Adopting commonly understood data formats, like Dublin Core or MARC

Interfaces

G - Optimising your metadata for reuse by third parties such as aggregation services and web search engines

H – Providing open and clearly documented mechanisms for accessing Applications Programming Interfaces (APIs)

I – Using your own APIs

[7] Readers should note minor re-casting of the headline descriptions in this report to take more explicit account of content as well as metadata publishing and discovery.

Service

J –Ensuring the currency and accuracy of your content and / or metadata

K – Supporting the discovery and / or data services you have developed

L – Measuring use of your content and / or metadata

Each step was self-assessed by projects on a 4-point scale that was converted into a score for visualization purposes, as follows:

4 = A expected outcome of our project (Green)

3 = A clear aspiration (Yellow)

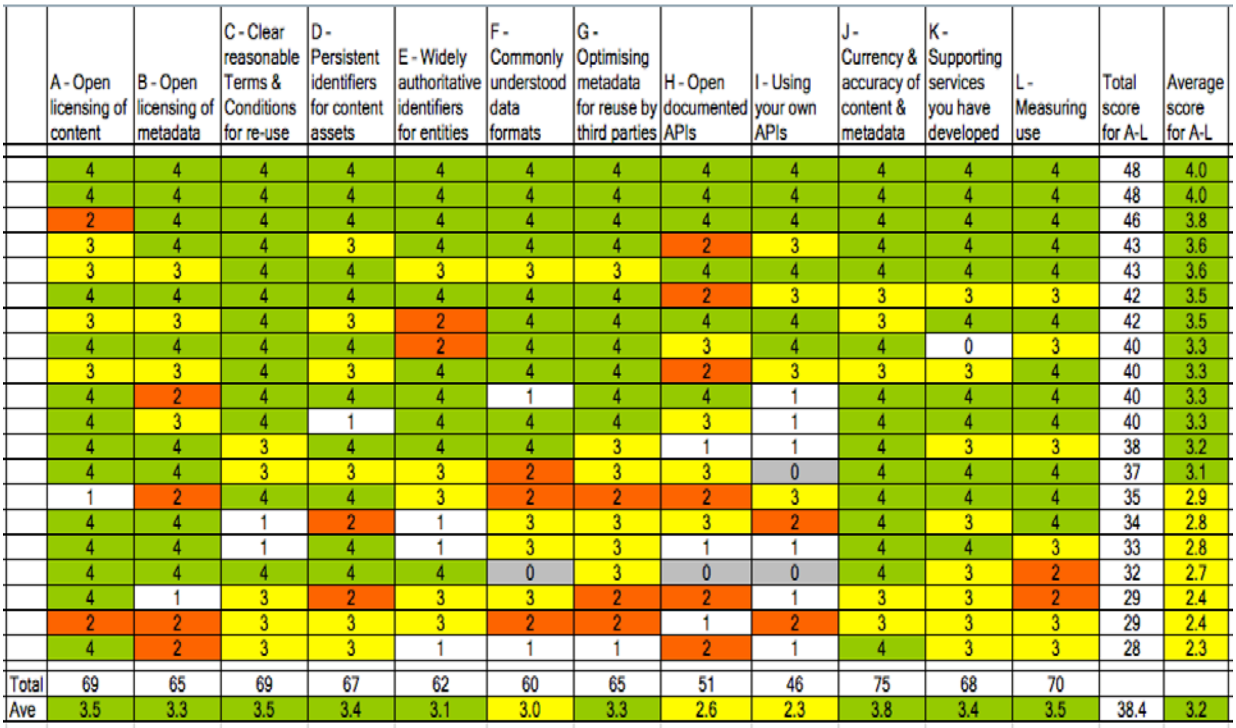
2 = A possible consideration (Red)

1 = Not relevant (White)

0 = Not understood (Grey)

3.2 – Responses

In order to identify patterns, the 20 project responses are ranked from the highest to lowest self-assessment for all categories in this 'carpet' visualization.



	A - Open licensing of content	B - Open licensing of metadata	C - Clear reasonable Terms & Conditions for re-use	D - Persistent identifiers for content assets	E - Widely authoritative identifiers for entities	F - Commonly understood data formats	G - Optimising metadata for reuse by third parties	H - Open documented APIs	I - Using your own APIs	J - Currency & accuracy of content & metadata	K - Supporting services you have developed	L - Measuring use	Total score for A-L	Average score for A-L
	4	4	4	4	4	4	4	4	4	4	4	4	48	4.0
	4	4	4	4	4	4	4	4	4	4	4	4	48	4.0
	2	4	4	4	4	4	4	4	4	4	4	4	46	3.8
	3	4	4	3	4	4	4	2	3	4	4	4	43	3.6
	3	3	4	4	3	3	3	4	4	4	4	4	43	3.6
	4	4	4	4	4	4	4	2	3	3	3	3	42	3.5
	3	3	4	3	2	4	4	4	4	3	4	4	42	3.5
	4	4	4	4	2	4	4	3	4	4	0	3	40	3.3
	3	3	4	3	4	4	4	2	3	3	3	4	40	3.3
	4	2	4	4	4	1	4	4	1	4	4	4	40	3.3
	4	3	4	1	4	4	4	3	1	4	4	4	40	3.3
	4	4	3	4	4	4	3	1	1	4	3	3	38	3.2
	4	4	3	3	3	2	3	3	0	4	4	4	37	3.1
	1	2	4	4	3	2	2	2	3	4	4	4	35	2.9
	4	4	1	2	1	3	3	3	2	4	3	4	34	2.8
	4	4	1	4	1	3	3	1	1	4	4	3	33	2.8
	4	4	4	4	4	0	3	0	0	4	3	2	32	2.7
	4	1	3	2	3	3	2	2	1	3	3	2	29	2.4
	2	2	3	3	3	2	2	1	2	3	3	3	29	2.4
	4	2	3	3	1	1	1	2	1	4	3	3	28	2.3
Total	69	65	69	67	62	60	65	51	46	75	68	70		
Ave	3.5	3.3	3.5	3.4	3.1	3.0	3.3	2.6	2.3	3.8	3.4	3.5	38.4	3.2

As set out in Section 2, projects need to prioritse their work and therefore some of these practical steps inevitably feature less strongly. The following headline observations therefore recognise that responses were necessarily determined by project scope, not least their primary focus on digitisation and content development, as well as representing some difficulties faced along the way:

» The principles identified by the Discovery programme were broadly understood, approved and adopted:

  - There were only 4 'not understood' returns out of 240 response cells.
  - Only three projects 'expected outcomes' in less than 4 of the 12 areas and reported at least 'clear aspiration' in less than 8

» Project responses and their emphases do not appear to be differentiated across the Programme Strands, which ranged in focus from large scale digitisation to surfacing assets in OERs

- Projects assessed especially highly in two categories:

    - Open Licensing (A, B, C) – with general commitment to Creative Commons licenses, though not necessarily CC0, and recognising that both content and metadata require licensing

    - Service (J, K, L) - notably in relation to maintaining currency and accuracy of content and metadata, though there was variable commitment to measuring use (L), perhaps made difficult by the frequent use of aggregators

- The lowest scoring projects (e.g. 29 or lower) typically assessed lower for a similar set of factors (E to I), all of which have technical and / or resource implications not directly related to content digitisation.
- More generally, core project objectives did not necessarily support prioritization of time consuming technical development and metadata classification; therefore it should be unsurprising though noteworthy that:

    - Projects assessed lowest for the technical aspects of reuse, especially relating to APIs (H, I)

    - Deployment of widely used and/or authoritative identifiers (E) was problematic, though some projects committed ambitiously to linked data; supporting comments indicated that scope / time available and domain specific vocabularies were major factors

Notwithstanding these self-assessments, Section 4 considers the actual manifestation of the projects on the public web at the end of the funded programme.

4. Early Web Assessment

4.1 - Approach

In addition to their self-assessment against the Discovery Principles (above), projects were subject to an external review of their web presence.

This exercise, undertaken at the end of the funded programme (September 2013), was based on the repeatable 'manual' measures of discoverability on the open web to be applied by the Spotlight on the Digital study (commencing Autumn 2013) to a much wider range of Jisc-related digitisation projects funded since 2002[8].

It should be noted that the Spotlight study will also apply automated 'webometric' tests to crosscheck these human findings, drawing on the Oxford Internet Institute Toolkit for the Impact of Digital Scholarly Resources (TIDSR)[9].

The assessment presented here, therefore, represents a snapshot early in the digital lifetime of these resources. This indicates generally strong initial achievements (such as Google ranking of the collections) and highlights distance to travel in terms of item level recognition and securing wider exposure through citation.

4.2 - Projects Assessed

At the time of this preliminary assessment, 19 out of 24 projects had enabled their digital assets to be accessed through a single URL.

Project Name	Actual Collection URL	Project Lead
3D Fossils Online	http://www.3d-fossils.ac.uk/	British Geological Society
Architectus	http://architectus.bcu.ac.uk/	Birmingham City
Board of Longitude	http://cudl.lib.cam.ac.uk/collections/longitude	Cambridge
Bombsight	http://www.bombsight.org/	Portsmouth
Broadside Ballads Online	http://ballads.bodleian.ox.ac.uk/	Oxford

[8] http://digitisation.jiscinvolve.org/wp/2013/09/09/spotlight-on-the-digital-how-discoverable-are-your-digitised-collections/
[9] http://microsites.oii.ox.ac.uk/tidsr/

BT Digital Archives	http://www.digitalarchives.bt.com/web/arena	Coventry
Contexts, Culture & Creativity	http://www.dance-archives.ac.uk/	Surrey
EngRich (aka Kritikos)	https://kritikos.liv.ac.uk/	Liverpool
Linking Parliamentary Records	http://liparm2.llgc.org.uk/en/home	King's College London
Manufacturing Pasts	http://www2.le.ac.uk/library/manufacturingpasts	Leicester
Manuscripts Online	http://www.manuscriptsonline.org/	Sheffield
Object based learning for HE	http://resources.jorum.ac.uk/xmlui/browse?value=OBL4HE&type=subject	Reading / University College London
Observing the 80s	http://blogs.sussex.ac.uk/observingthe80s/	Sussex
Old Maps Online	http://www.oldmapsonline.org/	Portsmouth
Open lives	http://humbox.ac.uk/3790/	Southampton
OVAM	http://www.onlineveterinaryanatomy.net/	Royal Veterinary College
UK Virtual Microscope	http://www.virtualmicroscope.org/	Open University
Virtual Microscope for Histology	**https://learn5.open.ac.uk/course/format/sciencelab/section.php?name=histology_microscope**	Open University
Zandra Rhodes Digital Study Coll'n	http://www.zandrarhodes.ucreative.ac.uk/p/welcome.html	University of the Creative Arts

4.3 - Collection Discoverability Tests

The following table, presented in alphabetic order of project name, presents results of six discoverability tests made at collection level in September 2013:

1. Is the page title well formed and informative?
2. Are there clear terms and conditions of use?
3. Is there a license for use, such as Creative Commons?
4. Does a search on the collection name yield a direct hit on Google p.1?
5. Does a search using 'sensible' related terms yield a hit on Google p.1?
6. Citations of the collection URL visible to Google in the '.ac.uk' domain?

The results are strong in most cases.

Project / Collection Name	Page title well formed?	Clear terms of use?	License for content use?	Collection title search ranking on Google p1	Collection found using "sensible" related terms?	Citations / links visible to Google in ac.uk domain
3D Fossils Online	Yes	Ok	CC (BYNCSA)	1	Yes	1
Architectus	Yes	Vague	CC (Unclear)	2	No	0
Board of Longitude	Yes	Clear	CC (BYNCSA)	2	Yes	1
Bomb Sight	Yes	Clear	CC (BYNCSA)	1	Yes	0
Broadside Ballads	Yes	Ok	CC (BYNCSA)	1	Yes	2
BT Digital Archives	Yes	Clear	CC (BYNCSA)	1	Yes	0
Contexts, Culture & Creativity	Yes	Ok	CC (BYNCSA)	1	Yes	0

EngRich	Yes	Clear	CC (BYNCSA)	1	Yes	3
Linking Parliamentary Records	Yes	(In hand)	(In hand)	No	No	4
Manufacturing Pasts	Yes	Ok	CC (BYNC)	1	Yes	3
Manuscripts Online	Yes	Clear	Aggr'n[10]	1	Yes	0
Object based learning for HE	Yes	Clear	JORUM & CC (BYNCSA)	No	No	0
Observing the 80s	Yes	Clear	CC (BYNCSA)	1	Yes	1
Old Maps Online	Yes	Clear	Aggr'n	1	Yes	8
Open lives	Yes	Clear	CC (BYNC)	1	Yes	0
OVAM	Yes	None	CC	1	Yes	1
UK Virtual Microscope	Yes	Clear	CC (BYNCSA)	3	Yes	1
Virtual Microscope for Histology	Yes	Clear	None	4	Yes	0
Zandra Rhodes Digital Study Collection	Yes	Clear	CC (BYNCSA)	1	Yes	3

[10] It is noted that aggregations fronting content controlled by and distributed across a variety of host systems face particular challenges in terms of achieving consistent licensing

4.4 - Observations on Licensing

The adoption of open Creative Commons licensing was a key requirement of this funding. Based on the table above, the following observations regarding how that licensing has been applied, are of interest. For further insights see Section 5.1.

- All collections had Terms and Conditions (T&C) of use and the majority were clear and linked directly to the selected Creative Commons licences
- T&C were somewhat weighty and therefore likely to be off-putting in a few cases where umbrella service / corporate T&Cs were linked
- The majority applied the Creative Commons 'CC BY-NC-SA' license granting reuse of content on the principles of Attribution (BY), Non-Commercial Use (NC) and Sharing-Alike (SA); see **http://creativecommons.org/licenses/by-nc-sa/3.0/**
- Two collections adopted CC BY-NC, not requiring Share-Alike, arguably realistically given the institutional governance under which educational content might be reused
- In a few cases, the CC license reference was not linked to the CC statement (for example on the CC site), which is an omission, especially after the effort of establishing open licensing
- Four collections (Boards of Longitude, Manufacturing Pasts, Open Lives, UK Virtual Microscope) decided to present licensing information at item level, thus allowing for variations; however they did not all make this approach clear in the site / collection level T&C
- A small number of collections were explicit about metadata licensing, as distinct from content; for example **http://cdm16445.contentdm.oclc.org/cdm/compoundobject/collection/p16445coll2/id/4551**
- Two services were acting as aggregation 'portals' for subsidiary collections and therefore open licensing of the external content was explicitly limited; however, in one case a standard CC license may have been applicable
- Two services mandated authentication before accessing open content, which can be a useful tactic so long as access is immediately granted (as in one case but not the other)
- Finally, a few sites made T&C / licensing into a feature of their home page as opposed to a footnote to be discovered; for example:
- **http://blogs.sussex.ac.uk/observingthe80s/** This was by no means obligatory but represents clear encouragement to the user.

4.5 - Item Discoverability Tests

The following table, which is presented in alphabetic order of project name, presents the results of three discoverability tests applied to a single randomly selected individual item or resource. The test criteria were:

1. Is the page title well formed and informative?
2. Does a search on the item name yield a hit on Google p.1?
3. Does a search using 'sensible' related terms yield a hit on Google p.1?

The results are less strong than at collection level in several cases. Whilst based on a single sample item, this potentially represents cause for concern based on the presumption that in many cases the student or researcher will know about the item level resource (e.g. a ballad) or an aspect thereof (e.g. a person or place name), as opposed to the collecting entity, which may be generic or otherwise focused.

Project Name	Example item	Page title well-formed?	Item title search ranking on Google p1	Can be found using "sensible" related terms?
3D Fossils Online	Fossil Information and Photographs - Coelogasteroceras dubius	No	No	No
Architectus	Image	Yes	No	No
Board of Longitude	Manuscripts - Letters from Longitude Commissioners to the Navy Board	Yes	1 - related page	Yes
Bomb Sight	Bomb Record - High Explosive Bomb at Petty Wales	Yes	1 - related page	Yes
Broadside Ballads	Image of manuscript - An excellent ballad of the mercers son of Midhurst	No	1	Yes
BT Digital Archives	Photograph – Cable at Long Witton	No	No	No
Contexts, Culture & Creativity	Images of choreography series - Blind Faith	No	No	Yes - related page
EngRich	Slideshow - Lecture 11: Reliability	No	No	No
Linking Parliamentary Records	Parliamentary record - HC Deb Monday, 19th March, 1923. vol 0161 c2065	No	1	Yes
Manufacturing Pasts	Newspaper Clipping – 'Lofty	Yes	1	Yes

Project Name	Example item	Page title well-formed?	Item title search ranking on Google p1	Can be found using "sensible" related terms?
	view of living'			
Manuscripts Online	* Note - This is an Aggregation			
Object based learning for HE	Transcript of interview with Ollie Douglas	Yes	1	No
Observing the 80s	Infographic - Sexuality in Thatcher's Britain	Yes	2	No
Old Maps Online	* Note - This is an Aggregation			
Open lives	Audio Interview with Jay G	Yes	1	Yes
OVAM	Photograph - Equine Radius And Ulna	Yes	1	Yes
UK Virtual Microscope	Meteorite sample image and factsheet - Adrar	Yes	No	Yes
V Microscope for Histology	Specimen - Temporal Lobe, Variant CJD	No	No	No
Zandra Rhodes Digital Study Coll'n	Image - Silk Coat	No	1	Yes

** Note – These criteria have not been applied where aggregation approaches rely on the varied strategies of source services for item exposure*

5. Discovery Principles: examples & appraisal of adoption

This section provides further detail about how the 2011-13 digitisation projects reported themselves to be addressing the Discovery principles. It draws verbatim on the commentary provided by projects (identified by the responding institution – for corresponding project names see Section 4.2) in their self-assessment survey responses, which identified practical pressure points and incentives, and reported deployment of a combination of the Discovery approaches.

The Wellcome Trust typified the underlying motivations in stating that, "the entire Wellcome Digital Library is aimed at supporting re-use of content, from use of creative commons licenses wherever possible, to providing user-friendly ways to find, view and download content."

Further Discovery case studies under each of these themes are published at **http://guidance.discovery.ac.uk**.

5.1 - Licensing

Here we consider the practical responses of content projects to open licensing:

- A – Adopting open licensing of content
- B – Adopting open licensing of metadata
- C – Providing clear and reasonable Terms & Conditions for reuse where it is not simply freely open

For explanation of A-C refer to *http://guidance.discovery.ac.uk/approach*

Open Content Licensing

The National Library of Wales decided "not to claim copyright of digitised images, and has determined to make this clear to users."

At **the University of Oxford,** Creative Commons licenses are under consideration for most Bodleian image assets, including those from the Ballads project.

The British Oceanographic Data Centre recognised "through working with the programme and other projects, that the OERs we create will be open content."

The British Geological Survey weighed the options available under Creative Commons to select the Share Alike Non-Commercial licence with attribution in order to satisfy the range of contributors.

Open Metadata Licensing

The University of Oxford determined that "all project metadata was to be released as Open Linked Data."

For the **University of Sheffield**, open metadata represents a 'win-win' opportunity: "Everyone we have approached, in Manuscripts Online as well as Connected Histories, recognises the value in sharing data within aggregation and federated search services - commercial content providers in particular. Once the licensing of subscription-based content has been agreed, we have found that commercial providers are very willing collaborators, primarily because these services drive users to their products whilst the subscription gateway protects their IP."

Achieving clarity with creators, contributors and providers

Birmingham City University recognised it is critical to "ensure content providers and partners are aware of the Open Licensing principles in advance; although this adds to the workload, a pre-publication approval process for how their (sometimes) repackaged content will be made available online provides re-assurance for rights holders."

Cambridge University Library made similar observations: "Open licensing of metadata is essential to our broader discovery strategy and our longer term ambitions regarding linked data. It was essential to be clear with those providing metadata that this was the case."

The University of Reading and University College London identified that such a process is also critical when working with OERs: "We learnt that despite a clear up-front policy that all resources created in this project would be released under a Creative Commons licence, creators still included content in their resources which had been originally released under unsuitable licences and failed to credit the originators." The project reflected that this points to the need for more extensive user training and the publication of clear guidelines for use of such as images taken from other sources.

The **University of Sussex** archive team reflected that "we initially anticipated that we already had permission to digitise and make available under our chosen licence, but discussion with (copyright expert) Naomi Korn helped us to realise we needed to seek individual permission from each Mass Observation contributor."

Reaching out in this way can be complex; for example, **the Royal Veterinary College** faced problems in developing consistent licences for institutions and commercial partners who were unfamiliar with Creative Commons.

Attaching Licensing Terms

The University of Southampton recognised the value of linking licensing terms directly with the content, stating that "at the outset, we decided that we would embed all license information on each item that we released. This has been time-consuming, but ultimately necessary, meaning that each published item has licensing information built in.

The University of Leicester emphasised the same lesson, advising to open publishers to "embed your chosen Creative Commons licences in the items themselves from the start - don't just rely on putting this information in the metadata".

Conclusions about Licensing

1. With clear explanation, fully open licensing (CCo) is often achievable, though Creative Commons variants such as Attribution may be important for content – see **http://creativecommons.org/licenses/**
2. Lack of awareness is likely to be a greater barrier than informed rejection
3. The effort of embedding in the individual items rather than broadly at collection level is strongly recommended

5.2 - Metadata

Here we illustrate the practical responses of content projects to the challenges of achieving a sufficiency of metadata:

- D – Establishing persistent identifiers for content assets, such as URIs
- E – Using widely authoritative identifiers for such as Place and Name
- F – Adopting commonly understood data formats, such as Dublin Core

For explanation of D-F refer to *http://guidance.discovery.ac.uk/approach*

Projects reflected that structured and standardised metadata is essential for enabling discovery and supporting reuse of content, whilst recognising that its cataloguing-style descriptions can become unfeasibly time consuming.

The Wellcome Trust reflected that "finding the balance between a suitable level of metadata, to enable efficient discovery, and the amount of time that can be feasibly dedicated to this is not easy."

The key to balancing these priorities lies in upfront tactical decisions and advance planning, which should be tested by trialling the processes involved.

Cambridge University Library advised that "metadata creation and processing takes a long time. You need to reflect this in your project plan, plus adopt any practices early that will make your life easier."

If metadata is to enable discovery, it should include the introduction of persistent URLs for referencing the original source where it remains hosted by the project itself.

The University of Reading and University College London highlighted the significance of "the development of a Persistent URL scheme meaning every object online can be accessed via a unique and permanent URL, enabling robust and simple republishing and citation."

Whilst potentially most exacting and time consuming, the adoption of widely used authorities offers significant discoverability benefits, especially in a linked data environment; authorities are likely to be an enduring choice, though metadata formats can be programmatically manipulated.

Cambridge University Library advised that "use of authorities for names, places, subjects etc with embedded identifiers is just as important (if not more so) than use of established metadata standards. It's comparatively easy to switch between metadata standards, very tricky to do the same with authorities and vocabularies."

Conclusions about Metadata

1. Persistent unique identifiers are critical at collection and item levels
2. There is a balance to be found between minimalist description and what might be called cataloguing
3. Whilst keyword tagging can bridge the gap, commonly used authorities (such as place, name or domain specific terms) offer significant benefits in terms of making connection within and beyond a collection

5.3 – Discoverability for machines

The Discovery guidance highlights technical means to ensure the resources are openly discoverable by machines, whether a search engine, harvester or a programmer's code.

- G - Optimising your metadata for reuse by aggregation services and web search engines
- H – Providing open and clearly documented mechanisms for accessing programming interfaces (APIs)
- I – Using your own APIs

***For explanation of G-I refer to** http://guidance.discovery.ac.uk/approach*

The impact of Search Engine Optimisation (SEO) is clearly recognised, though highly audience-specific resources may have less use for this level of diffusion.

The University of Bradford expressed the expectation that "ensuring that Digitised Diseases is indexed in search engines should ensure that users searching for a resource such as ours come across it."

The University of Sheffield alluded to the systematic processes and the benefits involved: "Regarding SEO, we have worked hard with Connected Histories to understand how Google explores our data. As a result, we now top-slice the top 100,000 person and place name entries in our search indexes and automatically generate searches against the data using these entries; we then expose the search results as static pages to help Google with its own indexing. This has really improved ranking and traffic."

The Royal Veterinary College envisaged a tight knit domain audience, thus concluding that "we don't imagine that SEO will be of much value."

As indicated in the discussion of Metadata, a greater investment of time is required to tap in to the amplification potentially offered by Linked Open Data, especially in introducing widely used authorities such as name and place and subject specific vocabularies.

Cambridge University Library assumed that "Our linked data work will make the content play better with search engines."

Whilst recognising that the development of programmatic interfaces (APIs) was not the core work of the content programme, a small number of projects undertook such developments.

The British Geological Survey calculated that "licence plus provision of API should facilitate re-use of content."

The University of Sheffield undertook to "document and make publicly available an API that provides access to our search indexes."

Conclusions about Discoverability for Machines

1. Search Engine Optimisation is tactically essential for resources to be discovered by non-specialist audiences and the payback for SEO effort is measurable
2. OAI-PMH is more specialist; it opens up content to multiple harvesters and is especially important in specific distributed communities
3. Provision of and commitment to sustain APIs is a more weighty undertaking that will be beyond the scope and expertise of many content-focused projects

5.4 - Amplification through distribution

Here we explore the efforts of projects to enhance discoverability by amplification (or 'diffusion').

Projects highlighted particular approaches to amplification, including embedding in other widely used resources, inclusion in recognised aggregations, and investment in linked data, typically underpinned by the use of persistent URLs.

Embedding in other resources, recognising the ripple effect of both global, local and highly domain specific amplifiers.

The University of Southampton encourages "wide scale deposit of related information in other places."

Cambridge University Library is "embedding as many links and hooks in Wikipedia as we can."

The University of Sussex is ensuring that "the materials are all available via the library catalogue, which should become Google searchable in the next year."

The Royal Veterinary College identifies that "links to WikiVet will be important."

Deployment through widely recognised and domain specific aggregations has been widely pursued; the range of broad reach destinations includes Connected Histories, Culture Grid (and onwards to Europeana), Humbox (and onwards through Xpert), Internet Archive, Jorum and Jisc Media Hub.

Where content is specialised, the habits of the target users should be clearly identified.

The University of Leicester advises projects to "work out what metadata aggregations you are going to contribute to and how you are going to do it - early on in your project."

Some projects need to target multiple destinations to reach their audience, which can be powerful but not without an overhead.

The University of Sussex has made the OER available "via various OER repositories such as Jorum and Humbox; the raw materials are also being deposited in Humbox and with ESDS Qualidata."

The University of Surrey has "distributed the material on JORUM with links to MERLOT, Humbox and WikiSource."

The University of Sheffield warns that "the problem with data aggregation and data sharing is the extent to which data has to be audited and re-engineered in order to fit new contexts … Successful shared data projects tend to have a limited number of datasets and engineer them for very specific contexts."

Aggregation involves either push to targets (data transfer initiated by the source) or pull initiated by interested parties (typically harvesting using the OAI-PMH protocol).

Push : **The British Oceanographic Data Centre** "will contribute metadata entries to several metadata catalogues to encourage data discovery – the SeaDataNet Common Data Index (CDI), the European Directory of Marine Environmental Data (EDMED), the MEDIN data discovery portal, the NERC Data Discovery Service and the National Aeronautics and Space Administration (NASA)."

Pull: **Birmingham City University** "allows OAI-PMH metadata harvesting / searchable by other open repositories."

Conclusions about Amplification through Distribution

1. Amplification can be achieved through a variety of push and pull tactics
2. Distribution across multiple aggregations is viable but comes with its own challenges in terms of effort and sustainability
3. Working to maximise impact in Wikipedia would be a primary target,potentially in collaboration with the Jisc Wikipedian Ambassador

5.5 – Amplification through user contribution

The possibility of engaging user contribution around specialised content presents an alternative to the technology and curator driven methods for enabling discovery and supporting reuse, which goes beyond the principles proposed by Discovery.

The prospect is enticing but should realistically be treated as a longer term strategic objective.

Cambridge University Library states that "enabling user contribution forms part of our broader plans for the Cambridge Digital Library."

The University of Sheffield recognises that "for Manuscripts Online this is unfinished business. Now that we have the core infrastructure in place we hope to explore a range of user-generated content / citizen science initiatives in order to build value and profile."

The projects identified a range of methods for eliciting user contributions, leveraging specialist and popular platforms as well as using methods of particular resonance with the resource.

Birmingham City University suggests that "the repository allows users to add their own keywords (like tag clouds). The web portal will encourage this when users see each asset entry."

The University of Southampton indicates that "HumBox allows users to comment on resources and to create their own collections with HumBox content."

The University of Sussex reports that "our web presence enables comment and we have created both Facebook and YouTube pages as well as depositing content in Google Drive and SoundCloud; these all allow users to engage with the material and with us."

The University of Surrey intends that "the Digital Dance Archive will support user scrapbooks."

Conclusions about Amplification through User Contribution

1. Understanding of the appetite for online user contribution relating to scholarly resources is at a relatively early stage
2. Popular social channels combine amplification with support for rating and review but not scholarly annotation and debate
3. Well-designed experimentation in this area should be welcomed

5.6 – Marketing

Projects demonstrated a strong sense of the value of dissemination, using a variety of digital signposts and targeted advocacy. Dissemination through social media has been seeded and driven by curators, especially using Twitter and Facebook. Social media also offer the prospect of viral amplification that might be triggered by user contribution (e.g. Recommendations such as Likes, Ratings, Reviews) and by resource presence and usage in public channels such as Youtube (as described by Sussex above).

More directed forms of digital marketing included use of QR Codes to spread the word within a known community, rather than as a viral mechanism.

The Royal Veterinary College "sent out QR codes linking to content … In particular to establish student annual projects for all partners."

Notwithstanding the potential of social media, the impact of OERs might be enhanced more directly by investment in systematic embedding within institutional teaching and learning.

University College London ensured that "reuse of the OERs will be supported by the role of a Teaching Fellow in Object Based Learning who is required to embed the use of digital resources in teaching and learning."

Jisc also animated content launches with digitally enabled Press Releases, drawing on the popular appeal of digitised material in areas such as UK history and science. These made demonstrable ripples through the web and the 'Twittersphere'.

The University of Coventry - BT Digital Archives

» Jisc PR - http://www.jisc.ac.uk/news/bt-digital-archives-to-celebrate-uks-telecoms-heritage-18-jul-2013

  Ripple - http://www.bbc.co.uk/news/uk-england-coventry-warwickshire-23363279

Cambridge University Library - Board of Longitude

» Jisc PR - http://www.jisc.ac.uk/news/the-longitude-problem-300-year-old-archive-opened-to-the-world-18-jul-2013

» Ripple - http://www.bbc.co.uk/news/science-environment-23514521

Conclusions about Marketing and Dissemination

1. Project planning should include provision for social media by designating Twitter hash tags and other repeatable tags
2. Dissemination drives are by nature 'bursty' and time-limited, typically focused on initial launch and this is even more the case in the digital realm
3. The logistics of continuity of dissemination cannot compete efficiently with maintaining No.1 position in Google

5.7 - Service

Discoverability goes hand in hand with sustainability – and especially keeping the resource up to date and assuring its quality. The Discovery principles highlight the importance of:

» J – Ensuring the currency & accuracy of your content and / or metadata

» K – Supporting the discovery and / or data services you have developed

» L – Measuring use of your content and / or metadata

The content programme projects represented a range of data lifecycles and curatorial situations with differing pressures and opportunities to enhance quality and to keep the content and the metadata up to date. The spectrum ranges from bounded historical sources to living topics which demand ongoing digitisation.

The Wellcome Institute recognises the London Medical Officer of Health reports as "a static resource."

The Royal Veterinary College, conversely, recognises sustainability of OVAM as "a real concern … [it] depends on institutions buying in to long term benefit of project."

A number of projects implemented or strengthened arrangements to systematise ongoing provider participation based on a variety of update cycles and methods.

The University of Sheffield has "licensing arrangements with all content providers to review their content every year"

The University of Portsmouth in the Old Maps service will have an "option to hide whole collections instantly if they wish to withdraw from the system"

The University of Oxford has ensured that "the partners have agreed to provide RSS feeds of updates to each resource."

Other projects hope to leverage user contribution in the form of edits and adaptation as well as reviews and comments, whilst recognising the associated implications for moderation and workflow.

Cambridge University Library suggests that "our long term strategy will rely on user editing / contribution directly via the interface."

Birmingham City University plans that "reviews and comments by users on resources [will be] moderated through the workflow. This will also identify any changes we might need to do - or updates to either metadata or content."

The British Oceanographic Data Centre intends that "OERs will be open and users will be able to adapt them to their requirements. We would ask that any alterations made be reported back to us so that we could improve our resources."

Conclusions about Service

1. When opening up content curators should assume expectations of continuity and consistency on the parts of users and of machines
2. Keeping up to date is a major consideration where the size or the detail collection is expanding
3. Jisc and other agencies can help by tracking and disseminating evolving techniques so curators can be best informed to maintain discoverability

6. Project Case Studies

6.1 - Board of Longitude

Context

The Board of Longitude was established in 1714 to administer a scheme to award prizes and grants to those attempting to solve the problem of determining longitude accurately while at sea, although later the work extended to other scientific endeavors. The Board of Longitude archive contains information about hundreds of submissions to the board as well as a wide range of other material including formal minutes to correspondence, maps and charts.

The archive is a substantial resource for studying the history & philosophy of science, as well as including information about many well known people, ships and places.

The project to digitise 65,000 pages of material from the archive has been undertaken by **Cambridge University Library** in collaboration with the **National Maritime Museum** and **an AHRC research project on the history of the Board of Longitude**. It was funded by the Jisc Digitisation Programme 2011-13.

Drivers and Benefits

Community Drivers
The Board of Longitude archive is being digitised in collaboration with the National Maritime Museum. As the archive contains many references to physical objects (ships, inventions, objects) - some of which are part of the museum's collection - there is a huge benefit from integration the archive (at Cambridge University Library) with the museum's collection.

The museum offers many resources for schools, and has created a rich set of **resources for schools based on the archive**.

The work of the Board of Longitude also has broad public appeal due to being the subject of a prizewinning novel and a television mini-series.

Global Value
The Board of Longitude archive contains material relevant to a global audience. As well as the global interest in the history and philosophy of science, the archive contains records that throw light on many other aspects of the period. For example the archive contains information relevant to the study of early European settlers in the Southern Hemisphere, which is of particular interest to researchers in Australia.

Institutional Drivers

The project to digitise the Board of Longitude archive is a collaboration between Cambridge University Library, the National Maritime Museum and an AHRC research project on the history of the Board of Longitude. The research project has created a huge amount of original content, which has been integrated with the digital resources. This integration between research content and digital objects adds considerable value to both projects, thus making the project an extremely good investment for the University of Cambridge and the library.

Learner Benefits

The University of Cambridge has existing courses in the History and Philosophy of Science at both Undergraduate and Postgraduate level that draw on the Board of Longitude archive. The digitisation of materials from the archive make the material immediately more accessible to students on these courses, and open up the possibility of the same materials being used in other courses both within the University of Cambridge and elsewhere.

Researcher Benefits

As the project to digitise the Board of Longitude archive is a collaboration with an existing **AHRC research project on the history of the Board of Longitude**, the use of the materials in research was a major consideration for the project. The process of digitisation in collaboration with researchers has been an enabler for research itself and has allowed researchers to understand the archive as never before. Sophie Waring, a graduate student working on the AHRC Board of Longitude project, said "What we've discovered has changed both the PhDs and doctoral level research in the project".

Perhaps the most obvious benefit to researchers is that materials from an archive that previously was relatively difficult to access is now freely available online.

The way the digital collection has been realised online with rich linking between relevant parts of the archive enables navigation and cross-referencing between parts of the archive and further connections with other collections both within and outside the University of Cambridge. This would be incredibly difficult or impossible in the physical collection.

Simon Schaffer (Principal Investigator, AHRC Board of Longitude project) and Sophie Waring (Graduate Student, AHRC Board of Longitude project) discuss the impact of the digitisation project in this video:

http://www.youtube.com/watch?v=fGsYphCa6QA

Practical Steps; Adopting Open Licensing

Content

Wherever possible images from the digitised Board of Longitude archive are available as JPEGs licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported License (CC BY-NC 3.0).

In cases where there are third party rights in relation to the digital objects (e.g. the family of a contributor to the archive has some rights in the content) then some negotiation has to be undertaken to agree under what rights the specific content will be offered.

Alongside the original digital content created through the digitisation of the Board of Longitude archive at Cambridge University Library there is also content from the National Maritime Museum collection, also licensed under a CC BY-NC 3.0 licence.

The digital archive also contains original content written as part of the related research project. While this content is technically categorised as 'metadata' it is clearly significant original content. In agreement with the contributors this has been licensed (along with the rest of the metadata) is licensed under a Creative Commons Zero (CC0) licence.

In the future it is possible that this written material is stored separately to the digital objects and associated metadata.

Metadata
Open licensing of metadata is essential to the Cambridge Digital Library broader discovery strategy and long term ambitions in relation to publishing and consuming linked data. In the case of the Board of Longitude Archive some of the information stored as 'metadata' is significant original writing contributed by external stakeholders (academics).

All metadata related to the Board of Longitude archive will be made available under a Creative Commons Zero (CC0) licence.

Having extremely clear communication with contributors on the licensing around the 'metadata' was extremely important, especially using a CC0 licence, which essentially puts the material into the public domain.

Using easily understood data models
The Board of Longitude archive is delivered as part of the Cambridge Digital Library, and as with the rest of the digital collections uses Encoded Archival Description (EAD) to store information about the nature and structure of the digital objects.

Making use of the **(Digital Archive Object Location)** allows the structure of the digital object to be stored in EAD without requiring the addition of METS as a wrapper for the digital object structure and metadata.

EAD is widely used in archives and is well understood within the archive community.

Deploying persistent Identifiers
Every digital object within the Cambridge Digital Library (including items from the Board of Longitude archive) has an http URI as identifier. Each page within a digital object similarly has an http URI as an identifier.

Cambridge Digital Library is interested in being able to offer identifiers for structural elements with objects (e.g. an identifier for each 'chapter' within a digitised book). However currently this is not possible and so page identifiers (e.g. first page of the chapter) have to be used in situations where this level of identification is required.

Establishing data relationships by re-using authoritative identifiers
For the Cambridge Digital Library the use of authorities with embedded identifiers for names, places, subjects and other entities, is just as important, if not more so, than use of established metadata standards.

Through established crosswalks it is often relatively straightforward to move metadata between standards, however it is challenging to do the same with authorities and vocabularies except where shared identifiers are used.

However, ensuring that appropriate identifiers are used take considerable effort. While the process is automated as far as possible, in many cases some level of human intervention is required to ensure people and places are linked to the correct external authorities. In some cases no appropriate 'authoritative' identifier is available.

External identifier schemes used include:

- National Maritime Museum identifiers
- VIAF IDs for People
- Getty Thesaurus IDs for Places
- Geonames IDs for Places (although this has proved less suitable for historical place names)
- Library of Congress IDs for Library of Congress Subject Headings

Optimising data for re-use
The Cambridge Digital Library supports several strategies for ensuring data can be, and is, re-used in other contexts and by other parties.

In the long term there is an ambition to use linked data as a mechanism of making all content in the Cambridge Digital Library available, including the Board of Longitude archive, and that this will enable a wide range of re-use. This may include using 'schema.org' to markup data within the digital library web interface to enable search engines to easily consume the relevant information.

In the shorter term work is being carried out to ensure that items and collections in the Cambridge Digital Library are indexed by search engines and to embed links to the collection in relevant Wikipedia articles.

Content from the Board of Longitude archive will also be contributed to relevant content aggregators. It is expected that content from the collection will be contributed directly to at least the following aggregations:

- Connected Histories
- Manuscripts Online
- 18th Century Connect

The possibility of contributing data to Europeana is also under consideration. Within the existing digital library interface there are straightforward links to:

- Download images
- Request reproduction rights
- Bookmark images
- Get a persistent link to the image for reference or sharing with others

Through these mechanisms users of the archive can easily re-use or reference content in their work.

Clear and documented APIs

The underlying data that is used to build the display of digital objects in the web interface is openly available in JSON format by adding '.json' to the end of the url used for viewing the object in the Cambridge Digital Library web interface. However, this access to the underlying data is not currently publicly documented.

Adopting widely understood data formats

The Board of Longitude archive uses several image formats of the digitised images for different purposes:

- TIFFs
- DZI (for zoom-able images in the object viewer)
- JPEG (for download)

All the data used to build the display of digital objects in the web interface is available in JSON format

Using your own APIs

The display of digital objects in the Cambridge Digital Library web interface is driven by underlying JSON data, which is used by a 'presentation layer'. The same JSON data is publicly available, although the library is not aware of any other users of this data.

As the presentation is separated from the content by this API, the library would be able to move to a different display mechanism in the future without requiring any change to the underlying data or systems. The same mechanism could potentially enable other partners to present the same content in a customised or branded viewer.

Collecting data to measure re-use
The Cambridge Digital Library interface uses Google Analytics to collect data on the usage of the collection via this route. However the library is also interested in other measures of re-use such as:

- Citation in books and articles
- Use in undergraduate or postgraduate courses
- Use by partners such as the National Maritime Museum

Collecting data beyond basic web analytics is challenging. For example while researchers may cite an item from the archive it is common practice to cite the 'original' physical item whether or not the digital representation was used.

Lessons Learned

- Metadata creation and processing can be extremely time-consuming, and should be factored into project planning
- Establish clear practices for metadata as early as possible in the process
- Ensure good communication with contributors on the licensing of any content they have contributed

Future Plans

- Contribution of content to major academic aggregations of relevant material
- Enabling user contributions around the content (as part of overall plans for the Cambridge Digital Library)

6.2 - Manufacturing Pasts

Context

The Manufacturing Pasts project led by the University of Leicester was a collaboration between the Beyond Distance Research Alliance, Centre for Urban History and the Library at the University of Leicester and the Record Office for Leicestershire, Leicester and Rutland to “create digital resources to enhance the learning and teaching of British industrial history during the second half of the twentieth century.” The project also created “a range of open educational resources based on the digitized resources which enable their practical use in the teaching of twentieth-century British history, planning and urban conservation.”

The Manufacturing Pasts project was funded under the Jisc Digitisation Programme 2011-13.

Drivers and Benefits

Global Value

With a focus on British Industrial History in the latter half of the 20th Century, and specifically viewing this through materials related to Leicester and the surrounding areas, the materials made available by Manufacturing Pasts are intended primarily for higher education lecturers and their students in the UK.

They are also relevant for schools and post-16 education. The materials are of direct interest to the local history community around Leicester and the Mayor of Leicester has expressed an interest in how the ‘story of Leicester’ can be told in different ways.

There is also an interest in the use of the resources created by the project in schools and post-16 education.

Institutional Drivers

As well as providing a rich set of digital resources for use within the university, the project and resources created provide a touch point with the local community, through the local history community, the Record Office for Leicestershire, Leicester and Rutland, and local businesses.

Such activity supports the institutional commitment to offering “an inclusive and accessible culture” and contributes towards the institutional aim of widening participation in higher education.

Manufacturing Pasts also relates directly to three of the aims of the University’s Learning & Teaching Strategy, namely, to:

- enrich and value the entirety of the learning experience for all students
- foster the critical intellectual development of all students through guided learning in a research environment
- provide an equivalent experience regardless of mode of learning or learning locations

Learner Benefits

The Manufacturing Pasts project was focused on producing resources relevant to learners, and teachers. The resources are already proving valuable and are being used in other institutions in teaching about the impact of de-industrialisation and urbanisation. The materials include oral histories that enliven the subject for learners, and materials such as staff newsletters and employee handbooks from the archives of local businesses that highlight aspects of social life such as how women were treated in the workplace.

For students moving beyond taught materials into research, Manufacturing Pasts provides engaging primary resources including factory plans and the oral histories and documents already mentioned above, together with support materials that – crucially – help them interpret these resources for themselves.

Researcher Benefits

The materials made available as the Manufacturing Pasts project are just one part of larger digital and print collections available from the University of Leicester and the Record Office for Leicestershire, Leicester and Rutland. Specific materials maybe of interest to researchers, especially those interested in the modern history of Leicester, and overall the materials may act as a showcase for the larger collections they are drawn from.

While the primary aim of Manufacturing Pasts was to make available materials relevant to taught students, as students move beyond taught course, especially 3rd year undergraduates and postgraduate students, they will find Manufacturing Pasts provides engaging primary resources including factory plans, oral histories and documents relevant to the social impact of industrialisation and urbanisation in 20th Century Britain.

Practical Steps; Adopting open licensing

Content

The content made available by the Manufacturing Pasts project was always intended for reuse in teaching in UK HE institutions and beyond. For this reason the content is licensed under a Creative Commons Attribution, Non-commercial licence (CC-BY-NC). Reuse of the content is a key measure of success, and the project wanted people to be able to reuse freely within an educational context. As the expectation is that resources will be contextualized and repurposed by those teaching using the resources, a 'non-derivatives' licence (e.g. CC-BY-NC-ND) was not seen as appropriate.

Where possible the project recommends content licences are embedded in resources, rather than just in external descriptions (metadata) of the resources – this is to ensure that as resources are reused, there is a good chance that licensing information continues to be available directly from the object. However, technical restrictions meant that the Manufacturing Pasts project was not able to implement this best practice in all cases.

Metadata

As the use and reuse of the resources is a key aim of the Manufacturing Pasts project, the metadata describing the resources is licensed as permissively as possible – using a Creative Commons Zero (CC0) licence, that puts the metadata itself, as far as is possible, into the public domain. The project clearly saw the metadata as a means to an end and wishes to see it reused as widely as possible on the basis this will drive the usage of the 'content', which as noted above is licensed under the more restrictive CC-BY-NC licence.

This permissive licensing of metadata is combined with a strategy to contribute the metadata records to a wide range of aggregators and search engines such as JORUM, the JISC Media Hub, Culture Grid (and from there to Europeana), Google, Wikipedia etc.

Each metadata record includes a statement on the licence that applies both to the resource described (CC-BY-NC), and to the metadata used to describe the resource (CC0).

Deploying persistent identifiers

The Manufacturing Pasts project recognises the need to have persistent identifiers for resources they make available. However due to technical issues with the system that hosts the resources (ContentDM) they have not been able to establish persistent identifiers for all the resources made available by the project.

Establishing data relationships by re-using authoritative identifiers

Due to the nature of the project many of the entities identified within the Manufacturing Pasts collection (places, corporations) are very specific to Leicester and the surrounding area.

The 'My Leicestershire History' project also based at the University of Leicester established standard forms for locations within Leicestershire, and these were reused by Manufacturing Pasts, providing consistency within the projects.

In both projects text strings were used to describe such entities, rather than establishing or reusing authoritative identifiers, this being in line with common practice in libraries and archives, and also supported by the systems used to create the digital archives.

Optimising data for reuse

The system used by the University of Leicester to host the materials made available by Manufacturing Pasts (ContentDM) is designed to expose content to Google with no additional work required locally. As a result of this the majority of traffic to the resources comes via Google.

ContentDM also supports the OAI-PMH protocol which is used widely by libraries to share metadata and related digital resources. Both Summon (a resource discovery system used by the University of Leicester) and Jorum (an aggregator of open educational resources for the UK HE sector) use this mechanism to aggregate the resources made available by Manufacturing Pasts.

Manufacturing Pasts is currently (March 2013) working towards contributing metadata to the Culture Grid (an aggregator of resources from UK Cultural Heritage institutions), which in turn will make the collection accessible via Europeana (a large aggregation of European Cultural Heritage resources).

As well as the dissemination of descriptive metadata as described above, Manufacturing Pasts have employed a PhD student to write articles for Wikipedia drawing on content from the Manufacturing Pasts project, as well as editing existing articles with relevant references to material where appropriate. The project has also used popular sites such as YouTube and Flickr to expose content in specific formats (video and images respectively) contributing directly to aggregations that are already widely used both within and outside the UK HE sector.

Clear and Documented APIs

The system used by the University of Leicester to host the materials made available by Manufacturing Pasts (ContentDM) has an API (http://www.contentdm.org/help6/custom/customize2a.asp) through which the collection can accessed, and also supports the widely used **OAI-PMH** protocol. Other aggregations and systems to which the resources are contributed also support APIs (e.g. Jorum API at http://www.jorum.ac.uk/powered/api-support; Summon API at http://api.summon.serialssolutions.com).

Adopting widely understood data formats

For the resources made available by Manufacturing Pasts an appropriate format was chosen depending on the type of resource – this included:

- JPEG
- MP3
- MP4

The project also created educational resources that combined one or more of the primary resources. These were made available in commonly used formats including:

- EPub – for combinations of text and image resources
- MP4 – for combinations of image and audio resources
- PDF – for timelines pointing to range of resources

For the metadata describing the resources, the Dublin Core Metadata Element Set was used - specifically the 15 properties in the “/elements/1.1” namespace, which are often seen as synonymous with ‘Dublin Core’.

Ensuring data currency and accuracy

Each metadata record used to describe resources made available by the Manufacturing Pasts project includes a ‘Contact’ statement which includes an email address for the library at the University of Leicester. As the metadata is distributed via a range of aggregators and search engines, this contact statement is included ensuring that when errors or issues are noted by those using or consuming the metadata there is a clear way of submitting a correction or update.

Where OAI-PMH is used to distribute the metadata or resources (e.g. by Jorum) if the resource or metadata is updated, this is automatically published via OAI-PMH which should ensure that aggregators using this mechanism always have an up to date copy of the metadata or attached resource.

Collecting data to measure use

The key mechanisms used to measure use of materials made available by Manufacturing Pasts are Google Analytics (embedded in the ContentDM platform used to host Manufacturing Pasts content) and usage reports for individual items directly available from ContentDM. The focus on the use of the content rather than any proxy measures makes the case for distributing the metadata widely obvious – as long as such metadata includes links back to the original content.

Lessons Learned

To aid practical re-use:

- Decide early on what licence you want to use by assessing what your objectives are
- Ensure licence terms are clearly indicated on the items themselves (not just in the associated metadata)
- Ensure licence terms are clear on any hosting sites
- Ensure items are easy to re-use technically
- Provide some practical guidance on how to re-use your resources (the project created and published a toolkit for this purpose)

Future plans

- Continue contributing metadata to selected aggregations
- Move all digitised collections to the same platform (ContentDM)
- Increase use of the resources created by local community groups in collaboration with our School of Historical Studies. This will include exploring their use on mobile devices.

6.3 - Manuscripts Online

Context

Manuscripts Online is an online resource which has indexed variety of online primary resources relating to written and early printed culture in Britain during the period 1000 to 1500. The indexed resources are provided by third party content providers (libraries, archives, universities and publishers) and include literary manuscripts, historical documents and early printed books. The indexed resources are searchable by keywords and for named entities which are identified during the indexing process. When search results are displayed textual snippets from primary sources are displayed to the user; however, the full text of resources can usually only be accessed by clicking through to the content owner.

Manuscripts Online is a collaboration lead by **the** Humanities Research Institute at the University of Sheffield and including the universities of Leicester, Birmingham, Glasgow, York and Queen's University Belfast. Manuscripts Online was funded under the Jisc Digitisation Programme 2011-13.

Drivers and Benefits

Learner Benefits

While the primary audience for the Manuscripts Online content is researchers, the project is extremely interested in the potential of the resources being used in teaching. A general call for interest in testing the Manuscripts Online site resulted in responses from across the globe, with a great deal of interest from the USA, where the popularity of Medieval Studies means the content being indexed by Manuscripts Online is of particular relevance.

Since the current focus of the site is research, the current interface is not optimised for teaching and learning. The API provided by Manuscripts Online could potentially be used to deliver something more targeted at this use.

Researcher Benefits

Researchers are the primary audience for Manuscripts Online. The project team believe it will "enable the HE research community (academics and postgraduates, within the UK and internationally) to address more effectively research questions such as the provenance of the Canterbury Tales manuscripts, the rise of English and the transformation of British society at this crucial period in our national narrative" :

http://www.jisc.ac.uk/media/documents/programmes/digitisation/manuscriptsonlineplan.pdf

The Manuscripts Online site enables collaboration and communication between researchers through the ability to add comments to search result items and by facilitating blogging of discoveries.

The project also aims to add value to the search process through the use of 'activity data' which will help researchers to discover relevant material which might otherwise not have come to their attention.

Practical Steps

Requiring clear reasonable terms and conditions

The Manuscripts Online site is based around the indexing of content from third parties. This means that a consideration of what Manuscripts Online and users of the site can do with the indexed content is at the heart of the process of adding content to the site. When a new resource is to be added to Manuscripts Online a content licence is drafted in collaboration with the content owner. This licence will cover taking acquisition of data, processing data and making the data available in the search engine. It will also include terms regarding what users of Manuscripts Online can do with regards the content from that provider – for example whether they can view site snippets.

The Licence covers 3rd parties using portions of the Manuscripts Online index, and so covers the use of the Manuscripts Online API by third parties.

Terms and Conditions of use are made available on the Manuscripts Online website:
http://www.manuscriptsonline.org/terms

Deploying persistent identifiers

The team responsible for Manuscripts Online have been producing web resources with persistent URIs for over 10 years, and so for materials hosted by the project, there are persistent, if not always user-friendly, URIs. While acknowledging the utility of more user-friendly URIs the team have been concerned that supporting such URIs could introduce an additional application tier that would require maintenance over time and may possibly be less sustainable than the URIs currently supported.

As well as providing persistent links to any materials, there is also an important requirement to support the ability to link to search results in a persistent manner. Learning from experience gained with the Old Bailey Online, Manuscripts Online supports a 'cite this page' option on every page, including search results, in order that researchers can easily find an appropriate and persistent URI for citation.

While not currently supported the team is also interested in persistent links to a specific set of search results (the current links simply re-run the same search which can lead to different results where additional data has been added between the original search and any subsequent uses of the search's URI).

Establishing data relationships by re-using authoritative identifiers

As many users of the Manuscripts Online site will want to discover resources then link through to the original content on an external site, the re-use of identifiers from third parties is very important. However, this comes with challenges as some of the content being linked to is not available freely and so even with the use of authoritative identifiers for third party resources links can fail due to authentication barriers.

Manuscripts Online does not currently use identifiers from external sources for named entities such as places or people.

Optimising data for reuse

Manuscripts Online aims to ensure content is well indexed by Google and other search engines. Building on lessons learnt from the Connected Histories project, Manuscripts Online plans to generate static pages for specific search results to ensure the site can be successfully indexed by search engines.

Clear and Documented APIs

Through previous projects, specifically the Connected Histories project, the Manuscripts Online team has a clear understanding of the challenges involved in creating a community around an open API. This experience suggests that APIs work best when focused on delivering very specific value over small amounts of data. The Manuscripts Online team are currently (10th March 2013) looking at running a workshop to work with potential users of an API to understand what such an API should support and how it might be used.

Using your own APIs

The Manuscripts Online public site is built using a web API which provides access to the underlying index. This makes the API essentially inseparable from the service. The project team believes that this approach makes it easier to develop, test and maintain the public interface because it is not tightly bound to the back end software. This is especially advantageous when different components are being worked on by different development teams, with a well documented API enabling each team to work independently but with a high degree of confidence in the outcome.

Any third parties using these APIs can have confidence that they will be well maintained and documented, as they are essential to the ongoing delivery of the Online Manuscripts site.

Collecting data to measure use

Manuscripts Online is using Google Analytics to collect usage information. Manuscripts Online relies on third parties to supply content to be indexed, with a charge to the content provider to have their content added to the index. This means that measuring and demonstrating use of the site, and 'click-throughs' to the full content is key. Content providers may also measure how contributing to the site affects traffic to their content, and at least one contributor to the Connected Histories project was keen to contribute more content because of the positive impact they measured.

Future plans

Embed Manuscripts Online within its native research community (medieval studies) through the following strategies:

- Implementing search engine optimisation to ensure the resources appear in response to appropriate search engine queries
- Ensure the resource is well linked from Wikipedia articles where relevant
- Aim for a 'critical mass' around the resource

Further resources will be added to Manuscripts Online beyond the period of JISC funding, with a view to enlarging the site's content and its value to research, by disseminating information about the site's purpose and value to the wider research community. Additional resources will be included in the site subject to a fee.

6.4 - Officer of Health Reports

Context

The Wellcome Library has created a major free online dataset covering public health in London from the mid-19th century to the late 20th century, based on the reports of the Medical Officers of Health in Greater London between 1848 and 1972.

The Medical Officers of Health reports include a wealth of information crossing boundaries between epidemiology, medical history, social history, local history and political issues. The creation of the digital collection has initially focused on those reports from Greater London as this covers a large population and has particularly comprehensive reports.

The digital collection includes 5,800 reports, the majority from the Wellcome Library collection and 580 supplied by the London Metropolitan Archives. The creation of the digital collection was funded under the Jisc Digitisation Programme 2011-13.

The full collection will be available from November 2013 and will provide:

- Free global access to the digitised MOH reports.
- Full-text searching and a searchable index of tables.
- Tabular statistical data in XML format.
- Improved catalogue searching for MOH reports.
- Interpretive information, including guidance on using the reports.

Reports are made available as they are digitised and can be accessed via the Wellcome Library catalogue e.g. http://search.wellcomelibrary.org/iii/encore/record/C__Rb1823998

Drivers and Benefits

Community Drivers

The Medical Officers of Health (MOH) reports are already used extensively. The Wellcome Library believe that the availability of the MOH reports for Greater London in digital form offers the opportunity to engage a wider range of users in the collection.

The digital collection will be of interest to a wider public audience including local record offices (specifically those in the Greater London areas covered by the MOH reports), family history researchers and schools.

Global Value

The Medical Officers of Health reports collection covers Wales, Scotland, Ireland and England, as well as including data from some British colonies. As such the collection is of interest to a global community of researchers and in particular there is interest from researchers in the USA.

While the current digitisation activity focuses on the reports from Greater London, this is still of significant interest globally as the reports cover a large and diverse population, and of course London itself is the focus of a wide range of research.

Institutional Drivers

In 2009 the Wellcome Library laid out a Transformation Strategy supporting a vision to be "the pre-eminent destination – physical and online – for anyone interested in exploring health in its cultural and historical contexts."
Strategic digitisation is a key part of this strategy, and the Wellcome library continues to develop its capacity to create, store and deliver digital content.

As such the digitisation of the reports of the Medical Officers of Health is part of the library's ongoing strategy and forms just one part of a large, and growing, digital collection.

Researcher Benefits

The Medical Officers of Health (MOH) reports collection is used extensively by researchers from both inside and outside the academic sector. The reports cover issues relevant to a wide range of disciplines – from epidemiology to political and social history.

The digitisation of a the MOH reports from Greater London makes a significant body of material more immediately available to interested researchers, and has the potential to reach a wider range of researchers.

The Wellcome Library would also like to go further than simple digitisation but also to extract tabular and other data from the reports and offer for download as discrete resources, making it easier for researchers to analyse the relevant statistical data.

Practical Steps; Adopting open licensing

Content

The entire Wellcome Digital Library is aimed at supporting re-use of content, from use of Creative Commons licenses wherever possible, to providing user-friendly ways to find, view and download content.

When content is digitised, the aim of the library is to make it available as openly as possible, taking into consideration any 3rd party rights in the content that may require some restrictions on the openness of the content.

In the Medical Officers of Health reports, the key 3rd party rights holders are London Borough's, which are the copyright holders for any of the reports that are still in copyright.

The Wellcome library approached each of the boroughs with rights in the content and requested the reports to be made available under an Open Government Licence (OGL), as this was a suitable licence for both textual/visual content and data such as that in tabular form in many of the reports. The OGL also had the advantage of being familiar to many of the stakeholders, some of whom would already be using it for other content or data publication.

The majority of stakeholders agreed to the OGL, but some preferred a licence that prohibited commercial use. In these cases the reports were made available to the Wellcome library under a Creative Commons Attribution-NonCommercial (CC-BY-NC) licence

The Wellcome library publishes all of the digitised content under a Creative Commons Attribution-NonCommercial (CC-BY-NC) licence. This is a consistent licence across the Wellcome library digital collections.

Metadata

A digital collection such as the Medical Officers of Health reports has a wide variety of metadata associated with it of different types. When agreeing standards for descriptive metadata, finding the balance between a suitable level of metadata to enable efficient discovery and the amount of time that can be feasibly dedicated to this is a challenge.

Bibliographic metadata describing the digitised items is stored in the library catalogue. There is also metadata from the digitisation process such as pagination data. Finally there is metadata describing access and licence conditions.

While much of this metadata is available freely – either from the library catalogue or in a JSON representation extracted from a METS file for a resource – there is no specific licence applied to the metadata.

Using easily understood data models

The Wellcome Library use well established standards for digital content and related metadata. The standards used include:

- MARC21 for bibliographic records
- ALTO for OCR generated content
- METS for descriptive, administrative and structural metadata

Within the web interface for viewing digital objects the Wellcome Library include data in Schema.org markup. This mechanism of embedding structured data within HTML is understood by search engines, including Bing, Google, Yahoo! and Yandex, making the data easily understandable to users outside the library and archive domains.

Establishing data relationships by re-using authoritative identifiers

'Place' is a key concept in the Medical Officers of Health reports. Each report is linked to the official name of the district that generated the report – which may be a parish or a borough, and of course may be an entity that has changed in terms of its geography or even ceased to exist since the publication of the report.

The Wellcome library use Library of Congress name authority files to apply consistent terminology to these entities, rather than using 'identifiers', this being in line with common practice in libraries and archives.

Optimising data for reuse

The bibliographic and related metadata describing the collection is fully incorporated into the Wellcome Library catalogue. From here the records are contributed onwards to **COPAC**, and aggregation of records from over 70 major UK and Irish libraries.

The Wellcome library is considering the possibility of contributing data from the collection to major aggregations including Europeana and the Internet Archive.

The pages that display the content are crawled by Google and other search engines, and use Schema.org markup to enable such search engines to extract structured data describing each digital object.

The Wellcome library will also create a 'microsite' (a set of dedicated webpages) that will expose the collection. These pages will be focused on researchers and the primary audiences for the collection and will also be crawled by search engines, such as Google, ensuring the presence of the collection in the major 'web search' destinations. The Wellcome library found that adding this markup had an immediate impact on how their digitised resources appeared in Google results, and a subsequent increase in traffic from Google to their site.

When content is viewed on the Wellcome Library site, an 'embed' option is offered to enable re-use of the content easily, while allowing the Wellcome library to continue to track usage, as the content is still delivered from the Wellcome Library, and has not been copied by a third party.

Clear and Documented APIs

The Wellcome library rely on APIs to deliver the content online, including an API to their library catalogue for the bibliographic description, and the delivery of data in JSON to the online 'player' that displays the digital resources.

These APIs are publicly accessible, but not promoted by the library currently.

Ensuring data currency and accuracy

The Wellcome library has a well-established long-term preservation strategy for digital content, and the digitised Medical Health Officers reports will be maintained or migrated to future formats and/or systems in line with this strategy.

The Wellcome library expects to eventually digitise the remaining 90% of the Medical Health Officers reports (i.e. those covering areas outside Greater London) and so expects the collection to grow over time.

However, once digitised the collection is a static resource, without any significant updates to metadata or content expected to individual resources once they have been digitised.

Using your own APIs

The Wellcome Library 'player' for digital content, which displays the items from the Medical Health Officers reports collection along with their associated bibliographic description, consumes data in JSON format. The same JSON is available publicly, but not currently promoted.

As the presentation is separated from the content by this API, the Wellcome library would be able to move to a different display mechanism in the future without requiring any change to the underlying data or systems.

Collecting data to measure use

The major mechanism for collecting data on use of the content within the Wellcome Library site is Google Analytics, which is used in many different ways across the site. While this delivers large amounts of rich data, there are also shortcomings, for example it is currently difficult to link Google Analytics data back to specific items within the collection, as the URLs used and reported on by Google Analytics are 'opaque' – i.e. it is not possible to tell to what item or collection they refer simply by looking at them.

Future plans

The Medical Officers of Health (MOH) reports collection is a single set of resources in a much wider strategic and ongoing digitisation effort at the Wellcome library. As such the future plans in the library are generally across all of its digital collections, and not limited to this one collection.

However, the library does have some aspirations that are directly relevant to the Medical Officers of Health reports collection including:

- Digitising the remaining 90% of the MOH reports collection
- The next phase of the Wellcome Digital Library will consider options for user contributions for all digital collections and crowd-sourcing options for specific collections including the MOH reports collection.
- Offer download of statistical data from the digitised reports for analysis, planned as part of a 'microsite' built around the collection
- Full text searching across the collection
- Better 'place' search including 'related places' and the addition of geo/polygon data for districts to records (geo-referencing data is already being added to catalogue records)

7. What works? Balancing theory, practice and reality

7.1 - Weaknesses

Commercial entities commit considerable ongoing efforts to ensuring their discoverability in the public web. To some extent, they set the pace in terms of discoverability and amplification, combining SEO with social media campaigns underpinned by persistence and functional tagging of assets. Their job is made easier than promotion of academic content because:

- Online channels are prioritised for investment as the principal routes to market
- Corporate mission and overlays such as pricing enable audiences to be more explicitly delineated
- Efforts can be linked to readily definable outcomes and business benefits

Curators of digitized scholarly and cultural resources very rarely find themselves in such a position, notwithstanding exceptional cases reported by such as the Rijksmuseum[11].

7.2 - Strengths

Notwithstanding the pressures to ensure resources play well in the arcane and changeable world of global search engines, curators and scholars may however have a number of factors working in their favour:

- Content – their content is typically unique and sometimes well known

[11] http://guidance.discovery.ac.uk/archives/casestudies/rijksmuseum

- Containers – there are reliable aggregation services where content can be deposited or mirrored, transformed and amplified
- Community – to a great extent, the community already exists that might communicate, recommend and cross-refer resources
- Commitment – some but not all resources, aggregations and portals have profile capable of attracting ongoing institutional commitment

7.3 - Threats

There remains however high likelihood that digital assets that are inappropriately named, poorly described and inadequately connected will disappear, becoming lost in digital space. Even for the most web-aware curators, there are particular challenges to be addressed:

- Discovery is not very useful without ease of access to and delivery of the content itself
- Legal rights pertaining to resources within a digitised collection are often complex with multiple parties involved
- Specialty brings an audience but potentially isolates a resource, especially in an increasingly cross-disciplinary environment
- Domain particularity needs to be balanced with openness to serendipity
- Aggregations bring their own issues of focus and sustainability
- Social marketing requires recurrent effort, suggesting a level of ongoing investment in staff dedicated to promoting the resource
- Machines, such as web robots, are particular and require precision

7.4 - Opportunities

It is clear from experiences in the Jisc 2011-13 digitisation programme (Sections 3 to 5) that a variety of fruitful practical responses to enhancing discoverability is accessible to curators and to projects whose primary focus is on content.

Whilst it is strongly recommended to consider each of the twelve Discovery Principles introduced and exemplified here, the experience of the 2011-13 digitisation projects emphasizes the importance of four foundational steps:

1. Establishing and publishing clear terms and conditions of use
2. Implementing persistent, resolvable, identifiers
3. Following current SEO practice
4. Developing a sustainability plan

These mission critical opportunities to enhance and sustain discoverability and to directly address the major threats are summarised as follows:

Establishing and publishing clear terms and conditions of use for all aspects of a digitised collection, including any descriptive metadata

The Manufacturing Pasts case study above describes how the project saw the descriptive metadata as one mechanism to drive usage of the digitised content they created. In order to ensure there was no barrier to this happening, they license the metadata as openly as possible, enabling all types of reuse, while maintaining the necessarily more restrictive licensing on the digitised resources.

The Cambridge Digital Library echoes this approach and in the Board of Longitude case study they describe how the open licensing of metadata as essential to their "discovery strategy".

Implementing persistent, resolvable, identifiers at every level you wish to make the resource addressable

Depending on the nature of the resource, this applies at collection level, item level, or even more granular levels such as 'chapter' or 'page'.

While ensuring that a resource continues to be discoverable in the long term is a huge problem area, moving into the area of digital preservation, providing persistent identifiers that can be used to locate the resource online aids discoverability in the short to medium term.

Providing persistent, resolvable, identifiers makes a resource easily 'citable' and ensures that those citations will continue to point to the resource in the medium term. The Manuscripts Online resource, described in the case study above, each page includes a 'Cite this page' link, which includes a persistent URL.

While there are some technical issues associated with providing persistent identifiers, it should be stressed that studies and discussions around persistent identifiers regularly come back to the conclusion that technology is not the problem, but rather persistence comes through the commitment of organisations and communities[12] [13]

Following good practice relating to search engine optimisation

[12] http://blogs.cetis.ac.uk/lmc/2010/02/09/jisc-persistent-identifier-meeting-general-discussion/
[13] http://www.ratswd.de/download/RatSWD_WP_2010/RatSWD_WP_147.pdf

There is a range of good practice available as regards 'search engine optimisation' (SEO), not least from Google[14]. SEO covers a wide range of practice, and continues to evolve. This can go from ensuring page titles are accurate and descriptive, to publishing sitemaps and robots.txt files, to adding markup to existing HTML pages.

In particular there is currently a focus from the major search engines on the use of 'schema.org' markup, which provides a mechanism of adding structured data to existing pages. The search engines use this to better index content and to add context to the information they crawl from the pages.

The Wellcome Library has recently implemented Schema.org markup in their 'viewer' page for digital objects, and found this had an immediate impact on the appearance of their resources in Google results (see Medical Officers of Health case study).

Developing a sustainability plan that can be executed by a core team

Unfortunately making content discoverable is not a one-off, nor a task that can be considered 'complete'. While underlying infrastructure may be established through up-front effort, the environment in which digitised collections exist is one subject to continual change.

As noted above, to ensure the persistence and resolvability or identifiers requires an ongoing commitment from the relevant organisations or communities. What qualifies as good SEO practice today may not be true tomorrow, as new routes of discovery and dissemination appear, and sometimes disappear.

Behind all of the four case studies in section 6 are teams with proven track records of providing sustainable services across long periods of time. Without this investment what is discoverable today, may become lost tomorrow.

[14] http://www.google.com/webmasters/docs/search-engine-optimization-starter-guide.pdf