Spotlight Web Assessment Report

David Kay, James Kay & Owen Stephens for Sero Consulting – March 2014

<http://digitisation.jiscinvolve.org/wp/?p=3001>

Table of Contents

1 - Headlines 1

2 - Assessment Areas 3

3 – Overall Scores 4

4 - Results for Individual Tests 6

5 - Related Spotlight Guide Topics and Tips 11

1 - Headlines

## Assessment Sample

The web assessment covered 217 digitised collections that received funding from AHRC Resource Enhancement, NOF Cultural Enrichment and Jisc programme across a range of universities and research institutions between 1998 and 2013.

## Undiscoverable Collections

Some collections become lost to the web over time (40 out of 217 tested, representing over £2m public investment). Explanations range from poor exposure to search engines, to the loss of web access to the content itself, to relocation elsewhere (within other local collections or aggregation services).

The headline reasons for failure in our sample were as follows: 26 cases where the Site no longer exists or is marked as closed without redirection; 8 cases where the assets have been incorporated elsewhere or the named collection cannot be identified within the site; 6 cases where the site still says it is beta / work in progress or site functionality is broken so the assets are not accessible.

The majority but not all ‘lost’ collections were linked to projects that took place before 2005; however, it is arguable that this is the effect of diminishing care over time rather than of technology, and therefore more recently digitised (and born digital) collections could follow the same pattern.

Whilst we were using the URL known to Jisc at the last review in 2009, there is a reasonable expectation that such movements should as a minimum be signposted from ‘old’ URLs. The Spotlight project therefore concluded that collections must be persistently available, with appropriate redirection where applicable.

## Undiscoverable Items

In most cases any individual user will be interested in finding specific items within a collection, and not in discovering the collection per se.

Content may be digitised thematically (driven by a research project or a funded initiative) thus falling under the umbrella of a collection. However, downstream use cases are likely to be driven by searches for the particular, or by themes that cut across collections. As a consequence in many cases collection groupings are not directly connected to later enquiries.

When searching for relevant resources, users are therefore likely to use terms relevant to the types of items they are searching for, rather than the collections that may contain such items. For example, to find the trial transcript of a murder in a specific London street, a possible set of search terms is:

“white cross street murder trial”

as opposed to:

“old bailey trials”

This is perhaps especially true for those users unaware of specific digital collections and using general search engines (such as Google) to find relevant resources.

The Spotlight project therefore concludes that users need items to be discoverable. See blog post at <http://digitisation.jiscinvolve.org/wp/2013/11/18/discovery-of-digitised-collections-vs-items/>

## The difficulty of measuring discoverability

Perhaps the true measure of ‘discoverability’ is how many people are accessing and using a digitised collection. The assessment carried out in the Spotlight project was unable to gather information on the actual use of collections, or what search terms had proved successful in locating an item within a collection. Such information would be available from log files and web analytics being collected in relation to individual collections.

Tests such as the ones described in this document at best offer only proxy measures for discoverability. Many of the tests described assess how far digitised collections follow established good practice for web publishing, and we can see that there are examples where good practice has been completely ignored but a Google search for an item by name, or by relevant search terms still finds the specific item within the collection easily. Anecdotally such cases seem linked to rare or unique items (e.g. a particular manuscript), while non-unique items (e.g. common objects, such a ‘spear’, where the context or description is what offers added value) can fare badly in terms of discoverability no matter what good practice is followed.

However, despite these limitations the assessments do highlight areas where good practice is not being followed, and so offer some insight into what actions could be taken to improve the discoverability of these digital collections.

# 2 - Assessment Areas

The web assessment undertook over 30 tests on each of the 177 websites for which the given URL resolved successfully, of which 9 were executed by software (including the Topsy service counts of social media interactions) and the rest manually.

The tests were selected to align with the patterns of user ‘discovery behaviour’ identified in the Spotlight literature review. The review report can be accessed at <http://digitisation.jiscinvolve.org/wp/2013/11/04/top-discovery-channels-in-online-user-behaviour/>.

Of the 14 behaviours identified through the literature, the following are most relevant to digitised content:

* General search engine
* Online library interface
* General web service (including Wikipedia)
* Google services
* Online research resource or database
* Recommendations from teachers
* Follow experts (including Twitter)
* Online social tools (including Twitter)
* Citation chaining

Based on this list, the web assessment focused on areas where the presentation of content on the public web (as opposed to local services such as the library interface and reading list recommendations) would be most likely to impact discovery – notably how well the collections and items play with search engines, Wikipedia and Twitter.

The results of the individual web assessment tests are set out in Section 4.

**Area 1 – Search Engine Discoverability (Collection & Items)**

1. Robots.txt file present?
2. Sitemap.xml file specified in robots.txt?
3. Item pages ‘crawlable’ by software?
4. URL well formed?
5. Page title present?
6. Page title well formed?
7. Page description present?
8. ‘Alt’ descriptions present for images?
9. Microdata present in item pages?
10. Title search ranking on Google page 1?
11. Can be found using "sensible" related search terms?

**Area 2 - Re-use (Collection & Items)**

1. Clear terms of use?
2. License for use?

**Area 3 - Wikipedia (Collection only)**

1. Collection has its own Wikipedia entry? Entry linked to the collection?
2. Collection directly referenced in another Wikipedia entry? Reference linked to the collection?
3. Number of website citations in Wikipedia

**Area 4 - Social Media (Collection only - Topsy statistics)**

1. Number of Tweets
2. Tweets from Influential Tweeters
3. Number of Tweets in the last years

**Area 5 - Basic usability (Items only)**

Items from Area 1 are measures of usability as well as search engine discoverability

1. URL well formed?
2. Page title present?
3. Page title well formed?
4. Page description present?
5. ‘Alt’ descriptions present for images?

# 3 – Overall Scores

## Overall Top Collections

Scored on these tests of collections and their items, using scales (0-5) where applicable, the ‘Top 12’ collections were as follows. As noted later, whilst there were other strong exemplars, the highest scorers typically performed best against the vast majority of the criteria.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Max 111 |  | Start | Fund | Lead |
| 94.5 | Archival Sound Recordings 2 | 2007 | Jisc | BL |
| 94 | Complete work of Charles Darwin online | 2005 | Jisc | Cambridge |
| 93 | Enriching the WW1 Poetry Archive | 2008 | Jisc | Oxford |
| 89 | Windows on Genius | 2011 | Jisc | Sussex |
| 89 | Bombsight | 2011 | Jisc | Portsmouth |
| 88.5 | Screen Online  |  | NOF | BFI |
| 83.5 | Medical Journal Backfiles | 2004 | Jisc | Wellcome |
| 83 | Observing the 80s  | 2011 | Jisc | Sussex |
| 81 | Opening Veterinary Access to Literature | 2011 | Jisc | RVC |
| 81 | Zandra Rhodes Digital Study Collection | 2011 | Jisc | UCA |
| 80.5 | Cabinet Papers 1915-1978 | 2007 | Jisc | TNA |
| 80 | Shakespeare Quartos Archive | 2008 | Jisc+ | Oxford |

## Overall Scoring Profile

The following graphs are constructed as follows:

* Vertical Axis – Score achieved by the collection / items for the tests described; the overall total available was 111 and differed for each of the four subsidiary areas graphed
* Horizontal Axis – The 177 collections tested, in the same order for each graph from overall best to overall worst score (left to right)
* The line graphed – The scores achieved by the collections, from overall best to overall worst score (left to right), thus showing relative performance for the tests graphed

The first graph represents the total web assessment scores for each of the 177 collections – from the top score of 94/111 (see table above) to the lowest of 20 points. Whilst the fall-off is broadly steady there is visibly a large number in the 50-60 scoring zone, where discoverability is demonstrated alongside plenty of scope for improvement - that group of 55 sites represents almost one third (31%) of the sample.

## Headline Category Scores

These four graphs illustrate differing performance patterns in particular areas of the assessment.



In every chart the 177 collections are displayed in descending order (left to right) of overall score (as per the previous illustration). The profile (steepness, spikes, etc) of each category graph therefore illustrates how collections perform differently in these particular areas of the web assessment.

Clockwise interpretation of the charts follows:

Collection score for web access (Scored out of 38) - This covers the manually assessed search engine discoverability tests (within Area 1 above).

* URL well formed?
* Page title present?
* Page title well formed?
* Page description present?
* Title search ranking on Google page 1?
* Can be found using "sensible" related search terms?

Collections generally score well / least badly here – as illustrated by the less steep fall-off of across the sample and the relatively less poor lower scores.

Two Items score for web access (Scored out of 44) – This covers the same tests from Area 1, but for items. The difference in scores between best and worst is greater than for collections. The fall-off is quite steady, rather than a sharp drop to a rump of stragglers.

Collection score for Wikipedia presence (Scored out of 14) - This covers all the Wikipedia tests (Area 3). It illustrates stronger engagement in the first two quartiles. However, apart from the very strong performance of some of the overall best scorers, the pattern is characterised by intermittently distributed spikes, representing those who do / do not work on Wikipedia citations. It is however noticeable that zero Wikipedia presence is concentrated in the final quartile.

For more on Wikipedia and discovery see <http://digitisation.jiscinvolve.org/wp/2013/12/05/what-wikimedia-can-do-for-digitised-content/>

**Collection score for social media presence** (Scored out of 15) – This covers the Twitter tests (Area 4). It illustrates a strong leading group followed by a sharp fall-off beyond the top 20+ overall performers. The pattern is broken by the spikes of around 10 collections that perform much better with Twitter than in their overall assessment position.

For more on social media and discovery see <http://digitisation.jiscinvolve.org/wp/2013/10/14/does-social-media-increase-discovery-of-digitised-collections/>

4 - Results for Individual Tests

## Discoverability (Collection & Items)

**Robots.txt file present?**

60% of collections had retrievable ‘robots.txt’ files

|  |  |  |
| --- | --- | --- |
|  | Present | Absent or unable to retrieve |
| Collection | 106 | 71 |

**Sitemap.xml file specified in robots.txt?**

0% of collections had a Sitemap.xml file specified in the ‘robots.txt’ files

|  |  |  |  |
| --- | --- | --- | --- |
|  | Specified | Not Specified | Absent or unable to retrieve |
| Collection | 0 | 106 | 71 |

Comment – It is possible, and common, to submit the location of Sitemap.xml files directly to Google and other search engines on a one by one basis and this may account for the absence of this information in any of the robots.txt files retrieved. The use of the robots.txt file to specify the location of the sitemap.xml is recommended as good practice alongside direct submission to the major search engines, as it allows other search engines to make use of sitemap.xml without the requirement to submit it directly to less well used search engines.

**Item pages ‘crawlable’ by software?**

85% of Items could be retrieved by web crawler software

|  |  |  |  |
| --- | --- | --- | --- |
|  | Crawlable | Not Crawlable | No URL given |
| Items | 150 | 6 | 21 |

URL well formed?

80% of collections had well formed URLs. In contrast, only 33% of item pages had well formed URLs.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Good | Ok | Poor |
| Collection | 142 | 19 | 16 |
| Items | 58 | 72 | 47 |

Comment – The ‘well formedness’ of URLs was subjectively assessed in relation to recommendations from Google and Sir Tim Berners-Lee. See <http://jisc.ac.uk/guides/make-google-searches-work-for-you#urls> for further information and examples.

Item Page Title present?

91% of retrievable item pages had a Title in the HTML <head>

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Present | Absent | Not Crawlable | No URL given |
| Items | 137 | 13 | 6 | 21 |

Comment - The presence of a Page Title is a very basic measure, and does not account for whether the page title is unique or descriptive of the page content. Whether the Page Title was ‘well formed’ in this sense was measured separately. Some item ‘pages’ were direct links to image or PDF files, and as such (i.e. non-html content) could not have a page title.

Page Title well formed?

98% of collections had well formed Page Titles. In contrast, only 47% of item pages had well formed Page Titles

|  |  |  |
| --- | --- | --- |
|  | Good | Poor |
| Collection | 173 | 4 |
| Items | 84 | 93 |

Comment - The ‘well formedness’ of page titles was subjectively assessed in relation to recommendations from Google which advise the use of unique and accurate page titles. See <http://jisc.ac.uk/guides/make-google-searches-work-for-you#titles> for further information and examples.

Item Page Description present?

19% of retrievable item pages had a description in the HTML <head>

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Present | Absent | Not Crawlable | No URL Given |
| Items | **28** | 122 | 6 | 21 |

Comment - The use of descriptions is recommended by all major search engines. This test simply checked for the presence of a populated <meta name=“description”> element in the HTML <head>, and did not attempt to assess whether the content of the element was a good description of the page content. Some item ‘pages’ were direct links to image or PDF files, and as such (i.e. non-html content) could not have a page description.

Percentage of images in page with ‘Alt’ descriptions present?

48% of retrievable item pages with images had ‘alt’ descriptions on three-quarters of the images on the page.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Unable to retrieve page/images | 0% | 1%-25% | 25%-49% | 50%-74% | 75%-99% | 100% |
| Items | 13 | 23 | 12 | 29 | 10 | **23** | **45** |

Comment - This test did not analyse the quality of ‘alt’ descriptions that were present. Some item ‘pages’ were direct links to image or PDF files, and as such (i.e. non-html content) could not have images with ‘alt’ text.

Microdata present in item pages?

1% of retrievable item pages contained microdata (structured data embedded in the HTML)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Present | Absent | Not Crawlable | No URL Given |
| Items | **2** | 145 | 9 | 21 |

Comment - The use of microdata is not yet well adopted across the web, and it is unsurprising that the use in the tested collections is so low. The two items that did have microdata present were from two collections hosted on the same digital library system, which is used by a number of UK institutions to host digital collections.

Title search ranking on Google page 1

For 95% of searches using a collection title searches, the relevant collection appears as the No.1 result in Google. In contrast, for 56% of searches using an item name or title the relevant item appeared as the number No.1 in Google.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Top 3 Hits | Lower on Page 1 | Not on Page 1 |
| Collection | 169 | 3 | 5 |
| Items | 100 | 8 | 69 |

Comment – Unique or rare items do better with Search Engines than might be suggested based on how far they follow recommended good practice for discoverability.

Can be found using "sensible" related search terms?

64% of collections were found easily using related search terms. In contrast, only 44% of items were found easily using related search terms.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Easily Found | ok | Not Found |
| Collection | 113 | 49 | 15 |
| Items | 77 | 34 | 66 |

## Re-use (Collection & Items)

Clear terms of use

52% of collections and similarly 48% of items had relatively clear (Clear/OK) terms of use.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Clear | ok | Unclear | None |
| Collection | 47 | 45 | 50 | 37 |
| Items | 47 | 38 | 50 | 42 |

License for use

20% collections and also 20% of items had some form (Creative Commons or other) of licence for use.

|  |  |  |  |
| --- | --- | --- | --- |
|  | CC | Other | None |
| Collection | 22 | 13 | 142 |
| Items | 22 | 13 | 142 |

Comment – Use of licenses, and especially Creative Commons, is a relatively recent practice and has typically not been retrospectively applied by curators. Where licenses are used they are appropriately rippled through and visible at the item level.

##

## Wikipedia (Collection only)

Collection Wikipedia article

7% of Collections had their own Wikipedia article, all of which linked back to the collection website

|  |  |  |
| --- | --- | --- |
|  | Yes | No |
| Collection entry | 12 | 165 |
| Linked back | 12 | 165 |

Comment – It is arguable that very few collections merit their own Wikipedia article and therefore that referencing within Wikipedia is a more important measure, as follows.

Collection directly referenced in another Wikipedia entry

74% of Collections were directly referenced in another Wikipedia entry, and 67% had references linked back to the collection website

|  |  |  |
| --- | --- | --- |
|  | Yes | No |
| Collection Reference | 131 | 46 |
| Linked back | 118 | 59 |

Number of website citations in Wikipedia

53% of Collections had one or more citations in Wikipedia

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 10+ | 6 > 9 | 1-5 | None |
| Collection Reference | 27 | 6 | 61 | 83 |

## Social Media (Collection only - Topsy statistics)

Number of Tweets

18% of Collections have had over 40 Tweets in total, but only 12% have had over 40 in the past year (Source – Topsy)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 81+ | 41>80 | 21>40 | 1>20 | None |
| Number of Tweets | 24 | 7 | 11 | 68 | 67 |
| Tweets in past year | 18 | 4 | 7 | 66 | 82 |

Comment – There was a significant divide between the highly active (see 80+ Tweets) and the very occasional and completely inactive (only 11 collections with 21-80 Tweets in the past year) – the pattern with using social media is one of extremes.

Influential Tweeters

18% of Collections have had over 40 Tweets in total, but only 5% had over 40 from Tweeters calculated as influential (Source – Topsy)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 81+ | 41>80 | 21>40 | 1>20 | None |
| Number of Tweets | 24 | 7 | 11 | 68 | 67 |
| From Influential Tweeters  | 6 | 2 | 4 | 58 | 107 |

Comment – A good proportion (25%) of collections that were active with Twitter engaged with influential Tweeters

## Area 5 - Basic usability (Items only)

All the tests that relate to basic usability are also related to other test areas and so are described above. The relevant tests are:

* URL well formed?
* Item Page Title present?
* Page Title well formed?
* Item Page description present?
* Percentage of images in page with ‘Alt’ descriptions present?

# 5 - Related Spotlight Guide Topics and Tips

The Spotlight Guide contains practical tips addressing issues relating to the tests described in this report.

See the full Guide at <http://jisc.ac.uk/guides/make-your-digital-resources-easier-to-discover>.

|  |  |
| --- | --- |
| Guidebook Tips | Tests |
| 1 - Use well structured URLs for collections and items | Manual |  |
| 2 - Use unique and accurate page titles in HTML pages | Manual | Auto |
| 3 - Use the description ‘meta’ tag in HTML pages |  | Auto |
| 4 - Ensure images have appropriate ‘alt’ text |  | Auto |
| 5 - Embed structured metadata for items into HTML pages using [Schema.org](http://schema.org/) |  | Auto |
| 6 - Use sitemaps to help search engines index content effectively |  | Auto |
| 7 - Use robots.txt to ensure search engines can index content effectively |  | Auto |
|  |  |  |
| 8 - Republish items on popular web services | Manual |  |
| 9 - Establish a presence for a collection on Twitter | Manual |  |
|  |  |  |
| 10 - Remove registration or authentication barriers where poss | Manual | Auto |
| 11 - Use appropriate licensing to enable reuse of resources in other contexts (such as CC-BY) | Manual |  |
| 12 - Publish metadata describing digitised resources under an extremely permissive licence (such as CC0) | Manual |  |