Jisc Spotlight on the Digital

Discoverability Tool Specification

Institution Name	National Library of Wales
Key Contacts	Prof Lorna Hughes (lorna.hughes@llgc.org.uk) Dr Owain Roberts (owain.roberts@llgc.org.uk) Paul McCann (paul.mccann@llg.org.uk)
Name of Tool	Discoverability Manager
Tool Summary	Enables institutions to manage their portfolio of digitised resources

Scope

What is the problem?

Since the first digitization projects were undertaken in the 1990s the institutions responsible for their outputs have had varying success in monitoring and enhancing the discoverability of their digitised resources. One reason for this may be that retaining good levels of discoverability requires constant monitoring and identification of the remedial actions required as the discovery environment changes and user discovery behaviours evolve. As the pace of digitization has increased many institutions have now found themselves responsible for a significant number of digitised resources. Consequently, the task of managing and ensuring discoverability of individual (and hence different) resources has become even more demanding. Managing and enhancing discoverability can require significant time and expertise as digitized resources will have been built on different technological platforms and contain different types of content. In some cases, institutions will have a significant 'back catalogue' of digitized resources but may lack the resources to:

1. Monitor and evaluate the discoverability each individual resource
2. Identify remedial action that will improve discoverability

How this tool will help?

This tool is aimed at larger institutions which have to manage a large number of resources (although it could be used by anyone with 1 or more digitized resources to manage). It will make the task of managing these resources easier by allowing easy monitoring of analytics and actions required to improve discoverability. However, actions to improve discoverability will still require the institutions to commit some of their own resources.

Doesn't this kind of thing exist already?

As part of the work of preparing this specification document some investigative work was undertaken to assess other relevant tools. Communication with the Digital Public Library of America (DPLA) revealed that their focus is on discovery tools and less on discoverability. No tools have been proposed for applications such as enhancing metadata, Google webmaster tools and sitemaps although these methods are considered.

Detailed Requirements

Definitions

Resource	A set of digitised content accessible via a specified URL
User	Staff member(s) at the institution responsible for managing the digitised resources
Discovery method(s)	The methods of discovery a consumer would use to discover content: general search engine (Google), library discovery products (such as Summon and Primo), library OPAC, aggregators
Consumer	Individuals that potentially want to consume the content provided in the form of the managed digitised resources
Module	An independent component of the tool that monitors discoverability of resources with respect to a specific method of discovery. Whilst the core part of the tool simply manages a list of resources these modules could be added/removed/developed in response to the way discover behaviours and methods evolve over time.

General Business Requirements

The user must be able to:

- access a list of the resources managed by his/her institution
- add and remove resources from the list of managed resources
- view resources deleted from the manager until they are purged by the user
- rank resources according to different 'methods' of discoverability
- choose which 'methods' of discovery they want to view and monitor for each resource
- obtain suggestions as to what actions should be taken to improve discoverability

The user or any 3rd party must be able to:

- develop new modules that can be easily incorporated into the tool
- modify, improve and customise modules

A Flexible Framework

It is proposed that this tool could be extended easily by the development of modules addressing areas such as:

- Search Engine Optimisation
- Data sharing
- Exposure in general web services e.g. Wikipedia
- Social Media exposure
- Linked data and citation compatibility checker
- MARC record creator for library OPACs – creates MARC records which could be loaded on to library management systems

- The Discoverability Playbook developed as part of the Jisc Spotlight Project
- Any other tools proposed as part of the Jisc Spotlight Project

Module Sustainability

One of the greatest challenges in developing modules will be the rapidly changing operating environment. There is a risk that that some modules may require regular updating e.g. if they are based on Google APIs (since these are known to change at short notice). This is why the tool must be flexible and allow new modules to be added and existing methods to be improved.

Module Requirements

This specification will outline the requirements for 2 modules:

1. Search terms module – for analysing the ranking of resources based on search terms
2. Data sharing module – for managing the process of sharing metadata and content

MODULE 1 - SEARCH TERMS

What will it do?
This module allows you to see if your content is appearing in the search engines, from ways users are already utilizing your site.

This module looks at what search terms people are using on the resource website to find content and drill down to resources that match their needs. These searches should reflect the type of content people are looking for when viewing that particular resource. Unlike a generic search engine, once a user is engaged enough to use the search facility within the site, they are already engaged enough to understand the type of content available and therefore searches within the site highlight the type of terms people are using to try and find the site.

The module will take the search terms used on the site and compare them to Google search results used to actually find the site. This then allows the resource owner to find out if these detailed search terms would have found the resource through a search engine.

As the resource owner makes changes to their site, the tool will be able to track if these changes are having an effect.

Why build this module?
There are numerous tools that deal with search engine optimisation and analytics. However, most of the available tools are expensive and heavily geared towards commercial marketing purposes. This tool will simplify the information provided and ensure that it is relevant to digitised resource managers.

Requirements

The resource owner must:

- have a resource with its own Search facility within the site
- have access to Google Analytics and webmaster tools accounts
- be able to add Google Analytics code to the search box

The module must be able to:

- automatically pull search keywords from the users website
- automatically pull search keywords used to find the site from Google Analytics
- automatically pull search keywords suppressed by Google in Google Analytics from Webmaster tools
- allow users to select Search Terms used ON the website to search terms used to FIND the site, and provide an average position for that search term and details on traffic
- allow the user to export the Search Terms to CSV to be used in other rank checker sites

MODULE 2 - DATA SHARING

What will it do?
This module will enable institutions to keep track of the way resource content and metadata are shared as a way of increasing exposure. This could include remote harvesting of metadata by resource discovery products, export of metadata to union catalogues and aggregators e.g. COPAC and Archives Hub. Very often these aggregators want the data in different formats and this tool will allow the user to keep track of where the data goes, how it is shared and how often.

Why build this module?

Sharing data is a good way of increasing exposure and discoverability of resources. However, it is often hard to keep track of where the data is going and when action is needed by the user to ensure the currency of the data held by aggregators.

Requirements

Users must be able to:

- add new targets for sharing i.e. aggregators/discovery products
- store information regarding the method of sharing (e.g. FTP, OAI-PMH)
- store information regarding timetable for refreshing the data and whether any intervention is required by the user
- notify the user when action is required to refresh the data or change technical configurations
- upload agreements signed with aggregators
- assign one uploaded agreement to multiple resources
- view number of referrals via the selected discovery product (possibly by integration with the search terms module)

Proposed Architecture / Technical Solution

GENERAL

- The tool must be a framework based on a modular and plugin design, allowing the addition of modules or extra functionality.
- The tool **could** utilise an existing framework or open source CMS with a modular/pluggable design to enable faster build and a large support community whilst opening up the tool to be used with other plugins as well as being fitted into tools already possibly in use. Wordpress, Joomla and Drupal would all provide a solid platform and utilise a pluggable structure that would fit with this tool.
- Browser-based interface, which should be built using modern web standards to work across devices.
- The tool and its plugins should be built in an open way preferably PHP or Python to allow use of existing GWT libraries
- The tool should use open standards with the project stored and versioned through a Git repository to allow others to use, add to and fork.
- Front end design and display should be handled through a template system to enable consistency across the modules used/created. Ideally the system should use a Model View Controller (MVC) pattern where possible.
- Plugins will be small data gathering tools that do the backend work such as talking to APIs, database etc. These plugins can then be used by multiple modules on the front end.
- Modules will do all the heavy lifting of handling user input, requests and authorization, issuing queries (via plugins), and transforming the results via the templating system to the end user

SEARCH TERMS MODULE

- The tool will contain a help area of HTML pages providing instructions on how to carry out a number of tasks, for example set up site search.
- The user will need to access to a Google Analytics and Google Webmaster Tools account
- The user must be able to add or update the Google Analytics code for their site.
- The user will need access to their webserver to place a html authorisation page.
- The user will need to configure Google Analytics to work with Site Search.
- The Tool will interact with the Google Analytics API[1] and will retrieve listings of search terms used on the site and stats related to these search terms.
- The tool will utilise the Google Webmaster API[2] and associated php/phython libraries[3] (until the API provides query data) to pull in Google search Query data. These libraries will download all the search queries used to find the website, the search position, number of results, impressions and clicks. This data is sourced from a separate resource because it is no longer available in Google Analytics due to increasing google searches using personalized.
- The module will then compare the search terms used to discover the resource with search terms used within the resource as laid out in the module requirements above.
- The webmaster tools API will also pull in 'Site Alerts' which google has flagged up which can be shown to the user as soon as they access the tool. They will also be able to create and update the sites listed in webmaster tools saving on the user having to visit multiple sites and also submit sitemaps for each site which have been created by other tools/modules.

[1] https://developers.google.com/analytics/devguides/reporting/
[2] https://developers.google.com/webmaster-tools/
[3] http://www.webpronews.com/google-webmaster-tools-api-to-soon-let-you-retrieve-search-queries-and-backlinks-data-2013-06

- The user will be able to add search terms not yet searched for on the site, or via google to see how the site is ranking for these terms anyway.
- The user will be able to discard/delete generic terms from their site search terms list.
- The tool will need to store the results in an open database so that results/positions can be tracked over time to gauge if an improvement has been made.
- The tool will automatically match up on site search terms with those found in off search queries resulting in site traffic.
- The tool will allow export of unmatched search terms, so that the position can be checked in other tools such as the SEOBOOK rank checker.

DATA SHARING MODULE

Since this module is in essence a record management module it will not require any detailed technical requirements beyond the ones detailed in the general section above. If referral data is supplied it will require integration with the Search Terms Module above.

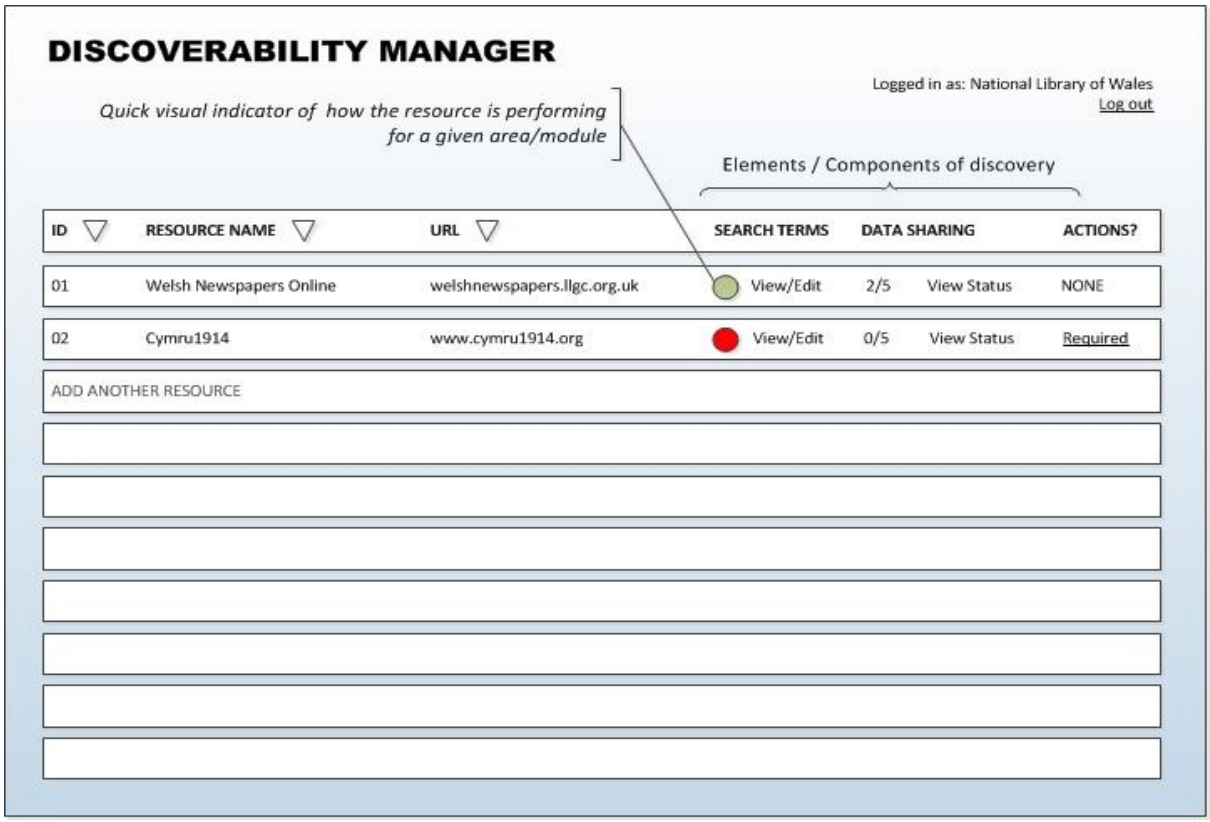
Expected Data Fields

FIELD NAME	DESCRIPTION
Data Targets	
Data Target ID	An unique ID for each data target added by the user
Name of Data Target	A label for the data target
Comments	A free text field
Resources	
Resource ID	An unique ID for each resource added by the user
Name of Resource	A label for the resource name
URL of Resource	The URL by which consumers can access the resource
Comments	A free text field
Sharing Agreement	**a one to many relationship exist between 'Data Targets and 'Resources'**
Agreement ID	An unique ID for each agreement added by the user
Data Target	The data target with who the agreement has been made (or is in the process of being made). Users should be able to choose from existing data targets set up in the main dashboard.
Resource(s)	The resources included in the agreement (users should be able to choose from existing resources set up in the main dashboard)
Agreement Signed	Yes/No (If Yes –option to upload a PDF of the agreement)
Comments	A free text field
Sharing Agreements	**will link to sharing instances (one to one relationship between 'Resources' and 'Data Targets')**
Method of sharing	e.g. FTP, OAI-PMH
Automatic updates?	Yes/No
Frequency of data/content updates	(Never/manual, daily, weekly, monthly, quarterly)
Comments	A free text field

User Interfaces

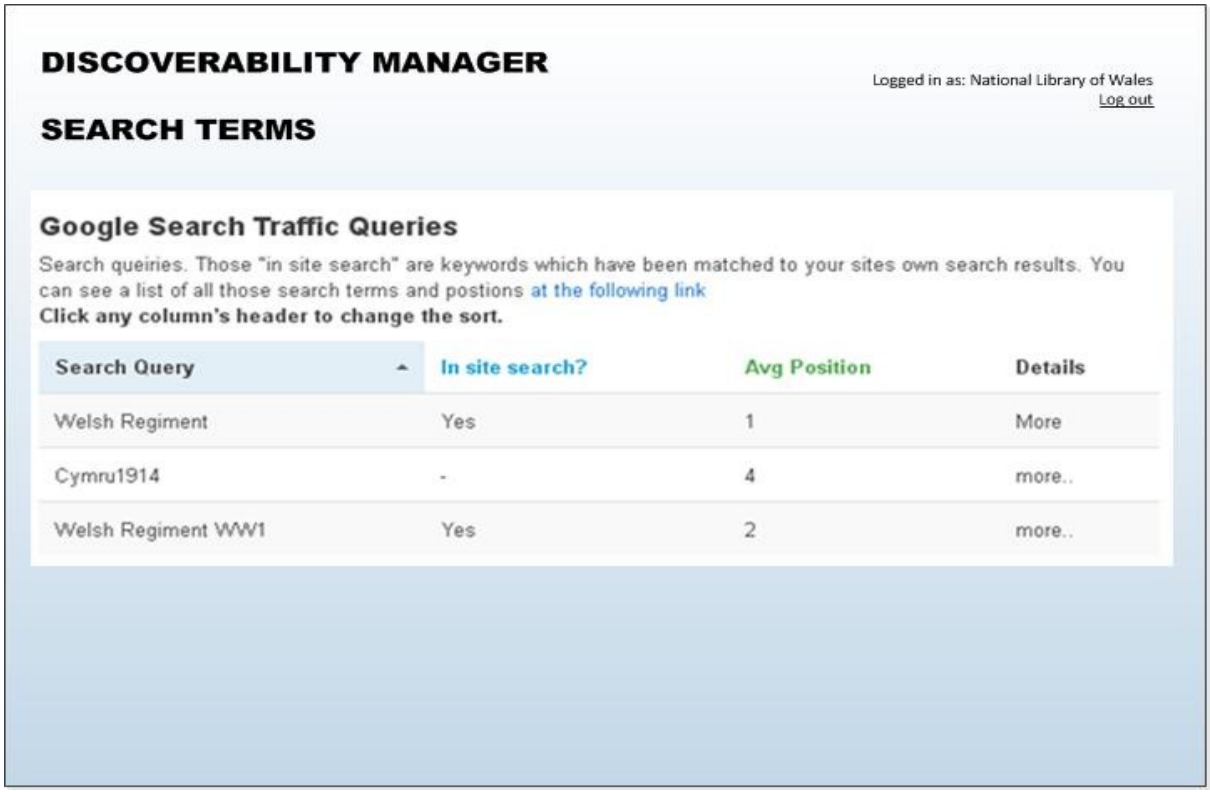
Components
It is envisaged that this will be a web-based tool comprising of the following pages:

Main Dashboard
This is where institutions can produce a canonical list of digitized resources they want to manage via the tool and monitor how discoverable their resources are for the methods addressed by the installed modules.

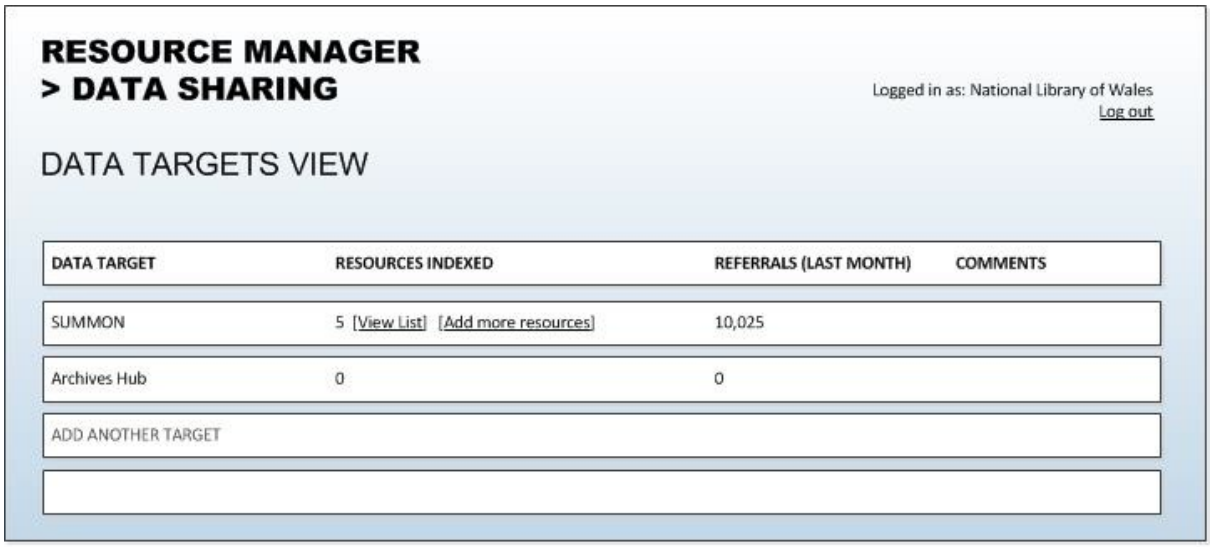


Module Pages
The content of these will be dependent on the functionality of the module selected.

MODULE 1 - SEARCH TERMS



MODULE 2 – DATA SHARING

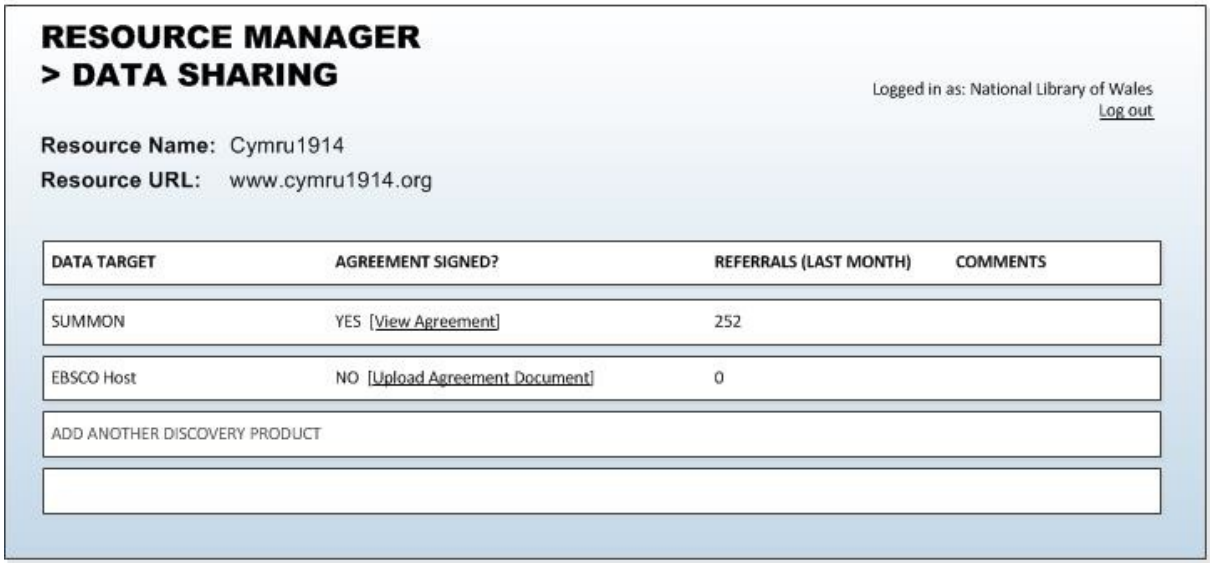


RESOURCE MANAGER
> DATA SHARING

Logged in as: National Library of Wales
Log out

DATA TARGETS VIEW

DATA TARGET	RESOURCES INDEXED	REFERRALS (LAST MONTH)	COMMENTS
SUMMON	5 [View List] [Add more resources]	10,025	
Archives Hub	0	0	
ADD ANOTHER TARGET			



RESOURCE MANAGER
> DATA SHARING

Logged in as: National Library of Wales
Log out

Resource Name: Cymru1914
Resource URL: www.cymru1914.org

DATA TARGET	AGREEMENT SIGNED?	REFERRALS (LAST MONTH)	COMMENTS
SUMMON	YES [View Agreement]	252	
EBSCO Host	NO [Upload Agreement Document]	0	
ADD ANOTHER DISCOVERY PRODUCT			

END OF DOCUMENT