discovery lessons

Senior Managers	
Metadata Creators	
Developers	✓
Suppliers	

Technologies

The Discovery programme has been funded by JISC to improve access to collections that support research and education. This document is part of a series that describes the lessons from 8 JISC projects funded under the Discovery programme in 2011 to explore open metadata for libraries, museums and archives. More information about the projects can be found at: http://www.jisc.ac.uk/whatwedo/programmes/inf11/infrastructureforresourcediscovery.aspx
The other documents in the series can be found at: http://discovery.ac.uk

Scope

There are a large, and ever increasing, number of technologies relevant to the manipulation and publication of data. This guide summarises some of the key technologies used by the eight Discovery projects, and other related work.

The Problem Space

This guide does not tackle a specific problem space, but rather attempts to present a snapshot of technologies and tools of relevance to the manipulation, publication and aggregation of metadata.

Outcomes

Data storage and related technologies

	Example Use	Related Tools
MongoDB A schema-less 'document-oriented' database, which stores JSON-style documents. MongoDB is designed to scale easily and offer high availability. http://www.mongodb.org	The Jerome project[1]	ElasticSearch, CouchDB[2]
ElasticSearch A schema-less data store, index and search engine, which uses the JSON data format. ElasticSearch is built on Lucene (see below), and designed to offer a fast, easy to configure and highly scalable search engine. http://www.elasticsearch.org	BibServer, part of the Bibliographic Knowledge Network toolset[3]	MongoDB, CouchDB, Lucene

1 http://jerome.blogs.lincoln.ac.uk
2 http://couchdb.apache.org
3 http://bibserver.okfn.org

	Example Use	Related Tools
Redis A key-value store, supporting keys containing strings, hashes, lists, sets and sorted sets. Redis keeps datasets in memory, with periodic writes to disk, making it extremely fast. http://redis.io	JISC Open Bibliography project for data analysis[4], and to enhance SPARQL results with geolocation data in a data visualisation[5].	
ARC2 + MySQL A PHP library for RDF. It supports the use of MySQL as an RDF triplestore. **ARC2:** https://github.com/semsol/arc2/wiki **MySQL:** http://www.mysql.com	COMET[6] used ARC2 + MySQL to host their linked data expression of bibliographic records from the University of Cambridge. While it proved easy to setup, speed and scalability were both issues.	OWLIM[7], 4Store[8], Virtuoso[9], the Talis Platform, Sesame[10]
Talis Platform A Software as a Service (SaaS) RDF triplestore http://www.talis.com/platform	SALDA[11], LOCAH[12]	ARC2, OWLIM, 4Store, Virtuoso, Sesame

Flexible Infrastructure

	Example Use	Related Tools
Amazon Web Services A set of services offering flexible, scalable, infrastructure. Specifically of note in this context, the 'Elastic Cloud Compute' (EC2) service offering deployment of virtual servers, and 'Simple Storage Service' (S3) offering disk space. http://aws.amazon.com	OpenART used AWS EC2 for easy deployment of virtual machines	Rackspace Cloud Hosting[13], Heroku[14], Google App Engine[15]

4 http://openbiblio.net/2010/11/18/characterising-the-british-library-bibliographic-dataset
5 http://benosteen.com/globe
6 http://cul-comet.blogspot.com
7 http://www.ontotext.com/owlim
8 http://4store.org
9 http://www.openlinksw.com/dataspace/dav/wiki/Main/VOSIndex
10 http://www.openrdf.org
11 http://blogs.sussex.ac.uk/salda
12 http://blogs.ukoln.ac.uk/locah
13 http://www.rackspace.co.uk/cloud-hosting
14 http://www.heroku.com
15 http://code.google.com/appengine

Information Extraction

	Example Use	Related Tools
OpenCalais An information extraction web service, including named entity recognition[16]. The Calais webservice accepts unstructured data as an input and returns entities, facts and events contained in the the data as RDF. http://www.opencalais.com	Open Metadata Pathfinder (OMP)[17]	DBpedia Spotlight, GATE[18]
DBPedia Spotlight An named entity recognition service based on data in DBPedia (linked data extracted from Wikipedia). http://dbpedia.org/spotlight	DiscoBro[19] (an entry to the Discovery developer competition[20])	OpenCalais, GATE

Indexing and Search tools

	Example Use	Related Tools
Sphinx An open source search server. It aims to offer very high speeds of both indexing and retrieval, and to be scalable up to indexing billions of documents. Sphinx can index from both SQL and NoSQL data stores as well as index files directly. http://sphinxsearch.com	The Jerome Project	Solr/Lucene, Zebra[21]
Solr/Lucene Open source indexing and search software. http://lucene.apache.org	As ElasticSearch is built on Lucene, by extension this is used by BibServer Solr/Lucene is used in a wide range of 'resource discovery' software and services including VuFind[22], Blacklight[23] and CultureGrid[24].	SolrMARC[25], VuFind, Blacklight, ElasticSearch, Sphinx, Zebra

16 http://en.wikipedia.org/wiki/Named_entity_recognition
17 http://openmetadatapathway.blogspot.com
18 http://gate.ac.uk
19 http://people.kmi.open.ac.uk/mathieu/about/discobro-discovering-linked-data-resources-while-browsing
20 http://discovery.ac.uk/developers/competition
21 http://www.indexdata.com/zebra
22 http://vufind.org
23 http://projectblacklight.org
24 http://www.culturegrid.org.uk
25 http://code.google.com/p/solrmarc

Data manipulation and transformation tools

	Example Use	Related Tools
MARC code libraries There are code libraries specifically for the reading, writing and manipulation of MARC records for a wide variety of programming languages. More information: http://wiki.code4lib.org/index.php/Working_with_MaRC	Comet used MARC/Perl[26] to create the 'Marc[27] to RDF triples conversion utility' and the 'Marc record sorter' tools . The Lucero project[28] used MARC4J[29] (a Java library) to convert MARC records to RDF	pymarc[30] (Python library for MARC), ruby-marc[31] (Ruby library for MARC), File_MARC[32] (PHP library for MARC)
XSLT XSLT is a language or mechanism to transform XML documents in various ways, including into HTML, plain text, RDF and new XML documents. An XSLT provides the rules for the transformation, which can then be applied to XML using an XSLT processor such as Saxon[33], Xalan[34] and xsltproc[35]. http://www.w3schools.com/xsl	Discovering Babel used XSLT[36] to transform information from the TEI XML format into Dublin Core LOCAH and SALDA both used XSLT to transform information from the EAD XML format commonly used for Archival description, to RDF	Xpath[37]
Google Refine Google Refine is a tool for "working with messy data". It can help clean up, transform and extend data. It also supports 'reconciling' (i.e. matching) entities from one dataset to other, remote, datasets. http://code.google.com/p/google-refine	OpenART[38] used Google Refine to manipulate information, and to connect parts of the data using the reconciliation feature.	Stanford Data Wrangler[39]

26 http://marcpm.sf.net/
27 http://data.lib.cam.ac.uk/code.php
28 http://lucero-project.info/lb
29 http://marc4j.tigris.org
30 http://pypi.python.org/pypi/pymarc
31 https://github.com/ruby-marc/ruby-marc
32 http://pear.php.net/package/File_MARC
33 http://saxon.sourceforge.net
34 http://xalan.apache.org
35 http://xmlsoft.org/XSLT/xsltproc2.html
36 http://blogs.oucs.ox.ac.uk/martinw/2011/08/03/discovering-babel-technical-issues
37 http://www.w3schools.com/xpath
38 http://yorkdl.wordpress.com/2011/11/02/openart-final-report
39 http://vis.stanford.edu/wrangler

Big data

	Example Use	Related Tools
Mahout A machine learning and data mining library designed to work with very large datasets. http://mahout.apache.org	The AEIOU Project used of Mahout for 'recommender systems'[40] The Library Innovation Lab at Harvard41 has been experimenting with analysing bibliographic records using Mahout and Gephi[42]	Hadoop[43], Gephi[44]

Recommendations

- Schema-less approaches to data storage are a good fit to some of the data issues in the library, archive and museum sectors
- Information Extraction software and services offer routes to link traditional metadata into other data sources on the web
- Issues of storing, manipulating and analysing data are common across many domains and sectors, so existing tools and technologies may well be useful to library, archive and museum data
- 'Big data' and 'Visualisation' technologies will become more relevant to libraries, archives and museums as they wish to build and exploit large aggregations of activity data and metadata

Key Discovery Projects

- Comet - http://cul-comet.blogspot.com
- Discovering Babel - http://blogs.oucs.ox.ac.uk/martinw/2011/03/23/discovering-babel
- Jerome - http://jerome.blogs.lincoln.ac.uk
- OMP - http://openmetadatapathway.blogspot.com
- SALDA - http://blogs.sussex.ac.uk/salda

Other Related Work

- Linked and Open Copac and Archives Hub (LOCAH) - http://blogs.ukoln.ac.uk/locah
  A project which addresses how Linked Data can support cross-domain discovery of L&A material and beyond
- Linking University Content for Education and Research Online (LUCERO) - http://lucero-project.info/lb
  A project to publish Linked Data at the Open University (including bibliographic data) and to provide guidance to those wishing to do the same
- AEIOU - http://www.wrn.aber.ac.uk/aeiou
  A project to increase the visibility and usage of Welsh academic research
- JISC Open Bibliography - http://openbiblio.net/tag/jiscopenbib
  Project to publish bibliographic metadata as Linked Open Data
- Bibliographic Knowledge Network tools - http://bibserver.okfn.org

40 http://www.wrn.aber.ac.uk/aeiou
41 http://librarylab.law.harvard.edu
42 http://danbri.org/words/2011/10/11/720
43 http://hadoop.apache.org
44 http://gephi.org