# Mashups and open data in libraries

**The ever increasing amount of digital information available on the web offers the opportunity to create new services and applications by combining, or 'mashing up', information from multiple sources. This paper describes the origins of mashups and how the concept spread from music and video to online data. The paper goes on to describe how data formats, licensing and retrieval mechanisms affect the reusability of data published online. It also examines how libraries act as both data publishers and mashup creators. Finally, the paper highlights the latest developments in the publication of open data by the library community, and resources available to those wishing to either publish or consume data online.**

***OWEN STEPHENS***
Owen Stephens Consulting

### What is a mashup?

The use of the term mashup to describe a 'fusion of disparate elements' dates back to at least the 19th century, but it was not until the late 20th century that it became common, initially with specific reference to music[1].

In music at least, the reuse of others' material in new contexts or new ways has a long history. Some works such as Rachmaninov's *Rhapsody on a theme of Paganini* (1934) explicitly took an existing theme and set it into a new context, using different instrumentation and playing with the basic structure of the piece to present something new.

However, it is with the advent of recording technology, and perhaps particularly the recording and publication of music in digital formats, that sampling particular recordings and reusing in new compositions became both easier and commonplace. Of particular note is the *Grey Album* by DJ Dangermouse, which is made up entirely of samples from the Beatles' *White Album* and JayZ's *Black Album*. The *Grey Album* attracted controversy and publicity as it sampled the source material without permission, and EMI subsequently requested those hosting the album to remove it due to the inclusion of material by the Beatles[2].

As tools and content in appropriate formats became more common, the number of mashups being produced increased substantially, both in their musical form and also across other media. Video mashups particularly became popular and easy to distribute or access via YouTube[3].

In my view, the factors that have encouraged mashups of cultural material are the availability of appropriate tools and source material in easy-to-manipulate formats. These factors apply equally to other forms of information and, in recent years, 'data mashups' combining data from online resources to produce a new view of existing information have become increasingly common. In particular, mashups that combined data with some geographic aspect with online maps started to expand rapidly from 2005 when Google Maps was first made available[4]. An early example of this is http://www.housingmaps.com/ which combined information from Craig's List (a classified advertising site) with Google Maps to offer a map-based interface to find properties to buy or rent.

While many mashups take advantage of a lightweight approach to deliver useful services quickly and easily, there are more substantial examples which use the same techniques to deliver high-profile services. Perhaps most notable are the BBC Music[5] and Wildlife Finder[6] web pages, which bring together content from the BBC, such as video clips or radio and television schedules, with data from a wide variety of data sources on the web, including Wikipedia, MusicBrainz (an online, crowd-sourced, 'music encyclopedia'[7]), WWF's Wildfinder[8], the IUCN's Red List of Threatened Species[9], the Zoological Society of London's EDGE of Existence programme[10], and the Animal Diversity Web[11]. One of the advantages the BBC team see in using sources such as Wikipedia and MusicBrainz is that BBC staff can contribute directly to them, updating and correcting information, treating the web as their content management system[12].

## Open and usable data

As noted above, the availability of data in an appropriate format is key to building mashups and other forms of reuse.

The term 'open data' is often used to mean specifically data licensed in such a way that can be reused by all, 'subject to at most the requirement to attribute and share-alike'[13]. However, some argue that we need a 'richer understanding of openness' which encompasses not just permissive licensing but, more broadly, the ease with which data can be used, taking into consideration aspects such as format and access mechanisms[14].

In this article 'open' will be used in the licensing sense above, but this should be seen as only one of three key factors that influence the ability to reuse data:

- the mechanisms supported for retrieving the data
- the format of the data
- the licensing of the data for reuse.

### *Mechanisms*

Typically, data used in mashups is accessed via an application programming interface (API). An API is essentially an interface to the data that can be used by a computer programme to retrieve or interact with the available data. Many popular online services offer such interfaces, for example Twitter, Flickr and many Google services, such as Google Maps and Google Book Search, offer APIs which are used by a wide range of services.

APIs often offer an easy way to interact with data from other sources in real time, and may include ways of not just retrieving data but also updating information. However, for data that does not change much, or at all, over time, and where it is appropriate to provide read-only data, a service may instead offer the ability to download data in a standard form, such as a spreadsheet or text file.

### *Formats*

There is a huge variety of types of data available on the web, and so there are large numbers of formats. I XML and JSON are two formats widely adopted. A key benefit of using these formats is the large number of programming tools available that can read data from them. While not as popular, a third format worth mentioning is RDF[15], which is widely used by the Linked Data and Semantic Web communities and has been used for an increasing variety of library data over the last few years[16].

When offering data for direct download texts formats called comma separated values (CSV) and tab separated values (TSV) are popular, as is the Excel spreadsheet file format.

Specialist formats may be used for specific types of data, the use of MARC for bibliographic data being one example of many.

Another approach to publishing data on the web has been to integrate structured data directly into web pages. This approach has the advantage of providing data via a single mechanism (the web page) while making it easily readable by both humans and software. While there are a number of different formats available to achieve this integration[17,18], the recent schema.org[19] initiative announced by Bing, Google and Yahoo! will significantly influence developments in this area.

Discussion of formats can be confusing because of the way format, structure and standards interact. A format often referred to when talking about mashups is RSS, which is a mechanism for delivering data in a standard XML format. A further example of how data formats and standards interact is MARC which is, in itself, a file format, but also a data standard. It is now relatively common to have MARC data expressed in XML[20], and recently there has been a proposal to also be able to express MARC in JSON[21].

Understanding and use of different formats varies considerably across different communities and perhaps the key question for those publishing data for reuse is who the target audience for the data is, and what are the most appropriate and useful formats for them. It is perhaps particularly relevant to libraries to note that the use of specialist data formats can significantly decrease the potential reuse of the data

### *Licensing*

As in the case of the *Grey Album* mentioned above, where mashups use material without permission, they can be the subject of legal action. Unfortunately it can often be unclear whether data can be reused freely or not and, if there are restrictions, what these are.

In order to remove uncertainty, data publishers may offer data under certain terms and conditions, or licences, which make explicit what can, or can't, be done with the data.

Some services may do this via a set of terms and conditions that a consumer must sign up to before they use the service. Service APIs (for example, Twitter[22]) often have separate terms and conditions to the main service, and these may cover many different aspects of use, not solely how data may be used.

In other cases, content or data is provided under a 'licence' which states any restrictions on using the content without requiring any action on the part of the consumer. Creative Commons[23] is probably the most widely recognized set of licences, with different licences covering different types of use. For example, the CC-BY licence[24] is designed to allow others to share and adapt the licensed content, for any purpose, while requiring that the originator of the content is attributed clearly.

Other licensing schemes, while not as widely adopted as Creative Commons, are available, perhaps notably the Open Data Commons licences[25] aimed specifically at data and databases rather than more traditional content.

Recently, there has been a significant push towards releasing publicly-funded data as open data. Notably, several governments, including the UK, have established 'data portals' such as data.gov.uk[26], which provide easy access to open government data. In the UK developments in this area have been championed by successive governments[27,28], with the release of mapping data by Ordnance Survey attracting particular attention[29]. As well as central government releasing data, local government and related bodies are also publishing increasing amounts of open, reusable data. The 'Live train map for the London Underground'[30] is a great example of a mashup built with data published by Transport for London and Google Maps.

## Mashups, open data and libraries

### *Libraries as data publishers*

Libraries have a long tradition of making data available for use and reuse by their members. Many online catalogues offer the ability to download records in a variety of formats and often offer programmatic interfaces to their data through mechanisms such as Z39.50 and SRU/SRW.

There is also evidence that library and bibliographic data is of interest to developers who create mashups and otherwise reuse data. In the last two years, the 'Young Rewired State' event in which developers aged 18 or under 'build apps using government data, and present them to press and government'[31] has included three projects directly related to libraries and book data, with the 'SocialLibrary'[32] project winning the 'Best App' award in 2010. When Warwickshire County Council ran a competition to build on a wide variety of open data they had published, one of the winning entries used an RSS feed of recent library acquisitions[33]. The recent Libraryhack competition[34] in Australia and New Zealand attracted 120 entries. All of these suggest there is significant interest in using the data that libraries have to create new and interesting applications, services and cultural objects.

However, libraries could do more in terms of offering both open and easily reusable data. The terms of use for data obtained from library catalogues are often unclear, and the formats offered are usually familiar only to those who specialize in library data.

A wide range of initiatives and projects globally have started to address these issues. In the UK, the JISC-funded Discovery initiative[35] has established Open Metadata Principles[36] which advocate the licensing of metadata using the Creative Commons Zero[37], or similar, licences which effectively put the data into the public domain. These principles have been signed by representatives from the British Library, the National Archives, the National Libraries of Scotland and Wales and a wide range of other UK institutions. Alongside this, JISC has funded the writing of a *Guide to Open Bibliographic Data*[38] (which I co-authored) and several projects[39] which have resulted in the release of a large amount of library, archive and museum metadata in reusable formats with permissive licensing.

At the same time the Open Bibliography Principles[40], released by the Open Bibliographic Data working group of the Open Knowledge Foundation (which also advocates the release of bibliographic data into the public domain), has attracted endorsements from the international library community and beyond.

While statements advocating the release of data under open licences may attract broad support, not all believe that bibliographic data can simply be shared freely. JISC Legal have produced guidance on the legal issues surrounding the use of bibliographic records[41]. The JISC Legal guide explores the variety of factors that might constrain the use of bibliographic data, such as copyright law and contractual agreements, and suggests a number of approaches to managing risks relating to using records and allowing others to do so[42]. Alongside its Open Metadata Principles, the Discovery initiative also provides guidance on licensing open data that recognizes fully 'open' data may not always be either possible or desirable[43].

The recent JISC-funded COMET project at the University of Cambridge took this work one step further and produced tools and workflows intended to help libraries analyze their data and so release it under appropriate licences[44]. The COMET project also entered into negotations regarding appropriate licences for the release of bibliographic data with a wide range of record suppliers[45].

Despite these issues, a large number of library organizations are now putting the principles of openess and reusability into practice: the British Library and the University of Cambridge in the UK, the hbz Union Catalogue in Germany; the Bibliothèque Nationale in France; the Library of Congress in the USA; the National Libraries of Australia and New Zealand; and many, many others.

### *Libraries as mashup creators*

As well as publishing data that can be used by others, libraries have been consuming data in mashup-type ways for some time. Syndetic Solutions (now owned by Bowker)[46] has been offering data enhancement to OPACs since before 2004, with book covers, reviews and other information being retrieved in real time and displayed alongside the relevant catalogue records.

Other services being used to deliver enhanced information about books available via library catalogues include Amazon, Google Book Search and LibraryThing for Libraries, while for journals, the JournalTOCs[47] API is being used to power a variety of mashups[48], although unfortunately terms for the reuse of the tables of contents (ToCs) are often unclear.

Many of these mashups bridge the gap between the library catalogue record and the actual content of books or journals. Perhaps one of the more striking examples of this is in the Belgium public library catalogue where data from the last.fm music service[49] is mashed up with catalogue data to provide a link between music items in the catalogue and recorded music online, as well as providing links to similar artists.

## Getting involved

Whether you are interested in contributing to the growing pool of reusable data, or exploiting it to build new applications or functions, there are many ways of getting involved. There are guides to open data[50], to licensing[51] and explanations of how to integrate descriptions of books into web pages so search engines can make use of it[52].

For building mashups there are books[53], presentations[54], tools[55], events[56,57] and beginners' guides[58,59].

With new types of data, such as user activity data from both circulating stock[60] and electronic resources[61] becoming available all the time, the potential for creating mashups in libraries has never been greater. What is needed is for creative thinking, professional knowledge and technical skills to be brought together to understand the potential of the data available and the services which could be offered.

## References

1. *Oxford English Dictionary*: http://www.oed.com/view/Entry/266403 (accessed 05 September 2011).
2. Chilling Effects: http://www.chillingeffects.org/fairuse/notice.cgi?NoticeID=1093 (accessed 05 September 2011).
3. Empire: http://www.empireonline.com/features/50-best-youtube-movie-mashups/ (accessed 05 September 2011).
4. Mashup Guide: http://mashupguide.net/1.0/html/ch01s02.xhtml (accessed 05 September 2011).
5. BBC Music: http://www.bbc.co.uk/music (accessed 05 September 2011).
6. BBC Wildlife Finder: http://www.bbc.co.uk/wildlifefinder (accessed 05 September 2011).
7. MusicBrainz: http://musicbrainz.org/ (accessed 05 September 2011).
8. WWF Wildlife Finder: http://gis.wwfus.org/wildfinder/ (accessed 05 September 2011).

9. IUCN Red List of Threatened Species: http://www.iucnredlist.org/ (accessed 05 September 2011).
10. EDGE of Existence: http://www.edgeofexistence.org/ (accessed 05 September 2011).
11. Animal Diversity Web: http://animaldiversity.ummz.umich.edu/site/index.html (accessed 05 September 2011).
12. Case Study: Use of Semantic Web Technologies on the BBC Web Sites: http://www.w3.org/2001/sw/sweo/public/UseCases/BBC/ (accessed 05 September 2011).
13. Open Definition: http://www.opendefinition.org/ (accessed 05 September 2011).
14. Metadata Aggregation Services: http://www.slideshare.net/paulwalk/metadata-aggregation-services (accessed 05 September 2011).
15. Quick Intro to RDF: http://www.rdfabout.com/quickintro.xpd (accessed 05 September 2011).
16. Library Linked Data: http://ckan.net/group/lld (accessed 05 September 2011).
17. RDFa for HTML Authors: http://www.w3.org/MarkUp/2009/rdfa-for-html-authors (accessed 05 September 2011).
18. Microformats: http://microformats.org/ (accessed 05 September 2011).
19. schema.org: http://schema.org/ (accessed 05 September 2011).
20. MARCXML: http://www.loc.gov/standards/marcxml/ (accessed 05 September 2011).
21. Diettante's Ball: http://dilettantes.code4lib.org/blog/2010/09/a-proposal-to-serialize-marc-in-json/ (accessed 05 September 2011).
22. Twitter developers: https://dev.twitter.com/terms/api-terms (accessed 05 September 2011).
23. Creative Commons: http://creativecommons.org/ (accessed 05 September 2011).
24. Creative Commons CC-BY: http://creativecommons.org/licenses/by/3.0/ (accessed 05 September 2011).
25. Open Data Commons Licences: http://opendatacommons.org/licenses/ (accessed 05 September 2011).
26. data.gov.uk: http://data.gov.uk (accessed 05 September 2011).
27. Number10: http://www.number10.gov.uk/news/pm-sets-ambitious-open-data-agenda/ (accessed 05 September 2011).
28. *The Guardian*: http://www.guardian.co.uk/technology/2009/nov/17/ordnance-survey-maps-online (accessed 05 September 2011).
29. *The Guardian Datablog*: http://www.guardian.co.uk/news/datablog/2010/apr/02/ordnance-survey-open-data (accessed 05 September 2011).
30. Live train map for the London Underground: http://traintimes.org.uk/map/tube/ (accessed 05 September 2011).
31. Young Rewired State: http://youngrewiredstate.org/ (accessed 05 September 2011).
32. SocialLibrary: http://rewiredstate.org/projects/sociallibrary (accessed 05 September 2011).
33. Hack Warwickshire: http://warwickshireopendata.wordpress.com/2010/07/09/never-mind-the-world-cup-here-are-the-hack-warwickshire-results/ (accessed 05 September 2011).
34. Libraryhack Entries: http://libraryhack.org/mix-mash-win/hack-entries/ (accessed 05 September 2011).
35. Discovery: http://www.discovery.ac.uk/ (accessed 05 September 2011).
36. Open Metadata Principles: http://www.discovery.ac.uk/businesscase/principles/ (accessed 05 September 2011).
37. CC0: http://creativecommons.org/choose/zero/ (accessed 05 September 2011).
38. Open Bibliographic Data Guide: http://obd.jisc.ac.uk/ (accessed 05 September 2011).
39. Infrastructure for Resource Discovery: http://www.jisc.ac.uk/whatwedo/programmes/inf11/infrastructureforresourcediscovery.aspx (accessed 05 September 2011).

40. Open Bibliography Principles: http://openbiblio.net/principles/ (accessed 05 September 2011).
41. Transfer and Use of Bibliographic Records: Guidance on Legal Issues: http://www.jisclegal.ac.uk/Projects/TransferandUseofBibliographicRecords.aspx (accessed 05 September 2011).
42. Transfer and Use of Bibliographic Records: Managing risk: http://www.jisclegal.ac.uk/Projects/TransferandUseofBibliographicRecords/Managingrisk.aspx (accessed 05 September 2011).
43. Licensing Open Data: A Practical Guide: http://discovery.ac.uk/files/pdf/Licensing_Open_Data_A_Practical_Guide.pdf (accessed 05 September 2011).
44. CUL-COMET Blog: http://cul-comet.blogspot.com/2011/07/where-exactly-does-record-come-from.html (accessed 05 September 2011).
45. CUL-COMET: Ownership of MARC21 Records: http://cul-comet.blogspot.com/p/ownership-of-marc-21-records.html (accessed 05 September 2011).
46. Syndetic Solutions: http://www.bowker.com/syndetics/ (accessed 05 September 2011).
47. JournalTOCs: http://www.journaltocs.ac.uk/ (accessed 05 September 2011).
48. Mashup of RSS feeds in a high-demanding research environment: http://www.journaltocs.hw.ac.uk/docs/conf/2010/emtacl10_santy.html (accessed 05 September 2011).
49. Last.fm: http://last.fm (accessed 05 September 2011).
50. Open Bibliographic Data Guide, ref. 38.
51. Licensing Open Data: A Practical Guide: http://discovery.ac.uk/files/pdf/Licensing_Open_Data_A_Practical_Guide.pdf (accessed 05 September 2011).
52. Spoonfeeding Library Data to Search Engines: http://go-to-hellman.blogspot.com/2011/07/spoonfeeding-library-data-to-search.html (accessed 05 September 2011).
53. Engard, N, *Library Mashups*, 2009, London, Facet.
54. Just Do IT Yourself: http://blog.ouseful.info/2011/04/04/just-do-it-yourself-my-uksg-presentation/ (accessed 05 September 2011).
55. Yahoo Pipes: http://pipes.yahoo.com/pipes/ (accessed 05 September 2011).
56. OCLC Developer News: http://www.oclc.org/developer/news (accessed 05 September 2011).
57. Mashed Library: http://www.mashedlibrary.com/ (accessed 05 September 2011).
58. I want to make a mashup: http://www.slideshare.net/garygre/i-want-to-make-a-mashup-but-i-dont-know-where-to-start (accessed 05 September 2011).
59. Putting Warwickshire Libraries on the Map: http://www.meanboyfriend.com/overdue_ideas/2010/04/putting-warwickshire-libraries-on-the-map/ (accessed 05 September 2011).
60. JISC MOSAIC Activity Data: http://ckan.net/package/jisc_mosaic (accessed 05 September 2011).
61. OpenURL Router Data: http://openurl.ac.uk/doc/data/data.html (accessed 05 September 2011).

■ **Owen Stephens**
**Owen Stephens Consulting**
**Tel: +44 (0121) 288 6936**
**E-mail: owen@ostephens.com**

The DOI for this article is 10.1629/24245. Click here to access original article via DOI:

**http://dx.doi.org/10.1629/24245**

For a link to the full table of contents for the issue of *Serials* in which this article first appeared, click here:

**http://serials.uksg.org/openurl.asp?genre=issue&issn=0953-0460&volume=24&issue=3**