Ender user development of digital collection mash-ups: A survey to assess the suitability of current infrastructure

Holley Long
Digital Initiatives Librarian
University of Colorado Boulder Libraries
Holley.Long@colorado.edu

Mash-ups, web applications that combine data from multiple sources to create a new service, have become common fixtures on libraries' web sites. Many libraries feature mash-ups that display branch information on a Google map or bring in Amazon book cover images into the catalog. As consumers of mashable data, these libraries are capitalizing on mash-ups to visualize information and create new resources. However, the way they could exploit this trend to greatest advantage is as providers of mashable data, by furnishing programmatic access to their own collections of digitized content and metadata. In "What is a mashup?" (Fichter, 2009), Darlene Fichter describes a few examples of mash-ups created by patrons using content furnished by libraries, including a new books virtual bookshelf employing data from an online catalog RSS feed and book cover images from Syndetic Solutions, and a Firefox extension called Bookburro that recognizes bibliographic data within a web page and offers related information such as local library holdings. These early efforts only begin to reveal the potential for user-developed digital collection mash-ups and highlight the need for services to support such activities.

This article will explore why it is advantageous for libraries to provide mashable digital collections for end-user development and discuss the three essential components of this service – digital content and metadata, programmatic access to the collections, as well as appropriate access controls and policies – which libraries either already offer or are well-positioned to do so. Providing this service will likely require that libraries retool technical infrastructure and policies to support direct access to the digital content and metadata. How much of an investment is needed? To answer this question, the author surveyed Association of Research Libraries (ARL, n.d.) members' digital libraries and other discovery platforms to determine if current infrastructure and terms of use policies can support user-developed mash-ups, and if not, what steps could be taken to fulfill this goal.

Benefits of a mashable digital collection service

Providing users with programmatic access to digital collections is a departure from libraries' conventional approach to delivering content. As Rich Gazan notes in his article on social annotations in digital libraries (Gazan, 2008):

> Though designed as systems for knowledge discovery, the majority of digital libraries operate from the traditional expert model. Subject experts create content, digital library experts provide access to it, and individual users consume it. Very few systems have been built with an architecture that encourages users to create content, associate it with collection items, or share their impressions with other users. Providing digital library users read-access to collections is the traditional finish line.

In other words, digital library infrastructure has been designed based on the premise that our users will be passive consumers of digital content. Shifting to a new model in which they can be co-creators will require an investment to change the traditional infrastructure. Nonetheless, the benefits of providing mashable digital collections far outweigh the costs since the provision of this service would not only advance libraries' goals to expand access to resources, but also increase the utility of their collections and foster user participation.

Traditionally, libraries have served digital collections from siloed systems isolated from the online environments that users frequent most often, and as a result, these collections are sometimes underutilized (Kalfatovic *et al.*, 2008). Providing APIs to the digital library software mitigates this problem by freeing the collections' objects and metadata from the platform's architecture, allowing developers to create mash-ups that serve as new access points for the collection. For example, larger subject-based collections can be created through mash-ups without the organizational and administrative overhead that more conventional efforts at collaborative collection-building require.

Encouraging users to mash-up library content is an excellent way to foster a sense of ownership. Users can tailor collections to specific information needs or create value-added services that have not traditionally been part of digital library platforms. Furthermore, inviting end-users to mash-up digital collections can expose the many layers of research value an information resource offers in ways that the collection managers may not immediately recognize. For example, a digital collection of wild flower drawings, cataloged in VRA Core 4, could be re-purposed in a botany mash-up to support scientific inquiry. When users create mash-ups in this manner, the collections can be understood and represented from a multitude of perspectives, which only increases their utility.

Three components of a mashable digital collection service

Establishing a service to facilitate end-user development of mash-ups affords several compelling benefits, but the three basic components of this service as currently offered – digital collections, programmatic access, and access management – are not all equally suited for this purpose. Libraries are obviously well equipped to supply the first requirement, digital collections and metadata. Libraries excel at selecting information-rich collections for digitization and creating high quality metadata according to nationally-recognized schema. As a result, their digital collections provide a richer base from which to develop mash-ups and the structured, standardized metadata more easily lends itself to use in mash-ups. Moreover, whenever possible, libraries have made a commitment to provide free long-term access to their resources, thus making their digital collections a more durable source of content for mash-ups. While many for-profit organizations that provide mashable data will give developers a free key to create mash-ups, there is no guarantee that access to the data will always be free or even available, as Google recently demonstrated with some uses of their Maps API (Mitchell, 2011).

The second prerequisite for establishing a mashable digital collections service, programmatic access to the digital content, can be achieved in a number of ways by leveraging established protocols and utilizing features already available in library software. The easiest and most common method for developing a mash-up is to take advantage of an API, defined as “a set of functions, procedures, or classes for accessing a web service … [that] … reveal[s] the underlying logic on which a service is built, its key resources, and the functions amenable to be performed from outside the site … ” (Biancu, 2009, p. 19) Many library platforms have APIs for staff use, so there is no technical reason why they could not be accessed by users to create digital collection mash-ups, as long as functions related to personal or sensitive data are restricted. Permitting this type of API usage would simply be a matter of lobbying vendors to make the functionality a public feature. However, until APIs are available for end-user development, RSS and Atom feeds – XML-based formats for syndicating news items and other web content – represent a more feasible entry point. Many

digital library platforms include RSS feeds for newly added items, and some next generation discovery layers offer them for search queries. These feeds are not only an excellent way to market digital library content and draw in new users (Breeding, 2009; Moffat, 2006), but also can be used by mash-up development tools like Yahoo! Pipes and Kapow Mashup Server as input data streams to generate mash-ups (Liu *et al.*, 2011). Moreover, linked open data and microdata, initiatives that have the potential to increase interoperability and semantically enrich digital collections, generate machine-readable data that can be mashed up. (Concordia *et al.*, 2010; Singer, 2009) Finally, existing library standards can be re-purposed for mashing up digital collections. For example, z39.50 and SRU/SRW originally were designed to support federated searching through a standardized query interface, but have been used by some libraries to supply catalog data or institutional repository content for mash-ups (Westram, 2011; Witt, 2009).

The third requirement for a mashable digital collections service consists of two parts: a terms of use agreement and a mechanism for controlling access to the mashable content. The terms of use agreement establishes a shared understanding between developers and data providers of the acceptable and prohibited uses of the data, while a mechanism for controlling access, such as an API key, allows the data provider to track and control specific instances of data usage. These measures are required to prevent developers from misusing the service and are common for publicly available APIs. For many libraries, this component may be the most underdeveloped of the three requirements. Digital collections' terms of use policies could serve as the foundation for a terms of service agreement with the advice of legal counsel; however, very few digital collections already have a well-defined policy. In fact, a study of digital library collections indicated that terms of use are frequently intermingled with copyright statements and lack standard terminology or uniform placement (Schlosser, 2009).

While libraries are familiar with managing access to subscription e-resources, the mechanisms needed to regulate access to their own mashable data are quite different. Popular APIs, such as Google Maps or Twitter, use an API key to track specific instances of data usage or to turn off access, should a violation of the terms of use occur. For this type of mechanism, (once the developer has agreed to the terms of service) the data provider issues a string of alphanumeric characters associated with the developer's account or the mash-up's domain. This key then must be inserted into the mash-up code in order to access data on the parent website, thus allowing the data provider to manage the traffic coming from mash-up sites to its server by tracking (and potentially capping) the number of requests coming from the developer's mash-up URL. Should a mash-up violate the terms of use, the data provider can then invalidate its key to deny access to the data, consequently shutting down the mash-up.

An investigation of ARL members' digital collections

Clearly libraries have rich digital content, the means for providing programmatic access to it, and some also have terms of use policies that can serve as the basis for access management. These components, however, are better suited to support conventional uses, and thus need to be reworked to establish an effective service based on mashable digital collections. One might wonder what resources are already in place to support this service and what work remains to be done?

Methodology

In order to explore these questions, the author surveyed Association of Research Libraries members' websites to determine which institutions already provide some form of programmatic access to their digital collections (see www.arl.org/arl/membership/members.shtml for a complete listing of ARL's member libraries and links to their websites). Rights statements and terms of use policies were also reviewed when available, to evaluate their coverage for references to use cases such as user-generated mash-ups. Specifically, the author began on a member library's homepage, browsing the site for digital collections. Most institutions employ more than one platform to serve digital collections: digital libraries, institutional repositories,

catalogs, and the next generation discovery layers that run on top of traditional online catalogs to enhance their functionality; a good faith attempt was made to evaluate all platforms to the extent that they could be easily identified from the library's website. Some libraries also contribute their collections to third-party websites that aggregate content such as Flickr or Internet Archive; these access points were also explored when they could be discovered from the libraries' websites. Collections that required special authorization and any functionality that was only available to logged-in users were not examined for this study.

For each collection, the author gathered data on the software platforms utilized; the available options for mashable access to the digital content and data; and terms of use or rights statements. This information was gathered based solely on an inspection of the libraries' websites, not spending more than one hour on any given library in order to accurately reflect the likely behavior of an end user searching for mashable content. A mash-up developer is not likely to contact a library or software vendor to inquire about APIs for a digital library platform, nor would he be likely to drill down deep into a library's website to find this information. Therefore, the findings of this study describe the options for mashing digital collections based on what is apparent on libraries' websites, but does not take into consideration unpublicized avenues for accessing mashable data.

Findings

The 126 library websites surveyed provided access to their digital collections via 280 platforms, categorized for this study as follows: 160 digital libraries, 69 institutional repositories, 24 next generation discovery interfaces, 15 catalogs, as well as 12 third-party content aggregators such as Flickr and Internet Archive. It is likely that more than 12 ARL libraries contribute digital collections to these platforms, but these means of access are not easily discoverable from their websites. The investigation uncovered that 21 of the 126 (17 percent) ARL libraries served digital collections from at least one platform that provides a means for mashing digital content. Some libraries had multiple platforms with programmatic access; for example, both a library's next generation discovery layer and institutional repository have RSS feeds. Fifty-four platforms provided RSS or Atom feeds, the majority of which were found in institutional repositories. Only 13 digital collections were served from platforms that provide publicly accessible APIs. Interestingly, one library disclosed on its website that their collection could be queried via z39.50. In total, there were 68 instances (24 percent) of some means for mashing the digital collections among the 280 platforms surveyed.

RSS feeds were, by far, the most prevalent way to gain access to the underlying structured data in these library systems. RSS or Really Simple Syndication is an XML format used to publish and distribute frequently updated content such as blog posts or news stories. In the library context, RSS feeds are often used to notify users when new materials are added to an institutional repository or transmit catalog search results. The majority of the feeds examined in this study included title, date, and author fields from the metadata, and only contained entries for recently added materials, which is reasonable for current uses; however, these limitations hamper users' ability to tap into all available metadata to mash-up an entire collection.

While RSS feeds are predominant, APIs are a more robust way to mash-up the contents of a digital collection. Some digital collection platforms, such as the two most common systems identified in this study ContentDM and DSpace, have APIs for staff use, yet none of the surveyed collections served from these platforms had a publicly accessible mechanism for mashing the data. Only one digital collection, *Chronicling America*, offered programmatic access to its collection from a locally-developed platform.

Chronicling America, a digital collection of nineteenth and early twentieth-century US newspapers created by the Library of Congress and funded by the National Endowment for the Humanities, was the only collection identified by the survey to realize the potential for an API intended for end users. In fact the website states that the developers “designed several different views of the

data […] [to] encourage a wide range of potential uses" (Library of Congress, n.d.). The platform was developed in Python using the Django Web Framework, RDFLib, Apache web server, and MySQL database and employs various protocols for providing access to their data and content. Specifically, the API is built on OpenSearch, an Amazon-developed protocol for standardizing search results; Cross-origin resource sharing (CORS), a W3C specification that establishes a method for cross-domain resource-sharing; JavaScript Object Notation with Padding (JSON-P); and stable URL patterns utilizing LCCNs, dates, issue numbers, edition numbers, and page sequence numbers. Unlike many commercially-supplied APIs, *Chronicling America* does not require a license agreement or key for development. Rather, the project site includes a general copyright and terms of use statement that authorizes noncommercial, educational, and research uses of the collection. (Library of Congress)

While *Chronicling America* is a good example of a digital collection that provides programmatic access to its content, it must be noted that not all libraries have the resources necessary to develop and maintain a similar platform. Still, this does not have to impede organizations that wish to open up their digital collections for mash-ups. Many cultural heritage institutions have posted their digital content to third-party sites like Flickr and Internet Archive, a trend that is apt to increase with the establishment of the collaborative initiatives Flickr Commons and Open Content Alliance. The literature suggests that most libraries contribute their digital content to these platforms to promote collections, increase their usage, and facilitate Web 2.0-style interaction (Michel and Tzoc, 2010; Kalfatovic *et al.*, 2008; Nogueira, 2010); yet they offer the additional benefit of robust public APIs for mash-ups. Lewis and Clark College's innovative project, accessCeramics, illustrates this potential. Watzek Library staff utilized the Flickr API to collect metadata and images from artists around the country for inclusion in a digital collection. In a case study on the project, Dahl and McWilliams (2009) suggest modeling future digital library platforms on Flickr, specifically citing its robust public-facing API as a critical component of the ideal digital asset management system.

Most of the collections with API access in the survey were served from Flickr or Internet Archive. Both sites supply RESTful APIs and return data in JSON and RPC XML; Flickr also has additional development options. The availability of an API was not readily apparent from the collections' web pages, an important point if the objective is to encourage users to develop digital collections mash-ups. In Flickr, information about the API can be found under the "Explore" menu tab by clicking on "App Garden". API access to Internet Archive collections is available through the Open Library Project, a related initiative to create an editable web page for every book ever published. In both cases, the presence of an API was not immediately obvious from the collections' pages; however, once that section of the website is located, the easy-to-use API documentation, sample code, and example applications foster mash-up development.

Providing programmatic access makes new uses of digital collections possible. If libraries aim to offer this service, they must revise terms of use policies and institute access controls to manage these new activities. Of the collections 59 percent surveyed for this study contained some type of terms of use or copyright statement, but they were usually minimal. The most frequent terms uncovered in this investigation gave blanket approval for non-commercial, academic, and personal-study usages. Many of these statements also required users to request permission to publish an item from the collection, specifically citing web pages as a form of publication. Obviously, these terms of use statements would not accommodate a mash-up service and need to be reconsidered for such a service.

API terms of use licenses from the commercial sector can serve as a template for libraries' policies. A review of several of these reveals some commonalities. The terms of service usually establish ownership rights and include standard exclusion of warranties and limitation of liabilities clauses. These agreements also prohibit mash-ups from duplicating or competing with the parent service

and require proper attribution, as well as adherence to branding guidelines, privacy, personal information, and intellectual property rights policies. Many of these terms of use prohibit charging for mash-ups developed from their APIs unless a special commercial license is obtained. These agreements also prohibit the caching of data and limit the number of transactions from the mash-up to the parent server in order to manage server traffic. Finally most licenses permit data providers to change the conditions of the agreement at any time and to terminate access to the API at their discretion.

These terms from commercial APIs can not only function as a template for libraries' policies, but also highlight some of the potential issues involved in making digital collections mashable. Once users can remix the content and data, libraries are no longer in complete control of how the resources are used or represented (Concordia *et al.*, 2010). Providing an API for digital collections could also pose security concerns. Moreover, successful APIs will increase traffic to the library's server, thus competing with the library's other services for bandwidth. While these issues warrant careful consideration, appropriate policies and technical infrastructure can mitigate the risks.

Conclusion

Opening up digital collections so that end users can mash-up the content into new resources is a radical shift from the traditional approach to delivery, but yields several worthy benefits. It encourages user participation, enhances collections' utility, and increases their findability. The three basic elements – digital collections, programmatic access, and access management – already exist in libraries, but the last two components have been designed for different purposes and therefore must be reworked to accommodate a mashable digital collection service. A survey of ARL libraries shows that only 17 percent of the institutions assessed currently serve their digital collections from at least one platform that offers some type of public-facing programmatic access. In the short term, libraries desiring to establish a mashable digital collection service can contribute to third-party content aggregators in order to leverage their APIs. However, the long-term goal should be to design digital collection systems with user-facing APIs. This investigation also found that more than half of the digital collections examined had terms of use and/or rights information, but they were minimal and crafted with traditional modes of publication in mind. Libraries can look to commercial APIs' terms of use licenses and access management techniques like API keys to model this component of the service. Clearly, a mashable digital collection service would require libraries to rework their infrastructure, but the barriers to entry are not insurmountable and the potential benefits are great.

References

Association of Research Libraries (n.d.), *Association of Research Libraries (ARL)*, available at: www.arl.org (accessed 15 April 2012).

Biancu, B. (2009), "*Behind the scenes: some technical details on mashups*", in Engard, N.C. (Ed.), *Library Mashups: Exploring New Ways to Deliver Library Data*, Information Today, Medford, NJ, pp. 19-34.

Breeding, M. (2009), "*Maximizing the impact of digital collections*", *Computers in Libraries*, pp. 32-4, April.

Concordia, C., Gradmann, S. and Siebinga, S. (2010), "*Not just another portal, not just another digital library: a portrait of Europeana as an application program interface*", *IFLA Journal*, Vol. 36 No. 1, pp. 61-9.

Dahl, M. and McWilliams, J. (2009), "*Flickr and digital image collections*", in Engard, N.C. (Ed.), *Library Mashups: Exploring New Ways to Deliver Library Data*, Information Today, Medford, NJ, pp. 181-94.

Fichter, D. (2009), "*What is a Mashup?*", in Engard, N.C. (Ed.), *Library Mashups: Exploring New Ways to Deliver Library Data*, Information Today, Medford, NJ, pp. 3-17.

Gazan, R. (2008), "*Social annotations in digital library collections*", *D-Lib Magazine*, Vol. 14 Nos 11/12, available at: www.dlib.org/dlib/november08/gazan/11gazan.html (accessed on 22 April 2012).

Kalfatovic, M.R., Kapsalis, E., Spiess, K.P., Van Vamp, A. and Edson, M. (2008), "*Smithsonian Team Flickr: a library, archives, and museum collaboration in web 2.0 space*", *Archival Science*, Vol. 8, pp. 267-77.

Library of Congress (n.d.), "*Historic American newspapers – chronicling America*", available at: http://chroniclingamerica.loc.gov/ (accessed on 22 April 2012).

Liu, Y., Liang, X., Xu, L., Staples, M. and Zhu, L. (2011), "*Composing enterprise mashup components and services using architecture integration patterns*", *The Journal of Systems and Software*, Vol. 84, pp. 1436-46.

Michel, J.P. and Tzoc, E. (2010), "*Automated bulk uploading of images and metadata to Flickr*", *Journal of Web Librarianship*, Vol. 4 No. 4, pp. 435-48.

Mitchell, T. (2011), "*Introduction of usage limits to the Maps API, Google Geo Developers Blog*", available at: http://googlegeodevelopers.blogspot.com/2011/10/introduction-of-usage-limits-to-maps.html (accessed on 22 April 2012).

Moffat, M. (2006), "*Marketing with metadata – how metadata can increase exposure and visibility of online content*", *New Review of Information Networking*, Vol. 12 Nos 1/2, pp. 23-40.

Nogueira, M. (2010), "*Archives in Web 2.0: new opportunities*", *Ariadne*, Vol. 63, November, available at: www.ariadne.ac.uk/issue63/nogueira/ (accessed on 22 April 2012).

Schlosser, M. (2009), "*Unless otherwise indicated: a survey of copyright statements on digital library collections*", *College and Research Libraries*, Vol. 70 No. 4, pp. 371-85.

Singer, R. (2009), "*Making your data available to be mashed up*", in Engard, N.C. (Ed.), *Library Mashups: Exploring New Ways to Deliver Library Data*, Information Today, Medford, NJ, pp. 35-49.

Westram, A.-L. (2011), "*The key to the future of the library catalog is openness*", *Computers in Libraries*, April, pp. 10-14.

Witt, M. (2009), "*Electronic dissertation mashups using SRU*", in Engard, N.C. (Ed.), *Library Mashups: Exploring New Ways to Deliver Library Data*, Information Today, Medford, NJ, pp. 277-89.