White Paper Report

Report ID: 100998

Application Number: HD5115110

Project Director: David Lester (dlester@umd.edu)

Institution: University of Maryland, College Park

Reporting Period: 9/1/2010-8/31/2011

Report Due: 11/30/2011

Date Submitted: 11/30/2011

White Paper

Report ID: 100998
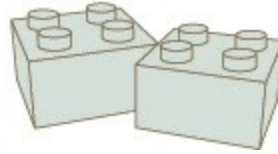
Grant Number: HD5115110

Project Title: MITH API Workshop

Project Director: David Lester (dlester@umd.edu)

Institution: University of Maryland, College Park

Date Submitted: November 30, 2011

Project Director: Dave Lester (dlester@umd.edu)
Website: http://mith.umd.edu/apiworkshop/

Workshop Overview
Funded by a Level 1 Digital Humanities Start-Up Grant, the MITH API Workshop took place at the
University of Maryland, February 25-26th, 2011, gathering digital humanities scholars and developers
interested in using Application Programming Interfaces (APIs) with leaders from industry who have
designed and deployed web-based APIs for their companies.

An API can be informally defined as a set of published commands that computer programmers can use in
their own code to interact with code that they did not write and to which they often have only limited
access. For example, an API is often provided to allow third party programmers to retrieve data from a
repository that they do not control. The value of this for a humanities scholar cannot be understated as
APIs allow access to data that may allow them to ask different questions that the application interface
allows and/or combine data from disparate repositories into a single archive to illuminate unknown or
unacknowledged connections. As such, the purpose of the workshop was to lay groundwork for the
integration of APIs into participant-driven digital humanities projects and to serve as a platform to
develop future ideas for how to share and access humanities data through APIs.

This white paper provides an overview of the workshop, identifies emergent themes addressed during the
workshop, includes briefs of invited presentations, and outlines possible next-steps.

Workshop Activities
At the February workshop, there were three distinct types of activities: invited presentations, informal
lightning talk presentations, and unconference sessions. Given the timeliness of the workshop topic, a
goal of the workshop was to reflect current questions being addressed in digital humanities projects by
pairing instructive presentations to frame relevant topics, with opportunities for the community of
participants to share their own work and experiences. Described in the original proposal as a "working
weekend", ideas from morning presentations often led to afternoon breakout sessions, with participants
drafting notes and ideas on Google Docs (made available through the workshop website).

The workshop began with a beginners overview of what APIs are, examples of existing APIs, and
principles behind their design. Later presentations including Mano Marks' presentation on Google Fusion
Tables and George Oates' Open Library presentation discussed not only how APIs offer a representation
of data, but also how important the data is itself. As much as the workshop discussion was about APIs and
their use, the conversation repeatedly returned back to questions about the underlying data that we're
creating in our projects and how we wish to use APIs to disseminate that data. A summary of invited
presentations and key points are included later in this white paper.

On both days, lunch included a designated period where participants could volunteer to present lightning
talks before the group. Each lightning talk presentation was limited to only four-minutes each, and
participants signed up at the event to pitch ideas, ask questions, or respond to ideas of the workshop. A
total of 18 lightning talks were given during the workshop.

Following lightning talks, workshop participants collaboratively developed a schedule of unconference sessions, to discuss and implement ideas related to APIs and morning presentations. A total of 12 different sessions took place in four separate rounds across the two days, ranging from *How to use an HTTP client* to *APIs for library data*. A complete schedule from the workshop, including titles of all unconference sessions and lightning talks is available as Appendix A.

The call for workshop participants was announced in December 2010, after launching an API workshop website, finalizing logistical details including time and place, and confirming the presence of invited presenters. The announcement was promoted through various social media including Twitter and Facebook, the Humanist listserv, and MITH's community mailing list, which includes hundreds of humanists and technologists from throughout the United States.

The API workshop generated interest in future events, and many participants commented that they hope to see the conversation about APIs in digital humanities projects remain active. In order to facilitate this communication, we have kept in place the public Google Group that was used for the workshop, prominently linking to it from the API Workshop website. The mailing list can continue to act as a forum for discussion and information sharing, as well as the #DHapi hashtag on Twitter.

Emergent Themes
Although the workshop was not thematically organized or divided into tracks, four notable themes emerged from the workshop that were addressed in presentations and unconference sessions: data mashups, API deprecation and sustainability, data contributions and enrichment via read/write APIs, and API design/REST principles. This section elaborates on these themes and examples discussed at the workshop.

Data Mashups
The emergence of APIs has facilitated the growth of "mashups": the combination of data from different sources into new forms. Mano Marks gave an invited presentation on Day One of the workshop about using data stored in Google Fusion Tables to create Google Maps mashups, exemplifying one of the most-common examples of data mashups on the web. Mano's morning presentation was followed-up by a hands-on workshop in the afternoon, where participants worked to learn more about the Google Maps API. Later during the workshop, an "API swap meet" unconference session took place, where participants discussed different APIs and services that may be useful to digital humanities projects, and shared links in a Google Doc now accessible from the workshop website.

APIs can facilitate different types of mashups discussed during the workshop. For connecting datasets, George Oates demonstrated how the Open Library links data sources by attaching multiple unique identifiers to resources (ISBN, WordCat ID, etc. Data can also be manipulated when passing it to an API, with a returned results set. For example, geocoder.us returns latitude of mailing addresses. Related to both the connection of data using APIs, and manipulation of data with APIs, visualization was one of the most-popular examples of mashups discussed; in particular, geospatial and temporal mashups were visuals that participants were attracted to.

API Deprecation and Sustainability
Adoption of Web APIs can create dependencies on third parties to provide critical infrastructure for one's project, posing significant issues if those APIs cease to exist. As an example of API deprecation, Raymond Yee sought to demonstrate the Amazon Wishlist API in his presentation, which provided access to data on Amazon.com that users would like to purchase, or have purchased for them in the future. Yee discovered that the API no longer existed and has been officially retired, since the time his book on API

mashups was published. Fortunately, shortly after the retirement of the Amazon Wishlist API, the open source community created a similar API that allows users to have similar access to data and operate their software using similar code. Had an application relied on the retired API, it would have likely ceased to function or experienced problems.

The Amazon Wishlist API is a cautionary tale of both the technical and social dimensions of APIs, and led to different discussions throughout the workshop about how APIs require a level of trust in a provider, ways that developers can assess the reliability of an API, and the likelihood that an API will be supported in the future. The state of support for an API can often be determined based upon the reputation of its provider, however in the case of Amazon its API was deprecated based upon there not being a viable business interest. If large companies choose to retire their APIs for economic reasons, how reliable can digital humanities projects that often operate on grant funding be? Demonstrated effort in the release of an API was often seen as an indication of commitment on the behalf of an API provider, including versioning APIs, documentation, as well as sometimes labeling the level of support provided. For example, Google makes distinctions between "experimental" APIs which can die on short notice, and "supported" APIs which require a three-year warning.

Data Contributions and Enrichment via Read/Write APIs
There was considerable interest in developing APIs for others to contribute and "give back" data, in addition to providing access to data. One of the most active unconference sessions was focused on transcribing handwritten documents that revolved the premise: would it be possible (and/or desirable) to create one Web site that could do the following: hosts the transcription interface, hosts the resulting transcriptions, and features an API allowing individual projects to grab those transcriptions? A series of highly visible crowd-sourcing projects currently exist including the UCL Bentham Project, War Department Papers, Open Library, and as future projects are developed they may benefit from the common approaches taken by these projects, rather attempting to re-implement similar features. In order to allow the crowd to interact with their data, an API (public or used internally by a web application) has been developed in each instance. Authentication is required in all instances of allowing users to write back to a data source, as well as maintainable of data revision histories.

Crowd-sourcing and participatory interactions are part of the value proposition for APIs in digital humanities: that by allowing greater access to content through standard methods, APIs can establish a platform for the enrichment of collections of data. While examples of crowd-sourcing typically focus on individuals contributing data (facilitated by an API and application), these same APIs act as a common way to interact with data and can be leveraged by automated processes. George Oates' talk about Open Library addressed how they are using automated "bots" to enrich Open Library data. Without the intervention of a human, bots perform tasks on data stored on Open Library's servers, and are given privileges to make changes to content. For example, Open Library's "ImportBot" imports book records that have been added to the Library of Congress into Open Library every Tuesday. Outside developers are encouraged develop their own bots to interact with their data, using their API.

In addition to further exploring the needs and uses of read/write APIs within digital humanities projects, the uses of bots on APIs is an interesting point of exploration. It may not be as likely that all projects adopt identical APIs (as discussed in the transcription session), but rather there may be specific tasks performed by bots that could act in commons ways across APIs and digital humanities projects.

API Design and REST principles
The design of APIs shared by participants varied in their purpose and scope, ranging from simple syndication of content to complex APIs including authentication and verbs. Different solutions may be suitable for different purposes, however "stealing shamelessly" may be the best way to initially design an API. As one of the most-popular and earlier APIs, Flickr is often used as an example. Peter Keane noted

that "imitation leads to interoperability"; the more we based APIs on the approaches of others, the closer we are to a common way of addressing data. One of the simplest standard approaches is to offer Atom feeds of data. Easily parsed and including link relations, Atom provides a format that can be reused by external applications. The Open Search standard formalizes the use of Atom by specifying its use in returning data from search queries on a website. Many popular open source software packages including WordPress and Omeka already support Atom, providing an out-of-the-box solution. Another unconference session focused on the REST Drupal module, which may provide a different platform for easily providing an API.

One of the most-common buzzwords related to APIs is REST, so an entire presentation was devoted to explaining the term and its related concepts. Peter Keane clarified REST is not a technology that we need to follow, but it's a term for a set of design principles that have been abstracted from what works well for APIS. The following principles are as follows: (1) everything is a Resource (bookmarks, orders, things, employees) (2) resources have Names -- URIs are the way we name things in HTTP (3) simple operations (GET,PUT,POST,DELETE) (4) resources have representations (html, rdf, json) (5) hypermedia is the engine of application state. A radical thing about REST is its simplicity, and how by using the Web and identifying resources with link relations, an API can be clear. Also, since REST uses HTTP, it's less fragile than other verb-based APIs.

Presentation Briefs
These briefs provide highlights, notes, and quotes of invited presentations at the MITH API Workshop. Edited videos of presentations are available on the MITH API Workshop website: http://mith.umd.edu/apiworkshop/

Raymond Yee (Founder, Data Unbound)
APIs: What they are and why they matter to digital humanities?

- Yee's talk explained APIs using easy-to-understand terms, and walked through several examples of mashups to offer ideas of how APIs can be used and connected.
- "The value of your data, when it is scattered throughout multiple databases and applications, grows if you can make it all work together. This value increases further when you leverage your information resources with the vast world of data on the Web. APIs can help you integrate data across your organization and beyond."
- Authentication is typically one of the more complicated parts of developing an API
- Yee shared a story of API deprecation that began a discussion about what constitutes a reliable source for an API
- Providing APIs and using APIs are not just technical matters. Everything from devoting reasons to create API, to deciding that as an organization must stand by an API, to deciding how much functionality will be allowed (read, write) affect the audience for users of the API
- Walk through of Flickr API, how it works, how data is accessed
- Yee demonstrated the mashup of an Amazon Wishlist with a Google Spreadsheet

Peter Keane (Software Developer/Librarian, The University of Texas at Austin)
REST in the design, use, and documentation of Web APIs

- Keane's talk explained REST, its principles, and how it's being implemented.

- "REST is more abstract than what people thing. It's not a concrete technology, but instead a design principle (and sometimes more difficult for people to understand). It's a set of ideas that are radical in their simplicity and minimalism"
- Keane walked through each of the five principles of REST:
    1. Everything is a Resource (bookmarks, orders, things, employees)
    2. Resources have Names -- URIs are the way we name things in HTTP
    3. Simple operations (GET,PUT,POST,DELETE)
    4. Resources have representations (html, rdf, json)
    5. Hypermedia is the engine of application state
- "Work with the grain of the web instead of against it"
- Demonstration of how REST allows single resources to have multiple representations, making it easier for humans or machines to reuse content. For example, the HTML content of a page could also be shared as an XML, JSON, or TXT file.
- Use of link relations makes it easier for machines to find resources.
- Every time that Flickr uses a verb in their API, they're creating a protocol that sits on top of HTTP, and therefore it is fragile.
- Several links and resources are suggested, including:
    - Roy Fielding's dissertation: "Architectural Styles and the Design of Network-based Software Architectures"
    - Restful Web Services by Leonard Richardson, Sam Ruby, David Heinemeier Hansson
    - RESTful Web Services Cookbook: Solutions for Improving Scalability and Simplicity by Subbu Allamaraju
    - Library of Congress: Authorities & Vocabularies: (http://id.loc.gov)
    - IANA registry for link relations

Mano Marks (Lead Geo Developer Advocate, Google)
Google Fusion Tables

- Marks' presentation focused on Google Fusion tables, and how it can be used as a platform to feed data into other tools and APIs including the Google Maps API to create visualizations of data.
- Google Fusion Tables allows users to manage large collections of tabular data in the cloud. Users can upload a spreadsheet, click on visualize (map, barchart, storyline, timeline), and you have a visualization that can be shared with people with no programming required. This data mashup can be embedded into a user's website.
- Data from Fusion Tables can be loaded into the Google Maps API, and Fusion Tables has its own API, including a SQL-like query language that you can get/post data to.
- Fusion Tables does not have authentication methods, however you can perform a direct query using public data
- Fusion Tables' lack of authentication methods began a discussion of what's required to establish authentication, and why many API providers are starting to require Oauth. Of the standards for everyone to adopt, OAuth is probably the most secure approach at this time.
- Several links and resources were suggested, including:
    - Google Faculty Research Awards
    - Open DataKit, "a suite of tools to help organizations collect, aggregate and visualize their data."

George Oates (Lead, Open Library)
Open Library

- Oates' talk described the Open Library project, how they make their data accessible, and their experience building a community around a read/write API.
- Similar to stamps on the back of a library book, Open Library allows you can see who has touched records, when, and what they did.
- Open Library has a RESTful API, they have developer documentation, and its code is open source.
- Many external sites are using Open Library resources, including the Trove search engine that has federated Open Library data into their search results. When there is a book in the public domain, a link to the digitized book from Open Library appears on the Trove website.
- OpenLibrary uses bots (pieces of software running, interacting with the website) that clean up and correct data, and encourages developers to write their own bots to interact with their read/write API.
- The Internet Archive's API was used by Dewey Music to develop a new interface for their music collection, providing an example of how powerful can be for reuse of content.
- Oates led a discussion about her idea of a "minimum viable record" for library book data, and how establishing such a standard would be helpful in establishing greater connections between data within a system like Open Library.

Mano Marks, prepared by Kirrily Robert (Community Lead, Freebase).
Google's Freebase

- Robert's talk was presented by Mano Marks, who stepped in to present her notes and presentation slides on her behalf. Several months before the workshop, Freebase had been acquired by Google. The presentation described Freebase, how it can be used to store linked data, and how data can be queried and repurposed using their API.
- Freebase is an entity graph of people, places, and things, built by a community. Data in Freebase is structured, allowing it to be queried, connected, and uniquely identified.
- Data in Freebase comes from external sources including Wikipedia, MusicBrainz, and Netflix, as well as developers that can write data to Freebase (adding, updating, or deleting content) using their API
- All Freebase data is licensed under Creative Commons attribution
- Identifiers in Freebase are like foreign keys into remote databases, and enable mashups. The more links to related data that a piece of data has, the more ways it it can be reused
- When importing data into Freebase, it will often need to be reconciled with what's already in Freebase. Matchmaker is a tool they've built to help reconcile imported data
- MQL: Metaweb Query Language is a SQL-like language that can be used to query Freebase, and requests can be sent through their API
- Google Refine can also be used to write to Freebase as an alternative to "mqlwrite"

Future Directions and Follow-up
Several potential follow-up ideas have emerged from the workshop, including similarly formatted workshops, training opportunities, the development or improvement of existing software to conform to principles of REST, and API hackathons. These ideas remain viable, and there is tremendous potential for

these ideas to be combined with related conversations around humanities data, including data publication, data curation, and linked data. Digital humanities API hackathon events have begun to appear, including an event focused on the Europeana API earlier this year, and there's potential to participate and facilitate similar events. MITH's immediate effort to follow-up the workshop has been in the early planning of a possible workshop thematically focused on APIs and museums. At the NEH-funded API workshop, there was a session on library data, but not specifically museums which this would provide.

We are optimistic that the workshop and its grant products have helped clarify topics of conversation about APIs within digital humanities. Several new NEH-funded projects now include APIs, including the Indiana Philosophy Ontology Project (https://inpho.cogs.indiana.edu/docs/) and Old Bailey (http://www.oldbaileyonline.org/obapi/), and the resources of the workshop website along with these examples of DH projects offering APIs provide energy to see substantial impact.

Appendix A. Workshop Schedule

Day One, September 25th, 2011	
8:30 – 9:00am	Breakfast and registration
9:00 – 9:30am	Dave Lester and Neil Fraistat Welcome and Introductions
9:30 – 10:15am	Presentation, Raymond Yee *APIs: What they are and why they matter to digital humanities?*
10:15 – 10:30am	Break
10:30 – 11:15am	Presentation, Peter Keane *REST in the design, use, and documentation of Web APIs*
11:30 – 12:00pm	Presentation, Mano Marks *Google Fusion Tables*
12:00 – 12:15pm	Break
12:15 – 1:30pm	Lunch and Lightning Talks • Ian Brown, *Making a Business Case for APIs* • Jason Clark, *Creating context for data* • Patrick Murray-John, *World views/perspectives built into APIs* • Michael Lascarides, *NYPL Recipes Website* • Jen Serventi, *NEH Office of Digital Humanities* • Ed Summers, *Open Search* • Chris Thatcher, *APIs at Library of Congress* • Jeff Mummert, *American Civil War Augmented Reality Project* • Neil Mansilla, *Push Notifications and real-time data via APIs*
1:30 – 2:10pm	Group discussion about presentations, and planning of day one unconference
2:15 – 3:15pm	First round, unconference sessions: • Room 7113: *Building/Using APIs for Social Networking Data* • Special Events Room: *Interested in Maps & API's?*
3:15 – 3:30pm	Coffee break, switch rooms
3:30 – 4:30pm	Second round, unconference sessions: • Room 7121: *How to use an HTTP client* • Room 7113: *Natural User Interfaces (NUI): "deep(er) exploration of how gesture-based interfaces transform what data is, can be."* • Special Events Room: *APIs for Library Data*
4:30 – 5:00pm	Reporting back of unconference sessions, wrap up with entire group

Day Two, September 26th, 2011	
10:00 – 10:30am	Breakfast
10:30 – 11:15am	Presentation, George Oates *Open Library Project*
11:15 – 12:00pm	Presentation, Mano Marks (prepared by Kirrily Robert) *Freebase*
12:00 – 12:15pm	Break
12:15 – 1:30pm	Lunch and Lightning Talks • Trevor Owens, *Recollection* • Seth Denbo, *Corpora Space* • Boris Capitanu, *SEASR/MEANDRE* • Piotr Adamcyzk, *Museum APIs* • Dot Porter, *TILE* • Kate Chapman, *Geocomons* • Damon Baker, *Cavewriting: spatial hypertext* • Patrick Murray-John, *Scripto* • Natalie Binder, *New audiences – rural broadband*
1:30 – 2:10pm	Group discussion about presentations, and planning of day two unconference
2:15 – 3:00pm	First round, unconference sessions: • Room 7121: *APIs in grant-funded projects* • Room 7113: *Hacking bibliographic metadata* • Special Events Room: *Transcription/Translation*
3:00 – 3:10pm	Coffee break, switch rooms
3:10 – 4:10pm	Second round, unconference sessions: • Room 7121: *Connecting Drupal to web APIs* • Room 7113: *Executable Scholarship* • Special Events Room: *Authoring a Collection API* • MITH Conference Room: *API swap meet*
4:10 – 4:30pm	Closing remarks and wrap-up

Appendix B. Participants and Presenters

Presenters

Name	Position	Institution
Keane, Peter	Software Developer/Librarian	University of Texas, Austin
Lester, Dave	Assistant Director	University of Maryland, MITH
Marks, Mano	Lead Geo Developer Advocate	Google
Oates, George	Lead	Open Library
Robert, Kirrily	Community Lead	Freebase
Yee, Raymond	Founder	Data Unbound

Participants

Name	Position	Institution
Adamczyk, Piotr	Associate Analyst	Metropolitan Museum of Art
Baker, Damon Loren	Assistant Professor	CUNY CityTech
Binder, Natalie	Technology Specialist	Florida State University/Jefferson County Public Library
Blanch, Kate	Administrator, Museum Databases	The Walters Art Museum
Bradley, Peter	Associate Professor of Philosophy	McDaniel College
Capitanu, Boris	Senior Project Coordinator	UIUC
Cayless, Hugh	Developer	NYU
Chandler, Zach	Academic Technology Specialist	Stanford
Chapman, Kate	Developer Community Lead	FortiusOne
Chen, Alex	Associate Professor	University of Maryland
Chun, Susan	Independent Researcher and Consultant / Faculty at Johns Hopkins University, Universita della Svizzerà Italiana	Project Manager / Museum Studies Faculty
Clark, Jason	Head of Digital Access and Web Services, Digital Initiatives Librarian	Montana State University Libraries
Denbo, Seth	Project Coordinator	MITH
Dickie, Grant	Developer	MITH
Farrell, Jenel	Digital Archivist	Minnesota Public Radio
Farrell, Robert	Director of Product Management	Gallery Systems
Flatscher, Markus	Project Editor and Developer	Rotunda/UVa Press

Gibbs, Fred	Project Manager and Faculty	CHNM, George Mason University
Gomez, Joshua	Digital Library Programmer/Analyst	George Washington University
Gordy, John	Web Manager	National Gallery of Art
Gorges, Boone	Developer	CUNY
Gurzick, David	Visiting Assistant Professor	Hood College
Hess, Kirk	Developer	Indiana University
Hooper, Wallace	Indiana University, Digital Library Program, History and Philosophy of Science	Project Manager, Programmer/Analyst, Chymistry of Isaac Newton Project
Klabnik, Steve	Developer	University of Pittsburgh
Kuhn, Jim	Head of Collection Information Services	Folger Shakespeare Library
Lascarides, Michael	Digital User Analyst	New York Public Library
Lawless, Ted	Library Applications Developer	Brown University
Lukehart, Peter M.	Associate Dean, Project Director	Center for Advanced Study in the Visual Arts, National Gallery of Art
Maron, Debbie	Zotero Community Lead	CHNM
Martin, Worthy	Acting Director, IATH and CS faculty	IATH, University of Virginia
Maza, Jim	CTO	Walters Art Museum
Mummert, Jeff	Social Studies Teacher, Adjunct Instructor	Hershey High School, York College of Pennsylvania
Murdock, Jaimie	Research Programmer	Indiana University
Murray-John, Patrick	Instructional Technology Specialist	University of Mary Washington
Muse, Eben	Deputy Head, School of Creative Studies and Media	Bangor University
Owens, Trevor	Digital Archivist	Library of Congress
Phillips, Mark	Assistant Dean for Digital Libraries	University of North Texas Libraries
Porter, Dot	Associate Director for Digital Library Content & Services	Indiana University
Posner, Miriam	Postdoctoral Fellow, Digital Scholarship Commons	Emory University
Poston, Michael	Database Applications Associate	Folger Shakespeare Library
Rieder, David	Associate Professor of English/program faculty, PhD in CRDM	NC State University
Safley, Jim	Web Programmer, Digital Archivist	CHNM

Schaefer, Sibyl	Digital Initiatives Metadata Librarian	University of Vermont
Sohn, Michael	Developer	Dumbarton Oaks Research Library and Collection
Summers, Ed	Information Technology Specialist	Library of Congress
Thatcher, Christopher	Digital & Web Initiatives Developer	Library of Congress
Varner, Stewart	Digital Scholarship Coordinator	Emory University, Woodruff Library
Williams, George H.	Assistant Professor of English	University of South Carolina - Upstate
Zacharski, Ron	Assistant Professor	University of Mary Washington

* institution and affiliation reflects the time of the workshop