Europeana Cloud WP1

So we've built it, but have they come? Investigating barriers and opportunities for API usage among the AHSS community

A joint NeDiMAH / Europeana Cloud Workshop
17th December 2014
The Hague, Netherlands

Dr. Jennifer Edmond and Vicky Garnett
Trinity College Dublin

Agiatis Benardou
DCU Athens

Abstract

The aim of this workshop was to deliver an event to demonstrate the potential for API usage to non-technical members of the eCloud and NeDiMAH key researcher cohorts and to gather further detail on perceived barriers and possible solutions. We invited researchers and developers to talk about their research practices; and non-technical researchers in Humanities and Social Sciences to tell us whether they find the potential interesting and/or the skills required too difficult.

Scientific Summary

There are plenty of researchers using cultural data, and it may well be that in some cases their systems of data capture do make use, wholly or in part, of an API service. But in our initial studies of digital workflow practices, most of the researchers we were able to identify really only cared about the data, and had no specific opinions about how that data was accessed: they cared about the electricity (data) and the lightbulb (results enabled by the data), but not the plug socket (API or other data transfer service).

The term API, referring as it does to the specific aim of connecting the lamp to the electricity, seems therefore to be *a priori* restricted to the use of and by developers.

As such, it seems the majority of users of cultural heritage APIs are still developers and computer scientists, although there is a small group of Humanities and Social Science researchers who are re-using the data they can obtain through a web-service or API. Who is doing the extracting of that data, however makes the difference. In the case of some exemplar digital humanists, they are making use of developers to make the calls to the APIs and obtain the data they need. They have the expertise to know what they can do with the data when they get it, but they don't 'dirty their hands' by writing the call to the API themselves. On the other hand, we have developers who not only write the code to call the API themselves, but also maintain the content for the API. We might call these people the 'data evangelists', as they showcase what can be done with a particular Cultural Heritage Institution's API and data.

But the potential future usage of APIs may not be reflected in the current patterns, as many of the current workflows scholars deploy engage similar functions and steps. This workshop therefore brought together and attempted to shed light on a full landscape of practice and possibility. As such, the event included perspectives of creators and developer/users of APIs, but also support services within the data-intensive humanities research lifecycle as well as those humanists reusing data themselves, with a specific focus on how they would acquire data and what they would want to do with it (that is, in most cases, what structure they would apply after download).

This workshop further investigated the workflows surrounding API use. In doing so, we were in a better position to determine the current state of the art of API use, the barriers, practices and justifications, and develop a workflow that non-technical as well as technically competent humanists can follow in order to obtain big data sets using web services and APIs.

Meeting Programme

Morning Session	
9.00am	Arrival
9.15am	Welcome from eCloud WP1 Leader, Agiati Benardou
9.15am	Context and Goals for the Day – What have we learned so far on API and web-service use? – Dr. Jennifer Edmond
10.00am	Adam Crymble - Digital histories - Data Mining
10.15am	Paul Rayson - Historical linguistics/Psychology
10.30am	Mark Sweetnam - 1641 Depositions/Cultura
10.45am	Coffee
11.00am	*Morning break out groups* - Rapid ideation session on humanities research 'data to knowledge' pipelines
12.15pm	Presenting the Outcomes
12.45pm	Lunch
Afternoon session	
1.45pm	**Basics of API** - what a call looks like - and what the data looks like. - What an API CAN'T do.
2.15pm	Gonzalo Parra, Work Package 3 - Europeana work
2.30pm	Dimitris Gavrilis, LoCloud/DCU Athena
2.45pm	Paul McCann, National Library of Wales
3.00pm	Questions for Developers
3.15pm	***Afternoon break out groups - Storyboarding use cases*** 3.15pm Decide on which one of the morning use-cases from each group provides the best use of the API, in terms of how it can pull, process, and present data 3.30pm Storyboard out use case of answering that research question using those tools, in a research environment.
4.30pm	Final Discussion about the process of developing prototypes
5.00pm	Concluding comments – close.

Storyboarding

Over the course of the day, the groups were asked to devise a series of research questions that they might want to answer. In doing so, the groups were asked to think of the data that would be required to answer their research question (be it a small or large part of the overall research), and how they might best obtain that information.

Once they had determined this, the groups were then asked to 'storyboard' the process by which they might obtain that information. This could be in the form of a website, an interface, or through coding.

Newspaper API Builder

The Linguistics group decided to work on an existing database of newspaper articles, and develop and interface that allowed the user to put a series of filters in place to generate a URL and/or the data output. These filters included the elements of a newspaper article that the search should concentrate on (e.g. headlines only, first sentences, by-lines, entire article), date of publication, date ranges, locations of publications, etc. Once all filters that are required have been applied, the URL and also is required, data output will then be shown in the right-hand side of the interface. This data output could be selected to appear in either: JSON, JSONP, XML or RDF.

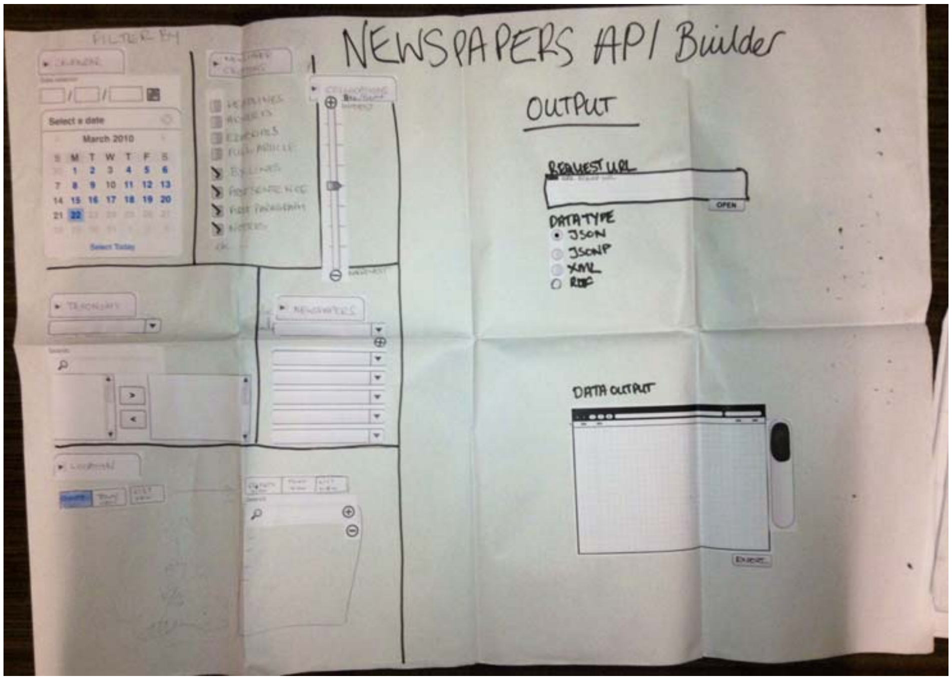


Figure 1 - "Newspaper API Request Builder" storyboard from the Linguistics group.

Image Clusters

The History group decided to look at the issue of how to combine all the different datasets that might be associated with images in collections. They devised an API that would visualise the output data in cluster formats. The filters that could be applied included visual information such as 'hair length' of the subject and gender.

Again, they used a split-screen approach, with the filtration occurring at the top, and the output occurring in a separate frame within the interface.

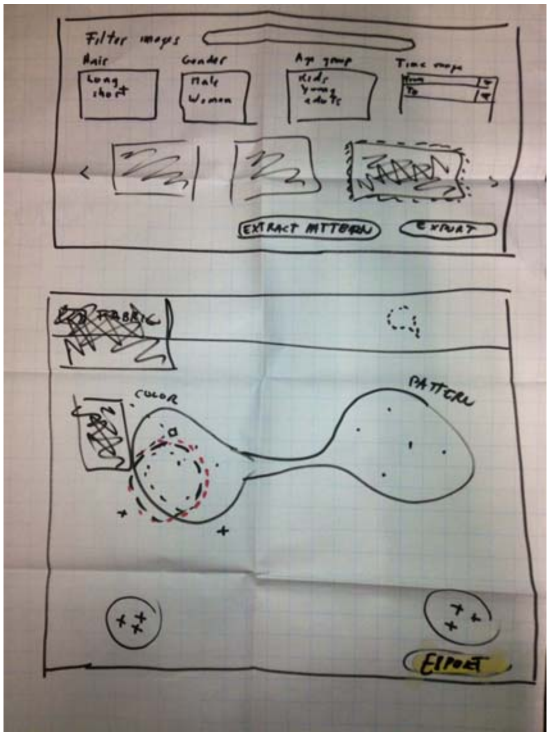
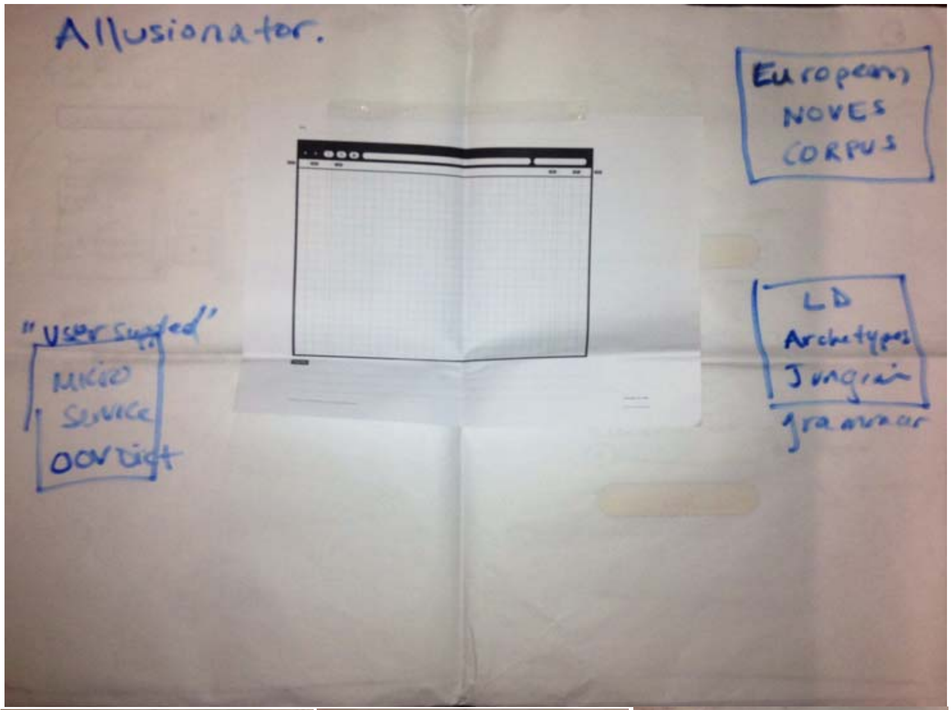
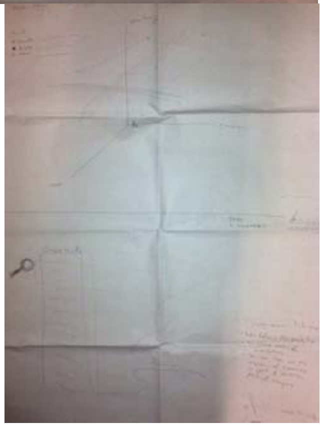


Figure 2 - Image clustering storyboard from the History group

The Allusionator

The Literature group decided to explore whether there were meaningful potential expansion avenues for tools like the Mallet software toolkit. The envisioned a two-layer approach to literary analysis, where by your model (which might be the result of topic modelling one or a small corpus of texts) was them applied as a template to a larger corpus. In this way, the user would be able to adjust the signal to noise ratio both in what they were looking for (granularity of the focus on the model) and where they were looking for it (size and composition of the corpus. By adjusting the controls on the model building kit, users would hopefully be able to gradually determine webs of interconnectedness among texts according to the proximity of the allusions contained in them. To assist in this, they would be supplied with an interface with multi-faceted filtering options, and visualising tools on a 3D plot diagram to show 'closeness' between texts.





Recognising the Barriers: Enabling the Participants

Aside from using the workshop to identify workflows for both technically proficient and non-technical humanists, the workshop also tried to ensure that the participants were able to use what they had learned during their time at the workshop as a launchpad to then go and put APIs into practice. In order to do this, we had to address the fundamentals of API use. In particular, the very basics of making a call to an API, and what that looks like. This small section of the day walked the participants through the process of making a call to an API. This included showing them tools such as code editors, the need for a particular syntax based on the coding language, writing the basic API call request, and how to ensure you get the data in the format you need it. We then showed what could be done with the data from the API call, and how it can be formatted into an excel spreadsheet for further analysis.

Unknown Unknowns and Very Clever Things

The response from the participants who had not previously seen this being done was positive, many of whom reported that while they were aware of coding and sites that could help to make an API call, they were completely in the dark about what kind of programme one might use to do that, and what you did with the resulting API call code. This, therefore, tapped into something that might often be overlooked by developers and data providers: the 'unknown unknowns' about data retrieval that are experienced by 'analogue' humanists. Many humanists who have not previously worked with digital data collections might be aware that Very Clever Things can be achieved using digital techniques, and may have seen those Very Clever Things in action, but they have no idea of the process involved in getting to that stage, and moreover, are not even aware of what they need to find out in order to give them this information.

Europeana therefore needs to look at how it might address these 'unknown unknowns' to the Analogue Humanist world in order to ensure that more of its potential users are able to get the most from the data collections on offer.

Acknowledgements

The organisers recognise the generous funding of NeDiMAH and the European Science Foundation for this workshop.