DELIVERABLE

Project Acronym:	**Europeana Cloud**
Grant Agreement number:	**325091**
Project Title:	**Europeana Cloud: Unlocking Europe's Research via The Cloud**

D3.2 Tools and services:

A set of tools and services for researchers that exploit Europeana content

Revision: v19

Authors:

Erik Duval (KU Leuven)
Gonzalo Parra (KU Leuven)
Sven Charleer (KU Leuven)
Anja Jentzsch (Open Knowledge Foundation Deutschland)
Hein van den Berg (Stichting VU-VUMC)
Giannis Stoitsis (ARIADNE Foundation)
Katerina Galani (ARIADNE Foundation)
Nikos Marianos (ARIADNE Foundation)
Andreas Drakos (ARIADNE Foundation)
Marnix van Berchum (KNAW-DANS / Utrecht University)
Maritina Stavrakaki (ARIADNE Foundation)

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	x
C	Confidential, only for members of the consortium and the Commission Services	

Revision History

Revision	Date	Author	Organisation	Description
V1	19.01.2015	Gonzalo Parra	KU Leuven	Adding the structure to include the work from Year 2
V2	19.01.2015	Andreas Drakos	ARIADNE	Update of Ariadne Finder section
V3	29.01.2015	Anja Jentzsch	OKDE	Input for Aruspix service
V4	29.01.2015	Gonzalo Parra	KU Leuven	Input in Music21 service, updates in Activity Stream section and minor edits
V5	30.01.2015	Hein van den Berg	VU	Corrections and minor edits
V6	30.01.2015	Marnix van Berchum	DANS-KNAW / Utrecht University	Minor edits on musicological parts
V7	01.02.2015	Erik Duval	KU Leuven	Consolidation, minor edits
V8	02.02.2015	Anja Jentzsch	OKDE	Input for adjusted TimeMapper
V9	13.01.2016	Sven Charleer	KU Leuven	Added Newspaper Exploration Tool
V10	14.01.2016	Anja Jentzsch	OKF DE	Update on Year 3 tools
V12	25.01.2016	Maritina Stavrakaki	ARIADNE	Update on AGRERI Discovery Microsite section and revision
V13	25.01.2016	Hein van den Berg	VU	edits
V14	26.01.2016	Anja Jentzsch	OKF DE	Formatting and update on eCloudDM
V15	29.01.2016	Anja Jentzsch	OKF DE	Layout and chapter structure
V16	29.01.2016	Sven Charleer	KU Leuven	Additional changes
V18	30.01.2016	Maritina Stavrakaki	ARIADNE	Updating and editing
V19	31.01.2016	Anja Jentzsch	OKF DE	Layout

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

D3.2 Tools and services

A set of tools and services for researchers that exploit Europeana content

Executive summary

This deliverable describes the tools developed in Europeana Cloud, which will be integrated with the Europeana Research Platform. Each tool is described, and references are included to the locations at which the software can be accessed.

This is the final version of the deliverable.

Table of Contents

List of Figures

1. Introduction

Work Package 3 (WP3) aims to develop services and tools that leverage Europeana content in the Europeana Cloud for researchers. During the first months, the work of WP3 focused on the development of personas, scenarios and use cases, in order to understand and analyse the user needs. This initial work on personas, scenarios and use cases was reported in Deliverable 3.1.

As described in the Description of Work (DoW), the work of WP3 proceeds in yearly cycles. Each cycle targeted a specific research community of users:

- Year 1 focused on the DM2E project (and more specifically the Wittgenstein archives at the University of Bergen) and the Axiom philosophy group at the VU University Amsterdam;
- Year 2 targeted a research community of musicologists that focus on Early Music;
- Year 3 focused on Digital Humanities researchers interested in the Europeana Newspaper Archive and on Agricultural researchers.

We developed and evaluated a prototype demonstrator that integrates a variety of tools relevant for the target audience. These tools are:

- ARIADNE finder: a personalised search micro site to help researchers search and find content related to their work coming from Europeana and other external sources; [5][1];
- TimeMapper: an integrated tool to visualise search results on a timeline and on an interactive map, so that users can further filter the content and get a better overview of the different resources found on Europeana;
- An Activity Stream, integrated with the other tools, to capture and present the different actions taken in this process (search, visualise, explore, annotate, download); [1][2],
- Aruspix, an optical music recognition (OMR) tool that scans early music prints and transcribes them using the MEI standard; [4][3],
- Music21, a toolkit for computer-aided musicology to parse, analyse and process encoded scores; [3][4].
- Newspaper Exploration Environment, a tool that facilitates visual exploration of digitised newspapers through multi-faceted filtering;
- eCloudDM, a tool for data mining texts from newspaper articles to augment the articles with named entities and topic tags;

[1] K. Makris, G. Skevakis, V. Kalokyri, P. Arapi, S. Christodoulakis, J. Stoitsis, N. Manolis, and S. L. Rojas. Federating natural history museums in natural europe. In Metadata and Semantics Research, pages 361–372. Springer, 2013.

[2] G. Parra, J. Klerkx, and E. Duval. Tinyarm: Awareness of research papers in a community of practice. In Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies, page 21. ACM, 2013.Shneiderman, B. 2008. Science 2.0. Science. Vol. 319, No. 5868, 1349-1350.

[3] L. Pugin, J. A. Burgoyne, and I. Fujinaga. Goal-directed evaluation for the improvement of optical music recognition on early music prints. In Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, pages 303–304. ACM, 2007.

[4] G. Haus and M. Longari. A multi-layered, timebased music description approach based on xml. Computer Music Journal, 29(1):70–85, 2005.

- the AGRERI Discovery Microsite: a personalized search microsite to help agricultural researchers search and find content related to their work coming from Europeana and other external sources.

In this document, we present a very short description of each of these tools, with a reference to where the software can be accessed on the web. An evaluation of these tools has also been carried out and is presented in Deliverable 3.3.

Although not detailed in this Deliverable, it is also interesting to note that the creation and testing of each of these tools also resulted in the creation of metadata, which WP3 will send back to Europeana Cloud for use by others. This fits the spirit of the Europeana Cloud project, which aims to allow users to contribute back to Europeana via the Europeana API.

2. The ARIADNE Finder

The first tool is the ARIADNE Finder, a personalised microsite that allows users to search and discover resources. The Finder searches predefined collections of datasets, as indicated by user input, and presents the results in a uniform way.

It comes with lightweight web-technologies (HTML, CSS, HTTP, Javascript, AJAX) in order to be easily embedded in sites and web-applications, without the need to make changes for matching the existing technologies of the application.

In the context of the Europeana Cloud project and WP3, prototypes of the Finder have been deployed to showcase how researchers can search and access resources coming mainly from Europeana, while still having access to all other tools.

Two versions of the finder were produced during the lifetime of the project, one for the Axiom philosophy group at VU Amsterdam[5] and one for a research community of musicologists[6]. These are described below.

First Prototype – Axiom Group

The first prototype of the Finder in WP3 was designed based on the needs of the Axiom philosophy group and is personalised in two main ways: 1) it is integrated with the web site of the Axiom group[7] and 2) is built on top of collections that have been requested by stakeholders. The main usage of the ARIADNE Finder is a faceted search interface that allows users to search and quickly filter the results. In addition, predefined categories that allow access over specific content (i.e. philosophers studied by the target audience) are also available.

Its development was led by discussions between the Axiom group and WP3, which allowed us to gather feedback concerning the collections to search, categories to use, and the facets that the stakeholders would like to use.

In order for the Finder to allow faceted search and uniform representation of the metadata from resources coming from different collections, we developed the Finder to use the existing ARIADNE infrastructure to store a repository with all the metadata. The resources and different collections

[5] http://greenlearningnetwork.com/axiom/

[6] http://greenlearningnetwork.com/cmme-finder/

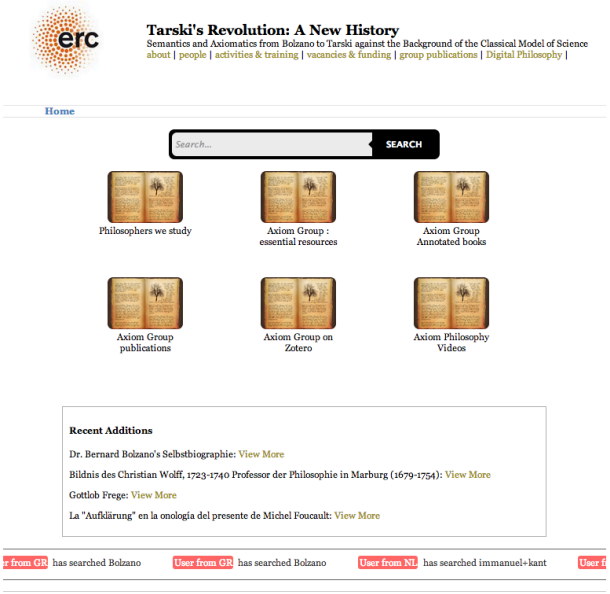
[7] http://axiom.vu.nl

stored in the repository were limited to Europeana and Google Books. In both cases, an API was used to filter thematically resources for the dataset.

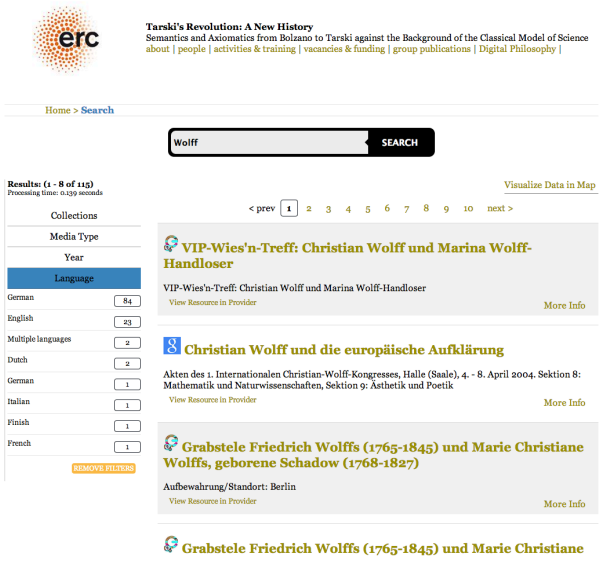
To provide a uniform representation and make the resources available through the Finder, we then transformed all files from their original scheme to an internal format. During this transformation procedure, metadata was also enriched.

Finally, we established the Finder as the main visible tool, on top of other tools with which it is integrated. On the first page of the Finder, the Activity Stream is integrated in the bottom screen as a floating message, while in the listing of the results a link to visualise the results in the TimeMapper is available. More information on these tools is available in the following sections.

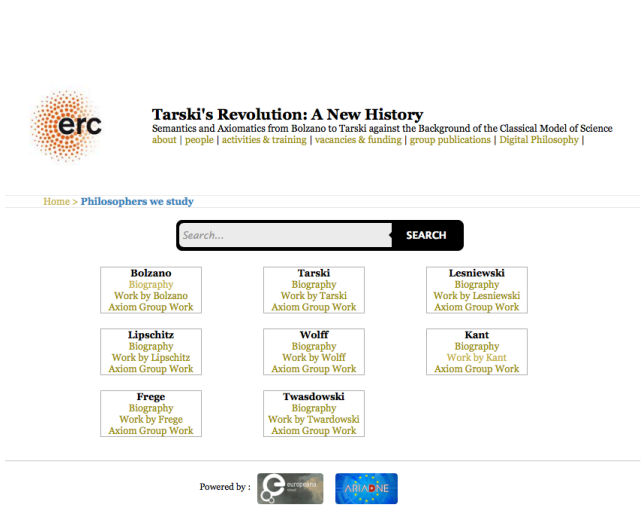
For the integration of the different tools, a number of technical issues had to be resolved, such as passing the POST search activities in the REST service of the activity stream, passing the JSON file to the TimeMapper, etc. Apart from online meetings and, when needed, bilateral communication, a specific WP3 discussion board was kept to discuss and resolve all technical issues.



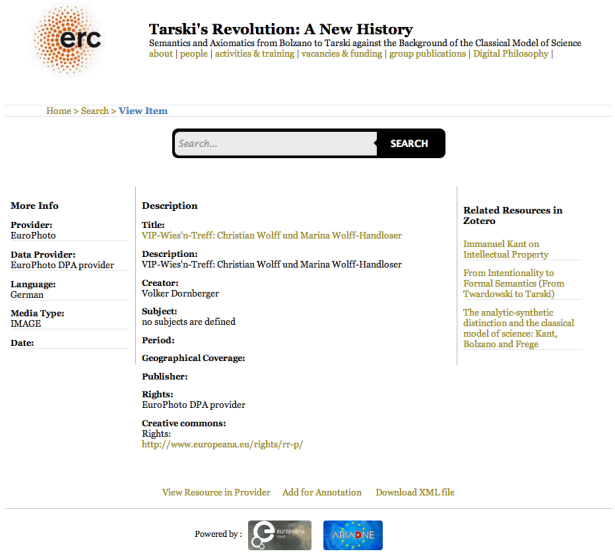
(a)



(b)



(c)



(d)

Figure 1: Screenshots from the first year version of the Ariadne Finder, showing the main page (a), listing results (b), customized categories (c) and view item (d)

The above figures show different screenshots from the Finder developed in Year 1. In Figure 1-a the main page of the Finder is shown, with the Activity Stream at the bottom. Figure 1-b shows the listing of the results after a search is executed, with the facets that can be used and the button to visualise the results in the TimeMapper. Figure 1-c shows the menu of the predefined customised categories, offering quick access to specific results. Finally, figure 1-d shows how a single result is presented.

The ARIADNE Finder for the Axiom philosophers group can be accessed at the following URL: http://greenlearningnetwork.com/axiom/

Second Prototype – Musicologists

During Year 2, the Finder's prototype was designed based on the needs of a group of researchers studying early music. Following same process as in Year 1, a series of meetings with WP3 partners and musicology researchers took place. Design feedback was collected and the appropriate requirements for additional collections of interest to integrate in the Finder were determined.

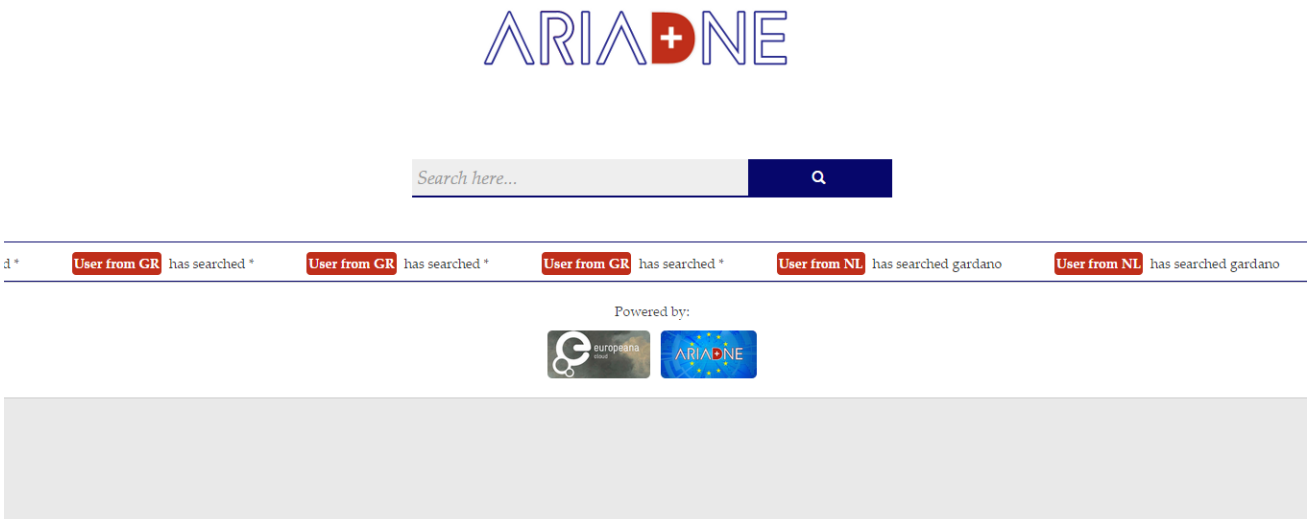
The second year version of the Finder includes a more simplified user interface, with no predefined categories available on the home screen. Instead, a list of four search facets (i.e. provider, media type, language, and year) is available on the screen where the search results are returned. The Finder includes resources coming from two different collections, Europeana and RISM[8]). The metadata was stored in the existing ARIADNE infrastructure and provided the user with a single search interface.

For the ARIADNE team, the integration of the RISM collection was a great challenge during the second year deployment. The data covered by the RISM collection (mainly scores) are

[8] http://www.rism.info

heterogeneous and quite different from the ones provided by Europeana. To allow the integration with the Finder and the visualisation of the results in a uniform way, transformation of the metadata to an ARIADNE internal format was required. A second obstacle of the RISM integration was related to the RISM project itself. The project had a very different time schedule than Europeana Cloud, and was still working on the process of providing open access data and linked data. This made the synchronisation of our efforts at many times difficult. During the development, quite often linking to the actual resource was rather complicated.

As in the first year, the Finder was used as the 'baseline' tool for the integration of the additional tools from WP3. Both the Activity Stream and the TimeMapper are integrated in the Finder to see the past activities (i.e. searches) and to visualise the results. When viewing a search result, the connection to Music21[9] is also available.



(a)

[9] http://web.mit.edu/music21/doc/about/index.html

(b)

(c)

Figure 2: Screenshots from the Ariadne Finder main page (a), listing results (b), view item (c)

In Figure 2 above, there are different screenshots from the Finder. Figure 2-a shows the main page of the Finder with the Activity Stream at the bottom. In Figure 2-b the listing of the results after a search is shown, with the facets that can be used. Figure 2-c shows how a specific result can be seen.

3. The TimeMapper

Europeana provides a variety of metadata for its resources. These might include images, geo-coordinates and time information. TimeMapper[10] is a data visualisation tool that can use this metadata to create timelines and timemaps using Google spreadsheets.

It can be used in combination with other tools, such as the ARIADNE Finder, in order to sift through through large amounts of records for specific searches. For example, the search for "Kant" returns 158 results.[11] We integrated TimeMapper in our tool chain to provide an interactive geo-spatial visualisation of these bibliographic metadata. This enables users to quickly navigate the metadata and to order resources on the basis of time and place of publication. By doing this they can easily identify various (types of) resources worth studying.

Figure 3 shows the TimeMapper when drilling down into resources that match the keyword "Kant"[12].

TimeMapper was adjusted for Europeana Cloud to use the JSON format defined by the ARIADNE Finder. It will furthermore be adjusted to meet the specific needs that were pointed out by our user group in a recent study (cf. D3.3), such as visualising different people and comparing their work based on geographical and temporal metadata.

TimeMapper is available under the MIT licence[13]. The tool can be accessed via the ARIADNE Finder's button labelled "Visualise data on map" for the Axiom philosophers' version and via the "View in TimeMapper" button for the musicologists' version.

For the second year deployment, more metadata was included into the TimeMapper. This allows documents by several authors to be displayed on a single timemap[14] (cf. Figure 4).

[10] http://timemapper.okfnlabs.org

[11] http://greenlearningnetwork.com/axiom/listing.html?query=kant#

[12] http://ecloud.okfn.de/timemapper/index.php?search=kant

[13] http://opensource.org/licenses/MIT

[14] http://timemapper.okfnlabs.org/anon/l2s4kd-early-music

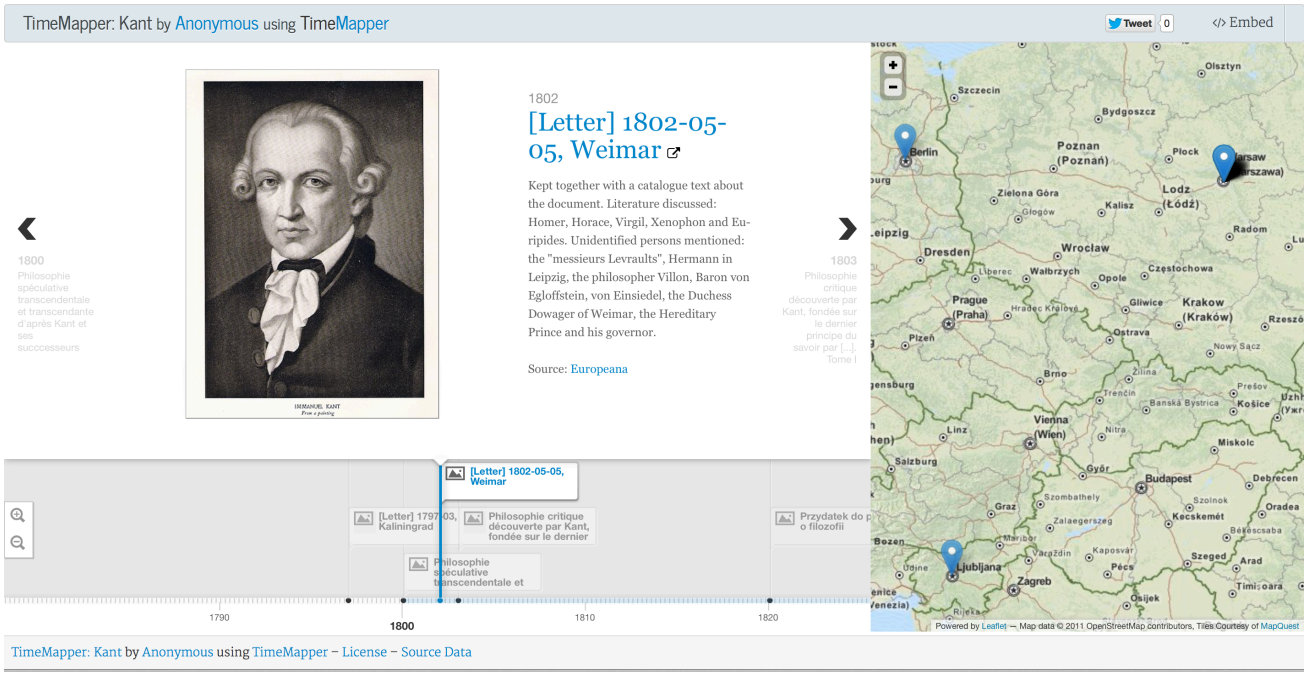


Figure 3: Screenshot of the TimeMapper showing resources matching the search term "Kant" in the ARIADNE Finder

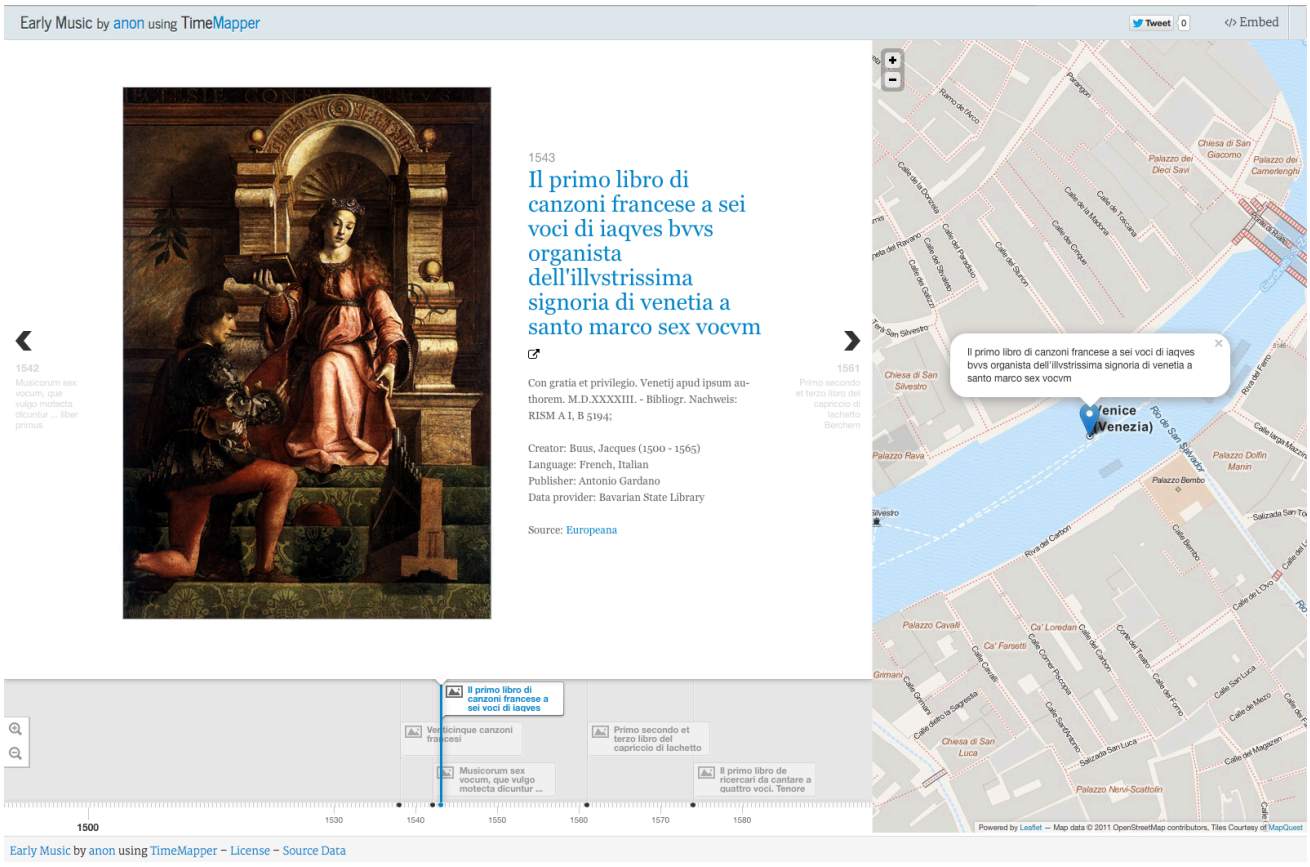


Figure 4: Screenshot of the TimeMapper showing resources that are published by Gardano based on the search results in the ARIADNE Finder

4. Activity Stream

Based on the community reading awareness provided by TiNYARM [1] and supporting the Science 2.0 idea of enhancing collaboration among researchers [2], we have deployed a web application called the "Activity Stream (AS)"; enabling researchers to share their work related activities with a community. Specifically, the application aims to aggregate "search" and "visualise" activities, and make researchers aware of what their peers are currently working on.

In the first prototype, the AS presents "searches" that have been performed using the ARIADNE Finder and terms that have been "visualised" using the TimeMapper, as seen in Figure 5.

The activities in the stream are structured as: *Actor* | *verb* | *(Object)*. For example, *User from GR* | *has searched* | *Bolzano*. For the second year, two activities were added to the activity stream: interpretation and processing. These represent the usage of the Aruspix[15] and Music21 components.

Figure 5: The Activity Stream main screen

The Activity Stream[16] is designed as a web application (using HTML and JavaScript) and deployed using the Google App Engine (GAE). Together with the terms used to perform a search or visualisation, a link to the tool showing the outcome of that action is provided. Also, to be able to provide users the flexibility to filter activities, the feature of tool grouping was added to the application. For instance, by clicking on the tool's name (e.g.: Finder or TimeMapper) you can get the stream of activities from that tool only.

[15] http://www.aruspix.net/index.html

[16] The Activity Stream for the Axiom philosophers group can be accessed at http://as-ecloud.appspot.com/ , and the Activity Stream of the second year can be found at http://as-ecm.appspot.com/

The Activity Stream allows us to digest different events sent from different tools (via REST services) used by researchers. It also provides the possibility to embed these on other software components. For example, the application supports RSS syndication as a passive form of notification system. Figure 6 illustrates the current activity sources and outlets.

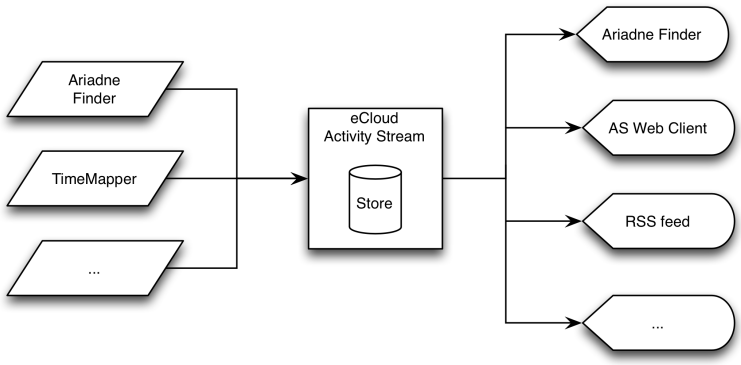


Figure 6: Information sources and outlets of the Activity Stream

5. Aruspix

Aruspix[17] is an optical music recognition (OMR) tool that scans early music prints, transcribes them and encodes them into the MEI standard[18].

While there are other OMR tools for new music available, Aruspix is the only tool to our knowledge being able to handle scores printed in the 16th and 17th centuries with movable typefaces. Such scores are often difficult to examine with existing superimposition and optical recognition software, as they present a number of specific layout and format problems and are quite often in a deteriorated state because of their age.

The printing techniques of that time mean that differences can exist between copies produced in the same print run, and comparison of these copies by superimposition can enable more accurate critical editions to be prepared. Digitising the scores through optical recognition can enable us to collate different editions regardless of layout, and is also useful in the preparation of digital music libraries, for example.

Aruspix is available under the GNU Public Licence.

Figure 7 shows the desktop application for Mac OS which allows for an interactive score transcription.

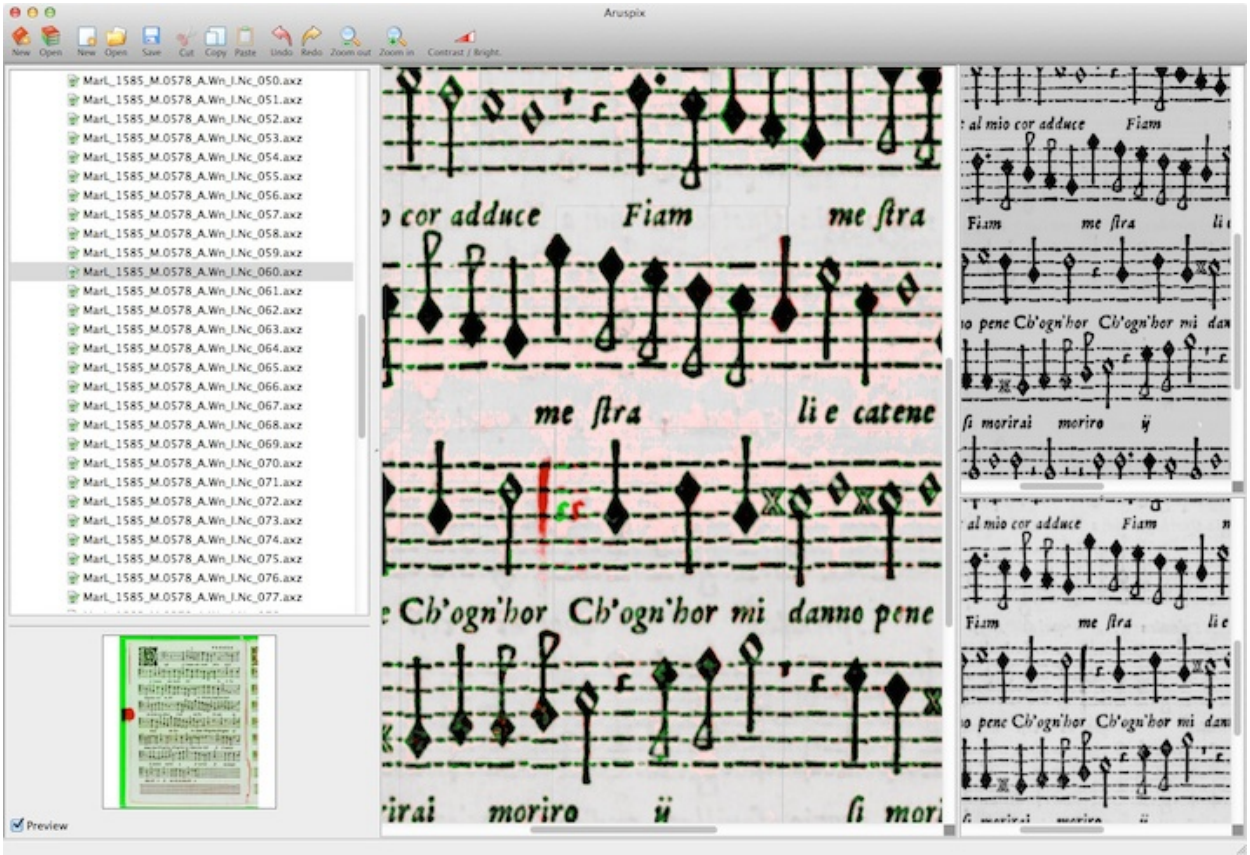


Figure 7: Aruspix desktop application

[17] http://www.aruspix.net

[18] http://www.music-encoding.org/exist/apps/mei/home

For Europeana Cloud, we use the command line version that automatically converts scores to MEI files in a page-wise fashion. We then need to combine the pages into a single score again. Moreover, the MEI version being used by Aruspix is a new and not yet standardised one.

Since Music21 (see next section) needs MEI as of the 2012 or 2013 specification, we wrote an XSLT program to transform the MEI files.

The command line version sends requested score transcriptions to the Music21 service for further analysis. Furthermore, it sends activity on transcribed scores to the Activity Stream.

6. Music21

Music21[19] is an open-source Python object-oriented toolkit for computer-aided musicology, developed at MIT, that allows music information, extraction and generation, together with music notation editing and scripting in symbolic (score-based) forms. The toolkit is able to import different formats, such as the MusicXML and Music Encoding Initiative (MEI) standards.

In the Year 2 of the Europeana Cloud project, we extended the Music21 web application module in order to provide parsing and processing requests to a Music21 installation running on a server. In the workflow, Music21 is used after the Aruspix service has created an MEI version of a score. With an MEI file, a specific set of actions becomes available to the musicologists: calculation of 'Parts and Measures', calculation of the 'Pitch ranges' and requesting the 'legal melodic intervals 'of a score. The main screen of the Music21 web interface is shown in Figure 8.

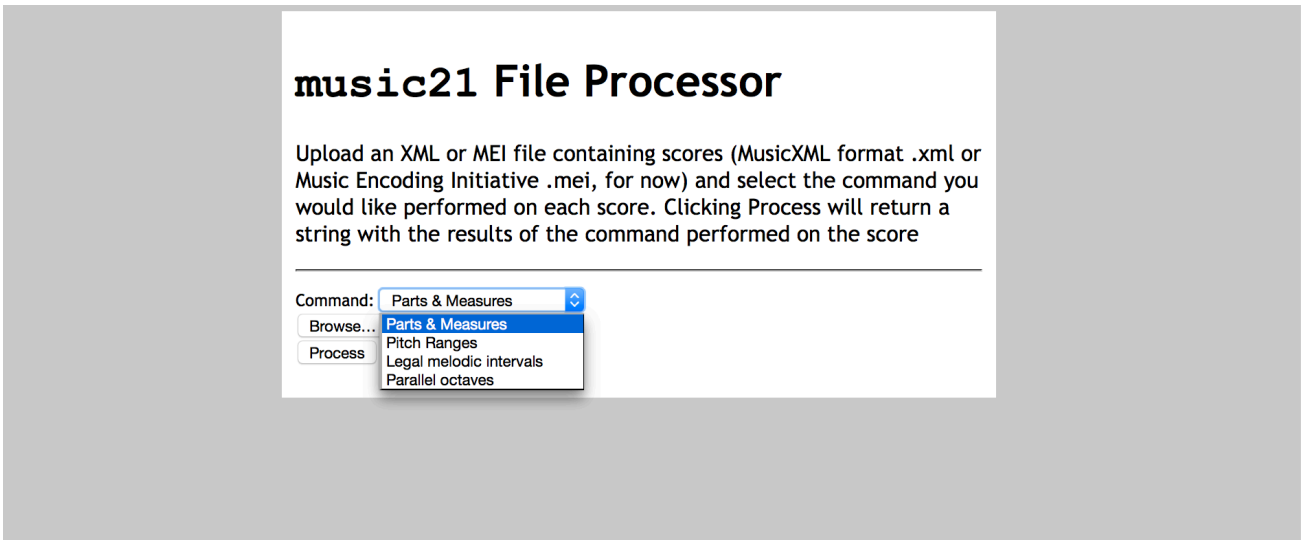
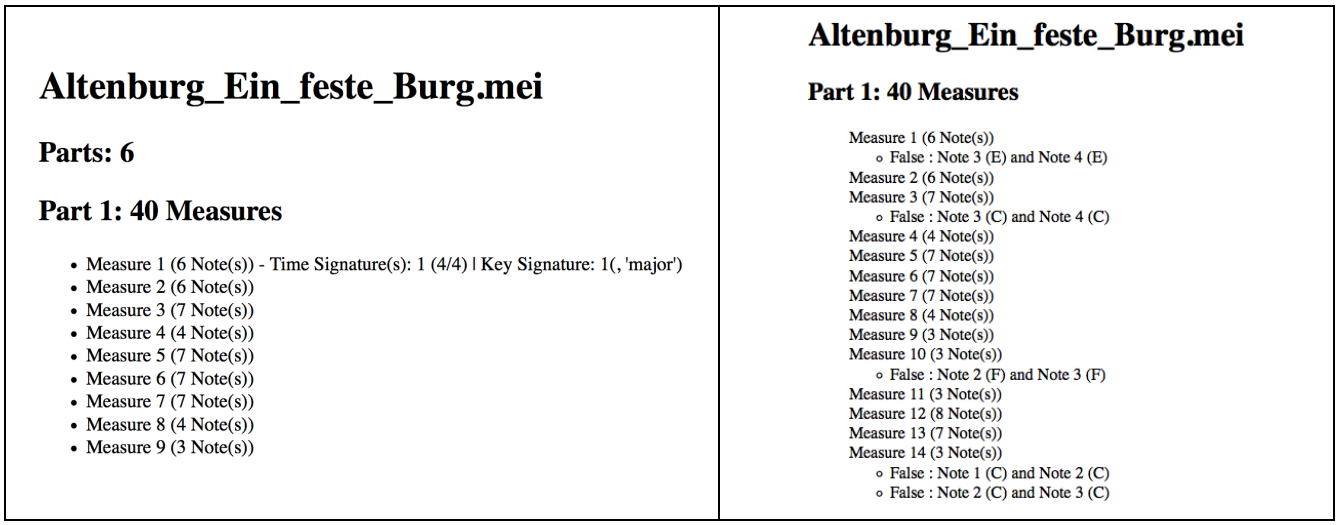


Figure 8: The Music21 web application

Figure 9 presents the different results obtained after processing the encoding of 'Altenburg - Ein feste Burg' encoding: parts, measures and notes and the legal melodic intervals of the measures inside the MEI.

[19] http://web.mit.edu/music21/



Altenburg_Ein_feste_Burg.mei

Parts: 6

Part 1: 40 Measures

- Measure 1 (6 Note(s)) - Time Signature(s): 1 (4/4) | Key Signature: 1(, 'major')
- Measure 2 (6 Note(s))
- Measure 3 (7 Note(s))
- Measure 4 (4 Note(s))
- Measure 5 (7 Note(s))
- Measure 6 (7 Note(s))
- Measure 7 (7 Note(s))
- Measure 8 (4 Note(s))
- Measure 9 (3 Note(s))

Altenburg_Ein_feste_Burg.mei

Part 1: 40 Measures

Measure 1 (6 Note(s))
- False : Note 3 (E) and Note 4 (E)

Measure 2 (6 Note(s))
Measure 3 (7 Note(s))
- False : Note 3 (C) and Note 4 (C)

Measure 4 (4 Note(s))
Measure 5 (7 Note(s))
Measure 6 (7 Note(s))
Measure 7 (7 Note(s))
Measure 8 (4 Note(s))
Measure 9 (3 Note(s))
Measure 10 (3 Note(s))
- False : Note 2 (F) and Note 3 (F)

Measure 11 (3 Note(s))
Measure 12 (8 Note(s))
Measure 13 (7 Note(s))
Measure 14 (3 Note(s))
- False : Note 1 (C) and Note 2 (C)
- False : Note 2 (C) and Note 3 (C)

Figure 9: The Music21 processed results

7. Newspaper Exploration Environment

The Newspaper Exploration Environment (NEE)[2021] is an interactive data visualisation tool to facilitate exploration of digitised newspapers through faceted search across coordinated multiple views (CMV) [6] and recommendations, creating an environment that supports both serendipitous discovery and targeted search.

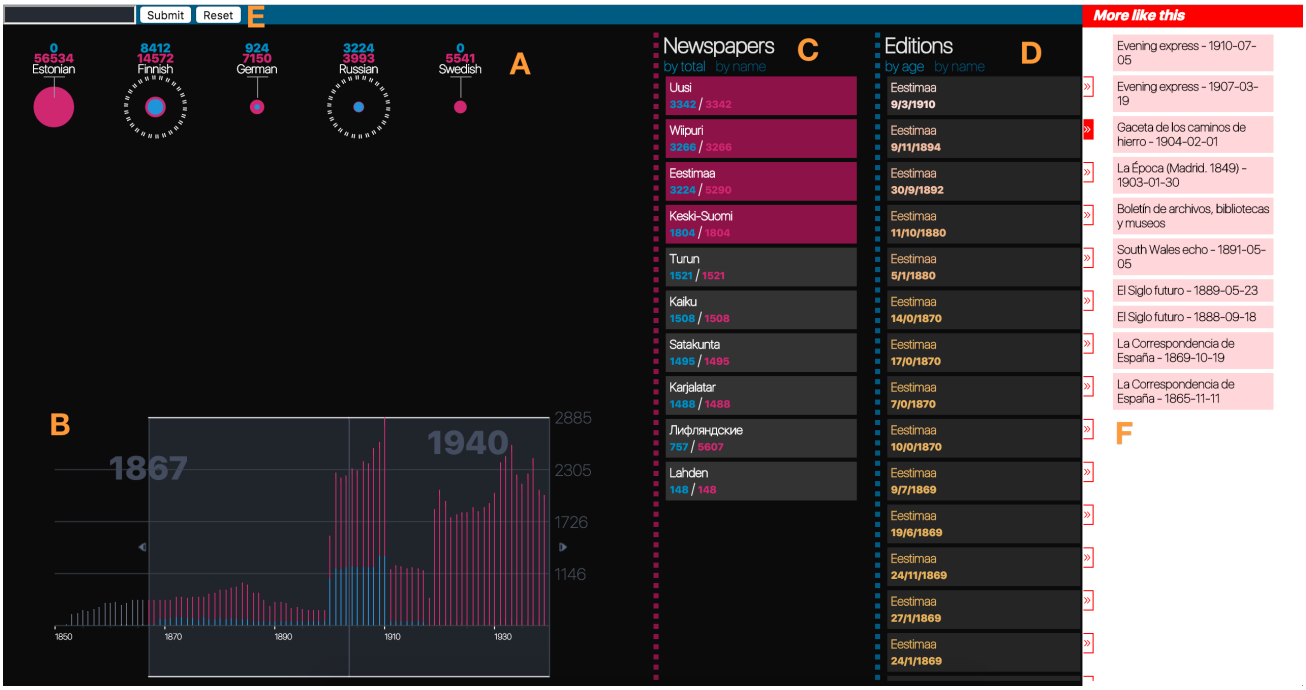


Figure 10: Newspaper Exploration Environment

Following the visual information-seeking mantra of Shneiderman [8]: "overview first, zoom and filter, then details-on-demand", researchers commence by taking a broad view of the data to orient themselves into the current dataset. This can then be followed by more specific investigations in parts of the datasets by using four filters (see Figure 10): a list of languages (A), a time-line widget to select years of publications (B), a list of newspaper titles and their number of issues (C), and the resulting list of issues (D). All widgets operate as coordinated views, i.e. a selection of a country will overlay the time-line with data from that selected language. Each visualisation widget follows the direct touch interaction approach to modify the selections.

The user can also search through the search input field (Figure 10 E). This can be either a word, or multiple words for an n-gram result. The column "More like this" (Figure 10 F) is populated through a custom recommender system: the red arrow next to a newspaper issue will load similar items which might be useful for the user.

[20] Link to NEE: http://daddi.cs.kuleuven.be/ecloud/login

[21] Feature overview video of NEE: https://www.youtube.com/watch?v=L0RrD1T-xmQ

Tapping/clicking a newspaper issue will open the Europeana Newspaper Archive website[22], providing direct access to the scans and full OCR text of the selected issue.

The frontend is developed using the web technologies Processing.js[23], JavaScript, HTML and CSS. To extend beyond single user desktop usage, methods were explored to support multi-user scenarios. With the support of TUIO.js[24], NEE has been deployed on multi-touch devices such as interactive tabletops (see Figure 11).

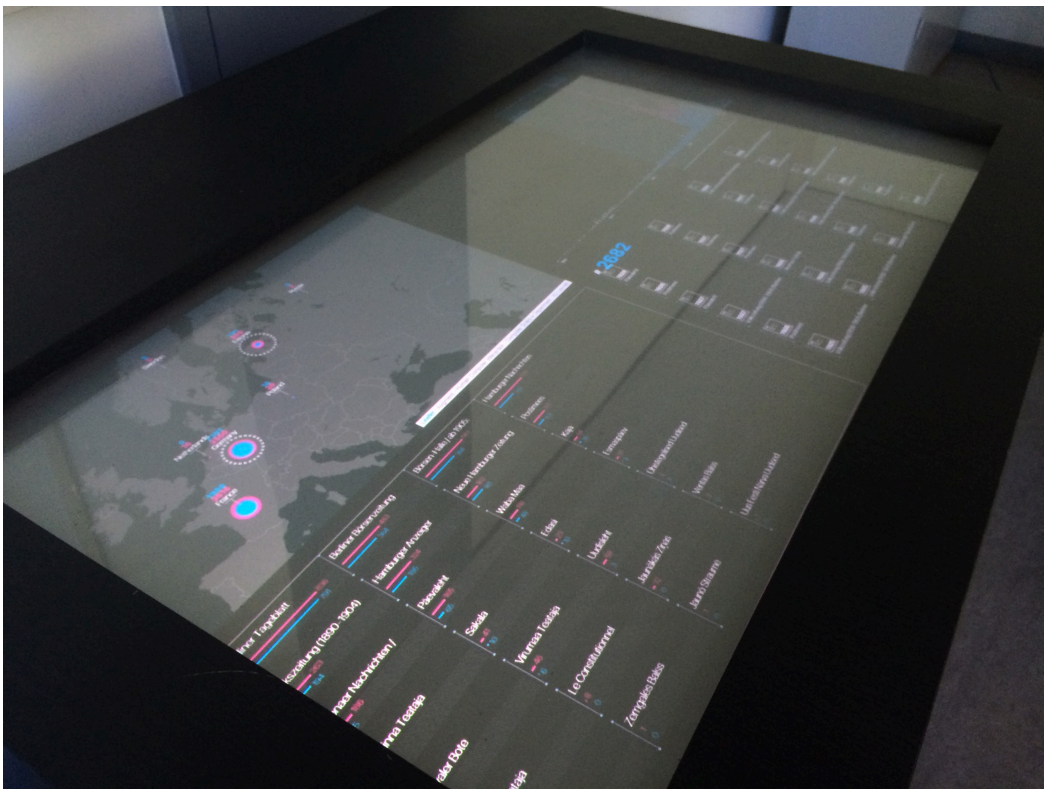


Figure 11: NEE running on an interactive tabletop for multiple users

Large interactive tabletops can already be found in public spaces such as museums, but are rare in the average research lab. As tablets become ubiquitous, combining multiple tablets to create a large faceted search interface becomes a feasible scenario [7] that could support a multi-user scenario similar to an interactive tabletop. A backend has been developed using Node.js[25] and Socket.IO[26] to enable per-device CMVs: each facet is visualised on a separate device, but creates a single CMV environment across multiple devices through live communication across web sockets[27].

[22] http://daddi.cs.kuleuven.be/ecloud/login

[23] http://processingjs.org/

[24] http://protium-labs.co/Tuio.js/

[25] https://nodejs.org/en/

[26] http://socket.io/

[27] Video of NEE's faceted search across multiple devices: https://www.youtube.com/watch?v=SO1cc6zXquU

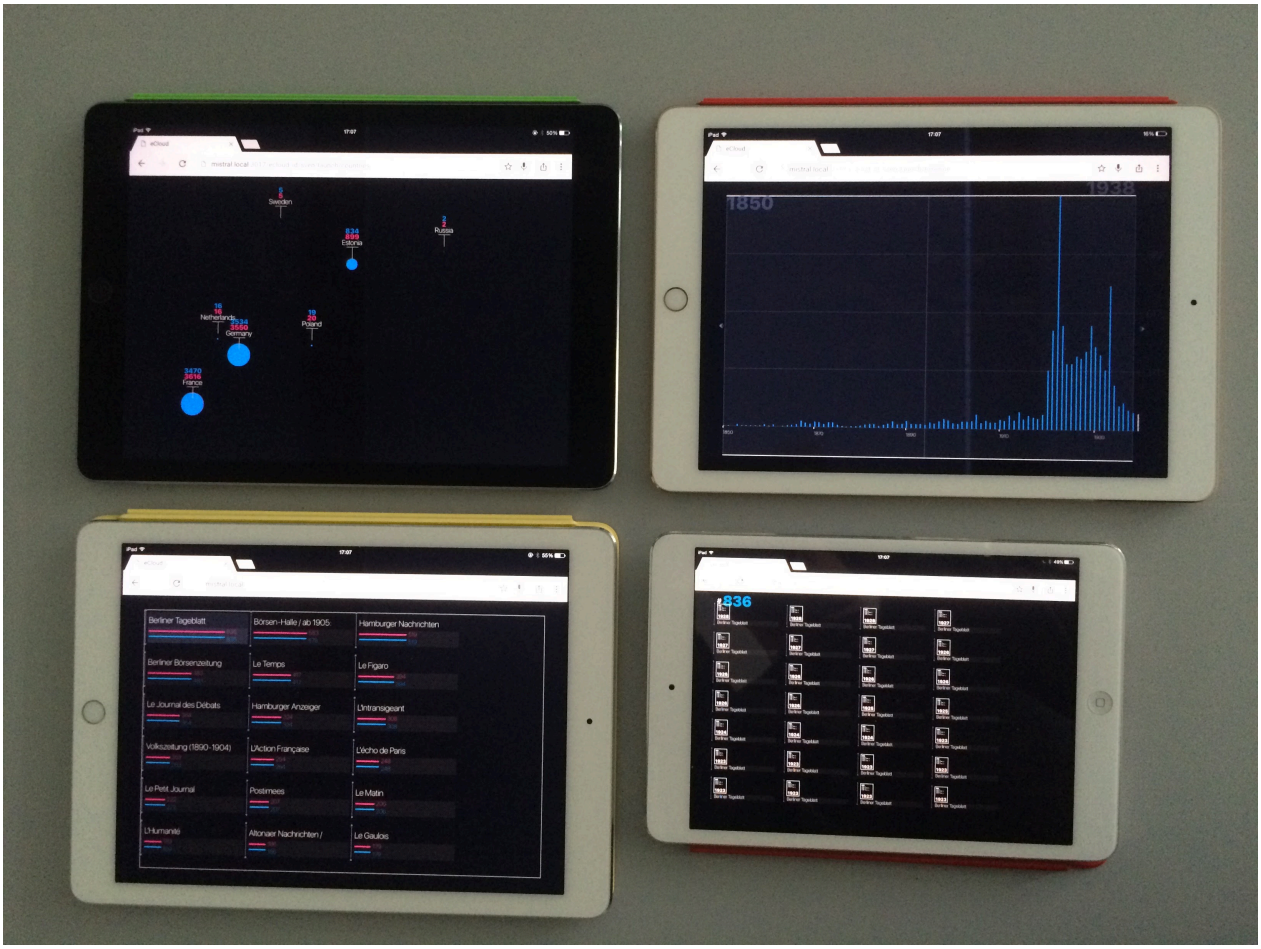


Figure 11: NEE running on multiple tablets while still providing a connected exploration environment

The Europeana Newspaper Archive API allows for access to 10 million digitised pages. Development of the API and NEE happened simultaneously, and the first version of NEE connected directly to the API. Afterwards, a data dump was created of a subset of digitised newspapers. This allowed the development and evaluations of NEE to continue independently from the API's progress. Requirements and limitations discovered in the API would not be blocking NEE's development, and could be reported to the API developers.

These findings resulted in the following changes:

- a new structure for the meta-data of the newspaper issues
- JSON for meta-data instead of XML
- on-going development of logical operators for combinations of different facets (e.g. (German OR French) AND (La Siècle OR Volkszeitung)
- date range query

8. eCloudDM – Data Mining the Newspaper Archive

The eCloudDM tool[28] is a data mining tool for the Europeana Newspaper Archive. It includes at the time of writing two libraries for named entity recognition and topic tagging. While this is already helpful to annotate and search newspaper articles from the Europeana Newspaper Archive, it can easily be used for any texts from Europeana.

eCloudDM is a Play application with a web interface to the articles from the Europeana Newspaper Archive and topics and entities mined from them.

Named Entity Recognition

In order to find relevant articles based on entities mentioned in them, we need to apply named entity recognition (NER) first.

For NER in our eCloudDM tool, we employ DBpedia Spotlight, which is an open source project developing a system for automatically annotating natural language text with entities and concepts from the DBpedia knowledge base. The input of the process is a portion of natural language text, and the output is a set of annotations associating entity or concept identifiers (DBpedia URIs) to particular positions in the input text.

The annotations generated by DBpedia Spotlight may refer to any of 3.77 million things in DBpedia, the Linked Data version of Wikipedia, out of which 2.35 million are classified according to a cross-domain ontology with 360 classes.

eCloudDM can determine named entities from texts in English, German, Dutch, French, Italian, Russian, Spanish, Turkish, Portuguese, and Hungarian.

eCloudDM offers a web interface to the Europeana Newspaper Archive and the named entities derived from newspaper editions that have OCR text versions available.

Figure 13 shows the March 10th 1944 edition of *Le Journal des Débats politiques et littéraires* and the named entities found on page 1. Entity counts are provided along with links to the corresponding DBpedia entities.

[28] http://ecloud.okfn.de/eclouddm/

Figure 13: Screenshot of eCloudDM: named entities for page 1 of a French newspaper from March 1944

Figure 14 shows DBpedia page for Pietro Badoglio, mentioned 3 times on the front page of the selected edition of *Le Journal des Débats politiques et littéraires*.

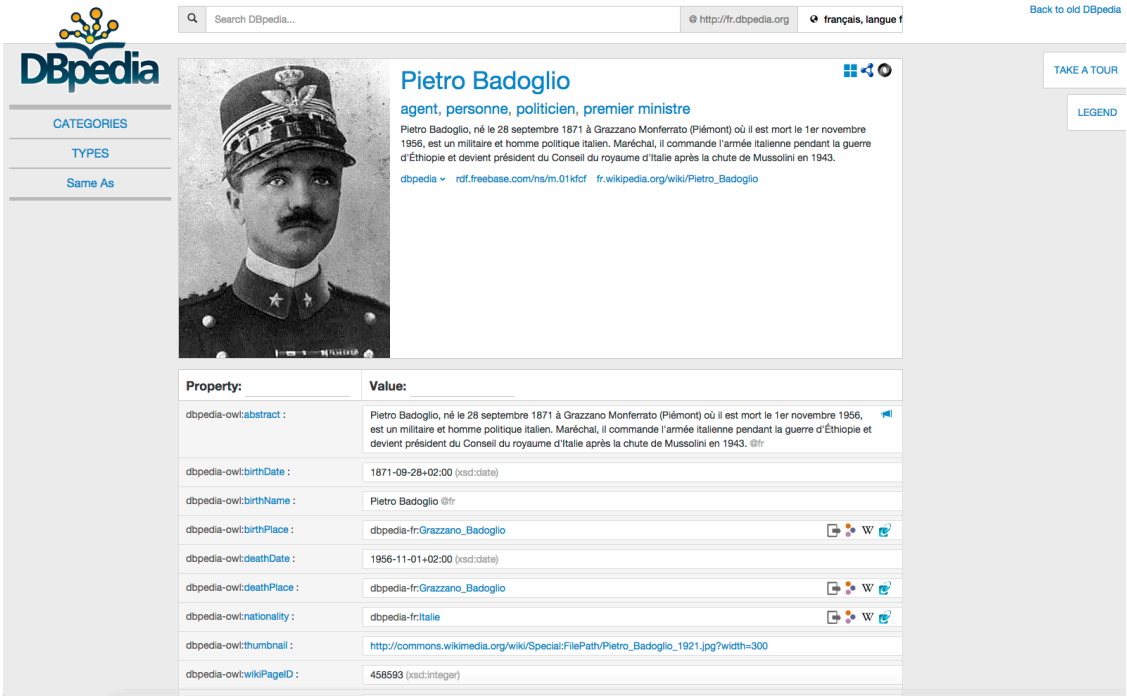


Figure 14: Screenshot of DBpedia view of Linked Data for Pietro Badaglio

Once named entities found in the newspaper articles can be stored to Europeana again, the newspapers can be easily searched via the eCloudDM website for certain entities.

Topic Tagging

Rather than finding documents through keyword search alone, we might first find the topic that we are interested in, and then examine the newspapers or texts related to that topic.

For example, consider using topics to explore the complete history of the New York Times. At a broad level some of the themes might correspond to the sections of the newspaper - foreign policy, national affairs, sports. We could zoom in on a topic of interest, such as foreign policy, to reveal various aspects of it. We could then navigate through time to reveal how these specific topics have changed, tracking, for example, the changes in the conflict in the Middle East over the last fifty years. And, in all of this exploration, we are pointed to the original articles relevant to the topics.

While more and more texts are available in Europeana and its Newspaper Archive, we simply do not have the human power to read and study them to provide the kind of browsing experience described above. To this end, machine learning researchers have developed probabilistic topic modeling, a suite of algorithms that aim to discover and annotate large archives of documents with thematic information. Topic modeling algorithms are statistical methods that analyze the words of the original texts to discover the topics that run through them, how those topics are connected to each other, and how they change over time. Topic modeling algorithms do not require any prior annotations or labeling of the documents—the topics emerge from the analysis of the original texts. Topic modeling enables us to organise and summarize electronic archives at a scale that would be impossible by human annotation.

In eCloudDM we employ latent Dirichlet allocation (LDA), which is the most successful topic modeling approach. Initially developed for both text analysis and population genetics, LDA has since been extended and used in many applications from time series to image analysis.

What are topics? A LDA algorithm is not given topics, so it must infer them from raw text. It defines a topic as a distribution over words

eCloudDM's LDA groups newspaper articles by topics and tags the topic bags with common words that allow users to detect the overall topic, e.g. sports, space exploration, and agriculture.

Topic modelling in eCloudDM allows users to find related articles and newspapers he might not have been aware of.

Ideally users can also change the topic name for each of the topic bags.

Once topic tags found in the newspaper articles can be stored to Europeana again, the newspapers can be easily searched via the eCloudDM website for certain topics.

Metadata Generation

One of the major goals of the Europeana Cloud project was allowing users to contribute back to Europeana via the Europeana API. While we tested a plethora of tools for integration into the Europeana Research Platform, we also create various metadata that can be send back for use by others.

When running eCloudDM on newspaper articles, we e.g. extract various named entities from the DBpedia classes person and place.

eCloudDM already creates valid metadata that can be pushed to the Europeana API. Ideally this is initiated by the users of our tools that decide which metadata is relevant to the article or text at hand.

This metadata will allow other users to find articles and other texts in Europeana faster and allows for better faceted searches that are based on selected entities or topics.

9. The AGRERI Discovery Microsite

The AGRERI Discovery Microsite is a personalised microsite that can be used by researchers in the field of agricultural economics to search and discover relevant resources. The microsite searches predefined collections of datasets based on user input and presents the results in a uniform way. It comes as a microsite, built with lightweight web-technologies (HTML, CSS, HTTP, Javascript, AJAX) in order to be easily embedded in sites and web-applications, without the need to make changes for matching the existing technologies of the application. The first prototype of the Finder in WP3 is designed based on the needs of the AGRERI agricultural researchers.

The Microsite is a personalized tool in the following two ways: it is graphically designed to be smoothly integrated with the web site of the AGRERI team[29] and is built on top of collections that have been requested by stakeholders.

The main usage of the AGRERI Discovery Microsite is a faceted search interface that allows users to search and quickly filter the results. In addition, predefined categories that allow access over specific content (i.e. agricultural economics) are also available.

The prototype has been designed and developed with the constant feedback from the AGRERI team in order to better catch and cover their needs. During development, a number of discussions were organised between the group and WP3 in order to gather feedback concerning the collections to search, categories to use, and the facets that the stakeholders would like to use.

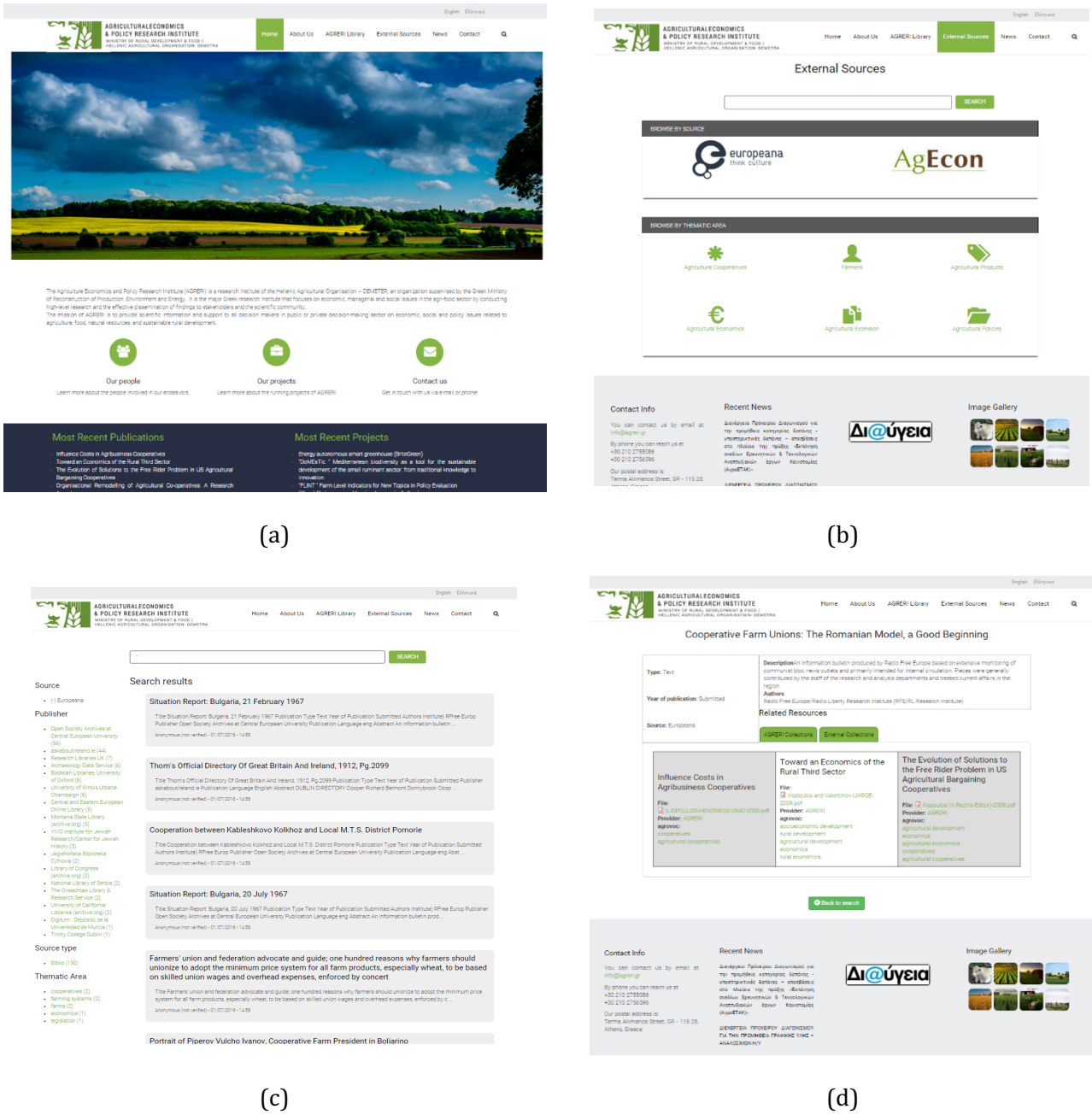
In order for the finder to allow faceted search and uniform representation of the metadata from resources coming from different collections, the Microsite uses the existing AGRERI infrastructure to store a repository with all the metadata. In the current development phase, the resources and different collections stored in the repository are the following:

- the AGRERI collection and collections from Europeana, the Europeana Newspaper Archive, the AgEcon repository[30] and FAO AGRIS[31].
- In all cases an API is used to filter thematically resources for the dataset.
- In order to provide this uniform representation and make the resources available throughout the Microsite a transformation process took place, where all metadata records were transformed from their original scheme to an internal format. During this transformation procedure metadata records have also been enriched.

[29] http://www.agreri.gr/

30 http://ageconsearch.umn.edu/

31 http://agris.fao.org



(a) (b)

(c) (d)

Figure 15: Screenshots from the AGRERI site (a), the AGRERI Discovery Microsite main page (b), listing results (c) and view item (d)

Figure 15 shows different screenshots from the AGRERI Discovery Microsite. In Figure 15-a the main page of the AGRERI site is shown. Figure 15-b shows the main page of the AGRERI Discovery Microsite. Figure 15-c shows the listing of the results after a search is executed, with the available facets. Finally, figure 15-d presents how a specific result can been viewed with the related resources from the different collections.

AGRERI Discovery Microsite architecture

The AGRERI Discovery Microsite is implemented using modules at three layers, namely a) the data Ingestion Layer, b) the repository layer and c) the front end layer (Figure 16).

The following modules and components were deployed in the context of the Europeana Cloud project in order to implement the AGRERI Discovery Microsite.

- **REST API client** that collects relevant content from Europeana. A custom client has been developed to collect the relevant Europeana content in a JSON based format through Europeana's REST API.
- **Data ingestion module** that imports the metadata in AGRIS AP format to the AGRERI site. This module can be used by the researchers to import new relevant content from Europeana. The data ingestion module includes the following components
    - **Transformation component** that transforms the data from Europeana format to AGRIS AP that is suitable for agricultural research
    - **Enrichment component** that enriches the Europeana metadata with information about the thematic category of the content
- **Data indexing module** that indexes the metadata so they are available in the AGRERI Discovery Microsite.
- **Front end modules** (AGRERI Discovery Microsite) site that allows the discovery of the content from Europeana and connects this content to the existing content for agricultural economics that the AGRERI institution holds.

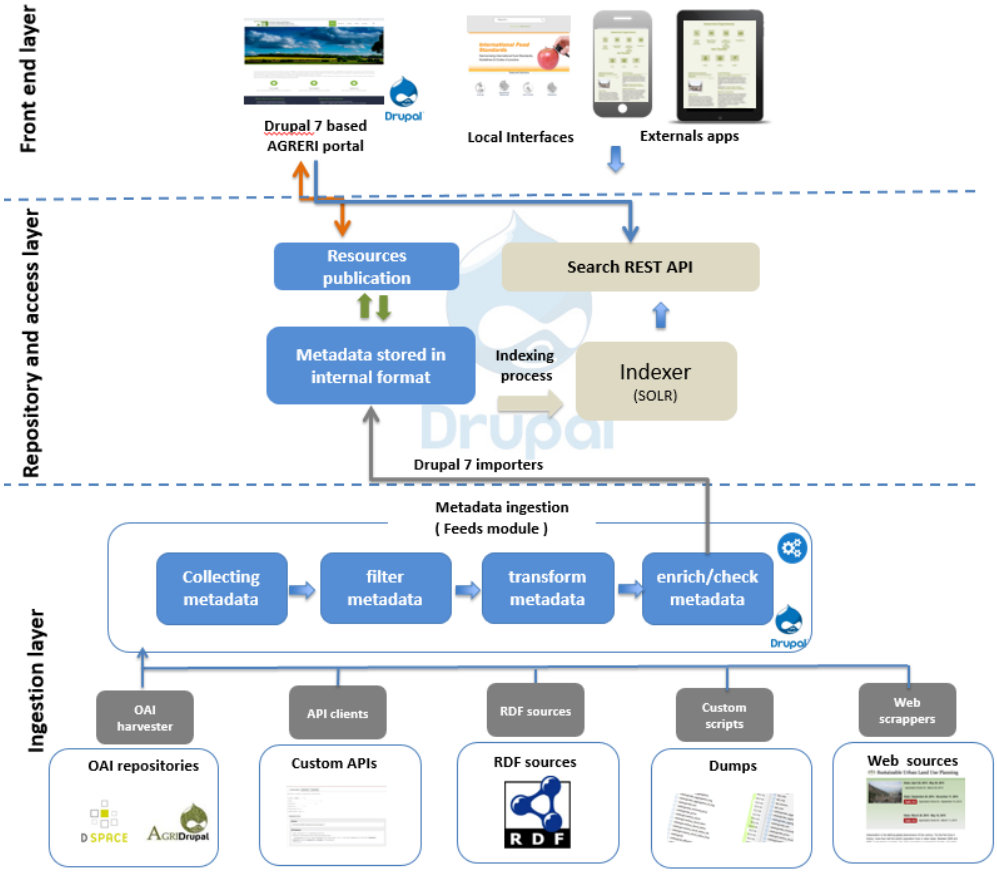


Figure 16. AGRERI Discovery Microsite architecture

Ingestion (Data import)

The AGRERI Discovery Microsite offers a data ingestion mechanism. The data ingestion mechanism is responsible for importing content and metadata records from multiple diverse sources that are publishing the metadata records through custom API (e.g. Europeana API), OAI-PMH protocol, RDF and dump files.

An importer can ingest content from raw files that are located to web apps directory or from a Web API through Rest Services and to be imported into AGRERI site. Also, before the data are stored in the database, the cleaning mechanism, if configured, removes unwanted content or curates it. For example, it can remove unwanted HTML characters. Another important functionality of the importer is the filtering mechanism that could be configured in order to filter out not relative content. For example, if a record is irrelevant to agriculture economics thematics, then this record will not be imported to the AGRERI Discovery Microsite.

In order to provide to the user all the resources of the AGRERI Discovery Microsite, the powerful Apache Solr 1.4 engine is used. Furthermore, the ingested content is indexed with Apache Solr and offered via this engine to the user search interface under a specific schema.

References

1. G. Parra, J. Klerkx, and E. Duval. Tinyarm: Awareness of research papers in a community of practice. In Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies, page 21. ACM, 2013.Shneiderman, B. 2008. Science 2.0. Science. Vol. 319, No. 5868, 1349-1350.
2. Cuthbert, M. S., Ariza, C. Music21: A Toolkit for Computer-Aided Musicology and Symbolic Music Data. In 11th International Society for Music Information Retrieval Conference (ISMIR 2010), August 9-13, 2010, Utrecht, Netherlands. pp. 637-642.
3. G. Haus and M. Longari. A multi-layered, timebased music description approach based on xml. Computer Music Journal, 29(1):70–85, 2005.
4. L. Pugin, J. A. Burgoyne, and I. Fujinaga. Goal-directed evaluation for the improvement of optical music recognition on early music prints. In Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, pages 303–304. ACM, 2007.
5. K. Makris, G. Skevakis, V. Kalokyri, P. Arapi, S. Christodoulakis, J. Stoitsis, N. Manolis, and S. L. Rojas. Federating natural history museums in natural europe. In Metadata and Semantics Research, pages 361–372. Springer, 2013.
6. Jonathan C Roberts. 2007. State of the art: Coordinated & multiple views in exploratory visualization. In Coordinated and Multiple Views in Exploratory Visualization, 2007. CMV'07. Fifth International Conference on. IEEE, 61–71.
7. Roman Rädle, Hans-Christian Jetter, Nicolai Marquardt, Harald Reiterer, and Yvonne Rogers. 2014. HuddleLamp: Spatially-aware mobile displays for ad-hoc around-the-table collaboration. In Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces. ACM, 45–54.
8. B. Shneiderman. 1996. The eyes have it: a task by data type taxonomy for information visualizations. In IEEE Symposium on Visual Languages. IEEE, 336–343.