



Europeana – Core Service Platform

DELIVERABLE

D3.2 Europeana Research distribution plan

Revision	1
Date of submission	1 September 2015
Author(s)	Alastair Dunning, Europeana Foundation
Dissemination Level	Public



Funded by
the Connecting Europe Programme
of the European Union

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision No.	Date	Author	Organisation	Description
1	31-8-2015	Alastair Dunning	Europeana Foundation	Final version

Statement of originality:

This milestone contains original unpublished work except where clearly indicated otherwise.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.

Executive summary

This document lays out the work to be done on behalf of the Europeana Research - Europeana's platform for connecting the data from the Europeana network with the academic community, particularly those working in the digital humanities. It outlines the long-term vision of Europeana Research and charts its specific work for the period of DSI1, until April 2016.

Introduction

Long-term vision of Europeana Research

For the scholar working in the digital realm, particularly those working in the digital humanities, the rich trove of digitised content from the cultural heritage sector offers an immensely valuable source. As scholars develop new methodologies for modelling, analysing and interpreting digitised data, corpora of digitised photographs, newspapers, oral histories, sound clips, movies, 3d reconstructions etc form valuable sources for putting these methodologies into practice.

Thanks to its privileged position within Europe's cultural heritage sector, Europeana can provide a significant role in connecting the academic community to the cultural heritage sector and the sources the latter can make available. In particular, Europeana can work with the digital humanities community, that is increasingly making use of large corpora of digitised datasets to inform new research work in various disciplines.

The essential aim of Europeana Research therefore is to increase the scholarly exposure to, and usage of, Europeana's open datasets.

There are various mechanisms by which this can happen:

1. By creating unique aggregations of content

Europeana's position as an aggregator from across Europe affords it a unique position. Europeana's potential to aggregate digitised content from numerous sources, and make it available as a unified, rights labelled dataset results in providing resources that no other European public organisation can offer.

Such aggregations provide several advantages - they provide scale (in that the content of several institutions is combined); they provide the opportunity for comparative analysis (as such corpora are drawn from several different states and cultures within Europe), and they work at a multi-lingual level. Additionally, they provide convenience. Researchers do not have to contact institutions individually, and to go through the complex process of harmonising data and the respective licences, but can work via one contact point.

Most importantly, such corpora provide a basis for new research questions to be asked. By aggregating content, Europeana can play a fundamental role in pushing the academic community towards new areas of discovery, investigation and insight.

2. By highlighting, promoting and disseminating collections valuable for academic work

While aggregated content provides unique sources for scholars, it is a time intensive task. It is impossible for Europeana to aggregate and provide corpora of data for all disciplines.

Nevertheless, Europeana can still fulfill an important role in highlighting and disseminating relevant collections (where the actual content is held at the original museum, library etc. rather than by Europeana) through the metadata it ingests. As these collections, partially stimulated by the Europeana Licensing Framework, become increasingly open and therefore available for re-use, there is more potential for their use within research.

Therefore, Europeana Research will provide simple resource discovery means for such collections, evaluating and describing the collections in terms familiar to researchers. Certain collections will also be highlighted, with their research potential analysed in greater depth.

The collections described in Europeana Research will also make use of the [NeDiMAH Methods Ontology](#) for describing scholarly practice in the digital humanities.

Finally, Europeana Research will award grants to research teams to work with content from Europeana (and its network) on specific research questions and tools. This could take various forms - a researcher-in-residence, small grants to researchers at specific universities, or small grants for rapid prototyping of research tools.

3) By disseminating Europeana Data via third parties

Given the multi-disciplinary nature of the humanities community, different intellectual disciplines have different tools and services that provide points of interaction. So rather than expect every user to visit the Europeana Research site, Europeana Research will actively work to embed its data into the tools of such third parties.

Particular attention will be paid to the related European research infrastructures, for example CLARIN (for the linguistics community) and DARIAH (for the digital humanities community), both of whom are part of this deliverable. But there exist other related projects and infrastructures they can utilise such data, for instance the European Holocaust Research Initiative, Pelagios and Cendari.

The exposure of Europeana data raises other issues, related to the measurement of impact, the branding and its citation, and they will also be explored.

4) By facilitating the links between the cultural heritage sector and the research community

Beyond the exchange and dissemination of data, Europeana Research can initiate and respond to strategic dialogue with other research infrastructures with overlapping aims. For example, the aggregation of data, licensing, research user requirements and appropriate content strategies are all topics that concern a variety of research infrastructures.

Other stakeholders can include content providers, publishers and new potential research projects. Each of them has an interest in how digitised data is being made available for research. Again, Europeana Research can play a role here, connecting with such groups, and advertising its areas of strength in the aggregation and licensing of content.

The long-term aim is to help Europeana refine its services for Research (as mentioned in the paragraphs above) and for others to have a more mature awareness of the position of Europeana within the research ecosystem.

Background to Europeana Research

Europeana Research was initially created under the Europeana Cloud project. Launched in April 2015, the website is at <http://research.europeana.eu>. Funds within Work Package 3, allocated to the Europeana Foundation, have been used to build the initial instance. Work Package 1 and Work Package 3 also use Europeana Research to disseminate information on their work.

Europeana Research in the 2015/6 Europeana DSI

Europeana has identified eight main goals for the initial year of Europeana Research. This covers the period from April 2015 to March 2016.

- **Exploring how full text and images from the Europeana Newspapers project can be made available for re-use**

The Europeana Newspapers project had made several million pages of text and images available in the public domain. We will be exploring how the content can be re-used by the digital humanities research community (and other relevant disciplines).

- **Building a full-text API based on Europeana Newspapers**

Based on the above newspapers content, Europeana Research will build a full text API and make it available for usage.

- **Pushing Europeana data in other research infrastructures**

Besides its newspaper collection, Europeana has around 45m metadata records related to cultural heritage. Europeana Research will explore how to share them with the research infrastructures.

- **Highlight other collections of interest to the research community**

Our sister service Europeana Labs makes similar collections available via an API. Europeana Research will highlight some of the most useful collections for re-use in research.

- **Continuing to gather user requirements for Europeana Research**

Europeana Research will continue to evolve. As part of the Europeana Cloud project, user requirements and feedback will be sought from research communities and infrastructures.

- **Developing a Content Strategy for Europeana Research**

As part of the Europeana Cloud project, a content strategy is being developed to guide the type of content that Europeana can make available to the research community.

- **Developing student modules for teaching usage of the Europeana data / API**

We will be working with the DARIAH network (Digital Research Infrastructure for the Arts and Humanities) to build teaching modules that explain the usage of the Europeana API
(*Note: This is pending following further discussions with DARIAH*)

- **Establishing an Editorial Board**

As part of Europeana Cloud, a small advisory board was established to guide Work Package 1's work on researcher needs. This will be expanded to a fully-fledged Editorial Board, that will oversee development of Europeana Research, in particular the grants programme.

Specific Tasks

1 . Placing Europeana data in CLARIN infrastructures

CLARIN will work with the Europeana Foundation to place Europeana data (from its newspapers corpus and its metadata set, and including data from The European Library) in the tools and services of the CLARIN infrastructure. This work will be undertaken by the German chapter of CLARIN (based at Berlin-Brandenburg Academy of Sciences and Humanities) and the CLARIN ERIC (base at University of Utrecht).

Task 1.1 - Incorporating Europeana Metadata

The following steps will be taken:

- Make an analysis of the metadata sets available through The European Library's OAI-PMH provider – resulting in [a first assessment](#) about the suitability of the collections for inclusion into CLARIN's Virtual Language Observatory (VLO).
- Run a test harvest and import into the VLO for a limited set of collections.
- Analyse the results in the VLO, ask feedback to CLARIN's VLO and metadata curation taskforce.
- Include the relevant metadata sets into the production instance of the VLO.
- Work on a CMDI (Component MetaData Infrastructure) profile for historical newspapers.

Task 1.2 - Incorporating Europeana Newspapers Corpus

- First manual inspection of [the provided samples](#).
- Describe a potential workflow to include this data in CLARIN's [federated content search](#).

The CLARIN federated content search (FCS) provides a way to search across different (sets of) corpora simultaneously. A central aggregator software provided by CLARIN integrates all available FCS endpoints via the SRU/CQL protocol sets. Internally, the FCS endpoints work as standardized interfaces to different and independent query engines at the individual sites. CLARIN offers all data providers to expose their endpoints via the FCS aggregator as long as they conform to the CLARIN SRU/CQL specifications. It translates common CQL queries into site-specific query languages, retrieves the full text search results from the local query engines and reports these results back to the central aggregator.

In order to technically integrate the Europeana data sets into the CLARIN FCS, the data sets have to be made available either via an existing FCS endpoint at one of the CLARIN centres or by Europeana providing its own endpoint(s) and local query engine(s). In order to make use of access right management as provided by CLARIN, external endpoints may want to additionally provide Shibboleth integration.

The integration of the historic newspaper data sets into the CLARIN FCS can be achieved at 2 levels:

1. At the full-text level: this does not require the data to be enriched (linguistically analyzed) and is mostly a matter of connecting the FCS endpoint library (Java API) provided by CLARIN to the local search system.
2. At a linguistically annotated level: each newspaper dataset would have to be fed through an analysis tool that splits up the texts into units (sentences, words, etc.) and adds information about these units (e. g. part of speech, lemma, modern lemma, etc.). This is far more labour-

intensive and is best performed at a place with the necessary linguistic expertise. The enriched collection can then be provided for search via the local centre's FCS endpoint.

As linguistic annotation depends on the accuracy of transcription and the adaptation of the linguistics tools to the text (text class, time period, etc.), the second option should be tested via a pilot. It remains to be seen if this can be done within the current project phase with limited time and resources. (Note: the most straightforward example from the provided sample, [Berliner Tageblatt](#), might already be available via the BBAW's [DTA](#).)

EF will explore the possibilities of embedding SRU/CQL protocols for its forthcoming Europeana Newspapers channel (due first half 2016). This will allow the text to be explored via FCS. Details of these investigations will be published so other holders of full-text in the Europeana network can learn of the potential.

Task 1.3 - CLARIN Communications

CLARIN will communicate the work undertaken by the following means

- An article on the clarin.eu and europeana blog that can be featured in the monthly Newsflash.
- The two sets of data (newspapers data, TEL metadata) can be featured as resources at [the CLARIN portal](#)
- Various announcements can be made to the national consortia via the National Coordinator Forum. They can then communicate it via their national channels.

Task 1.4 - Measuring Impact

CLARIN will evaluate the impact of the disseminations of the Europeana data in the following ways

- Measure the number of included metadata records in the VLO.
- Measure number of visits to the VLO section with the Europeana/TEL metadata.
- Defining a common data citation metric between CLARIN and Europeana. This will help measure impact in citation in the long term. This will include guidelines on how to cite:
 - a dataset (e.g. newspaper corpus) made available via Europeana
 - a Europeana dataset enriched and deposited at a CLARIN centre

2. Highlighting Europeana Collections

Task 2.1 - Surfacing Collections on Europeana Research

Currently, Europeana Research displays featured collections that highlight some of the data Europeana has aggregated, and its usefulness for academic research. These are curated via a manual process, whereby contributors to Europeana Research select relevant collections and add their own descriptions / reviews of the collections.

It is intended to increase the scale of this process so that a much larger number of collections is made available for researchers. In particular, it will focus on how Europeana collections can be segregated according to discipline (eg archaeology, modern history, literature) to allow for specific datasets to be advertised to different disciplines.

The Bolt CMS, currently used for Europeana Research, can be expanded to incorporate specific collections descriptions.

To achieve this, the following tasks will need to be undertaken

- Decide upon content policy for inclusion of collections (especially concerning openness of data)
- Decide upon metadata fields and thesauri for describing collections
- Add metadata to a sample of collections data
- Upload to Bolt
- Receive feedback from Europeana Research Advisory board
- Continue process of adding data on collections to Bolt

Additionally, much data in Europeana is not labelled as specific collections. Adding new data to Europeana Research would also involve developing new descriptions

Task 2.2 - Commissioning Featured Collections / Use Cases

Besides the standard collections, European Research will also feature commissioned blog posts and articles. Planned work to be commissioned includes

- 3 or 4 blog posts on User Generated Content for Research
- Digital Humanities in the GLAM sector; some case studies
- Use Cases of how academics have made use of Europeana and GLAM collections for teaching or research

These blog posts will be commissioned by AD and MS, with assistance from the members of Europeana Cloud Work Package 1.

Task 2.3 - Content Strategy Research

Within WP1 of Europeana Cloud, there is a specific task looking at potential content areas where Europeana could fruitfully aggregated data to as to provide a critical mass of data for reuse by researchers. Such aggregations of data may also be helpful in deciding upon future Europeana Channels.

This involves analysis existing content within the topic, assessing the tools and functionality available for analysing that topic and assessing the feasibility of Europeana focussing on that area in the future. Four or five content topics have already been chosen. Via the Europeana DSI a further topic will be included and explored by MS (from ONB).

The final document, to be delivered by WP1 of Europeana Cloud, will be available in January 2016.

Task 2.4 - Devise Editorial Board

An editorial board for Europeana Research will be devised in Q4 of 2015. This will be an expanded version of the current Europeana Research Advisory Board, developed as part of the Europeana Cloud project. The terms of reference for the board will be reviewed and updated. An

initial meeting will be held to help scope work of DSI2. Potential invitees for the Editorial Board are named at the end of this document.

3. Building and Exploiting a Newspapers API

Task 3.1 - Building the Newspapers API

Based on the aggregation of newspapers content developed in the Europeana Newspapers project, AF and RG will develop, document publish the full-text API. This is scheduled for the end of August 2015. AD will oversee and contribute to this task.

Task 3.2 - Publicising the Newspapers API

Based on the Europeana Research Communications Plan (developed as part of Europeana Cloud project), AD and NvS will publicise the Newspapers API (from August 2014 onwards) Until the newspapers are moved from the TEL to the Europeana Cloud servers (scheduled for early 2016) , the service will be limited to avoid too much stress on the TEL servers.

Task 3.3 - Exposing the Newspapers Dataset

AD and NF will explore various opportunities for exposing the newspapers dataset. The key task (to be completed by September 2014) is to make the full text available for reuse via the TEL servers, following feedback from potential re-users. This will also involve liaison with WP3 of Europeana Cloud, that is reusing the dataset in their tools development. Other opportunities (such as the EU DAT Call for Data Pilots) will also be explored.

4. Working with DARIAH

Task 4.1 - DARIAH Contribution to Europeana Research Editorial Board

DARIAH contribution to Editorial Board for Europeana Research; DARIAH will scope a nominee after a call for a volunteer from its broad network, VCCs and working groups. This method is to ensure that the appropriate expertise and commitment are secured.

Task 4.2 - Communicating Europeana to DH Community

DARIAH will focus on aiding in the broadcasting of the Europeana API and Europeana Research, that is, encouraging and supporting its wider usage and embedding within DARIAH structures and activities, direct and indirect. DARIAH will do this using multiple channels:

4.2.1 Via its centrally sponsored workshops and the summer school, as well as through the various digital humanities classes that are organised through the VCC2 activities, as well as, for example, by means of Erasmus+, including DARIATeach and The DARIAH Reference Curriculum (DARIAH-RC), whose aim is to “strengthen alliances and foster innovative teaching and learning practices among members of...DARIAH” (<http://dariahre.hypotheses.org/>)

4.2.2. This would be augmented by informing and re-engaging those who have contributed to the Digital Humanities Course Registry (<http://www.clariah.nl/en/dodh/course-registry>) about the availability and usefulness of Europeana Research, especially in its aspect of Task 2.5- Developing student modules for teaching usage of the Europeana data / API

4.2.3 Through the talks and presentations given by the DARIAH CEO, to audiences both internal and external to DARIAH.

4.2.4 News items and further information will be made available on the DARIAH web site and published in its regular newsletter.

4.2.5 By working with raising even greater awareness of Europeana Research amongst and via the DARIAH Working Groups, especially those who are looking involved in text analysis and semantic technologies, in terms of linking concepts and ideas with ontologies, as well as fostering new research methodologies via extending technical methods, and measuring the applied value of resources such as Europeana Research. Examples would include the Working Groups on:

- Lexical Resources
- Digital Annotation
- Community Engagement
- Digital Methods, Practices and Ontologies (DIMPO)
- Impact factors and Success Criteria
- Guidelines and Standards (GiST)

4.2.6 In relation to 4.2.1, 4.2.2 and 4.2.5 DARIAH will encourage and support focus on the areas of experimentation, usage and feedback on the Europeana API as part of the DH classes and events it runs both centrally, and promotes and sponsors via affiliated bodies, where appropriate, and on what student and teacher participants do with the Europeana API, and how they would want it to develop. Any such feedback will also be part of the short report specified under 4.4, where any experimentation could further the better connection of tools and datasets.

Task 4.3 - GLAMs and DH: Recommendations for providing access to content

This task will provide a report (to be delivered by March 2016) on how GLAMs and Europeana can better provide access of their content to researchers, exploring, technical (cf. EAD, TEI, EDM), hosting and licensing issues. The aim would be to provide a succinct but incisive document to enable greater understanding of the techniques and methods needed on the part of GLAMs to provide material that is 'Digital Humanities' ready. It will also explore how GLAMs, research support in universities and research centres, and academics could aid each other more effectively in the goal of opening up and utilizing such material in research projects, tools development, and so forth.

This would be guided alongside collaboration with others, starting with Europeana, but also as the OpenGlam network, LIBER (Ligue des Bibliothèques Européennes de Recherche – Association of European Research Libraries) and CERL (the Consortium of European Research Libraries), and would further the explorations of recent reports on opening up GLAM material to Digital Humanists.

Task 4.4 - Connecting DARIAH Services to Europeana Dataset

Relevant working groups from will explore tools from DARIAH services and their possible usage in the context of the Europeana dataset. This will includes tools from Cendari and EHRI; a short report (for March 2016) will suggest how the tools and datasets could be better connected.

In-kind contributions will be made alongside the contribution made to DARIAH-Europe by the Europeana DSI

Appendix 1

Potential Invitees for Europeana Research Editorial Board

Marianne Ping Huang, Vice Dean for Education, Faculty of Arts University of Aarhus
Toma Tasovac, Director, Center for Digital Humanities in Belgrade,
Dr Leif Isaksen, University of Southampton
Seamus Ross, Incoming Professor of Digital Humanities, University of Glasgow
Christophe Verbruggen, Director, Department of Digital Humanities, University of Ghent
Patrik Svensson, Director of HUMlab, Umeå University
Milena Žic-Fuchs, Croatian Academy of Sciences and Arts
Franciska De Jong, Executive Director, Clarin

Draft Terms of Reference for Europeana Research Editorial Board

- Developing and steering the ongoing Europeana Research strategy
- Devising, Executing and Promoting the Europeana Research Grant awards
- Acting as ambassadors for Europeana Research, forging and extending links with key stakeholders (policy makers, funding bodies, other infrastructures)
- Creating recommendations for future content to be ingested and tools to be developed by Europeana generally
- Steering and contributing to Europeana Research editorial work (descriptions of open collections, short essays)

Potential Infrastructures Working Group

(remit still requires better definition)

Daan Broeder, EU-DAT
Peter Solagna, European Grid Infrastructure
Dieter Van Uytvanck, CLARIN ERIC
tba, DARIAH