europeana

Europeana DSI 2– Access to Digital Resources of European Heritage

MILESTONE

MS2.2: Results and Impact of Sharing Europeana Data with CLARIN

Revision	10
Date of submission	31 August 2017
Author(s)	Twan Goosen, Dieter Van Uytvanck (CLARIN ERIC), Nuno Freire (INESC-ID)
Dissemination Level	Public

Co-financed by the European Union
Connecting Europe Facility

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision No.	Date	Author	Organisation	Description
1	2017-07-26	Twan Goosen	CLARIN ERIC	First partial draft
2	2017-07-27	Nuno Freire	INESC-ID	Commented version
3	2017-08-02	Twan Goosen	CLARIN ERIC	First complete draft
4	2017-08-14	Twan Goosen	CLARIN ERIC	Updated draft, for review before internal submisison
5a	2017-08-14	Nienke van Schaverbeke	EF	Review and comments
5b	2017-08-14	Dieter van Uytvanck	CLARIN ERIC	Proofreading
6	2017-08-14	Twan Goosen	CLARIN ERIC	Updated draft after pre-submission proofreading. Also added screenshots.
7	2017-08-14	Nuno Freire	INESC-ID	Proofreading
8	2017-08-15	Twan Goosen	CLARIN ERIC	Version for final review before submission
9	2017-08-15	Twan Goosen	CLARIN ERIC	Version submitted for internal review
10	2017-08-30	Twan Goosen	CLARIN ERIC	Final version, updated tables

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This milestone contains original unpublished work except where clearly indicated otherwise.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

The sole responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein.

‘Europeana DSI is co-financed by the European Union's Connecting Europe Facility’

Table of Contents

Introduction

One of the lines of action of Europeana is to facilitate research on the digitised content of Europe's galleries, libraries, archives and museums, especially for the social sciences and humanities (SSH). This work is conducted in the scope of Europeana Research, where issues affecting the research re-use of cultural heritage data and content (such as licensing, interoperability and access) are addressed. Europeana Research's distribution plan[1] describes a long-term vision, in which "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 SSH 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." The distribution plan enumerates a number of mechanisms towards reaching this objective:

1. creating unique aggregations of content;
2. highlighting, promoting and disseminating collections valuable for academic work;
3. disseminating Europeana Data via third parties;
4. facilitating the links between the cultural heritage sector and the research community.

The task that this document reports on focuses on the third mechanism in this list. More concretely it describes efforts that can be taken on the part of research infrastructures and specifically the method and results of such efforts carried out by CLARIN (Common Language Resources and Technology Infrastructure). On basis of this experience, we have formulated a set of recommendations for the future development and maintenance of Europeana's Digital Service Infrastructure as well as a number of future tasks and potential extensions that we expect to contribute to the further adoption of Europeana data within CLARIN and by the research community as a whole, which are also presented in this report.

Requirements for cultural heritage data in research infrastructures

Different areas of research obviously have different requirements when it comes to the nature of source data and the way in which it is disseminated. In the present context, we focus on the requirements that we know apply to CLARIN's target community - that is SSH and more specifically scholars using digital language related resources (mainly text and speech data) in conducting their research. However, we believe that many of these requirements can be generalised to the broader SSH research community and to some extent to research based on digitally available resources in general.

The core requirements for digital resources for research can be characterised by the "FAIR" principles: such data must be Findable, Accessible, Interoperable and Re-usable[2]. Applying these to the case of disseminating Europeana data through CLARIN's infrastructure, we can extend these with a set of concrete requirements, which are described in the following sections.

[1] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI/Deliverables/europeana-dsi-d3.2-europeana-research-distribution-plan.pdf
[2] https://www.force11.org/group/fairgroup/fairprinciples

Resource findability

Findability depends on identification and description of resources. CLARIN standardises on identification by means of persistent identifiers (PIDs)[3], although other types of URI based identifiers are also accepted in the infrastructure. Resource descriptions must be provided using Component Metadata (CMD)[4] or in a format that can be converted. CLARIN harvests metadata from its centres and external sources and indexes these to make them discoverable in a unified way through the Virtual Language Observatory (VLO)[5]. Although CMD is based on schema flexibility, the presence of certain properties is required, and the inclusion of several other salient properties is highly recommended. The most important, or ‘core’ metadata fields contributing to discoverability are reflected by the ‘facets’ to which values from aggregated metadata records are mapped in the VLO; in addition to title information and a textual resource description these include content language, resource type, file format and licence.

Requirements: persistent and unique resource identification; description by means of a supported metadata standard providing a good coverage of ‘core’ metadata properties.

Resource accessibility

Metadata should always be publicly available, and provide pointers to resource content, ideally by means of PIDs (see previous section). Resource content should be accessible in its ‘raw’ form, i.e. not (only) presented through a ‘viewer’ application, or other non-standard method of access. This is an important requirement for enabling machine processing of resources (see “Machine processability”).

To deal with protected resources and services, CLARIN set up a *Service Provider Federation* (SPF)[6], connected to various *Identity Federations*, which allows academics from many EU countries to authenticate using a single identity managed by their “home” organisation. Since SPF membership is tied to CLARIN membership, this approach cannot be applied to services and resources outside CLARIN. Therefore, *external* resources such as those coming from Europeana and its data providers, in practice can only be accepted if access to these does not require any form of authentication. Europeana’s content strategy[7] is aligned with this requirement.

Requirements: publicly accessible metadata and data; access to raw data.

Resource interoperability

CLARIN standardises on CMD for metadata. Other metadata standards are only supported in its infrastructure through conversion (a conversion facility for Dublin Core and OLAC formats is embedded in the infrastructure). In terms of resource data, CLARIN supports a much wider range of standards and in fact only *recommends* (rather than mandates) a set of standards to be used depending on the area of application[8]. However, in practice only a relatively small number of formats are broadly supported by tools available within CLARIN. Therefore, meaningful interoperability can only be achieved if data is provided in common, open formats such as ASCII or UTF-8 encoded plain text, PDF documents, JPEG images or MPEG encoded audio and video. Such resources must be referenced from the metadata in the method prescribed by the used standard (CMD in the case of CLARIN), as is the case for links to subordinate metadata records.

[3] See http://hdl.handle.net/11372/DOC-30
[4] https://www.clarin.eu/cmdi
[5] https://vlo.clarin.eu
[6] https://www.clarin.eu/content/service-provider-federation
[7] http://pro.europeana.eu/publication/content-strategy
[8] https://www.clarin.eu/content/standards-and-formats

Requirements: metadata provided in a supported format or converted to such a format; links to resources in a broadly supported data format.

Resource reusability

For data to be useful to scholars there must be no legal restrictions towards using the provided data for academic research purposes. There must be a clear and explicit (referenced) statement indicating the rights and conditions associated with the resource. CLARIN prefers the usage of licence URIs for this purpose, such as those available and commonly used for the Creative Commons licences (e.g. *https://creativecommons.org/licenses/by/4.0*).

Requirements: explicit statement of usage conditions that allow for reuse in a research context.

Additional requirements

Machine processability

Enabling secondary services and tools to process ‘raw’ data, either as a discrete resource or in batch, is an important requirement for CLARIN and its community. This requirement deviates from the needs of an ‘ordinary’ user that might be better served by easy-to-use embedded tools and viewers. For a large part, this requirement will be fulfilled if the provided resources adhere to the FAIR principles and the extensions to these detailed above. One additional requirement is the availability of technical metadata. Services processing resources and infrastructure components linking resources and suitable services generally benefit from, or in many cases even require *a priori* technical information about a data object such as media type and file size.

OAI-PMH

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[9] is a widely-adopted standard for the exchange of metadata records, on which many infrastructures including CLARIN base their metadata processing pipeline. Therefore, resource metadata must be made available by means of a reliable and well-performing OAI-PMH endpoint that disseminates all relevant information and links to the data.

Full text availability for text documents

Although different types of resources can be of interest to SSH researchers, including photographs, videos and scans of manuscripts, many scholars within CLARIN’s community are primarily interested in text documents with full text content available – either as plain text, embedded in PDF document or encoded in any other standardised way. Quite some tools currently provided by CLARIN centres operate on text content. Therefore, in practice we consider it a ‘soft’ requirement for text resources to be available with full text. Other factors, in particular relevance and potential for processing in other ways, may override this requirement, hence we consider it a soft requirement. In cases where no transcription of a digital source is available, the use of CLARIN tools (e.g. OCR or automated speech recognition) might be an opportunity to work towards the fulfilment of this requirement.

[9] http://www.openarchives.org/OAI/openarchivesprotocol.html

Implementation

A work plan[10] was created, outlining the tasks to be carried out in order to implement the integration of Europeana data into CLARIN's infrastructure. To summarise, these tasks comprise the selection of data sets relevant to CLARIN; inclusion of metadata obtained from Europeana into the VLO; description of a number of demonstration cases linking Europeana data to tools provided by CLARIN centres; adaptation of the involved CLARIN infrastructure components to an increased strain in terms of required storage, processing power and memory.

Data selection, harvesting and dissemination

At the time of writing, Europeana offers about 775 data sets that are available both over OAI-PMH and via its REST APIs[11]. The type of content, number of records and status (e.g. completed, undergoing updates or disabled) vary strongly among these. During DSI-1, CLARIN carried out an assessment of data sets available from The European Library (TEL), and selected a number of sets potentially relevant to CLARIN. Since (most) TEL metadata has been migrated to Europeana's provider in the meantime, this earlier selection could no longer be used 'as is', but was used as input for a more recent assessment based on the full list of available Europeana data sets (of which TEL sets constitute about half), based on the requirements laid out in this report (see "Requirements for cultural heritage data in research infrastructures"). The table below lists the data sets selected after assessment. Note that the actual selection harvested by CLARIN will be dynamic and will be re-evaluated regularly and therefore is subject to change.

DATA SET ID	CONTENT	RECORD COUNT
9200366_AG_EU_TEL_A0641_NEWSPAPERS_SLOVENIA	Newspapers (OCR)	47739
9200360_AG_EU_TEL_A0639_NEWSPAPERS_LUXEMBOURG	Newspapers (OCR)	64271
9200384_AG_EU_TEL_A0613_NEWSPAPERS_ONB	Newspapers (OCR)	629498
9200301_Ag_EU_TEL_a0611_Newspapers_Finland	Newspapers (OCR)	24207
92076_Ag_EU_TEL_a0497_DutchBooksOnline	Books (OCR)	9484
08804_Ag_EU_ETravel_DebBooks	Misc printed (OCR)	68
2021006_Ag_FI_NDL_ephemera_tb	Misc printed (OCR)	760
92068_Ag_Slovenia_ETravel	Misc printed (OCR)	697
92099_Ag_EU_TEL_a1080_Europeana_Regia_France	Manuscripts	167
2022402_Ag_RO_Elocal_arhivele	Manuscripts	194
2022411_Ag_RO_Elocal_audioinb	Speech recordings	30

CLARIN has carried out various experiments with harvesting data sets from Europeana's OAI-PMH provider during the course of DSI-2. CLARIN's OAI-PMH harvested could easily be configured to harvest Europeana metadata. However, technical issues were encountered various times during DSI-2. These are described in more detail in the section "Results". At times when the provider was available, it appeared to be capable of providing several millions of records in a single harvesting session, albeit at a relatively slow pace of about 1-2 million records per 24 hours (depending on the exact data sets included). Adaptations to CLARIN's harvesting software have been made that enable 'incremental' harvesting, which would allow for a substantially increase to the efficiency by which metadata can be harvested, as only changes (deltas) are transferred using this mechanism. However, it is currently not supported by Europeana's OAI-PMH endpoint. Therefore, CLARIN will set up a schedule to harvest metadata from Europeana

[10] https://www.clarin.eu/sites/default/files/EuropeanaDSI-2task2.6.3workplan.pdf

[11] This number refers to the data sets listed in both the "Provider and datasets" API and the OAI-PMH endpoint (via the 'ListSets' verb). At the time of writing, another ~500 additional data sets are listed in the API and ~1500 in the OAI-PMH endpoint – see http://labs.europeana.eu/api/provider and http://oai.europeana.eu/oaicat.

with a relatively high interval – initially each data set will not be harvested more than once per month. The performance will be monitored and as soon as it improves and/or incremental harvesting support has been implemented, the harvesting frequency could be increased for a higher degree of synchronisation between the two infrastructures.

Metadata conversion

While metadata of any format can be retrieved over OAI-PMH, only *CMDI records*, i.e. records adhering to CLARIN's implementation of the CMD model, can be imported into the VLO. To also support other common metadata schemas, CLARIN's OAI-PMH harvester has been equipped to perform conversions by means of XSLT stylesheets. Applying this mechanism to incorporate EDM metadata required the creation of a CMD profile for EDM in the CMDI Component Registry[12] and its corresponding implementation in XSLT[13]. The XSLT stylesheet allows the conversion of the RDF/XML EDM records available in Europeana's OAI-PMH provider, into CMDI records.

The conversion mechanism, described in detail in the README file bundled with the conversion stylesheet[14], is utilised to achieve a mapping of all information contained in EDM records according to Europeana's implementation of the model at the time of writing[15]. The conversion method can be summarised as a 'crosswalk' that imposes a partial hierarchy onto the originally flat (but linked) EDM data, which fits the hierarchical nature of CMDI metadata and thus provides a context for objects such as times, places and actors that can be interpreted by CLARIN services such as the VLO.

A number of unimplemented or unsupported features are omitted in the mapping. This is the case for support for additional EDM *profiles*, such as the Europeana Sounds Profile and the Technical Metadata Profile, although a number of properties defined in the latter are included in the mapping (specifically *ebucore:hasMimeType* which provides the media type of a web resource).

Some information that is not explicitly expressed in the original EDM content gets extrapolated in the conversion process. In particular, media type information currently is not disseminated through the metadata provided by Europeana via OAI-PMH[16]. Given the importance of such information for machine processing (see Requirements for cultural heritage data in research infrastructures), logic was added to infer the media type based on the provided resource URI for a limited set of relevant, often occurring and broadly supported groups of media types such as PDF and plain text documents.

Record hierarchies[17] are ignored and not represented in the converted metadata records. There is a potential of representing record hierarches present in Europeana's data structure as metadata hierarchies as supported in CMD, but this needs to be further investigated.

The XSLT stylesheet has also already been applied by Europeana's R&D team, in its ongoing data pilot with EUDAT[18], a research data infrastructure that has partial support for CMDI metadata.

[12] https://www.clarin.eu/componentregistry
[13] Available via https://github.com/clarin-eric/metadata-conversion
[14] https://github.com/clarin-eric/metadata-conversion/blob/master/edm-cmdi/README.md
[15] See https://github.com/europeana/corelib/wiki/EDMObjectTemplatesProviders
[16] For a detailed report on this, see https://office.clarin.eu/v/CE-2017-0996-CLARIN-Europeana-OAI-report_final.pdf
[17] See http://labs.europeana.eu/api/hierarchical-records
[18] https://eudat.eu/communities/enriching-europeana-newspapers

Aggregation and presentation

The converted metadata was included as input to the VLO's importing mechanism, which extracts, post-processes (i.e. sanitises and normalises) and indexes information before it is made available in the VLO's front end. Two times a week, the full set of available metadata (sourced from CLARIN centres and external sources including Europeana) is imported 'freshly' into the VLO to ensure the inclusion of all recently harvested metadata and the application of the latest version of the mapping and normalisation logic and definitions. Metadata content is mapped to the VLO's fields and facets based on the semantic annotations in the definition of the metadata profiles (so-called 'concept links'). Because existing concepts have been used in the definition of the CMDI/EDM profile, no adaptations to the VLO's mapping definitions were necessary in order to achieve a good mapping of the records sourced from Europeana. Licenses and rights statements are normalised into URIs and 'licence categories'; the definitions for this normalisation were extended to include the various rights statements occurring in Europeana records (i.e. the http://www.europeana.eu/rights and https://rightsstatements.org URIs). After applying several performance enhancements to the VLO's importing mechanism, which are described in the next section, it was capable of importing several millions of converted EDM records within a timespan of several hours. Once an import completes, all included records are immediately visible and searchable in the VLO's web front end. There, users can search for data based on all information provided in the metadata records, directly access linked resources and investigate processing options by means of the CLARIN Language Resource Switchboard (LRS)[19].

Infrastructure component adaptations

At the onset of DSI-2, CLARIN regularly harvested close to a million metadata records from CLARIN centres and other selected providers in order to present these and make them findable in the VLO. Harvesting sessions are scheduled to run several times per week and take between several hours and a full day, depending on which providers are harvested from[20]. Importing all harvested metadata into the VLO takes roughly two hours depending on the exact number of metadata records and their size and complexity. Incorporating Europeana data into this workflow entails a substantial increase in the amount of data that needs to be collected and processing; assuming that 1.5 million Europeana records will be included, this comes down to an increase of ~175% in terms of file count and, based on our experiments, a ~450% increase of data size measured in bytes, as (converted) Europeana metadata records are generally larger than the average metadata record currently processed by CLARIN. Therefore, we foresaw scalability issues and included measures to prepare the involved infrastructure components for the increased load on available storage, memory and processing power. The OAI-PMH harvester used by CLARIN has been adapted to support incremental harvesting as mentioned earlier in this section, anticipating support for such functionality from Europeana's OAI-PMH provider. The VLO's importer has been adapted, mainly through better exploitation of parallel processing capacities on CLARIN's servers, to yield better performance and now is capable of importing several millions of converted EDM records in a few hours rather than 24 hours or more. This allows us to schedule reasonably frequent harvests (once per month) and include the results in VLO import runs without having to make significant adaptations to the update frequency of other data sets.

[19] See CLARIN-PLUS deliverable D2.5 https://office.clarin.eu/v/CE-2016-0881-CLARINPLUS-D2_5.pdf
[20] The exact schedule can be found at https://www.clarin.eu/faq/what-update-schedule-metadata-vlo

Results

Harvesting and conversion

The CMD profile for EDM and the stylesheet were implemented in parallel over several iterations and the result has now been published as a 'toolkit' for converting EDM to CMDI. We have integrated the conversion stylesheet into CLARIN's OAI-PMH harvester and executed a harvest of 11 selected metadata sets representing various collections from Europeana, resulting in a total of 780 thousand successfully retrieved and converted, schema valid records. These records were subsequently imported into an instance of the VLO, which allowed us to inspect the result from an end-user's perspective as the import process entails the mapping of selected data categories to so called search facets as well as various post-processing steps.

Combined, a complete, successful harvest and import of the selected data sets takes roughly 48 hours with the current state of the software and infrastructure, both on Europeana's and CLARIN's end (including optimisations carried out during DSI-2, as described in "Infrastructure component adaptations"). The number of EDM records of potential interest to CLARIN's community is at least several times higher. We believe that improvements to Europeana's OAI-PMH provider could make such an expansion feasible.

As mentioned, technical issues with Europeana's OAI-PMH provider were encountered at various times during DSI-2. These have been reported to Europeana's technical staff and were resolved but then later resurfaced multiple times. Although we did manage to regularly harvest a large number of records, during several periods within DSI-2, the provider was unable to serve any records at all. We therefore conclude that the OAI-PMH provider currently is not sufficiently reliable. In this regard, it has to be noted that the OAI-PMH provider is not advertised by Europeana as being ready for production purposes[21]. In "Recommendations with respect to the technical infrastructure" we argue that high priority should be assigned to the development and maintenance of a stable provider.

Content evaluation

We evaluated the content as presented in the VLO after harvesting and importing based on the requirements for resources for SSH scholars in general and the CLARIN community in particular. **Findability** is good, with each document identified through a unique URI - although these are not in a standardised PID format. Descriptions are good, as all records carry values that can be mapped to the most *title* and *description* fields as well as search facets *language*, *collection*, *resource type*, *subject*, *country* and *organisation*. The quality of the metadata itself is generally good, but we frequently encountered incorrect content language information – in most cases this appeared due to the application of a 'default' language to an entire dataset that in fact was multilingual. Resource **accessibility** is reasonably good, as raw data can be accessed for all selected data sets, although in most cases there are no direct links to the data files from the metadata and in some cases getting to these files requires a rather elaborate sequence of actions. Raw data is provided in broadly supported file formats such as PDF and MP3, which, combined with CLARIN compatible metadata, leads to good **interoperability**. Europeana adopted RigthStatements.org, which is favourable in relation to **reusability.** In terms of the actual right statements actually applied, accordance with the requirements varies from very good for data sets that offer data under a *CC0* licence to problematic for resources that carry a "Copyright not evaluated" rights statement[22], which is common within newspaper collections. Other resources offered are under copyright, which potentially limits reusability within an academic

[21] http://labs.europeana.eu/api/oai-pmh-introduction: "Currently, the Europeana OAI-PMH Service is in beta"
[22] http://rightsstatements.org/page/CNE/1.0/

context. Of the additional requirements, the data sets score well on **full text availability** (where applicable, i.e. excluding for example audio resources). However, in many cases the full text content can only be accessed via a landing page and manually triggering a download option, sometimes requiring a rather long sequence of actions. This strongly reduces the suitability of such data for **machine processing**. Moreover, media type information is only available for those resources where it could be derived from the resource URI, and other technical metadata is not available at all.

Regarding *other data sets* provided by Europeana, i.e. those that have not been selected to be processed by CLARIN yet but are of potential interest, no detailed analysis has been carried out. However, on basis of our general evaluation we conclude that findability is good in general owing to the presence of unique identifiers and overall good metadata descriptions. Accessibility and reusability are often good, although for some data providers, raw data is hard to find or not accessible at all. Full text availability is generally not as good as it could be, and technical metadata often is not available at all, reducing the potential for machine processing.

Demonstration cases

A goal set within CLARIN's task in DSI-2 was to describe a number of cases that demonstrate a fully functional processing pipeline taking resources that have been provided to CLARIN by Europeana and are findable via the VLO as input to tools performing natural language processing tasks (or other tasks relevant to SSH scholars) that are also available within CLARIN. The Language Resource Switchboard (LRS) plays a central role in such demonstration cases as it bridges resource discovery in the VLO on the one hand and tool discovery and invocation on the other hand. Given a resource of a specific type and with a specific content language (for example a PDF file with German text content), a user of the VLO can invoke the LRS which then presents a list of connected tools that can process such a resource in a specified way. As resources of many types in many different languages are available through Europeana and different tools available within CLARIN support different types of resources and content languages, there is a large potential for interesting demonstration cases. The following table describes a number of resource – tool pairs that can currently be combined using the VLO and the LRS. See Appendix A for a set of screenshots that illustrate a typical processing workflow.

RESOURCE	TOOL	PROCESSING
"Extract uit het register der resolutien…"	Weblicht / Alpino	Dependency parsing (PDF, Dutch)
"USSR Intourist"	Voyant Tools[23]	Text analysis (PDF, multilingual)
"Zur See mit s/s Aranda"	Weblicht / German Named Entity	Named entity recognition (German)
"Description d'une nouvelle espèce de petite gazelle…"	Weblicht / TreeTagger	Part-of-speech tagging (French)

Technical issues

While implementing the integration of Europeana data into the CLARIN infrastructure, we encountered a number of technical issues. These were reported to and discussed with Europeana's technical team. For completeness, we provide a brief summary of the most salient issues we encountered.

[23] Voyant Tools (Sinclair, Stéfan and Geoffrey Rockwell, 2016) is not provided by a CLARIN centre but it is connected to the Language Resource Switchboard, and is capable of processing plain text files and PDF documents with embedded text.

- The OAI-PMH provider proved to be unstable at times. Large responses were sometimes delivered incompletely, as a result of which harvests could not be completed or not carried out at all.
- Harvesting large numbers of records OAI-PMH provider takes a fairly long time. Harvesting a million records took roughly 10 hours throughout our experiments. As we expect to harvest up to a few million records on a regular basis in the mid to long term, this leads to scheduling issues. Incremental harvesting, which would resolve these issues, is currently not supported by the OAI-PMH provider.
- Metadata retrieved using the OAI-PMH provider appeared to be out of sync with metadata retrieved from Europeana's REST APIs, which generally was more complete and up-to-date. Metadata that fits special profiles, such as technical metadata, is not present in the EDM representation returned by the OAI-PMH provider.

On basis of our experience, we have compiled a set of recommendations regarding the technical infrastructure that are mostly related to these issues. They are enumerated and detailed in the section "Recommendations with respect to the technical infrastructure".

Recommendations

The following recommendations for the Europeana Digital Service Infrastructure are based on CLARIN's experience with integrating Europeana data and services into its infrastructure and the results described above. The first set of recommendations relates to metadata and data content, while the second set describes suggested improvements to the technical infrastructure.

Recommendations with respect to metadata and data content

1. Provide technical metadata in all disseminated EDM records
   - The EDM profile for technical metadata[24] allows for the inclusion of several properties required for most machine processing purposes. Many data providers implement this profile, making such data available e.g. via Europeana's REST APIs. However, most records do not contain this information. Moreover, these properties are not included when metadata is retrieved over OAI-PMH.
2. Provide direct links to raw (machine processable) resources
   - Although many data sets provide access to high quality digital objects, very often these are not directly accessible in a machine processable ('raw') form through a URI included in the metadata, and in many cases not accessible at all – i.e., only (a link to) a viewer is provided. Europeana's content strategy[25] mentions that links to digital objects "can point directly to the digital object or to a website or viewer where this digital object is shown in context." However, for research purposes direct links to objects are strongly preferred over interactive viewers. This could be emphasised in the content strategy.
   - For CLARIN and its community, direct links to full text resources (e.g. plain text files or PDF files with embedded text content) are of interest in particular.
3. Direct attention and/or curation efforts to including correct content language information
   - For CLARIN users and many SSH scholars in general, content language is an important data facet. However, several collections (in particular newspaper collections and general library collections) contain resources in several languages but appear to have a default language that is applied to all records.
4. Offer more homogeneous data sets or subsets thereof

[24] http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_profiles/Technical%20Metadata%20properties_20150217.pdf

[25] http://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana%20Content%20Strategy.pdf

- Many data providers offer good quality data that is well described and accessible. In some cases, however, the records are ‘bundled’ into large, highly diverse data sets[26]. For example, manuscripts, maps and musical scores may all be found in a single data set. As data sets are the highest level of granularity by which records can be retrieved over OAI-PMH (other than at the individual record level), and manual selection of records within the data set is generally not feasible, infrastructures using this method of retrieval will have to decide whether to include data at the data set level. If a data set is too heterogeneous in terms of the nature of its content, inclusion of the full set might lead to a lot of ‘noise’ in the downstream aggregation. Infrastructures may therefore choose to exclude this data set, at the cost of missing out on potentially very interesting resources.
- We understand that there may be technical reasons underlying the lack of more fine-grained, homogeneous data sets. However, for harvesting parties such as CLARIN, it would suffice to be able to harvest subsets based on criteria such as content type or full text availability.

Recommendations with respect to the technical infrastructure

These recommendations mainly arise from issues experienced while setting up the integration between Europeana and CLARIN – see “Technical issues” (page 11).

5. Ensure a stable and well performing OAI-PMH provider
   - OAI-PMH forms the basis for integrating Europeana (meta)data into the CLARIN infrastructure and this is very likely to be the case for other potential infrastructures looking to do the same. Therefore, it is vital that the Europeana’s OAI-PMH provider supports reliable and swift harvesting by infrastructures.
6. Synchronise metadata properties to the OAI- PMH provider
   - Metadata provided over OAI-PMH should be as complete and up-to-date as metadata available through e.g. the APIs, as this is the primary means of access to metadata for infrastructures such as CLARIN.
7. Support incremental (or ‘selective’) harvesting
   - Information regarding the last date and time of change or deletion can be used to allow retrieval of metadata change sets in addition to ‘full’ harvests[27]. This allows for more regular synchronisations and less strain on both the providing and harvesting infrastructures.
8. Provide Dublin Core metadata
   - The OAI-PMH protocol and implementation guidelines[28] mandate support for Dublin Core (DC) metadata. Although not as expressive as EDM or, e.g., CMDI, it is broadly accepted and most infrastructures should be able to adopt DC metadata with little effort. Currently, DC representations of Europeana metadata cannot be retrieved from its OAI-PMH provider.

Conclusions and future work

During DSI-2, CLARIN has implemented a functioning integration of Europeana data into its infrastructure that is stable and sufficiently performing for use in production with the potential to serve as a template for similar integrations into other research infrastructures. There is also room for enhancement and extension. In addition to the recommendations for Europeana DSI listed

[26] See e.g. the following data set from the Bavarian State Library: http://www.europeana.eu/portal/en/search?q=europeana_collectionName%3A(9200386_Ag_EU_TEL_a1194_BSB)&view=grid

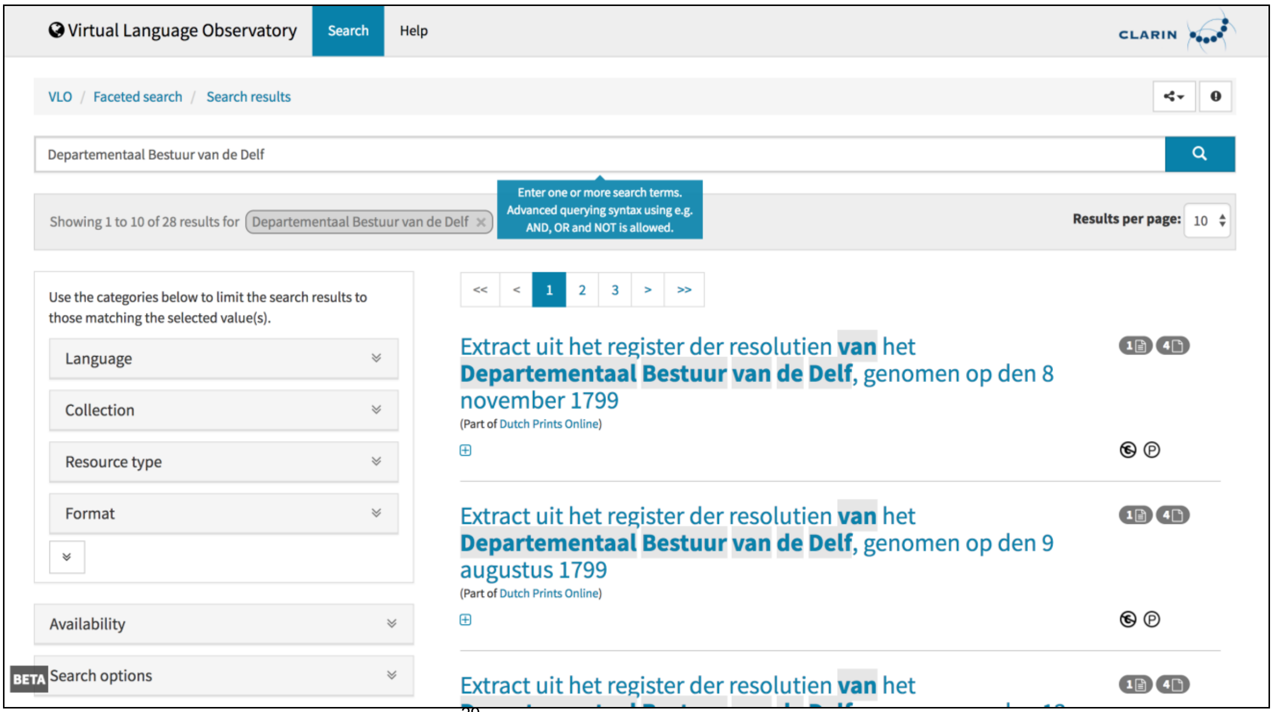
[27] As described in detail in section 2.7 of the OAI-PMH protocol version 2.0: https://www.openarchives.org/OAI/2.0/openarchivesprotocol.2015-01-08.htm#SelectiveHarvesting

[28] https://www.openarchives.org/OAI/2.0/guidelines-repository.htm#MinimalImplementation-DC

above, we have also identified concrete actions that CLARIN can take to further the dissemination of Europeana data to its community:

- Harvest and index new data sets as they become available and/or are adapted to meet the requirements; in particular, the various Europeana sounds data sets are of interest.
- Support additional EDM profiles in the conversion to CMDI.
- Integrate data and metadata available by other means than through Europeana's OAI-PMH provider for selected data sets, in particular those in the Newspapers collection.
- Exploit community specific communication channels to call attention to Europeana data within CLARIN's infrastructure, and to Europeana and its resources and services in general.
- Gather feedback from the CLARIN community regarding Europeana and its data, and assess usage and impact of the implemented (and extended) integration.

Appendix A: Screen shots for a demonstration case



Virtual Language Observatory | Search | Help | CLARIN

VLO / Faceted search / Search results

Departementaal Bestuur van de Delf

Enter one or more search terms. Advanced querying syntax using e.g. AND, OR and NOT is allowed.

Showing 1 to 10 of 28 results for Departementaal Bestuur van de Delf ✕

Results per page: 10

Use the categories below to limit the search results to those matching the selected value(s).

Language

Collection

Resource type

Format

Availability

BETA Search options

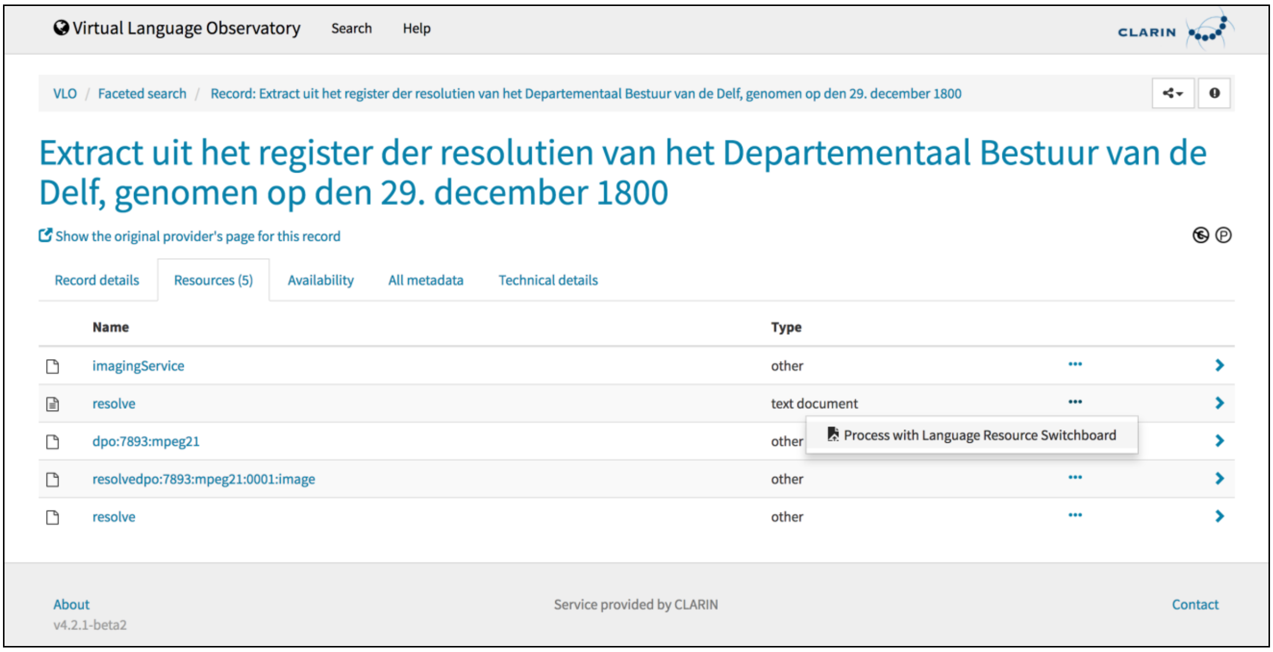
<< < 1 2 3 > >>

Extract uit het register der resolutien **van** het **Departementaal Bestuur van de Delf**, genomen op den 8 november 1799
(Part of Dutch Prints Online)

Extract uit het register der resolutien **van** het **Departementaal Bestuur van de Delf**, genomen op den 9 augustus 1799
(Part of Dutch Prints Online)

Extract uit het register der resolutien **van** het

Figure 1: Finding resources in the VLO[29]



Virtual Language Observatory | Search | Help | CLARIN

VLO / Faceted search / Record: Extract uit het register der resolutien van het Departementaal Bestuur van de Delf, genomen op den 29. december 1800

Extract uit het register der resolutien van het Departementaal Bestuur van de Delf, genomen op den 29. december 1800

Show the original provider's page for this record

Record details | Resources (5) | Availability | All metadata | Technical details

Name	Type
imagingService	other
resolve	text document
dpo:7893:mpeg21	other
resolvedpo:7893:mpeg21:0001:image	other
resolve	other

Process with Language Resource Switchboard

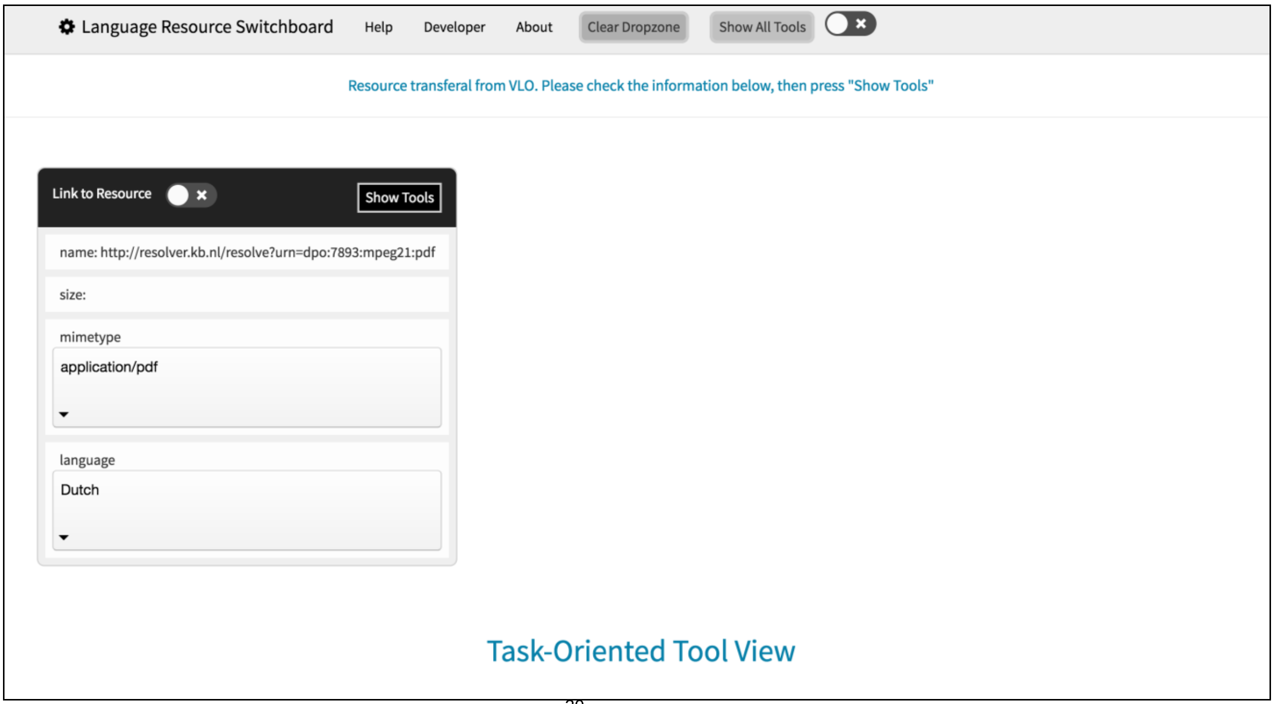
About
v4.2.1-beta2

Service provided by CLARIN

Contact

Figure 2: Selecting a resource for processing

[29] https://vlo.clarin.eu



Resource transferal from VLO. Please check the information below, then press "Show Tools"

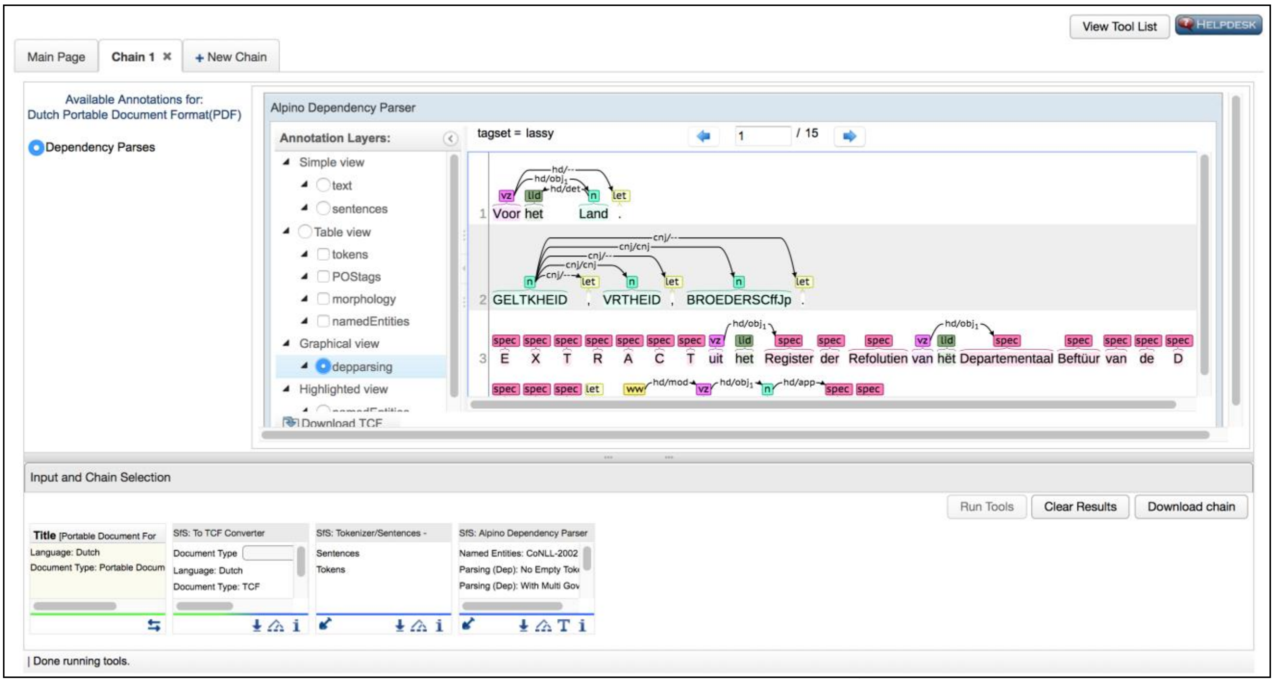
name: http://resolver.kb.nl/resolve?urn=dpo:7893:mpeg21:pdf

size:

mimetype: application/pdf

language: Dutch

Figure 3: Finding processing options in the LRS[30]



1 Voor het Land .

2 GELTKHEID , VRTHEID , BROEDERSCffJp .

3 E X T R A C T uit het Register der Refolutien van het Departementaal Beftüur van de D

Title [Portable Document For	SfS: To TCF Converter	SfS: Tokenizer/Sentences -	SfS: Alpino Dependency Parser
Language: Dutch Document Type: Portable Docum	Document Type Language: Dutch Document Type: TCF	Sentences Tokens	Named Entities: CoNLL-2002 Parsing (Dep): No Empty Toke Parsing (Dep): With Multi Gov

| Done running tools.

Figure 4: Sample result of processing a PDF resource in Weblicht[31]

[30] https://www.clarin.eu/switchboard

[31] https://weblicht.sfs.uni-tuebingen.de