Opening Digitized Newspapers Corpora: Europeana's Full-Text Data Interoperability Case

Nuno Freire
INESC-ID, Lisbon, Portugal
nuno.freire@tecnico.ulisboa.pt

Antoine Isaac
Europeana Foundation, The Hague, The Netherlands
Vrije Universiteit Amsterdam, The Netherlands
antoine.isaac@europeana.eu

Twan Goosen
CLARIN ERIC, Utrecht, The Netherlands
twan@clarin.eu

Daan Broeder
KNAW Humanities Cluster, Amsterdam, The Netherlands
daan.broeder@di.huc.knaw.nl

Hugo Manguinhas
Europeana Foundation, The Hague, The Netherlands
hugo.manguinhas@europeana.eu

Valentine Charles
Europeana Foundation, The Hague, The Netherlands
valentine.charles@europeana.eu

Abstract

Cultural heritage institutions hold collections of printed newspapers that are valuable resources for the study of history, linguistics and other Digital Humanities scientific domains. Effective retrieval of newspapers content based on metadata only is a task nearly impossible, making the retrieval based on (digitized) full-text particularly relevant. Europeana, Europe's Digital Library, is in the position to provide access to large newspapers collections with full-text resources. Full-text corpora are also relevant for Europeana's objective of promoting the usage of cultural heritage resources for use within research infrastructures. We have derived requirements for aggregating and publishing Europeana's newspapers full-text corpus in an interoperable way, based on investigations into the specific characteristics of cultural data, the needs of two research infrastructures (CLARIN and EUDAT) and the practices being promoted in the International Image Interoperability Framework (IIIF) community. We have then defined a "full-text profile" for the Europeana Data Model, which is being applied to Europeana's newspaper corpus.

2012 ACM Subject Classification Applied computing → Annotation; Applied computing → Document metadata; Applied computing → Digital libraries and archives

Keywords and phrases Metadata, Full-text, Interoperability, Data aggregation, Cultural Heritage, Research Infrastructures

Digital Object Identifier 10.4230/OASIcs.LDK.2019.22

Funding *Nuno Freire*: This work was partly supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2019, and by the European Commission under contract number 30-CE-0885387/00-80.e.

2nd Conference on Language, Data and Knowledge (LDK 2019).
Editors: Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski; Article No. 22; pp. 22:1–22:14

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Cultural Heritage Institutions (CHI), such as libraries and archives, hold collections of printed newspapers of the past centuries. These are valuable resources for historians, linguists and other researchers working in Digital Humanities. The retrieval of printed newspapers' content based on metadata only is a task nearly impossible, however. Cultural Heritage Institutions usually describe the series of a newspaper publication (typically known as "title level" description) and its individual publications ("issues") in their catalogs, but no description of individual articles. The typical use of the catalogs of newspapers is thus only to retrieve issues by date of publication, as there is no detail for effective retrieval of the content at finer-grained levels.

The wide interest in newspapers and the challenges they pose for retrieval has motivated CHIs to prioritize the digitization of their newspapers collections. CHIs also realized that the retrieval of newspapers' content based on machine readable full-text is particularly important, given the unavailability of article level descriptions in the catalogs. Accordingly, CHIs have also sought to apply Optical Character Recognition (OCR) during the digitization process.

Our work addresses the general problem of the retrieval of newspapers in the context of aggregations of digital Cultural Heritage (CH) resources, in particular that of Europeana. Europeana seeks to facilitate the use of resources from and about Europe. It enables access to objects via its Collections portal,[1] which supports all official languages of Europe, and its open APIs enable third-party applications. Europeana is based on metatada provided by its CHI partners and presently holds metadata from over 3,700 CHIs.[2] Providing access to newspapers is relevant to Europeana's mission, especially for promoting the re-use of CH resources for research. Europeana indeed also aims to facilitate research, especially for the digital humanities, via its Europeana Research initiative.[3] This initiative seeks to address issues related to, e.g., licensing, which affect the research re-use of CH metadata and content. In particular, it has identified research re-use of newspapers resources as a key use case, as well as an area with strong system and data interoperability challenges.

Digitized newspapers are Europeana's first case of aggregation and distribution of full-text CH resources. Europeana's systems have relied so far on metadata and links to digitized resources at partners' sites. The Europeana Data Model (EDM) [7] allows it to perform scalable aggregation of (and access to) references to digital representations of CH artifacts with rich context metadata. EDM follows the Linked Open Data principles [1]. An important aspect of EDM is its flexibility and genericity: it can be easily mapped to other (CH) data models and extended [3]. This makes it a potential base for the interoperability of full-text resources within the Europeana ecosystem.

This paper presents how we have tested this assumption by trying to extend EDM to cater for interoperability of full-text CH corpora. The first aim of our work is to support a centralized search engine and rich user interfaces. But we have also investigated the issue of interoperability of full-text between Europeana and research infrastructures (EUDAT and CLARIN). Our work focuses on Europeana and research use, but we claim it has impact on other application contexts, as we sought to align with the generic International Image

[1] https://europeana.eu
[2] https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI/Deliverables/europeana-dsi-d1.2-amount-of-data-partners-and-outreach-to-major-institutions.pdf
[3] https://research.europeana.eu

Interoperability Framework (IIIF).[4] IIIF is a family of specifications that were conceived to facilitate systematic reuse of image resources in digital repositories maintained by CH institutions. It specifies several HTTP based web services covering access to images, the presentation and structure of complex digital objects composed of one or more images, and searching within their content. IIIF's strength resides in the presentation possibilities it provides for end-users. We present related work on digitized newspapers and the use of CH data in research infrastructures in Section 2. Section 3 presents the exploratory work conducted by Europeana, EUDAT and CLARIN, and the interoperability requirements derived from it. Section 4 presents our EDM extension for full-text, and Section 5 concludes.

2 Related work

Several initiatives exist worldwide with similar target user groups and use cases as Europeana, with respect to aggregation of digitized newspapers. The organizational structure and technical interoperability context of Europeana are quite different, however. For example, Chronicling America,[5] a national aggregation of newspapers in the United States of America, gathers its corpus from the digitization conducted under the National Newspaper Digitization Program. The direct relation of Chronicling America with the digitization process, results in more homogeneous metadata and full-text content to provide access to.

Europeana Newspapers [6] was an earlier project from the Europeana community, which aggregated metadata and full-text content in a portal that, while currently hosted by Europeana, sits on a completely disconnected platform. The project established interoperability by defining a METS/ALTO profile [11], but its application was restricted to the project and did not spread to other CHIs afterwards.

The IIIF Community has conducted similar work to ours in establishing a generic representation of full-text associated with images for the IIIF Presentation API. We participate in a IIIF Newspapers Community Group that gathers IIIF community members working with digitized newspapers. The IIIF representation patterns strongly inspired our work. These, however, are quite generic and the connection with (descriptive) metadata is rather loose in the IIIF presentation API, which relies on linking to document using models like EDM for representing fully-fledged metadata. Furthermore, directly relying on IIIF APIs is an obstacle for the metadata providers who cannot deploy IIIF services for their content.

Regarding interoperability with research infrastructures, related work in CH digitized resources and OCR full-text includes Herbadrop [5]. This initiative works with resources from museums and botanical gardens, which own collections of plant samples with detailed annotations from botanists. Herbadrop has worked with the EUDAT CDI;[6] as part of a data pilot [5].

Finally, some CHIs provide data to CLARIN,[7] in particular university libraries. CLARIN aggregates CH resources in a similar process to Europeana's but uses a different metadata format [4]. Regarding full-text corpora within CLARIN, we observe a prevalence of the Text Encoding Initiative (TEI) format[8] next to plain text content in terms of support by existing tools and also in published research. TEI usage within the Europeana Network is limited: it is only present in CHIs that focus on supporting researchers. Plain text content is often not provided by CHIs.

[4] https//iiif.io
[5] https://chroniclingamerica.loc.gov/
[6] EUDAT Collaborative Data Infrastructure; https://www.eudat.eu/eudatcdi
[7] Common Language Resources and Technology Infrastructure; https://www.clarin.eu/
[8] TEI – Text Encoding Initiative; https://www.tei-c.org/

3 Needs for interoperability with Research Infrastructures

Europeana is interested in investigating how research data infrastructures can facilitate the research use of CH resources. By leveraging on research infrastructures that operate at a European level and across scientific disciplines, it hopes to reach researchers from all scientific disciplines, without having to work with many national and domain-specific research infrastructures or providing its own. We describe here the efforts on the Europeana Newspapers corpus conducted with two infrastructures: CLARIN and EUDAT. This corpus was aggregated from 11 CHIs during the Europeana Newspapers project. It contains metadata descriptions, digitized images and full-text of 911 newspaper titles that, in total, comprise over 11 million pages [6], in multiple languages and scripts. We present, in this section, the interoperability challenges identified and what we did to tackle them.

3.1 Interoperability with CLARIN

CLARIN is a federation of language data repositories, service centers and centers of expertise. CLARIN aggregates metadata and makes the underlying resources discoverable and usable within research workflows. It allows researchers to carry out natural language processing tasks by invoking processing tools directly from its generic user interface. Establishing good interoperability between Europeana and CLARIN can help fitting a large number of CH resources into CLARIN's supported workflows. It will open up new applications for CLARIN's processing tools and promote research incorporating CH resources.

CLARIN carried out a first analysis of the Europeana Newspapers corpus in 2015, establishing goals and a ground for connecting the two infrastructures and full-text interoperability. Later, we sought to address the interoperability issue for metadata [9]. The two infrastructures use specific metadata models: EDM for Europeana and the Component MetaData Infrastructure (CMDI) for CLARIN [4]. Interoperability is achieved via CLARIN's metadata conversion mechanisms, based on a CMDI profile for EDM.[9] Europeana's metadata for Newspapers and other datasets can thus be made available within the CLARIN systems.

The desirable level of interoperability between the two infrastructures has not been achieved, however. The newspapers full-text corpus, although partially discoverable within CLARIN, cannot yet be processed by CLARIN's tools in research workflows. The following requirements for how metadata and full-text content are made available by CHIs were noted and greatly influenced our work on extending EDM for exchanging full-text content:

- Direct links to content files – when CHIs only expose links to websites or viewers in the metadata aggregated by Europeana, the files cannot be processed by CLARIN (and others).
- Technical metadata – information like media type and file size are essential for automated processing workflows and highly desirable for discovery
- Language of the content – most natural language processing tools are language dependent, making the language information carried in CH metadata essential.

3.2 Interoperability with EUDAT

EUDAT is a European infrastructure of integrated data services devoted to scientific and research data storage and life cycle management. It has been developed in close collaboration with over 50 research communities spanning across many different scientific disciplines

[9] Available in CLARIN's component registry: `https://catalog.clarin.eu/ds/ComponentRegistry/#/?itemId=clarin.eu%3Acr1%3Ap_1475136016208`

such as Life Sciences, Humanities, Earth Sciences and Physics, with more than 20 major European research organizations, data centres and computing centres involved. Many of these collaborations are carried out as data pilots providing test-beds that vary in disciplines, communities, project group sizes and technological maturity. Europeana conducted a data pilot with EUDAT that consisted in a case study on the Europeana Newspapers corpus [5]. The general goal was to investigate how EUDAT data services can facilitate the use of CH resources for research purposes. The questions laid out at the start of the data pilot were:

- How can the resources be discovered?
- How can the resources be shared in practical ways for researchers?
- How can advanced computation be applied to these CH datasets?
- How can the resources and datasets be cited and referenced in research?
- How can the CH institutions re-use the outcomes of research?

An evaluation of the available EUDAT services was conducted, using the newspapers corpus as case study. The two infrastructures were successfully interconnected and EUDAT fulfilled the expectations for making the corpus available to researchers and for computational processing. The persistent identification of EUDAT resources also met the citability requirement. The EUDAT service did not scale to the dimension of the corpus, but only due to an underestimation of the required computational capacity during the pilot [5]. Beyond the full-text corpus case study, interoperability was also trialled for metadata-based discovery of CH datasets. Both infrastructures have common underlying technologies that facilitate interoperability, including on modelling full-text, since EUDAT is developing its semantic annotation service based on the W3C Web Annotation Data Model,[10] which is a key component of the EDM extension we are going to present in the next section.

4 Building a full-text profile for the Europeana Data Model

A profile for representing full-text in EDM is a key requirement for achieving a sustainable interoperability framework for full-text CH corpora in Europeana. It has potential applications in full-text aggregation, indexing, user experience and data re-use. This section presents the context, requirements and the EDM full-text profile.

4.1 Context and requirements for designing the data model

Based on the corpus of full-text newspapers, the case studies with research infrastructures and recommendations from the earlier Europeana projects [6, 2], we have identified these requirements:

- The availability of full-text must be stated explicitly in the metadata.
- The representation of full-text should be compatible with the representation of the newspapers' structure (issue, page, article, etc.) in the descriptive metadata.
- The representation of full-text must allow the specification of the language and script of the text, and it should allow this specification to be done at several levels of granularity of the text (e.g. for a paragraph, for a word, etc.).
- URLs to views of the digital objects must be explicitly stated in the metadata.
- Multiple full-text resources must be referenced via direct URLs.
- Resources requiring a protocol to be served need to be clearly identifiable.

[10] https://www.w3.org/TR/annotation-model/

- When more than one full-text resource is associated with a digital object, it should be possible to represent their part-whole relationship.
- When more than one full-text resource is associated with a digital object, it should be possible to represent their sequential order.
- When a full-text resource is available as a fragment of text, the URI or the literal identifying the specific text fragment may be provided in the data.
- When a full-text fragment is available, the image area it refers to should be identified (via coordinates).

The IIIF community has suggested to publish textual representations of (part of) images, such as transcriptions, using annotations from the W3C Web Annotation model (WA). Annotations are included in the IIIF "manifests"[11] of the newspapers, as a list of annotations, each one referring to a portion of the full-text and indicating its corresponding position in the image of a page. Representing full-text as annotations seems the best solution as it can support simple scenarios such as the positioning of a text fragment on an image as well as more complex ones like OCR correction.

This approach, besides its community traction, is compatible with the Linked Data vision and fits well Europeana's use of annotations for other purposes [10]. One of the cases that has recently emerged in Europeana is indeed the representation of manual transcriptions of content.[12] As meeting the requirements of these related cases in similar ways is extremely desirable, we decided to follow the IIIF Community approach. Our modelling exercise thus becomes one of fitting into EDM a representation of the full-text content of newspapers as annotations on the images of newspapers' pages.

4.2 EDM extension addressing the initial full-text requirements

Our extension of EDM for representing full-text follows the recommendations of IIIF (in its coming version 3) and WA. Full-text is represented as the body of an annotation that has as target an image, as illustrated in Figure 1. We model the image as an `edm:WebResource` (the usual EDM approach) and the text itself as a new proposed subclass of `edm:WebResource`, `edm:FullTextResource`.[13] Figure 2 illustrates the simplest case. Annotations are modeled using WA's `oa:Annotation` class and `oa:hasBody` and `oa:hasTarget` properties. Annotations used for representing full-text must have the property `oa:motivatedBy` with the value `edm:transcribing`, distinguishing them from Europeana annotations used for other motives, as well as following IIIF's latest best practices[14] (NB: we omit it from our figures for readability reasons).

The extension supports two levels of detail for associating the full-text with the image: with and without its position within the image. The text can also be provided by value (a plain literal) or by reference (as a URI, and/or as a selection/extract from another text resource). The following sections present the details of these options.

[11] IIIF manifests are "the overall description of the structure and properties of the digital representation of an object."; `http://iiif.io/api/presentation/2.0/#primary-resource-types`

[12] Cf. Europeana's initiative on transcribing WWI-related content; `https://transcribathon.com/`

[13] The full-text comes as `rdf:value` for the `edm:FullTextResource`, using WA's "embedded text" pattern (`https://www.w3.org/TR/annotation-model/#embedded-textual-body`) with a type independent from the resource's being used in an annotation, unlike WA's `oa:TextualBody`.

[14] Cf. IIIF API issue 1258: `https://github.com/IIIF/api/issues/1258`

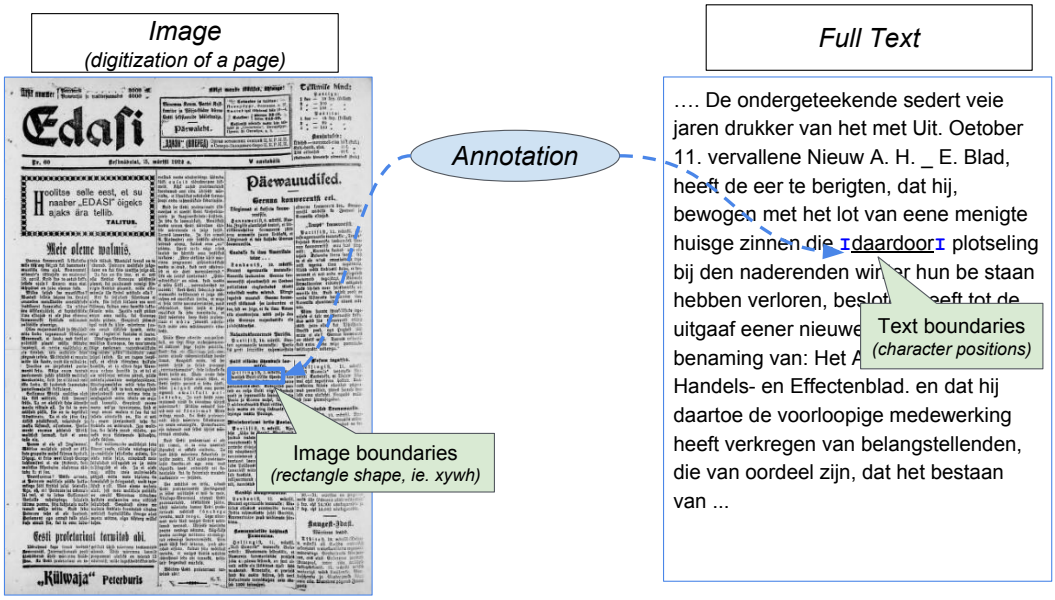


Figure 1 General principles for full-text annotations in the EDM extension.

4.2.1 Full-text without position

In the simplest case, illustrated in Figure 2, full-text is associated with an image without any information about the position of the text within the image.

4.2.2 Full-text associated with fragments with a position in the image

An earlier analysis of newspapers corpora [2] has shown that full-text is sometimes represented as several fragments of text, each referring to a specific area of an image (an article, a specific line in the text or a word). In this case, the full-text fragment is accompanied with coordinates indicating its position on the image.

To support this requirement, we introduce in the model the `oa:SpecificResource` that "is used in between the Annotation and the body or target, as appropriate, to capture the additional description of how it is used in the annotation" [9]. An `oa:FragmentSelector` is applied as selector within the `oa:SpecificResource` to restrict the original target (the `edm:WebResource`) to the specific area to which the text, or fragment, corresponds. Figures 3, 4 and 5 show examples of this solution.

In Figure 3, the `edm:FullTextResource` consists of a fully-fledged resource that corresponds to a paragraph whose position is indicated by the `oa:FragmentSelector`. Note that for rectangle areas, coordinates in the `oa:FragmentSelector` must follow the Media Fragments W3C recommendation and be the subject of a `dcterms:conformsTo` statement referring to `http://www.w3.org/TR/media-frags/` (not shown in the figure).

4.2.3 Full-text selections represented as fragments with a position in the image

Figure 4 shows how more details – in this case, the position of a particular word – can be specified for the association between full-text and images. The area is indicated using the pattern already seen in Figure 3, but the paragraph fragment that corresponds to the

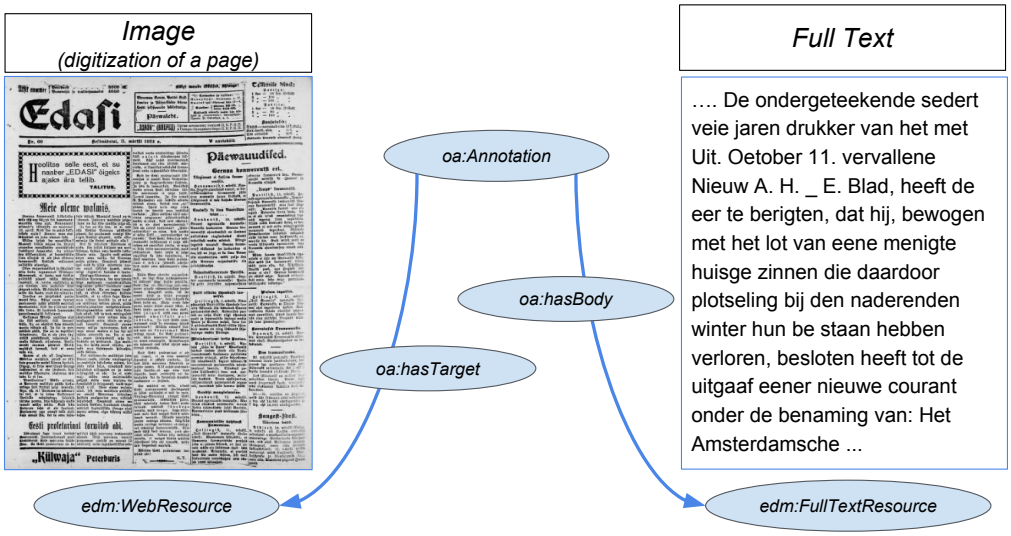


Figure 2 Full-text without position information.

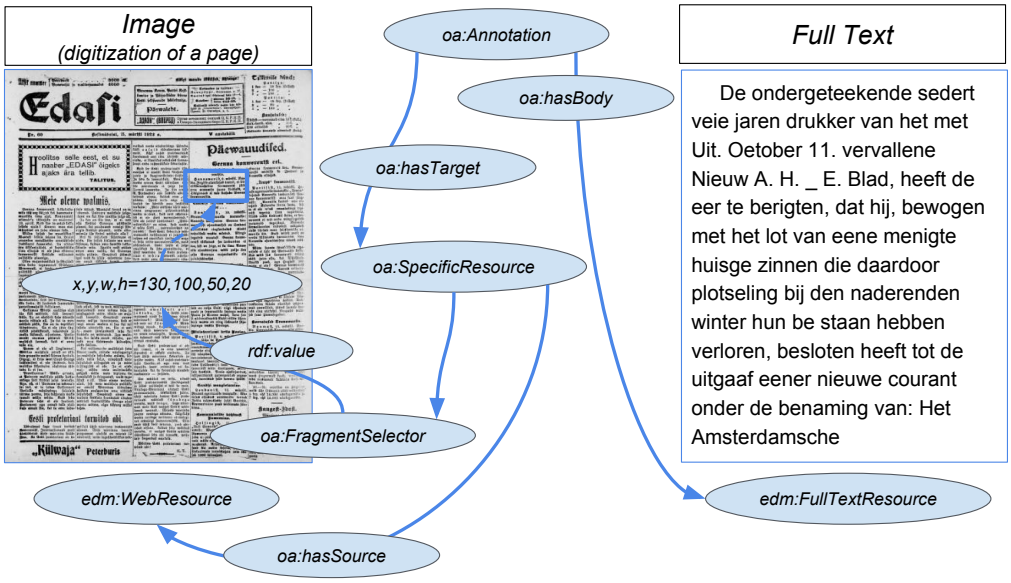


Figure 3 Full-text resource with position on the image.

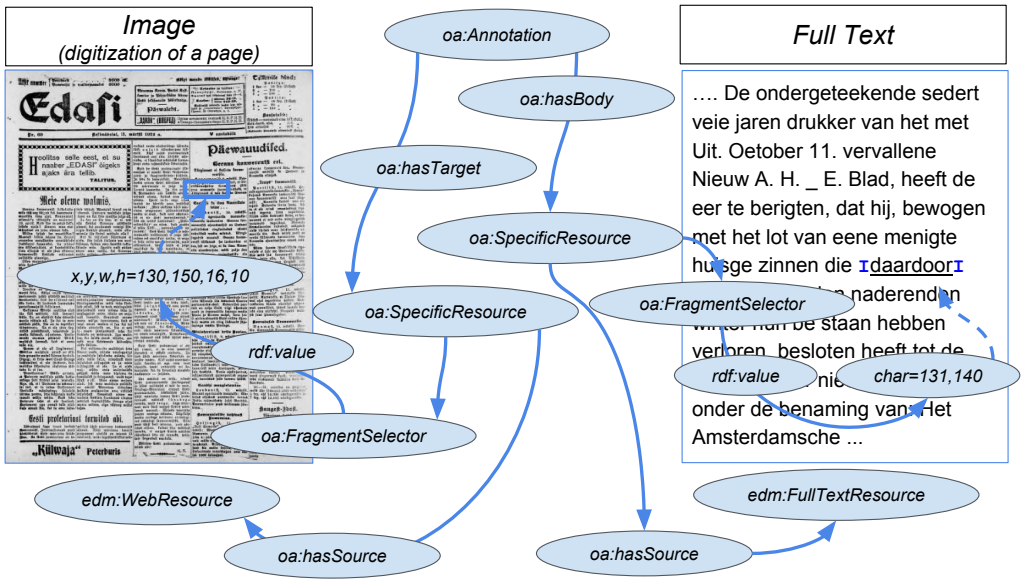


Figure 4 Full-text fragment with position on the image using *oa:FragmentSelector*.

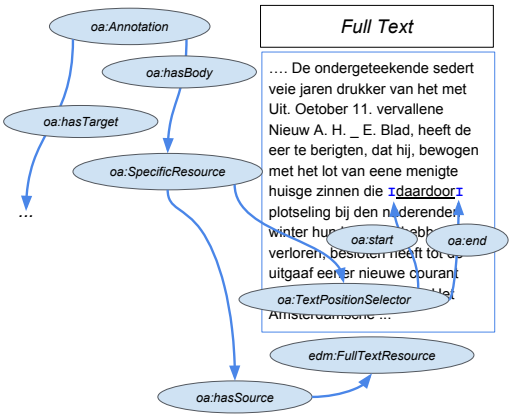


Figure 5 Full-text fragment with position using *oa:TextPositionSelector*.

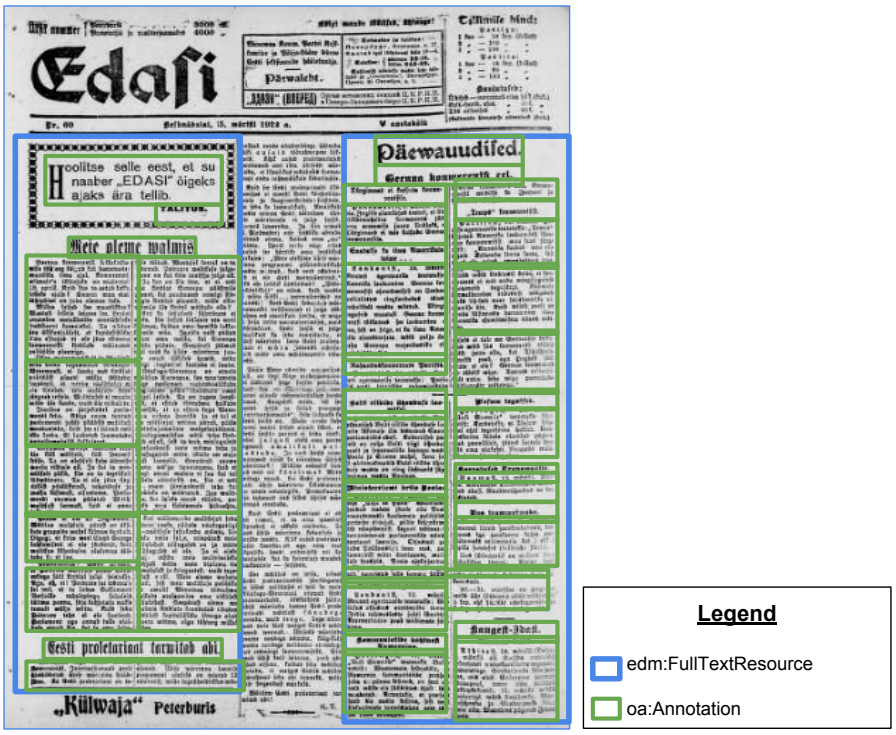


Figure 6 Representing the logical structure of articles and paragraphs of full-text with *edm:FullTextResource* and *oa:Annotation*.

word in the full-text is also given: an `oa:SpecificResource` is created to represent how the textual body of the annotation is derived from another resource. An `oa:FragmentSelector` resource describes the range of text by recording the first and last characters' positions within the source. The `oa:FragmentSelector` must follow RFC 5147 and be the subject of a `dcterms:conformsTo` statement referring to `http://tools.ietf.org/rfc/rfc5147` (not shown in Figure 4). Note that the WA model offers alternatives for representing fragments: e.g., for text fragments, the data from Figure 4 can also be represented using an `oa:TextPositionSelector`, recording the start and end positions with specific properties (see Figure 5). We have decided for now to be flexible in what Europeana will accept, opening the possibility to use equivalent WA selectors. But we will seek to normalize the data we publish, i.e. retaining only one of the options – yet to be discussed with the community.

4.2.4 Logical structure of the full-text

Some digitization efforts apply segmentation techniques to detect the independent sections (such as articles) within a newspaper page. Our EDM extension allows representing the different sections in the full-text. First, text of different levels can be represented as different `edm:FullTextResources` connected across levels using Dublin Core `dcterms:hasPart` and `dcterms:isPartOf` properties. EDM allows this for any digital representation, and this pattern can be used in particular between a newspaper file that contains several pages (images) and the image of each page. In this case, however, text is duplicated across levels. An alternative is to represent the logical structure via the organization of `edm:FullTextResources` and `oa:Annotations`. Our extension assumes that each `edm:FullTextResource` can reflect

a section within a page and act as grouping for all related `oa:Annotations`. Figure 6 shows a newspaper page where two `edm:FullTextResources` represent two articles in the page. It also highlights how (targets of) `oa:Annotations` represent the paragraphs within each `edm:FullTextResource`.

4.2.5 Specifying the language of the full-text

The profile allows the indication of language of the full-text at several levels of detail. At the most general level, the language indicated in the data for the original cultural object (using Dublin Core's `dc:language` property on EDM's `edm:ProvidedCHO` resource[15]) can be seen to apply to the whole full-text as well. Our profile assumes that when a (sub-component of) the full-text does not specify its language, then it inherits the language from the higher levels of its hierarchy. This pattern enables to represent cases when a word in one language is present within a text in another language. But there can be different languages, or a data publisher may prefer to express precise information that does not depend on implicit "propagation" rules between levels in the data. Therefore, the language may be specified at the level of any `edm:FullTextResource`, using an RDF language tag on the `rdf:value` of the resource or the `dc:language` property.[16] At the finest level of detail, languages may be specified on the `oa:SpecificResource` referring to text fragments. Figure 7 illustrates using `dc:language` on the `edm:FullTextResource` and the `oa:SpecificResource`.

4.3 Application of the profile

At this time, the EDM full-text profile is already applied at production level. Europeana has converted the Europeana Newspapers corpus to the EDM full-text profile, therefore, the profile has been applied to more than 11 million pages of newspaper full-text transcriptions, in multiple languages and scripts. Since this corpus originates from data providers from different countries using different practices for digitisation, we see this application as evidence that the model can accommodate the different ways of structuring full-text in digitised objects.

Europeana has also made significant steps implementing the full-text profile in its systems. It has adapted its data infrastructure to support the ingestion of full-text according to the profile (no support for full-text existed previously in Europeana).

Regarding indexing and retrieval of full-text EDM data, Europeana has completed a first version of its solution, which combines the joint retrieval of resources described by metadata only, with resources with full-text and metadata.

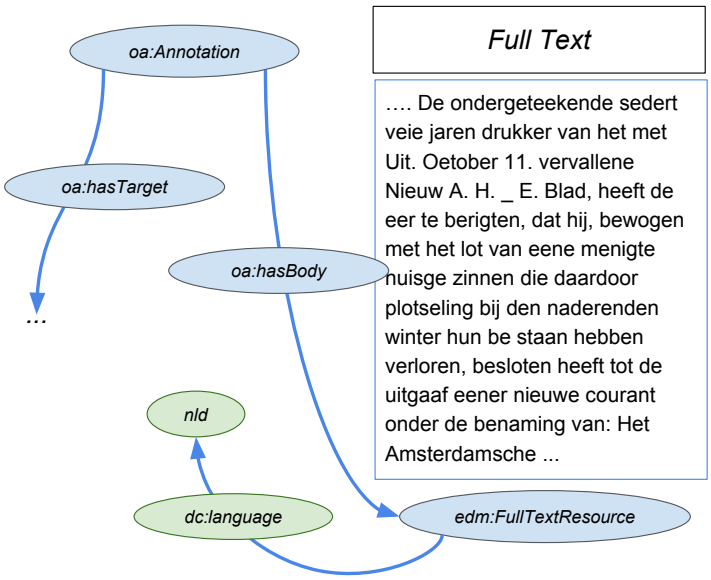
It has completed a first version of indexing and search services, which provides retrieval of full-text resources. This first version is not yet integrated with the main search systems of Europeana (that works only on metadata), but the first steps have taken place for investigating a solution for accomplishing a joint search system.

On top of this, Europeana's final products are a portal and an API. The portal is specialised for the newspapers corpus[17] and provides a user-interface based on full-text retrieval and the association, via image coordinates, between digitised images and the

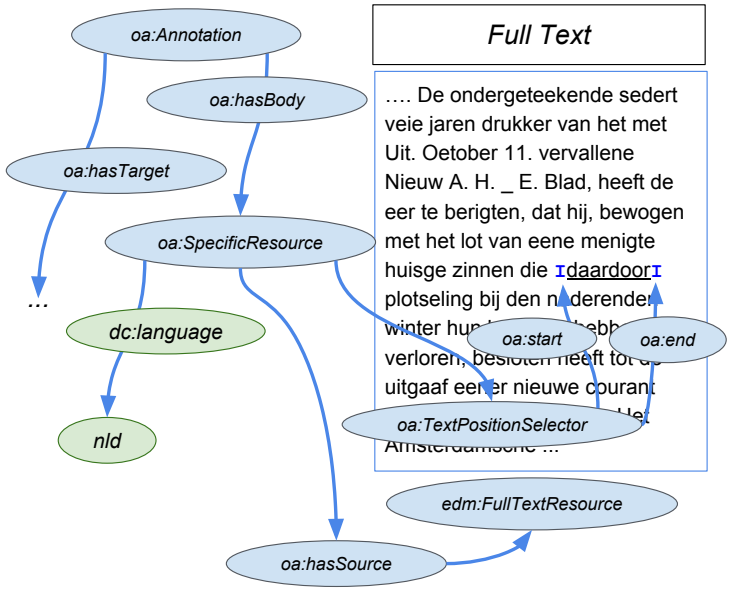
[15] `ProvidedCHO` stands for "Provided Cultural Heritage Object". It is the original object that is described. It may be either a physical object (painting, book, etc.) or digital-born object.

[16] Here again there are two equivalent modeling alternatives: the "traditional" RDF one (already used in EDM and one preferred by the WA model. We intend to accept both and publish both in parallel, but this choice is still open to community feedback.

[17] `https://www.europeana.eu/portal/en/collections/newspapers`



(a) for the whole *edm:FullTextResource*



(b) for a piece of text in isolation (i.e., a word)

Figure 7 Specification of the language of the text.

transcription. This interface uses the full-text to down to word-level detail (when word level coordinates have been recorded during digitisation and OCR). The API service now available for newspapers[18] complements the existing Europeana API with functionality specialised in full-text search and access, including making the full-text available according to the IIIF Presentation API - where the IIIF output is generated from the EDM representation. This improves Europeana's capacity to promote data re-use of CH content through research infrastructures and other target user groups.

5 Future Work and Conclusion

Europeana's investigations in exploring its newspapers full-text corpus with research infrastructures has provided valuable input for making CH corpora better discoverable, accessible, machine processable and citable in research contexts. The requirements identified for research usage of CH full-text corpora support several aspects of the current strategy of Europeana towards improving data quality and direct access to the media contents of CH digital objects [8].

The currently aggregated full-text corpus of Europeana Newspapers has not grown since the end of the Europeana Newspapers project, and an aggregation process based on the ALTO profile was not possible to establish in a sustainable way at Europeana, due to its high technical complexity for adoption by data providers, and also for aggregators. The new model, being based on EDM and following the IIIF Community approach is expected to lower the technical barriers to establish a sustainable full-text aggregation process.

In the near future, our EDM full-text profile is going to be used as the basis to resume the aggregation processes of full-text newspapers content across the Europeana Network. In parallel, we will update the EDM full-text profile, by devising a more precise approach to the modeling alternatives that the current version allows – we have already begun to actively seek feedback from the IIIF Newspapers community. We will also tackle new requirement that could emerge during its adoption: for example, some Europeana stakeholders have voiced interested in an explicit representation of the granularity of the full-text (page, article, paragraph, line, word).

Regarding the re-use of CH full-text data for research, CLARIN is starting an assessment of the applicability of the full-text content, as disseminated by Europeana, to its infrastructure and the connected tools in the context of various typical research use cases, covering resource discovery, retrieval and processing. On basis of the findings of this assessment, we expect to be able to fine-tune the full-text profile and the content APIs on the side of Europeana, and adapt the exploitation of Europeana's services by CLARIN accordingly, so as to achieve a broad integration of large volumes of full-text content with real-world applicability for the social sciences and humanities research communities.

References

1 Timothy Berners-Lee. Linked Data Design Issues. W3C-Internal Document, 2006.
2 Valentine Charles, Nuno Freire, Hugo Manguinhas, Peter Vos, and Glen Robson. Recommendations for enhancing EDM to represent digital content. Technical report, Europeana Cloud D4.4, 2016.
3 Valentine Charles and Antoine Isaac. Enhancing the Europeana Data Model (EDM). Technical report, Europeana V3.0, 2015.

[18] https://pro.europeana.eu/data/newspapers-getting-started

4 CMDI Taskforce. *Component Metadata Infrastructure (CMDI) Component Metadata Specification Version 1.2*, 2016.

5 Pascal Dugenie, Nuno Freire, and Daan Broeder. Building new knowledge from distributed scientific corpus: HERBADROP & EUROPEANA: Two concrete case studies for exploring big archival data. In Jian-Yun Nie, Zoran Obradovic, Toyotaro Suzumura, Rumi Ghosh, Raghunath Nambiar, Chonggang Wang, Hui Zang, Ricardo A. Baeza-Yates, Xiaohua Hu, Jeremy Kepner, Alfredo Cuzzocrea, Jian Tang, and Masashi Toyoda, editors, *2017 IEEE International Conference on Big Data, BigData 2017, Boston, MA, USA, December 11-14, 2017*, pages 2231–2239. IEEE Computer Society, 2017. doi:10.1109/BigData.2017.8258174.

6 Alastair Dunning, Alena Fedesenka, Anastasia Gasia, and Markus Muhr. Report on newspapers data aggregated by The European Library. Technical report, Europeana Newspapers D4.5, 2015.

7 Europeana Foundation. *Definition of the Europeana Data Model v5.2.8*, 2017.

8 Europeana Foundation. *Europeana Publishing Guide v1.5*, 2017.

9 Twan Goosen, Dieter Van Uytvanck, and Nuno Freire. Results and Impact of Sharing Europeana Data with CLARIN. Technical report, Europeana DSI-2 MS2.2, 2017.

10 Sergiu Gordea, Hugo Manguinhas, Antoine Isaac, Valentine Charles, Maarten Brinkerink, Alessio Piccioli, and Breandán Knowlton. Modelling and exchanging annotation for Europeana projects. In *Semantic Web in Libraries Conference 2015*, 2015.

11 Günter Mühlberger. METS ALTO Profile (ENMAP). Technical report, Europeana Newspapers D5.2, 2014.