DELIVERABLE

Project Acronym:	Europeana Libraries
Grant Agreement number:	270933
Project Title:	Europeana Libraries: Aggregating digital content from Europe's libraries

D4.3 Report on how the full-text content will be made available to Europeana

Version 1.0

Authors:

Valentine Charles	The European Library / Europeana
Robina Clayphan	The European Library / Europeana
Sally Chambers	The European Library
Nuno Freire	The European Library
Andreas Juffinger	The European Library
Gilberto Pedrosa	Instituto Superior Ténico, Lisbon

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	**Public**	**P**

Revision History

Revision	Date	Author	Organization	Description
0.1	02.12.11	Nuno Freire	The European Library	Initial draft
0.2	05.12.11	Nuno Freire	The European Library	Draft
0.3	07.12.11	Nuno Freire Valentine Charles	The European Library / Europeana	Draft
0.4	12.12.11	Nuno Freire Valentine Charles Gilberto Pedrosa	The European Library / Europeana / IST	Added executive summary Replaced the use of edm:incorporates with dcterms:hasFormat Draft
0.4	19.12.11			Version sent for review to the appraisal group
1.0	28.12.11	Valentine Charles	The European Library	Added comments from the appraisal group Final version
1.0	29.12.11	Andreas Juffinger	The European Library	Adoptions in presentation and document outline

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise.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

1 Executive Summary

The European Library holds a centralized full-text index containing over 24 million pages of text made available by national libraries during the TELplus[1] project. Building on the experience in this and other projects, *The Europeana Library* is building a sustainable full-text aggregation infrastructure and establishes standards for full-text interoperability between data providers.

This document refines the requirements and describes the lessons learned from providing full-text to Europeana. The current interoperability framework of Europeana is based on the exchange of metadata about arbitrary digital objects. The specific nature of full-text and the various ways of how full-text can be made available, full-text metadata should be explicitly incorporated into the Europeana Data Model (EDM).

Most of the requirements for representing full-text resources are met by EDM. For representing digital objects with full-text, it is required that a new class is added to EDM, and to define its relations with other EDM classes. The additional class *FullTextResource* is a subclass of the available class *InformationResource*. It allows the representation of the individual full-text content, separately from the end-user access versions, which are modelled in EDM in the *WebResource* class.

The *FullTextResource*'s are included in the digital object metadata by *hasFormat* relations from *WebResource*'s to *FullTextResource*'s. This allows the aggregators to access the full-text, and to provide the end-user with access to the views of the digital object separately.

The current version of the XML schema for EDM requires changes to accommodate the representation of full-text resources. Besides the addition of the class *FullTextResource*, the properties *dcterms:hasFormat* and *edm:isNextInSequence* must be allowed for use within Web Resources. Regarding the property *edm:isNextInSequence*, its use with *FullTextResource*'s must be included in the schema definition for FullTextResources.

Currently, parts of the full-text corpora, already aggregated by *The European Library*, are being processed for delivery to Europeana. A total of 3,198,859 pages have been made available in ESE to Europeana in December 2011.

2 Introduction

The European Library holds a centralized full-text index containing over 24 million pages of text. National libraries have made the full-text available during the TELplus project, with the purpose of indexing them and thus improving the overall search services. The full-text comes from various optical character recognition (OCR) activities across Europe.

This outcome of the TELplus project provided the first practical experience attempting to create a centralized index of full-text from different national libraries. Building up on this experience, Europeana Libraries is conducting further work in order to build a sustainable full-text aggregation infrastructure, and establish standards for full-text interoperability between data providers and aggregators.

This document addresses the data exchange requirements for this full-text aggregation infrastructure, representing one further step towards establishing full-text standards in the context of Europeana.

[1] http://www.theeuropeanlibrary.org/portal/organisation/cooperation/telplus/

More specifically, it addresses the requirements for the data exchange between data providers and an aggregator, in order to support a centralized full-text search and retrieval service.

This work was conducted by studying the full-text corpora currently aggregated by *The European Library*. These corpora are presented in Section 3, along with an analysis of the full-text representation approaches typically used by libraries. Section 4 presents the requirements for a data model that could support the type of full-text aggregation being addressed in Europeana Libraries. A data model addressing these requirements, and based in the Europeana Data Model, is proposed in Section 5. The current status of the delivery of full-text to Europeana is presented in Section 6. Section 7 concludes and presents the future work for the project.

3 Analysis of the Aggregated Full-Text Corpora

The European Library's aggregated full-text resources studied for the analysis presented originate from 14 national libraries. The complete description of these corpora is presented in Table 1. This amount of content and diversity provides a good basis to analyse the full-text data models in use by different libraries.

Country of origin	Material type	Pages	Temporal coverage	Languages
Austria	Newspapers, governmental material	534.000	1862 – 1925	German
Czech Republic	Books, newspapers	2.579.511	1800 – 1989	Czech, German
Estonia	Newspapers, journals	713.933	1821 – 1940	Estonian
France	Books, periodicals	8.242.908	1650 – 1930	French (some others)
Hungary	Periodicals, newspapers, journals, books, monographs, pamphlets	237.914	1590 – 1992	Hungarian, Latin, English, German
Iceland	Newspapers, journals	5.727.149	1773 – 2002	Icelandic, Faroese, Greenlandic
Latvia	Newspapers, books	195.075	1900 – 1952	German, Latvian
Lithuania	Newspapers	125.477	1904 – 1940	Lithuanian
Norway	Books, journals	1.600.000	By authors dead for more that 70 years	Norwegian (others)
Poland	Newspapers, books	436.198	Before 1939	Polish, German, Czech, Ukrainian, Belarusian, Yiddish
Slovakia	Newspapers	185.000	Before 1918	Slovak, Hungarian, German
Slovenia	Newspapers, books, journals	328.502	1500 – 1945	Slovenian
Spain	Newspapers, books	3.033.525	17th – 19th Century	Spanish
Sweden	Newspapers, books, journals, printed ephemera	253.653	Until the 20th century	Swedish

Table 1 Full-text aggregated by *The European Library*

The first major distinction between full-text materials, with significant impact on the modelling of the full-text object, is whether the object is born-digital or has been digitised. While on born-digital the file format typically contains all the text along with its visual attributes, in digitised objects libraries adopted different types of solutions to represent text and visual attributes. Digitized full-text objects are the result of a two-step process. Firstly, digitization creates the images from the analogue object, and secondly OCRing is performed on the images extracting the text. The outcomes of this process are two sets of files, which are then combined in different ways to form the full-text object. Libraries adopted different solutions to combine images and textual information. These solutions generally fall into three kind of strategies. Table 2 presents these three main strategies and their main characteristics.

	Description	Example formats and standards
Single file	These solutions use specific file formats, which enable both the images and the text to be combined.	PDF, RTF
Files combined	The image and the text files exist separately, and are combined via external structural metadata that represent the complete logical and physical structure of the object (e.g. chapters, sections…)	Images: JPEG, PNG, a.o. Text: text/plain, HTML, TEI Metadata: METS
Files combined with coordinates	Similarly to above, but with coordinates of every term in the according image.	Images:JPEG, PNG, a.o. Text: text/plain, HTML, TEI Metadata: METS/ALTO

Table 2 Typical solutions for full-text representation

The data model to aggregate full-text object for Europeana must accommodate these three general solutions, given that a crosswalk between them is extremely complex and not practicable. However, given the full-text aggregation restrictions, which will be presented in the following section, end-user access to the digital objects must be done in the data provider websites.

Another relevant aspect to mention is the descriptive metadata creation practices used by libraries regarding full-text materials. For some of these materials, libraries do not create detailed descriptive metadata, either because it may be too expensive or because the retrieval of the digital object is based on the full-text and not on metadata. In these cases the descriptive metadata associated with a full-text object may consist only in a title or a date, and one *isPartOf* relation with another descriptive metadata record. These cases are frequent for newspapers issues, where each issue is represented as an independent full-text object, which may have only one metadata element with the *issued* date and the relation with another metadata record describing the newspaper.

Summing up, the data model for full-text must enable the representation of the URLs of the text files, and their corresponding access URLs for end-user views, or alternatively, the URL for access to the entry point of the digital object. Metadata associated with full-text objects may be extremely simple and require referencing to other metadata records.

4 Requirements Analysis

The full-text aggregation infrastructure being developed in Europeana Libraries considers two types of actors:

- Data providers, which hold the full-text objects and provide end user access to them.
- Aggregators, which harvest the full-text objects and provide search and retrieval services.

The full-text aggregation infrastructure being developed in Europeana Libraries, assumes the existence of several full-text providing libraries, whose full-text objects are aggregated by *The European Library*, and further made available to Europeana. In the context of this document, we consider the full-text requirements for *The European Library* and Europeana to be the same, so they are both referred to as aggregators.

The full-text aggregation within the Europeana environment must operate under the following system constraints:

C1) Digital objects are not persisted by an aggregator.
C2) End-user access to the digital objects happens at the data provider side.
C3) The exchange of data about the full-text contents should be based on the current interoperability framework of Europeana.

Due to the constraints C1) and C2), the aggregator may only provide a service based on textual versions of the digital object, which may differ from the actual version of the object that is made available to end users in the data provider's websites.

The constraint C3) implies that the exchange of data about the full-text contents must be based in the metadata about the digital objects made available to Europeana, therefore full-text metadata should be incorporated in ESE and in the future into the Europeana Data Model (EDM).

From the analysis of the existing full-text corpora, it is known that the textual versions of the digital objects may be structurally different from the access versions. For example, the access version may have page information and word coordinates within the page (such as a PDF file) while the textual version made available to the aggregator may consist in a plain text file that contains all the same textual content as the PDF but no page information and word coordinates are available.

For this scenario, we define the following requirements for a data model to support the exchange of full-text contents between a data provider and an aggregator:

R1) The availability of full-text must be stated explicitly in the metadata.
R2) The URLs to the views of the digital objects must be explicitly stated in the data.
R3) The full-text resources must be referenced via direct URLs to one or more files.
R4) When more than one full-text resource is associated with a digital object, it should be possible to represent their sequential order.
R5) URLs to access specific parts of the digital objects (for example, to a section or page) may be provided in the data.
R6) Metadata about full-text digital objects may be hierarchical, requiring referencing to other metadata records.

Although in some cases, the URL for accessing a digital object is enough to recognize the existence of full-text content, this is not possible for all file formats, therefore this must be made explicit in R1). For example, a PDF or HTML file may contain full-text contents or images or both. In order to allow the full-text harvesting software to make the decision to process a digital object, separate URLs specifying the full-text files must be provided. The same argument justifies requirement R2), since in some cases, the full-text files are created only for indexing, making them unusable for the end-users, therefore information exchange and presentation must be separated.

The full-text file formats supported, in context of requirement R3), are not totally clear at this point of the project. These will be defined in a later stage of the project, when all the technical and technological constraints of the full-text aggregation infrastructure are known.

When interacting with full-text digital objects, the information needs of the end-user may be restricted to logical parts of the digital object, such as a section, or a page. Requirement R5) reflects this type of information need refining the linking between the central search system at the aggregator, and the data provider websites which serve the content to end-users. These URLs must be associated with units of text provided as URLs, as stated in requirements R3). The sequential order of the full-text is relevant for end users and the aggregator to perform information extraction on the full-text, therefore the need for R4).

Requirement R6) originates from those cases where the metadata associated with full-text objects may be extremely simple and require referencing to other metadata records in order to allow the end-user to evaluate the relevance of the digital object to his information needs.

5 Proposal for an EDM Based Implementation

The current interoperability framework of Europeana is based on the exchange of metadata about the digital objects. Therefore, it makes sense to build on the existing framework to allow the exchange of digital object data to enable full-text aggregation. For this reason, full-text metadata should be incorporated into EDM. This section presents a proposal to use EDM to fulfil the requirements for data exchange for full-text aggregation, and the main design decisions taken. This section assumes that the reader is familiar with EDM, which at the time of writing is in version 5.2.2[2].

5.1 Proposed Model

This is restricted to the subset of EDM that is relevant to the full-text requirements presented in the previous section. Any parts of the EDM not mentioned are applicable to full-text objects in the same way as they are applicable to any other digital object.

Most of the requirements for representing full-text resources are already met by EDM. Nevertheless, for representing digital objects with full-text, it is required that a new class is added to EDM, and to define its relations with other EDM classes. Figure 1 presents the classes and relations that would allow the representation of full-text resources in EDM, and meet all the requirements presented in the previous section[3].

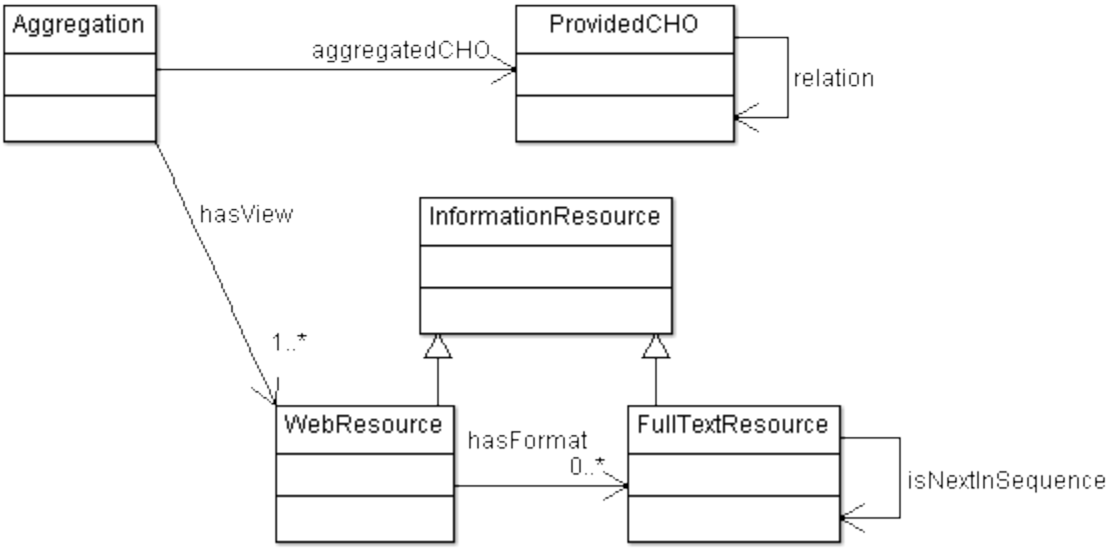

Figure 1 – Subset of the EDM relevant for full-text

[2] http://pro.europeana.eu/web/guest/edm-documentation

[3] For readability purposes, we omit the namespaces of the EDM classes and properties, since all which are used in this document have unambiguous names within EDM.

The additional class *FullTextResource* is a subclass of *InformationResource*. It allows the representation of the individual full-text content, separately from the end-user access versions, which are modelled in EDM in the class *WebResource*.

The *FullTextResource'*s are included in the digital object metadata by *hasFormat* relations from *WebResource*'s to *FullTextResource*'s. This allows the aggregators to index the full-text, and to provide the end-user with access to the views of the digital object. Table 3 describes how the requirements for full-text are met by this EDM based model.

Requirement	Description
R1 – When full-text is available for a digital object, it must be explicitly stated in the data.	This requirement is met by the existence of *FullTextResource* class, which explicitly states the existence of full-text content.
R2 – The URLs to the views of the digital objects must be explicitly stated in the data.	The *WebResource* class, when associated with an aggregation through a *hasView* relation meets this requirement. The *WebResource*'s URIs explicitly state how to access the digital object's views.
R3 – The full-text resources can be referenced via direct URLs to one or more files with textual content.	This requirement is met by the URI of the *FullTextResource*. The existence of multiple *FullTextResource*'s for a digital object is met by allowing a *WebResource* to have more than one *hasFormat* relation with *FullTextResource*'s.
R4 – When more than one full-text resource is associated with a digital object, it should be possible to represent their sequential order.	The *isNextInSequence* relation between two *FullTextResource*'s meets this requirement.
R6 – URLs to access specific parts of the digital objects (for example, to a section or page) may be provided in the data.	This requirement is met by the *hasFormat* relation between *WebResource*'s and the *FullTextResource*'s. The *WebResource* provides the view URL for every *FullTextResource* it has a *hasFormat* relation with.
R7 – Metadata about full-text digital objects may be hierarchical, requiring referencing to other metadata records.	This requirement is met by EDM through the possible types of relations between *ProvidedCHO*s, in particular the *isPartOf* and the *hasPart* relations.

Table 3 Requirements Matrix

5.2 Element Definitions

In scope of the “Definition of the Europeana Data Model elements” document[4], the proposed model for full-text requires a new element definition for the class *FullTextResource*, which is presented in Table 4.

Class name:	FullTextResource
Namespace	<to be assigned>
URI	<to be assigned>
Label	Full-text Resource
Definition	InformationResources that have at least one WebResource and an URI
Subclass of	edm:InformationResource
Obligation & Occurrence	The relation between Web Resources and the Full-text Resources is zero to many.
Example	A full-text resource containing the full-text of a book in machine-readable form.
Rationale	When full-text is available for a digital object, it must be represented in specific class, since in some cases, the Web Resource that provides the view of the digital object to the end-users, does not carry the full-text in machine-readable form. The full-text resources are associated with a Web Resource by the use of dcterms:hasFormat properties from the Web Resource to the Full-text Resource. The sequential order of Full-text resources within a cultural heritage object should be represented with edm:isNextInSequence.

Table 4 Definition of full-text resource

The property *edm:isNextInSequence* is defined in EDM v5.2.2 as applicable to any *rdf:resource*, therefore its use with *WebResource*’s and *FullTextResource*’s does not require any additions or modifications to EDM.

5.3 Impact on the EDM XML schema

The current version of the XML schema for EDM[5] would also require changes to accommodate the representation of full-text resources. Besides the addition of the class full-text resource (as described in Section 5.2), some further changes are required, since the XML schema defines heavy restrictions on the use of EDM classes and properties.

The required changes involve the properties *dcterms:hasFormat* and *edm:isNextInSequence*. In the current version of the XML schema for EDM, the property *dcterms:hasFormat* cannot be used in *WebResource*’s, therefore this use needs to be allowed within the definition of *WebResource*’s.

For the property *edm:isNextInSequence*, its use with *FullTextResource*’s must be included in the schema definition for *FullTextResource*’s.

6 Current Status of Full-text Delivery to Europeana

Since a final version of the EDM XML schema for XML representation of the metadata is not yet established, the full-text corpora aggregated by *The European Library* is being made available to Europeana in Europeana Semantic Elements (ESE), extended with an element from *The European Library*’s XML namespace, which holds the URLs for the full-text contents (*tel:fullText*).

[4] The EDM Specifications V5.2.2 are available at: http://pro.europeana.eu/web/guest/edm-documentation

[5] EDM xml schema available at: http://www.europeana.eu/schemas/edm/

The current status of the full-text delivery to Europeana, in terms of digital objects and full-text pages is presented in Table 5. At the time of writing of this document, December 2011, a total of 3,198,859 pages in full-text have been made available to Europeana.

Country of origin	Digital objects provided[6]	Pages provided	Total Pages
Austria			534,000
Czech Republic			2,579,511
Estonia	40,637	713,933	713,933
France			8,242,908
Hungary			237,914
Iceland			5,727,149
Latvia	21,190	195,075	195,075
Lithuania			125,477
Norway	7,242	1,600,000	1,600,000
Poland	2,964	436,198	436,198
Slovakia			185,000
Slovenia			328,502
Spain			3,033,525
Sweden	58,277	253,653	253,653
Total:	130,310	3,198,859	

Table 5 Full-text objects made available for Europeana

7 Conclusion and future work

This document presented the data exchange requirements for the full-text aggregation infrastructure for the Europeana context, which is being developed in Europeana Libraries. Most of the requirements for representing full-text resources are met by EDM, but one additional class is necessary to clearly represent full-text contents in EDM.

Currently, the full-text corpora, already aggregated by *The European Library*, are being processed for delivery to Europeana. A total of 3,198,859 pages are available to Europeana. This delivery is being made in ESE, until an implementation of the EDM based data model is established in *The European Library* and Europeana.

The proposed model is also an input for the full-text harvesting software being developed in work package 4, where it will be implemented.

[6] Empty field indicates that the collection is not yet transformed into ESE.