DELIVERABLE

Project Acronym: **E-ARK**

Grant Agreement Number: **620998**

Project Title: **European Archival Records and Knowledge Preservation**

DELIVERABLE DETAILS

DELIVERABLE REFERENCE NO.	**D5.3**
DELIVERABLE TITLE	**E-ARK Pilot DIP Specification**
REVISION	**2.0**

AUTHOR(S)	
Name(s)	Organisation(s)
Alex Thirifays (main author) Anders Bo Nielsen Ann-Kristin Egeland Kathrine Hougaard Edsen Johansen Phillip Tømmerholt	Danish National Archives (DNA)

Janet Delve Richard Healey	University of Brighton (UoB) University of Portsmouth (UPHEC)
István Alföldi Zoltán Lux	National Archives of Hungary (NAH)
Anja Paulič Jože Škofljanec Gregor Završnik	National Archives of Slovenia (NAS)
Kuldar Aas	National Archives of Estonia (NAE)

REVIEWER(S)	
Name(s)	Organisation(s)
Alex Thirifays Anders Bo Nielsen Tommy Balle	Danish National Archives (DNA)
Kuldar Aas	National Archives of Estonia (NAE)
Jože Škofljanec	National Archives of Slovenia (NAS)
Andrew Wilson	University of Brighton (UoB)
Janet Delve	University of Brighton (UoB)

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

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Submitted Revisions History

Revision No.	Date	Authors(s)	Organisation	Description
2.0	2016-05-01	Alex Thirifays Clive Billenness Janet Delve	DNA UoB	1st submitted version

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This deliverable contains original unpublished work except where clearly indicated otherwise.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

Table of Contents

Tables

Figures

1 Executive Summary

The primary aim of this deliverable is to present the *pilot* version of the E-ARK Dissemination Information Package (DIP) formats. The secondary aim is to describe the access scenarios in which these DIP formats will be rendered for use. The deliverable is a revision of the deliverable D5.2 E-ARK DIP Draft Specification, which was published in June 2015.

The pilot version of the DIP formats will be used in the E-ARK pilots that are specific to Access. Feedback from these pilots will be taken into account to amend the specification of the formats where necessary. The result of these alterations will result in the final DIP format specifications that are due in January 2017.

To frame the setting in which the DIPs will be processed, this deliverable briefly reiterates the high-level illustration of the Access flow described in deliverable D5.2. Conversely, the Access use cases from the same deliverable are too detailed to be repeated here.

Once the stage has been set, the ***DIP reference format*** as well as the ***DIP representation formats*** that have been conceived to preserve Content Information within the DIP reference format will be described.

Firstly, a description of the high-level requirements will be delivered in concordance with the four essential purposes of the DIP: to render content information; to manage Access Rights Information; to ensure Authenticity; and to enable exchange of information packages.

This is followed by a description of the ***DIP reference format***, which essentially consists of a specification of its physical folder structure and a semantic description of the three core metadata categories (structural (METS); preservation (PREMIS); and descriptive (EAD)).

The specification of the DIP reference format is followed by the specifications of the ***DIP representation formats*** for Electronic Records Management Systems (ERMS); for Simple File-System Based Records (SFSB); for relational databases; for data warehouses; and for geodata.

Access to archival records is largely dependent on the workflow, the use cases, and the access scenarios in which Access Software is deployed to render content information and associated metadata. Therefore a description of these **access scenarios** follows each of the sections that define the specific DIP representation formats.

2 Purpose and Method

The purpose of this deliverable is to enable the specification of the final E-ARK Dissemination Information Package (DIP) format[1],[2],[3] as well as set out the requirements needed for the development of Access Software[4]. This development will happen during the pilot period through the iterations that will trial the pilot version of the DIP format and the Access Software. It is important to underline the fact that this deliverable is the *pilot* DIP specification, not the *final* DIP specification[5].

The DIP reference format[6] represents the recommended practice for interoperable DIPs and can be applied across different Access Software and access systems. As such this format can in the future be supported as the default output format for preservation systems.

The DIP representation[7] formats are content specific implementations of the DIP reference format and offer examples of content information type[8] specific scenarios.

The current document is an official deliverable (D5.3) and has been developed by the partners of the E-ARK project. It is mainly based upon another deliverable (D5.2[9]), but also on other existing work and requirements that have been identified employing both a bottom-up and a top-down approach.

The bottom-up approach identified relevant requirements by investigating the common specification[10]; analysing best practices[11] and user needs[12]; examining the E-ARK SIP[13] and the E-ARK Archival Information

[1] Technical terms from OAIS, PREMIS, E-ARK, etc. can be found in Chapter 5 Glossary and will be available in the E-ARK Knowledge Center: http://kc.dlmforum.eu./home. The first time a term from the glossary is encountered in this deliverable (Executive Summary excluded) a definition of it will be provided in a footnote.

[2] The Dissemination Information Package is an Information Package, derived from one or more AIPs, and sent by Archives to the Consumer in response to a request to the OAIS. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf

[3] All OAIS terms are capitalised.

[4] A type of software that presents part of or all of the information content of an Information Object in forms understandable to humans or systems. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf

[5] Cf. Milestone 10, "Final release of E-ARK formats and tools", which is due in month 36 (end of the E-ARK project, January 31st 2017).

[6] The DIP reference format refers to the E-ARK container format which is conceived to store the content information and its associated metadata. It is to a large extent built on the E-ARK Common Specification (http://www.eark-project.com/resources/specificationdocs/50-draftcommonspec-1) and the E-ARK AIP Format (http://www.eark-project.com/resources/project-deliverables/53-d43earkaipspec-1).

[7] A representation is the set of files, including structural metadata, needed for a complete and reasonable rendering of an Intellectual Entity. For example, a journal article may be complete in one PDF file; this single file constitutes the representation. Another journal article may consist of one SGML file and two image files; these three files constitute the representation. A third article may be represented by one TIFF image for each of 12 pages plus an XML file of structural metadata showing the order of the pages; these 13 files constitute the representation. Source PREMIS: http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf, p.8.

[8] Content Information Types are the data types for which format specifications have been created, cf. Electronic Management Systems (ERMS), Simple File-Based Systems (SFBS), databases, and geo-data.

[9] D5.2 E-ARK DIP Draft Specification, http://www.eark-project.com/resources/project-deliverables/31-d52

[10] Internal E-ARK deliverable: E-ARK Draft Common Specification http://www.eark-project.com/resources/specificationdocs/50-draftcommonspec-1. The common IP specification for E-ARK IPs conceived to constitute a common basis for the E-ARK SIP, AIP and DIP Specifications.

Package[14] (AIP)[15]; querying the pilot sites[16]; as well as scrutinising metadata elements from various metadata standards[17].

The top-down approach consisted of creating high-level workflows and descriptions of the generic steps in the whole Access[18] process. Below is an illustration of the very top level of those workflows, which is detailed elsewhere[19]. The high-level illustration of the E-ARK Access process encompasses four main steps:

1. "Search & Order Management" where the Consumer[20] can search for; identify; and order information packages of interest, using a Finding Aid[21].
2. "DIP Preparation" where the IP is prepared for the end-user[22], for example by migrating an AIP into a DIP;
3. "DIP Delivery" where the DIP is delivered to the end-user via a suitable Graphical User Interface[23] (GUI);
4. "DIP Management" where the DIP is either deleted or sent to a permanent or temporary DIP storage.

[11] D3.1 E-ARK Report on Available Best Practices, http://www.eark-project.com/resources/project-deliverables/6-d31-e-ark-report-on-available-best-practices.
[12] D5.1 E-ARK GAP report between requirements for access and current access solutions http://www.eark-project.com/resources/project-deliverables/3-d51-e-ark-gap-report.
[13] D3.3 E-ARK SIP Pilot Specification, http://www.eark-project.com/resources/project-deliverables/51-d33pilotspec
D3.3 E-ARK SMURF, http://www.eark-project.com/resources/project-deliverables/52-d33smurf .
[14] An Archival Information Package, consisting of the content information and the associated Preservation Description Information (PDI), which is preserved within an OAIS. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf
[15] D4.3 E-ARK AIP Specification, http://www.eark-project.com/resources/project-deliverables/53-d43earkaipspec-1.
[16] D2.3 Detailed Pilots Specification, http://www.eark-project.com/resources/project-deliverables/60-23pilotsspec.
[17] Internal E-ARK deliverable: E-ARK DIP Format Requirements - not published, but available on request.
[18] The OAIS Access functional entity contains the services and functions which make the archival information holdings and related services visible to Consumers. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf
[19] For the detailed BPMN models of the Access flow, http://www.eark-project.com/resources/project-deliverables/5-d21-e-ark-general-pilot-model-and-use-case-definition. For both detailed illustrations and descriptions of the Access flow, see D5.2, http://www.eark-project.com/resources/project-deliverables/31-d52
[20] The role played by those persons or client systems, which interact with OAIS services to find preserved information of interest and to access that information in detail. This can include other OAIS's, as well as internal OAIS persons or systems. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf
In E-ARK "Consumer" is an umbrella term that designates all users of archival holdings, thus both internal users, cf. archivists, and external users, cf. end-user.
[21] A type of Access Aid that allows a user to search for and identify Information Packages of interest. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf
[22] The end-user designates an external user who seeks content information in archival holdings.
[23] A Graphical User Interface (GUI) is a graphical interface to a program on a computer. It takes advantage of the computer's graphics capabilities to make the program easier to use.

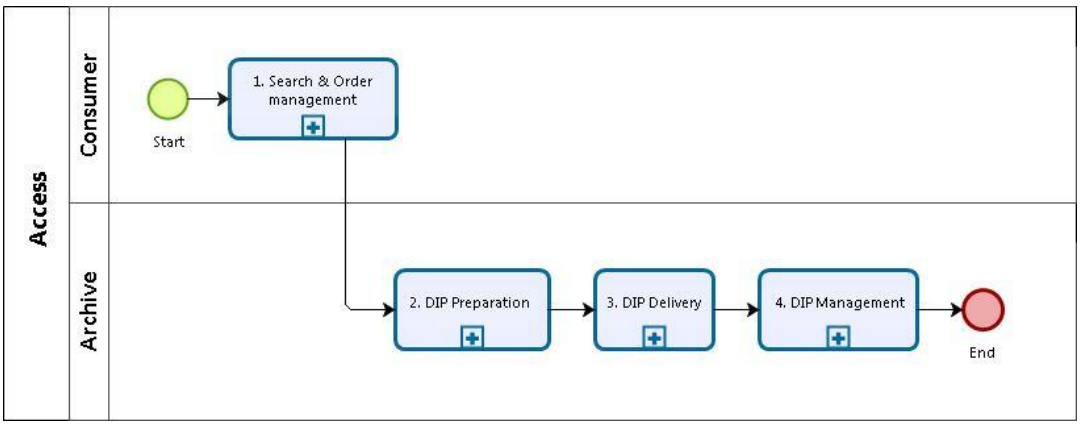

Figure 1 – Overview of the E-ARK Access process

This work contributed to reaching a common understanding between the archivists of the project as well as defining the scope of the Access activities that need to be underpinned by tools developed in E-ARK. Subsequently, the identification of these generic process steps enabled the creation of use cases that have:

1. served as communication platforms between archivists and developers and thus been used to facilitate the creation of a deployment environment facilitating agile development where short feedback cycles quickly rectify potential misconceptions; and
2. completed the identification of the specific requirements of the DIP format and of the Access Software.

Where appropriate the use cases were enhanced with acceptance criteria that in essence define quality goals (how will the product satisfy the user?).

The two approaches are complementary and were adopted to ensure that all requirements were taken into account in the development of the E-ARK DIP format and the Access Software.

In addition, feedback has been gathered from various sources such as virtual and physical E-ARK working group meetings, workshops, Advisory Board meetings, presentations at conferences, etc.

The target audiences of the present deliverable are:

1. The digital preservation practitioners of the archives that need to process an interoperable E-ARK DIP format;
2. The intended users of content information[24] in E-ARK DIPs. This can be both end-users (external users) and archivists (internal users);
3. The European Commission since this is an official deliverable of the E-ARK-project;

[24] The content information is the result of the association between the Content Data Object and its Representation Information, and this is what the Consumer requests from an archive. Together with the preservation description information, which includes reference information, provenance information, context information, access rights information and fixity information, the content information makes up an IP. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf

4. The developers of the E-ARK-project who develop tools that process the format.

3 Requirements for the E-ARK DIP format

This chapter presents the most crucial requirements of the E-ARK DIP specification. All other requirements of the E-ARK IP specifications can be found in the relevant documents, but by virtue of the inheritance principle, they also apply to the DIP format.

The DIP requirements have mainly been derived from five sources:

1. The E-ARK common specification which presents generic requirements for all E-ARK IP specifications
2. The SIP specification that focuses on facilitating the transfer of information from production systems.
3. The AIP specification that focuses on requirements that allow for long-term preservation.
4. The user needs identified in a previous study (D5.1).
5. The requirements identified from the Access use cases (D5.2).

Thus the requirements of the common specification, the SIP Format and the AIP Format have been evaluated in regard to their applicability in access scenarios[25]. The user needs and use cases have been used to draw out requirements missing in the three previously mentioned formats.

The resulting high-level requirements in this document have been grouped into four categories that are essential to Access:

1. Representation Information[26]
2. Access Rights Information[27]
3. Authenticity[28]
4. Exchange[29]

The requirements listed below are therefore requirements that cover these areas, and only these. This does not mean that the DIP format has been stripped of the requirements that cater for long-term preservation or other purposes, because, as mentioned, the DIP inherits only the appropriate aspects of both the SIP and

[25] Access scenario is used as a term to describe the environment, the DIP and the Access Software which altogether are used to render content information and associated metadata.
[26] Representation Information is metadata that that transforms a Digital Object into an Information Object and thereby making it understandable by a human being. It consists of Semantic and Structure Information. Source OAIS: http://public.ccsds.org/publications/archive/650x0m2.pdf
[27] The information that identifies the access restrictions pertaining to the content information, including the legal framework, licensing terms, and access control. It contains the access and distribution conditions stated within the Submission Agreement, related to both preservation (by the OAIS) and final usage (by the Consumer). It also includes the specifications for the application of rights enforcement measures. Source OAIS: http://public.ccsds.org/publications/archive/650x0m2.pdf
[28] The degree to which a person (or system) regards an object as what it is purported to be. Authenticity is judged on the basis of evidence. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf
[29] Refers to the DIP as an exchange format, and as such it is essential that it is possible to transfer DIPs, for example between a repository and various Access environments.

the AIP. For example, the DIP does not need to fulfil specific submission integrity or long-term preservation requirements. Therefore these requirements have been left out of this document and only apply to the SIP and AIP specifications respectively. Also, the Access Software requirements that guide the software development and allow for special tools to process the DIP formats[30] are not listed here, but are available on Redmine[31].

Some of the requirements would also be well suited to be placed in a category called 'Generic requirements', but since they all also have specific purposes, they have been categorised below. Lastly, note that some of the requirements overlap between categories.

3.1 Requirements for Representation Information

Representation Information allows the content information and associated metadata to be rendered and understood. It is therefore crucial for the reuse of any archival material and represents a key requirement category for access.

1. The DIP format **MUST** allow for the inclusion of Representation Information and any supplementary metadata that ensure the rendering, the understandability, and the usability of the DIP.

2. The structure and the metadata of the DIP **MUST** be human understandable and accessible with simple text-processing tools in case no dedicated tools are available to process it.

3. The DIP format **MUST** include an overview of the structure and its content.

4. The DIP format **MUST** allow for the generation of an overview of the structure, the metadata and the contents of the package.

5. The DIP format **MUST** include sufficient metadata for the content information of the DIP to be correctly understood.

6. The DIP format **MUST** ensure that efficient search and navigation in the DIP is possible.

7. The DIP format **MUST** allow for designated access tools to cater for an automatic rendering process of the DIP in appropriate GUIs.

8. The DIP format **MUST** allow for the identification of the DIP representation format of its data.

9. The DIP format **MUST** allow the indication of the Access Software with which it is currently to be accessed.

[30] There are several DIP format specifications depending on the nature of the content information, i.e. geodata has one specification, databases another, etc.

[31] https://e-ark-redmine.magenta-aps.dk/projects/wp5/issues. Access can be granted upon request.

3.2 Requirements for Access Rights Information

Access restrictions are inherent to a large part of existing archival material, and in many countries all archives are restricted by default, so it is crucial that the E-ARK project enables access control over both the content information and associated metadata[32].

1. The DIP format **MUST** allow for the inclusion of any information that is needed to ensure that access restrictions can be observed and administered. This includes, but is not restricted to:
    a. Information about the sensitivity of the data (i.e. personal data, business oriented restrictions etc.).
    b. Information about copyright restrictions.
    c. Information about specific reuse conditions (for example, if use of the content is only allowed in specific physical locations for a limited time period).
    d. Information about any additional restrictions.

2. The DIP format **MUST** allow for the tracking of events that change access restrictions.

3. The DIP format **MUST** allow for the specification of access rights for:
    a. The whole DIP;
    b. Metadata;
    c. Individual intellectual units and/or computer files.

4. The DIP format **MUST** specify Access Rights Information in a clear way, in both human and machine readable forms.

5. The DIP format **MUST** ensure that access restrictions are described and the description located in the package in a way which is uniquely understandable and discoverable by Access Software

3.3 Requirements for Authenticity

The Authenticity of archival records[33] is another key aspect to access, and it must be possible to accept a digital "object as what it is purported to be"[34] for purposes of research, legislative purposes, or other.

1. The DIP format **MUST** allow for the inclusion of any information that is needed to do manual and automatic Authenticity checks.

2. The DIP format **MUST** allow for the inclusion of any relevant metadata about its structure and content.

[32] D2.2 Legal Issues Report: European Cultural Preservation in a Changing Legislative Landscape, http://www.eark-project.com/resources/project-deliverables/33-d22-legal-issues-report-european-cultural-preservation-in-a-changing-legislative-landscape.

[33] Materials created or received by a person, family, or organization, public or private, in the conduct of their affairs that are preserved because of the enduring value contained in the information they contain or as evidence of the functions and responsibilities of their creator. Source Society of American Archivists http://www2.archivists.org/glossary/terms/a/archival-records#.VyB5VXqd9iN

[34] OAIS, page 1-9.

3. The DIP format **MUST** include information about the author and the time of its creation.

4. The DIP format **MUST** allow for including / logging information about any changes done to the IP during Ingest (SIP), Archival Storage (AIP) and Access (DIP)[35].

5. The DIP format **MUST** include information which allows for its validation and authentication.

6. The DIP format **MUST** allow for the inclusion of information about its DIP status[36].

3.4 Requirements for Exchange

The DIP is also an exchange format, and as such it is essential that it is possible to transfer DIPs, for example between a repository and various access environments.

1. The DIP format **MUST** allow for the use of any information and mechanisms that are needed to allow for the transfer of the DIP between archives and users and archives and archives

2. The DIP format **MUST**, to the largest possible extent, use internationally recognised and standardised metadata schemas.

3. The DIP metadata **MUST** allow for the validation of the structure and content of any information package in terms of integrity, fixity and syntax.

4. It **MUST** be possible to split the DIP in case it is too big to be carried on one media or in one message (for online exchange).

5. It **SHOULD** be possible to globally uniquely identify any DIP. It is recommended that the identification mechanism implemented at the repository provides for global uniqueness in identification of information packages in order to support a wider range of interoperability scenarios (for example, joining multiple repositories).

[35] Such a log is not relevant for many users, but if the user has the need to prove the Authenticity then certainly (s)he needs to be able to do so. As such the "MUST" addresses the possibility of including a log and not that all DIPs really MUST include such a log.

[36] The E-ARK DIP can have three statuses: See DIP_0, DIP_u and DIP_p, cf. the Glossary.

4 DIP Format Specification

The DIP format is the last in sequence of the three IP formats defined in the OAIS reference model. The two others, the SIP and the AIP format, have - in their E-ARK context - already been defined, as noted above. All three formats use the same common specification in order to ensure consistency and compatibility between them.

The definition of an E-ARK DIP[37] is that it corresponds to an IP which is ready to be processed by its designated Access Software; if it is not suited for automatic processing and rendering by its designated Access Software, it is not a DIP. This is a very generic, but handy, definition. To be more specific, it is important to state that since the process of AIP to DIP transformation is often not simple and the IP might go through many different steps of transformation, E-ARK has adopted a more pragmatic take on what the DIP is:

1. The IP which is sent (or is ready to be sent) to the user or access environment;
2. All E-ARK DIPs are supported by tools, i.e. are machine-readable and possible to be automatically rendered by relevant Access Software.

Firstly, we will explain the difference between the reference format and the representation formats ("4.1 DIP reference format vs. DIP representation formats"). This followed by a description of the AIP to DIP workflow ("4.2 Formats and Tools").

In section "4.3 The reference DIP" we will provide a description of the DIP format itself understood as a container with no content, i.e. the DIP reference format. This section will highlight the changes that occur as a consequence of the AIP to DIP conversion.

Lastly, the sections included in "4.4 Specifications for DIP representation formats and description of pertaining access scenarios" will describe how the DIP reference format should be applied for specific content information types, namely by defining the DIP representation format specifications. Currently there is a small selection of content information types, but everybody is welcome to create new type-based specifications as long as the requirements of the reference DIP are met.

4.1 DIP reference format vs. DIP representation formats

The DIP Format Specification has two layers: a generic layer which is not content information dependent, and a layer which is.

In reality, a DIP without its content information has no raison d'être. However, it makes sense to describe what it would look like because there are a number of traits that make it differ from the AIP and that are content information type agnostic. For example the simple fact that when a DIP has been created from one or several AIPs this needs to be recorded somewhere (in PREMIS) regardless of the content information type of the IP.

[37] A DIPu, cf. Glossary.

As opposed to the theoretical existence of the reference DIP, the DIP that holds content information is very real, and each content information type therefore has to be specified. The illustration below shows the Access perspective of the relation between the different specifications of the E-ARK project[38].

Generic	**Common specification →**	**Common specification**				
	IP reference formats →	**SIP, AIP, DIP**				
Content information	**DIP representation formats →**	**SMURF ERMS[39]**	**SMURF SFSB**	**SIARD[40]**	**SIARD; OLAP[41]**	**GML[42]; GeoTIFF[43]**
	Content information types →	**ERMS[44] & case files**	**SFSB[45]**	**Database[46]**	**Data warehouse[47]**	**Geodata[48]**

[38] The content information types do not correspond to an E-ARK format specification, but the E-ARK' Project's formal way to categorise the information types for which IP formats have been created.
[39] Semantically marked up records formats. SMURF is an IP format for ERMS systems and SFSB (simple file-system based records) conceived by the E-ARK project.
[40] Software Independent Archiving of Relational Databases. IP format for databases. Currently there exist three versions: SIARD1.0, SIARDDK and SIARD2.0.
[41] In computing, online analytical processing, or OLAP, is an approach to answering multi-dimensional analytical (MDA) queries swiftly. OLAP is part of the broader category of business intelligence, which also encompasses relational database, report writing and data mining. Typical applications of OLAP include business reporting for sales, marketing, management reporting, business process management (BPM), budgeting and forecasting, financial reporting and similar areas, with new applications coming up, such as agriculture. Source Wikipedia https://en.wikipedia.org/wiki/Online_analytical_processing
[42] The Geography Mark-up Language: the XML grammar defined by the Open Geospatial Consortium (OGC) to express geographical features. GML serves as a modelling language for geographic systems as well as an open interchange format for geographic transactions on the Internet.
[43] GeoTIFF is a public domain metadata standard which allows georeferencing information to be embedded within a TIFF file. The potential additional information includes map projection, coordinate systems, ellipsoids, datums, and everything else necessary to establish the exact spatial reference for the file.
[44] Electronic Records Management System is a type of content management system and refers to the combined technologies of document management and records management systems as an integrated system.
[45] Simple File-System Based Records. Simple file-system based records: records that contain simple file-system based folders or files, including those originating from content and data management systems, such as SharePoint, that are not based on true file systems. They address the submission of computer files or folders from the file Producers rather than from an ERMS. They require manual enrichment with additional descriptive metadata.
[46] A database is an organised collection of data. It is the collection of schemas, tables, queries, reports, views and other objects. Source: Wikipedia: https://en.wikipedia.org/wiki/Database
[47] In computing, a data warehouse (DW or DWH), also known as an enterprise data warehouse (EDW), is a system used for reporting and data analysis, and is considered as a core component of Business Intelligence [1] environment. DWs are central repositories of integrated data from one or more disparate sources. They store current and historical data and are used for creating analytical reports for knowledge workers throughout the enterprise. Examples of reports could range from annual and quarterly comparisons and trends to detailed daily sales analysis.
[48] Geodata is information about geographic locations that is stored in a format that can be used with a geographic information system (GIS). Geodata can be stored in a database, geodatabase, shapefile, coverage, raster image, or

Table 1 - Overview of the relation between content information types and the different specifications of the E-ARK project

The concept of "DIP representation format" is introduced in the last row of the figure, and the concept equates to the proper E-ARK sub-format for the different content information types. If an AIP e.g. consists of data which originated from an ERMS, the E-ARK sub-format or the "DIP representation format" will be that found in the SMURF ERMS.

The sections that describe the specific DIP representation formats will, in addition to the specifications themselves, include descriptions of the access scenarios in which the DIPs are rendered. These descriptions focus on the concrete implementations and uses of each content information type. It is important to remember that the access scenarios described are examples of use. As such, an archive may use QGIS[49] to render GML files, but other applications may also be used to accomplish the same task.

4.2 Formats and Tools

In order to understand the content information types and their relationships to other formats and tools of the E-ARK project, a table illustrating the information process is helpful: each row shows the format transformations that each content information type undergoes on its way to user consultation as well as the tools that are used to perform these transformations.

Content information types	AIP representation formats	DIP creation tools[50]	DIP representation formats	Access Software
Databases	SIARDDK[51]	DBPTK[52]; RDBMS[53]	SIARDDK	RDBMS; DB[54]-Viewer
	SIARD1.0	DBPTK; RDBMS	SIARD1.0	RDBMS
	SIARD2.0	DBPTK; RDBMS	SIARD2.0	RDBMS; DB-Viewer

even a dbf table or Microsoft Excel spreadsheet. Throughout this section we use the abbreviation "geodata" for geographical data.

[49] A Free and Open Source Geographic Information System. http://www.qgis.org/en/site/

[50] When 'Generic AIP2DIP converter' is indicated in this column it means that a local preservation system is employed to select and extract a specific AIP representation format from the AIP and then creates the corresponding DIP. In E-ARK, this can be the Repository of Authentic Digital Objects (RODA - http://www.roda-community.org/), the ESSArch Preservation Platform (EPP - http://epp.essarch.org/) or the E-ARK Web - https://earkdev.ait.ac.at:8443/cas/login?service (access can be granted upon request).

[51] DK is an abbreviation for "Denmark" – and this SIARD version a Danish variant of the original Swiss *SIARD1.0* format.

[52] The Database Preservation Tool Kit is a piece of software which, from an Access perspective, enables the loading of a SIARD file into an RDBMS http://keeps.github.io/db-preservation-toolkit/. It is developed by KEEP SOLUTIONS which is a partner of the E-ARK project http://www.keep.pt/en

[53] A relational database management system (RDBMS) is a computer software application that interacts with the user, other applications, and the database itself to capture and analyse data. A general-purpose RDBMS is designed to allow the definition, creation, querying, update, and administration of databases.

[54] Database.

Content information types	AIP representation formats	DIP creation tools[50]	DIP representation formats	Access Software
		DBPTK; MDDBMS[55]	OLAP Cube	MDDBMS
Data warehouse	SIARD 2.0	DBPTK; MDDBMS	OLAP Cube	MDDBMS
ERMS	SMURF ERMS	Generic AIP2DIP converter	SMURF ERMS	ERMS-Viewer
SFSB	SMURF SFSB	Generic AIP2DIP converter	SMURF SFSB	SFSB-Viewer
Geodata	GML[56] (Vector)	Generic AIP2DIP converter	GML	QGIS / GeoServer
	GeoTIFF (Raster)	Generic AIP2DIP converter	GeoTIFF / GML Frame	QGIS / GeoServer

Table 2 - Transformations that each content information type undergoes in the Access process

The ***content information types*** refer to the data types for which the E-ARK project is developing tools and formats. The content information types are ERMS systems, Simple File-System Based Records, databases and geodata. Databases may or may not contain binary files (e.g. a pdf); the ERMS systems are records management systems that always contain binary files and that may or may not be MoReq compliant; geodata are found in the Vector file format or Raster graphics; and lastly, the Simple File-System Based Records originate, for example, from a writer's hard drive or a politician's inbox, and can be anything form a Word file to an EML file.

The ***AIP representation formats*** refer to different technical representations of the intellectual entity[57] preserved in a repository. Each of these is contained in a representation folder of an IP. A simple example is that an AIP holds one representation of the intellectual entity for example the originally submitted .docx file and another representation of the same intellectual entity, for example a .pdf file.

The ***DIP* creation tools** correspond to the tools that are needed to create a DIP format from the AIP. Note that the table only indicates AIP to DIP conversion tools on the content information type level - not on the reference DIP level, cf. footnote 50.

The ***DIP representation formats*** designate the formats that have been prepared for Access and that are ready to be rendered by Access Software.

[55] A MultiDimensional DBMS is a particular kind of *RDBMS* that is specifically geared towards OLAP (in fact MDDBMS is often used co-terminously with OLAP).
[56] Geography Markup Language.
[57] A set of content that is considered a single intellectual unit for purposes of management and description: for example, a particular book, map, photograph, or database. An Intellectual Entity can include other Intellectual Entities; for example, a Web site can include a Web page; a Web page can include an image. An Intellectual Entity may have one or more digital representations. Source PREMIS http://www.digitizationguidelines.gov/term.php?term=intellectualentity

The ***Access Software*** lists the tools that render the different DIP representation formats to the Consumer.

4.3 The reference DIP

To a very large extent, the DIP is similar to the AIP from which it is created. However, the DIP is subject to a number of changes that are necessary in order to fulfil its purposes, which are described in Chapter 3 Requirements for the E-ARK DIP format.

First of all, the DIP looks like the AIP: The reference DIP replicates the structure of the AIP from which it is derived. It also inherits all the metadata as well as the intellectual entity of the AIP, regardless of any format migrations that may have occurred during the AIP-DIP conversion process.

Secondly, the DIP is different from the AIP: The DIP allows for example for the inclusion of new DIP representation formats, which are more user-friendly. It also allows for the updating of the metadata as well as for the addition of new metadata elements. Representation Information, which is required for rendering and understanding the intellectual content, is also added, and as a direct consequence, there may be a need for new folders and files, for example within the 'Documentation' folder.

4.3.1 DIP Data Model and Physical Folder Structure

The physical structure of the E-ARK DIP must comply with the principles outlined in the E-ARK common specification. The basic E-ARK information package structure as presented in the common specification is seen in Figure 2 below. It is encapsulated in a ZIP or a TAR file, and the top-most folder carries the unique name of the DIP. The DIP consists of a "METS.xml" file that reflects the structure of the DIP and provides an inventory of its content (Packaging Information[58]). The "metadata/" folder holds any metadata files (e.g. PREMIS, EAD), and the 'Representations' folder contains any number of representations of the content. Any number of necessary folders and files may be placed inside these folders, and optionally a "schemas/" folder and a "documentation/" folder may be introduced.

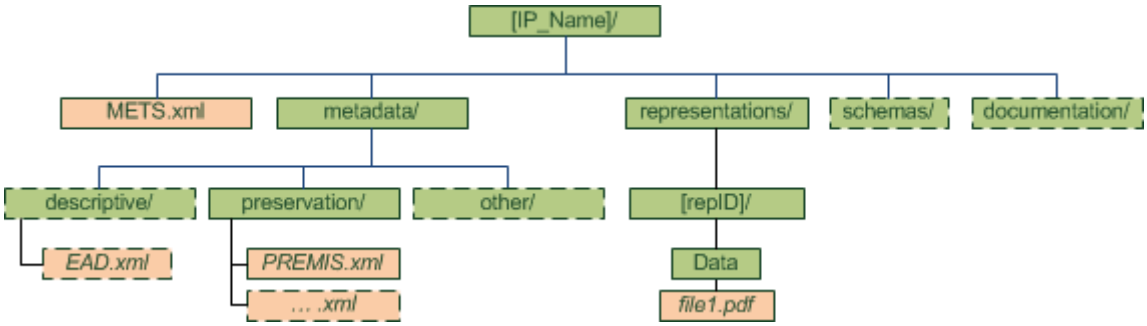

Figure 2– Minimum E-ARK IP structure requirements

The diagram above represents the minimal DIP structure. However, a more complex structure is sometimes required in order to be able to describe and render the content information properly. The specific

[58] The information that is used to bind and identify the components of an Information Package. For example, it may be the ISO 9660 volume and directory information used on a CD-ROM to provide the content of several files containing content information and Preservation Description Information. Source: OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf

requirements for the potentially more complex reference E-ARK IP structures are detailed in the common specification[59] and will not be reiterated here.

4.3.2 Metadata in the DIP

The description of the metadata builds upon the existing common, SIP and AIP specifications. The description is divided into three sections, which correspond to the three core metadata categories: structural[60] (METS[61]); preservation[62] (PREMIS[63]); and descriptive[64] (EAD[65])).

Metadata elements that are necessary for Access are highlighted below[66]. These Access specific metadata elements have been identified using the method which has already been described[67]. As such they have,

[59] Cf. "Introduction to the Common Specification for Information Packages in the E-ARK project", chapter 4.2 Full structure of the E-ARK Information Package, pages 24-25.

[60] Structural metadata describes the physical and/or logical structure of digital resources; it expresses the intellectual boundaries of complex objects and can be used to describe relationships between an object's component parts. Structural metadata is commonly used to facilitate navigation and presentation of complex items by defining structural characteristics such as pagination and sequence. And, like METS, can be used to aggregate related metadata. Source http://www.library.illinois.edu/dcc/bestpractices/chapter_11_structuralmetadata.html
The standard that E-ARK recommends for structural metadata is METS

[61] Metadata Encoding and Transmission Standard. The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language of the World Wide Web Consortium. The standard is maintained in the Network Development and MARC Standards Office of the Library of Congress, and is being developed as an initiative of the Digital Library Federation. Source http://www.loc.gov/standards/mets/

[62] Preservation metadata is an essential component of most digital preservation strategies. As an increasing proportion of the world's information output shifts from analog to digital form, it is necessary to develop new strategies to preserve this information for the long-term. Preservation metadata is information that supports and documents the digital preservation process. Preservation metadata is sometimes considered a subset of technical or administrative metadata. Source https://en.wikipedia.org/wiki/Preservation_metadata
The standard that E-ARK recommends for preservation metadata is PREMIS.

[63] The Preservation Metadata: Implementation Strategies. The PREMIS Data Dictionary for Preservation Metadata is the international standard for metadata to support the preservation of Digital Objects and ensure their long-term usability. Developed by an international team of experts, PREMIS is implemented in digital preservation projects around the world, and support for PREMIS is incorporated into a number of commercial and open-source digital preservation tools and systems. The PREMIS Editorial Committee coordinates revisions and implementation of the standard, which consists of the Data Dictionary, an XML schema, and supporting documentation. Source http://www.loc.gov/standards/premis/

[64] Also named Descriptive Information in OAIS: The set of information, consisting primarily of Package Descriptions, which is provided to Data Management to support the finding, ordering, and retrieving of OAIS information holdings by Consumers. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf
The standard that E-ARK recommends for descriptive metadata is EAD.

[65] Encoded Archival Description. A non-proprietary de facto standard for the encoding of Finding Aids for use in a networked (online) environment. Finding Aids are inventories, indexes, or guides that are created by archival and manuscript repositories to provide information about specific collections. While the Finding Aids may vary somewhat in style, their common purpose is to provide detailed description of the content and intellectual organization of collections of archival materials. EAD allows the standardization of collection information in Finding Aids within and across repositories.
http://www.loc.gov/ead/eadabout.html

[66] The final DIP specification, which is due in January 2017, will contain references to all metadata schemas and metadata sample files (thus for EAD, METS and PREMIS).

roughly speaking, been derived from the four main requirements categories listed in Chapter 3, Requirements for the E-ARK DIP format, namely Representation Information, Access Rights Information, Authenticity, and exchange. It is important to mention that the access tools have also contributed to the identification of access metadata elements

Access metadata elements that are specific to each DIP representation format will be described in the sections 4.4.1; 4.4.2; 4.4.3; and 4.4.4 that pertain to those formats.

4.3.2.1 Use of METS in an E-ARK Dissemination Information Package (DIP)

The central component of the E-ARK DIP is METS in the form of one or more METS XML files (and the METS XML Schema.)

A DIP must include one and only one METS file in the root folder of the package, named “METS.xml” and referred to as the "main root METS".
When the full structure (see the Common Specification) is used the package needs to include one additional “METS.xml” file in the root folder for each representation. These files will be referred to as “representation root METS” in the rest of this document.

When the IP is segmented (due to size) into IP segments the main root METS files (in the main IP) will refer to the representation root METS file(s) in the first segment of the representation(s).
The representation root METS file(s) in the first segment of the representation(s) will refer to the representation root METS in the following segments.

The representation root METS file(s) refer to the representations folder METS files.
Only the representations folder METS files refer to the data files.

The main requirement for METS files in an E-ARK Information Package is that these need to follow the official METS Schema version 1.11.

This chapter is structured according to the core METS elements: METS root element, header, dmdSec , amdSec, fileSec and structMap.

In each of these sections we explain in a concise way limitations imposed by the E-ARK implementation when compared to the official METS documentation. Also, differences between creating a root METS file and representation METS file are described.
The difference between the E-ARK AIP specification and the E-ARK DIP specification is described after each element and attribute in the "DIP diff "row.

4.3.2.1.1 Use of the METS element (<mets>)

The purpose of the METS element is to establish the container for the information being stored and/or transmitted, which is held within the sections of the METS file.

[67] Briefly in Chapter 2 Purpose and Method, and in detail in D5.2.

The *xsi:schemaLocation*[68] attribute of the METS element *<mets>* must refer to all necessary XML schemas. In the case of the recommended use of the "schemas" folder all schemas need to be referred to by relative path (for example: "schemas/mets.xsd" in the case of the main root METS.xml file and "../../schemas/mets.xsd" in the case of the representation level METS.xsd file).

The specific requirements for the root element and its attributes are described in the following table:

<element ATTRIBUTE DIP diff**	***Use in root folder METS.xml***	***Use in representations folder METS.xml***
<mets	mandatory by def.	mandatory by def.
ID	Optional, no further requirements	Optional, no further requirements
DIP diff	*If used it SHOULD be changed to another ID than in the AIP*	*If used it SHOULD be changed to another ID than in the AIP*
OBJID	Mandatory. Must be the same as the name or ID of the package (the name of the root folder). The OBJID must meet the Common Specification requirement of being unique at least across the repository	Mandatory. Must be the same as the ID of the representation (the name of the representation folder). The OBJID must meet the Common Specification requirement of being unique at least within the package
DIP diff	*new ID for DIP*	*new ID for DIP*
LABEL	Optional, if used should be filled with a human-readable description of the package	Optional, if used should be filled with a human-readable description of the representation
DIP diff	*If used it SHOULD be changed to another label than in the AIP*	*If used it SHOULD be changed to another label than in the AIP*
TYPE	Mandatory. The TYPE attribute must be used for identifying the OAIS type of the package (DIP) and the content type of the package (ERMS, RDB, SFSB, mixed). The value has to be expressed according to the following rule: *<OAISType>:<ContentType>.* Example: "*DIP:database"* Please note that the next version of the E-ARK IP specification will include specific vocabularies for the values of the TYPE attribute	Mandatory. The TYPE attribute must be used in a similar way as for the root METS file with the exception that instead of the OAIS type the first part of the attributes value is a fixed string "representation". Example: "representation:database"
DIP diff	*OAISType: DIP_0, DIP_u or DIP_p* *ContentType: SFSB, RDB, ERMS, GEODATA*	*none* *representation:* *ContentType: SFSB, RDB, ERMS, GEODATA*

[68]xsi stands here for the common namespace prefix of the schema at URL http://www.w3.org/2001/XMLSchema-instance

<element ATTRIBUTE DIP diff	*Use in root folder METS.xml*	*Use in representations folder METS.xml*
PROFILE	Mandatory. The PROFILE attribute has to be filled with the URL of the official E-ARK METS Profile. As this is not yet available the placeholder value to be used is “http://www.eark-project.com/METS/IP.xml”	Not used
DIP diff	*none*	*none*

Table 3 - METS element <mets>

4.3.2.1.2 Use of the METS header (<metsHdr>)

The purpose of the METS header section is to describe the METS document itself, for example information about the creator of the IP.

The requirements for the <metsHdr> element, its sub-elements and attributes are presented in the following table:

<element ATTRIBUTE DIP diff	*Use in root folder METS.xml*	*Use in representations folder METS.xml*
<metsHdr		
ID	Optional, no further requirements	Optional, no further requirements
DIP diff	*none*	*none*
ADMID	Optional, referring to the appropriate administrative metadata section if available	Optional, referring to the appropriate administrative metadata section if available
DIP diff	*none*	*none*
RECORDSTATUS	Optional, no further requirements	Optional, no further requirements
DIP diff	*none*	*none*
CREATEDATE	Mandatory, the date of creation of the package	Mandatory, the date of creation of the package
DIP diff	*new date – the creation of the DIP*	*new date – the creation of the DIP*
LASTMODDATE	Mandatory if relevant (in case the package has been modified)	Mandatory if relevant (in case the package has been modified)
DIP diff	*new date for DIP_u and DIP_p*	*new date for DIP_u and DIP_p*
<agent	The metsHdr must include at least one agent describing the software which has been used to create the package (TYPE=”OTHER”	Optional, no further requirements

<element ATTRIBUTE DIP diff	*Use in root folder METS.xml*	*Use in representations folder METS.xml*
	ROLE="CREATOR" OTHERTYPE="SOFTWARE"). Description of all other agents is optional.	
DIP diff	*name of the AIP2DIP software*	*name of the AIP2DIP software*
<altRecordID	Optional, no further requirements	Optional, no further requirements
DIP diff	*none*	*none*
ID	Optional, no further requirements	Optional, no further requirements
DIP diff	*none*	*none*
TYPE	Optional, no further requirements	Optional, no further requirements
DIP diff	*none*	*none*
<metsDocumentID	Optional, E-ARK recommends the value to be the same as OBJID.	Optional, E-ARK recommends the value to be the same as OBJID.
DIP diff	*same value as the DIP OBJID*	*same value as the DIP OBJID*

Table 4 - <metsHdr>

4.3.2.1.3 Use of the METS descriptive metadata section (<dmdSec>)

The purpose of the METS descriptive data section is to refer to the files containing descriptive metadata.

All descriptive metadata[69] must be placed as separate files into the metadata/descriptive folder and referenced using the <mdRef> element.

If the package includes multiple versions of the same metadata (as an example an EAD file created by the submitting entity and another version updated by the archives) these must be presented as separate <dmdSec> occurrences. In this case we also recommend using the STATUS attribute of the <dmdSec> element with values "current" or "superseded".

<element ATTRIBUTE DIP diff	*Use in root folder METS.xml*	*Use in representations folder METS.xml*
<dmdSec	Mandatory to include exactly one <dmdSec> which refers to all root folder descriptive metadata files	Mandatory to include exactly one <dmdSec> which refers to all representations folder descriptive metadata files
DIP diff	*none*	*none*
ID	Mandatory, identifier must be unique within the	Mandatory, identifier must be unique

[69] Also named Descriptive Information in OAIS: The set of information, consisting primarily of Package Descriptions, which is provided to Data Management to support the finding, ordering, and retrieving of OAIS information holdings by Consumers.

<element ATTRIBUTE DIP diff	*Use in root folder METS.xml*	*Use in representations folder METS.xml*
	package	within the representation
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*
STATUS	Mandatory, must include one of the two values “SUPERSEDED”; “CURRENT”	Mandatory, must include one of the two values “SUPERSEDED”; “CURRENT”
DIP diff	*none*	*none*
<mdRef	All references to the metadata files should be made using the XLink href attribute and the file protocol using the relative location of the file. This requires, in turn, the usage of the XLink type attribute with the value “simple”.	All references to the metadata files should be made using the XLink href attribute and the file protocol using the relative location of the file. This requires, in turn, the usage of the XLink type attribute with the value “simple”.
DIP diff	*none*	*none*
ID	Mandatory, identifier must be unique within the package	Mandatory, identifier must be unique within the package
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*
LOCTYPE	Mandatory, always using value “URL”	Mandatory, always using value “URL”
DIP diff	*none*	*none*
CREATED	Mandatory, used according to the official METS guidelines	Mandatory, used according to the official METS guidelines
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*
CHECKSUM	Mandatory, used according to the official METS guidelines	Mandatory, used according to the official METS guidelines
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*
CHECKSUMTYPE	Mandatory, used according to the vocabulary presented in the official METS schema	Mandatory, used according to the vocabulary presented in the official METS schema
DIP diff	*none*	*none*
MDTYPE	Mandatory, used according to the vocabulary presented in the official METS schema	Mandatory, used according to the vocabulary presented in the official
DIP diff	*none*	*none*

<element ATTRIBUTE DIP diff	*Use in root folder METS.xml*	*Use in representations folder METS.xml*
MIMETYPE	Mandatory, used according to the official METS guidelines	Mandatory, used according to the official METS guidelines
DIP diff	*none*	*none*
SIZE	Optional	Optional
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*

Table 5 - <dmdSec>

4.3.2.1.4 Use of the METS administrative metadata section (<amdSec>)

The purpose of the METS administrative data section is to refer to files containing this type of metadata.

Due to Common Specification requirement 3.2 (any Information Package should separate different types of metadata) all preservation metadata must be stored outside the METS.xml file and referenced by using the <mdRef> element and thus not embedded (i.e. the use of <mdWrap> element is not allowed).

The METS <amdSec> element must include references to all relevant metadata files located in the folder "metadata/preservation". This means also that the root level METS.xml file must refer only to the root level preservation metadata and the representations level METS.xml file must refer only to the representations level preservation metadata.

The E-ARK Information Package requires having all administrative metadata described in a single <amdSec> element (i.e. not repeatable).

The specific requirements for the <amdSec> element, its sub-elements and attributes are presented in the following table:

<element ATTRIBUTE DIP diff	*Use in root folder METS.xml*	*Use in representations folder METS.xml*
<amdSec	Mandatory to include exactly one <amdSec> which refers to all root preservation metadata files	Mandatory to include exactly one <amdSec> which refers to all representation preservation metadata files
DIP diff	*none*	*none*
ID	Mandatory, identifier must be unique within the package	Mandatory, identifier must be unique within the package
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*
<amdSec/	Mandatory to include one <digiprovMD> element	Mandatory to include one <digiprovMD>

<element ATTRIBUTE DIP diff	*Use in root folder METS.xml*	*Use in representations folder METS.xml*
<digiprovMD	for each file in the “metadata/preservation” folder.	element for each file in the “metadata/preservation” folder.
DIP diff	*none*	*none*
ID	Mandatory, identifier must be unique within the package	Mandatory, identifier must be unique within the package
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*
GROUPID	Optional, no further requirements	Optional, no further requirements
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*
ADMID	Optional, no further requirements	Optional, no further requirements
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*
CREATED	Not used	Not used
DIP diff	*none*	*none*
STATUS	Mandatory, must include one of the two values “SUPERSEDED”; “CURRENT”	Mandatory, must include one of the two values “SUPERSEDED”; “CURRENT”
DIP diff	*none*	*none*
<amdSec/ <digiprovMD / <mdWrap	Not used	Not used
DIP diff	*none*	*none*
<amdSec/ <digiprovMD / <mdRef	All references to the metadata files should be made using the XLink href attribute and the file protocol using the relative location of the file. This requires, in turn, the usage of the XLink type attribute with the value “simple”.	All references to the metadata files should be made using the XLink href attribute and the file protocol using the relative location of the file. This requires, in turn, the usage of the XLink type attribute with the value “simple”.
DIP diff	*none*	*none*
ID	Mandatory, identifier must be unique within the package	Mandatory, identifier must be unique within the package
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*

<element ATTRIBUTE DIP diff	*Use in root folder METS.xml*	*Use in representations folder METS.xml*
LOCTYPE	Mandatory, always using value “URL”	Mandatory, always using value “URL”
DIP diff	*none*	*none*
CREATED	Mandatory, used according to the official METS guidelines	Mandatory, used according to the official METS guidelines
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*
CHECKSUM	Mandatory, used according to the official METS guidelines	Mandatory, used according to the official METS guidelines
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*
CHECKSUMTYPE	Mandatory, used according to the vocabulary presented in the official METS schema	Mandatory, used according to the vocabulary presented in the official METS schema
DIP diff	*none*	*none*
MDTYPE	Mandatory, used according to the vocabulary presented in the official METS schema	Mandatory, used according to the vocabulary presented in the official
DIP diff	*none*	*none*
MIMETYPE	Mandatory, used according to the official METS guidelines	Mandatory, used according to the official METS guidelines
DIP diff	*none*	*none*
<amdSec/ <techMD	The use of <techMD> is not recommended. Instead, detailed technical metadata should be included into or referenced from appropriate PREMIS files	The use of <techMD> is not recommended. Instead, detailed technical metadata should be included into or referenced from appropriate PREMIS files
DIP diff	*none*	*none*
<amdSec/ <rightsMD	Optional. E-ARK recommends including a simple rights statement which describes the overall access status of the package (as an example with values: *open, closed, partially closed, not known*). However, the exact schema and element is up to individual implementations to decide	Optional. E-ARK recommends including a simple rights statement which describes the overall access status of the package (as an example with values: *open, closed, partially closed, not known*). However, the exact schema and element is up to individual implementations to decide
DIP diff	*none*	*none*
<amdSec/ <sourceMD	Optional, no further requirements	Optional, no further requirements

<element ATTRIBUTE DIP diff	*Use in root folder METS.xml*	*Use in representations folder METS.xml*
DIP diff	*none*	*none*

Table 6 - <amdSec>

4.3.2.1.5 Use of the METS file section (<fileSec>)

Use of the METS <fileSec> element is highly recommended by E-ARK (though not mandatory). It should describe all files within the package which have not been included in the <amdSec> and <dmdSec> elements. For all files the location and checksum need to be provided. Therefore the main purpose of the METS file section is to serve as a "table of contents" or "manifest" (the latter is the term that E-ARK is using) and allow validating the integrity of the files included in the package.

The main requirement of the E-ARK IP specification is that the file section (<fileSec> element) of both the root folder and representations folder METS files include at least one file group (<fileGrp> element). This so-called "E-ARK file group" sshould follow the requirements below:

- The file group should be defined by a single <fileGrp> element
  - It is mandatory to use the USE attribute with a fixed value of either "E-ARK files root" (for the root folder METS file) or "E-ARK files representation [representation ID]" (for the representations folder METS file)
  - *Example: <fileGrp USE="E-ARK files root">*
  - Each of the structural components (i.e. documentation, schemas, data) should be described by its own nested <fileGrp> element
  - The value of the USE attribute of the nested <fileGrp> element should reflect the name of the folder (i.e. USE="documentation"; USE="data"; USE="schemas");
- The data files of a representation should be described only in the representation METS. The root METS file should still include a <fileGrp> for each representation but only describe the representation METS file in it;

The specific requirements for elements, sub-elements and attributes are listed in the following table:

<element ATTRIBUTE DIP diff	*Use in root folder METS.xml*	*Use in representations folder METS.xml*
<fileSec	Recommended to include one <fileSec> element into each METS file	Recommended to include one <fileSec> element into each METS file
DIP diff	none	none
<fileSec/ <fileGrp	Recommended to include one E-ARK defined <fileGrp> element. Implementers are welcome to define and add additional file groups necessary for internal purposes.	Recommended to include one E-ARK defined <fileGrp> element. Implementers are welcome to define and add additional file groups necessary

<element ATTRIBUTE DIP diff	*Use in root folder METS.xml*	*Use in representations folder METS.xml*
		for internal purposes.
DIP diff	none	none
ID	Mandatory, identifier must be unique within the package	Mandatory, identifier must be unique within the package
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*
USE	Mandatory, value must be "E-ARK files root"	Mandatory, value must be "E-ARK files representation [representation ID]
DIP diff	*none*	*none*
<fileSec/ <fileGrp/ <fileGrp/	The main <fileGrp> element includes additional nested <fileGrp> elements, one for each folder of the package (except metadata described in <amdSec> and <dmdSec>)	The main <fileGrp> element includes additional nested <fileGrp> elements, one for each folder of the package (except metadata described in <amdSec> and <dmdSec>)
DIP diff	*none*	*none*
ID	Mandatory, identifier must be unique within the package	Mandatory, identifier must be unique within the package
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*
USE	Mandatory, value must be the same as the name of the folder (schemas, documentation, data, etc)	Mandatory, value must be the same as the name of the folder (schemas, documentation, data, etc)
DIP diff	*none*	*none*
<fileSec/ <fileGrp/…/file	Each file within the folders described by <fileGrp> elements by one <file> element	Each file within the folders described by <fileGrp> elements by one <file> element
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*
MIMETYPE	Mandatory	
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*
USE	Optional, no further requirements	
DIP diff	*none*	*none*
CHECKSUMTYPE	Mandatory, values according to the official METS guidelines	

<element ATTRIBUTE DIP diff	*Use in root folder METS.xml*	*Use in representations folder METS.xml*
DIP diff	*none*	*none*
CREATED	Mandatory	
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*
CHECKSUM	Mandatory	
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*
ID	Mandatory, must be unique across the package	
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*
SIZE	Mandatory	
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*
<fileSec/ <fileGrp/…/ <file/ <FLocat	The location of each file must be defined by the <FLocat> element using the same rules as for referencing to metadata files. All references to files should be made using the XLink href attribute and the file protocol using the relative location of the file. The XLink type attribute is used with the fixed value “simple”. The LOCTYPE attribute is used with the fixed value ”URL”	The location of each file must be defined by the <FLocat> element using the same rules as for referencing to metadata files. All references to files should be made using the XLink href attribute and the file protocol using the relative location of the file. The XLink type attribute is used with the fixed value “simple”. The LOCTYPE attribute is used with the fixed value ”URL”
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*

Table 7 - <fileSec>

4.3.2.1.6 Use of the METS structural map (<structMap>)

The purpose of the METS structural map section is to provide an overview of ALL components of an E-ARK Information Package. It also links the elements of that structure to associated content files and metadata. It is a mandatory and ultimate means to define the full structure of the package – including metadata, representations, schemas, documentation and user added components and folders. In other words, E-ARK tools will count on the information available within the <structMap> element as the primary means of identifying all components of the package. As such it is the most crucial component for the validation of any E-ARK Information Package and must always be present.

The E-ARK Information Package requires the inclusion of one structural map according to the principles described below. However, implementers are welcome to define additional structural maps for their internal purposes by repeating the <structMap> element. These additional structural maps are not exploited in E-ARK tools.

The most crucial requirements for the E-ARK mandated structural map are as follows:

- The <structMap> element has a mandatory attribute LABEL which has the fixed value of "E-ARK structural map". The LABEL attribute is used to distinguish the E-ARK structural map from any other, user-defined, structural maps. As such, we can also derive the requirement, that any user-defined structural maps must not use the LABEL value "E-ARK structural map";
- The internal structure of the structural map (expressed by hierarchical elements) follows the E-ARK physical structure as described in chapter 4, thereby grouping together metadata, representations, schemas, documentation and user-defined folders,
    - All elements must use the attribute LABEL with the value being the name of the folder (as an example "metadata")
- The structural map in the root METS file
    - Lists all files in all folders with the exception of the content of the representation folders
    - Lists all representations (as separate elements)
    - Lists only the appropriate METS file using the <mptr> element as the content of the representation
- The structural map in the representation METS file lists all files within the representation with no exceptions

The specific requirements for elements, sub-elements and attributes are listed in the following table:

<element ATTRIBUTE DIP diff	***Use in root folder METS.xml***	***Use in representations folder METS.xml***
<structMap	Each METS file needs to include exactly one <structMap> element which is used exactly as described in this table. Institutions can add their own custom structural maps as separate <structMap> elements next to it	same req. as root
DIP diff	*none*	*none*
ID	Optional, if used must be unique within the packagesame req. as root	same req. as root
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*
TYPE	Mandatory, value must be "physical"	same req. as root
DIP diff	*none*	*none*

<element ATTRIBUTE DIP diff	*Use in root folder METS.xml*	*Use in representations folder METS.xml*
LABEL	Mandatory, value must be “E-ARK structural map”	same req. as root
DIP diff	*none*	*none*
structMap/div	Each folder (and sub-folder) within the package must be represented by an occurrence of the element. Please note that sub-folders must be represented as nested div elements.	same req. as root
DIP diff	*none*	*none*
ID	Mandatory, identifier must be unique within the package	same req. as root
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*
ORDER	Not used	same req. as root
DIP diff	*none*	*none*
ORDERLABEL	Not used	same req. as root
DIP diff	*none*	*none*
LABEL	Mandatory, value must be the name of the folder („metadata“, „descriptive“, „schemas“, „representations“, etc). The LABEL value of the first element in the package is the ID of the package	Mandatory, value must be the name of the folder („metadata“, „descriptive“, „schemas“, „data“, etc). The LABEL value of the first element in the package is the ID/name of the representation
DIP diff	*none*	*none*
DMDID	No specific requirements	same as root
DIP diff	*none*	*none*
ADMID	No specific requirements	same as root
DIP diff	*none*	*none*
TYPE	No specific requirements	same as root
DIP diff	*none*	*none*
structMap/div/…/div/fptr	If the folder which is described by the element includes computer files these must be referenced by using the <fptr> element. The only exception is the description of representations (see below for the use of <mptr>). The <fptr> child elements <par>, <seq> and	Inside the representations folder METS file <fptr> element is used to reference all files within the representation with no exceptions. The <fptr> child elements <par>, <seq> and <area> must not be used.

<element ATTRIBUTE DIP diff	*Use in root folder METS.xml*	*Use in representations folder METS.xml*
	<area> must not be used.	
DIP diff	*none*	*none*
ID	No specific requirements	same as root
DIP diff	*none*	*none*
FILEID	Mandatory, must be the ID used in the appropriate <file> or <mdRef> element	same as root
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*
CONTENTIDS	No specific requirements	same as root
DIP diff	*none*	*none*
structMap/div/div/ mptr	In the case of describing representations within the package (i.e. representations/representation1) the content of the representations must not be described. Instead the of the specific representation should include one and only one occurrence of the <mptr> element, pointing to the appropriate representation METS file. The references to representation METS files must be made using the XLink href attribute and the file protocol using the relative location of the file. The XLink type attribute is used with the fixed value “simple”. The LOCTYPE attribute is used with the fixed value ”URL”	Not used
DIP diff	*same req. as for AIP - i.e. new value for DIP*	*same req. as for AIP - i.e. new value for DIP*
ID	Not used	same as root
DIP diff	*none*	*none*
CONTENTIDS	Not used	same as root
DIP diff	*none*	*none*

Table 8 - <structMap>

4.3.2.2 Use of PREMIS in an E-ARK Dissemination Information Package (DIP)

PREMIS is mainly a standard that caters for long-term preservation and technical usability, which for example is used to facilitate a range of preservation strategies including migration and emulation.

From an Access perspective, PREMIS especially satisfies the needs pertaining to the recording of Representation Information, but also to internal needs for managing rights information as well as Authenticity - including the tracking of preservation and access actions, for example the AIP to DIP conversion.

The E-ARK project recommends the use of PREMIS for preservation metadata, because it is the de facto standard in this field. From an Access perspective it is practical to state in a formalised and consistent way how the Access Software should behave and where it should look when dealing with different pieces of information, such as which representation formats are included in the DIP. Therefore all E-ARK Access Software assumes the availability of PREMIS metadata according to the specification below.

E-ARK has identified no semantic unit extensions to the PREMIS standard. The E-ARK project therefore recommends the use of PREMIS version 3.0[70] as is. In addition, PREMIS 3.0 is considered to be more pertinent for Access than its predecessor (PREMIS version 2.2) because it is more flexible in regards to the recording of Representation Information.

Below we have highlighted the semantic units that need to be implemented according to E-ARK Access specifics.

4.3.2.2.1 Metadata regarding Representations and Access Software

In PREMIS, a representation is a "set of files, including structural metadata, needed for a complete and reasonable rendition of an Intellectual Entity."[71] In E-ARK Access, as already mentioned, the DIP representation formats are SMURF ERMS, SMURF SFSB, SIARD1.0, SIARD2.0, SIARDDK, OLAP, GML, and GeoTIFF. In PREMIS, a representation is indicated using the semantic unit "1.1 objectIdentifier".

It is important to emphasise that the E-ARK project has neither created specifications nor tools for specific file formats, but only for the aforementioned DIP representation formats.

Hence, the Access Software developed by the E-ARK project does guarantee the rendition of the E-ARK representations, but not of specific file formats contained in an E-ARK representation. As an example, the SMURF ERMS could contain several file formats unknown to the E-ARK ERMS Viewer[72], even though this is unlikely, because archives generally make sure that the number of file formats that they preserve is limited and their use widespread.

Everything needed to describe the representation is already taken care of in the AIP, and the only other piece of information needed from an Access perspective is information about rendition (Representation Information). To enable rendition, three pieces of information are needed in PREMIS: One identifying the representation to be rendered; one identifying the software to enable this; and one establishing a relationship between the two.

[70] http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf
[71] Cf. PREMIS 3.0 page 8, https://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf
[72] GUI conceived by the E-ARK project to view ERMS systems.

The below descriptions therefore show how to:

1. Describe which DIP representation format is contained in the DIP (description 1);
2. Describe which piece(s) of Access Software are needed to render a specific DIP representation format. Several pieces of software may indeed be needed (description 2);
3. Describe the relationship between the DIP representation format and its Access Software (description 3).

Description 1 - The recording of representation formats

In order to describe the specific DIP representation format the semantic component "1.5.4 format" is used. The semantic component "1.5.4 format" is conceived to describe which file format a certain file or bitstream has. In the E-ARK project, the requirement is to record information about the DIP representation format, and not file formats. However, since the DIP representation format can be rendered using a specific piece of software, these DIP representation formats are treated as file formats (compound objects[73]).

Therefore, the E-ARK project has opted for using the semantic component <format> to describe the DIP representation format. An example is:

```
<object xsi:type="file">
  <objectIdentifier>
    <objectIdentifierType>DIP representation format</objectIdentifierType>
    <objectIdentifierValue>Database</objectIdentifierValue>
  </objectIdentifier>
  <objectIdentifier>
    <objectIdentifierType>repository</objectIdentifierType>
    <objectIdentifierValue>uuid:35c870ee-da2b-4s2c-8700-g5148a0e8g5g</objectIdentifierValue>
  </objectIdentifier>
  <objectCharacteristics>
    <format>
      <formatDesignation>
        <formatName>SIARD</formatName>
        <formatVersion>2.0</formatVersion>
      </formatDesignation>
    </format>
  </objectCharacteristics>
</object>
```

Figure 3 – PREMIS example of DIP representation format

Note that in order to describe the DIP representation format it is required that two objectIdentifiers for the "file" object must be present in PREMIS.xml:

objectIdentifier 1:

- the mandatory 1.1.1 objectIdentifierType-element must have the value "DIP representation format".

[73] The closest we get in PREMIS to defining the E-ARK representation formats is that they are compound objects: »Digital Object composed of multiple Files: for example, a Web Page composed of text and image Files«.

- the mandatory 1.1.2 objectIdentifierValue-element must have the content information type as its value. Possible objectIdentifierValues within E-ARK scope are: “ERMS, “SFSB”, “Database”, “Data warehouse”, “Geodata”, but other content information types can be implemented.

objectIdentifier 2:

- the mandatory 1.1.1 objectIdentifierType-element must have the value “repository”.
- the mandatory 1.1.2 objectIdentifierValue-element must have the UUID for the representation folder as its value.

Description 2 - The recording of Access Software

In PREMIS 3.0 a description of an environment has become an object itself, so that both non-environmental objects and environmental objects exist. Access Software is therefore an environmental object which per default is an intellectual entity[74].

The semantic unit “1.9 environmentFunction” is conceived to describe the environment object(s) with which the non-environment object is rendered. The E-ARK Access Software for rendering the different E-ARK DIP representation formats is:

1. ERMS and SFSB Viewer[75] (for SMURF ERMS and SMURF SFSB);
2. MDDBMS (for OLAP);
3. DBPTK, RDBMS and DB Viewer[76] (for databases);
4. QGIS and Geoserver (for GML and GeoTIFF).

Since it is not always possible to render the DIP representation formats with one piece of Access Software, it is necessary to model software dependencies and sequences between several pieces of software in PREMIS. As an example the following Figure 4 and Figure 5 entail descriptions of two pieces of software (DBPTK and RDBMS) that are needed to render a specific DIP representation format (SIARD). Additionally the second piece of software (RDBMS) to be used can be chosen from several products (here PostgreSQL and MySQL):

[74] See http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf, p.251.
[75] GUI conceived by the E-ARK project to view Single File-Based System Records.
[76] The Database Viewer is a GUI conceived by the E-ARK project to view and analyse databases.

```
<object xsi:type="intellectualEntity">
  <objectIdentifier>
    <objectIdentifierType>local</objectIdentifierType>
    <objectIdentifierValue>DBPTK</objectIdentifierValue>
  </objectIdentifier>
  <environmentFunction>
    <environmentFunctionType>software</environmentFunctionType>
    <environmentFunctionLevel>1</environmentFunctionLevel>
  </environmentFunction>
  <environmentFunction>
    <environmentFunctionType>software application</environmentFunctionType>
    <environmentFunctionLevel>2</environmentFunctionLevel>
  </environmentFunction>
  <environmentDesignation>
    <environmentName>DBPTK</environmentName>
    <environmentVersion>2.4.1</environmentVersion>
    <environmentDesignationNote>Documentation at github.com/eark-project/software/DBPTK</environmentDesignationNote>
  </environmentDesignation>
</object>
```

Figure 4 – PREMIS example of software 1: DBPTK

```
<object xsi:type="intellectualEntity">
  <objectIdentifier>
    <objectIdentifierType>local</objectIdentifierType>
    <objectIdentifierValue>RDBMS</objectIdentifierValue>
  </objectIdentifier>
  <environmentFunction>
    <environmentFunctionType>software</environmentFunctionType>
    <environmentFunctionLevel>1</environmentFunctionLevel>
  </environmentFunction>
  <environmentFunction>
    <environmentFunctionType>software application</environmentFunctionType>
    <environmentFunctionLevel>2</environmentFunctionLevel>
  </environmentFunction>
  <environmentDesignation>
    <environmentName>PostgreSQL</environmentName>
    <environmentVersion>9.1.21</environmentVersion>
    <environmentDesignationNote>Documentation at http://www.postgresql.org/docs/</environmentDesignationNote>
  </environmentDesignation>
  <environmentDesignation>
    <environmentName>MySQL</environmentName>
    <environmentVersion>5.7</environmentVersion>
    <environmentDesignationNote>Documentation at http://dev.mysql.com/doc/</environmentDesignationNote>
  </environmentDesignation>
</object>
```

Figure 5 – PREMIS example of software 2: RDBMS

Description 3 - The recording of the relation between the representations and the Access Software

In order to establish a connection between the DIP representation format to be rendered and the Access Software to render it, it is necessary to use the semantic unit “1.13 relationship”.

The relationship element can bind both non-environmental objects together with environmental objects and it can bind environmental objects together with other environmental objects. In the case of the presented example the file object from Figure 3 will have a relationship to the intellectual entity in Figure 4. See Figure 6 below how this is done.

```
<object xsi:type="file">
  <objectIdentifier>
    <objectIdentifierType>DIP representation format</objectIdentifierType>
    <objectIdentifierValue>Database</objectIdentifierValue>
  </objectIdentifier>
  <objectIdentifier>
    <objectIdentifierType>repository</objectIdentifierType>
    <objectIdentifierValue>uuid:35c870ee-da2b-4s2c-8700-g5148a0e8g5g</objectIdentifierValue>
  </objectIdentifier>
  <objectCharacteristics>
    <format>
      <formatDesignation>
        <formatName>SIARD</formatName>
        <formatVersion>2.0</formatVersion>
      </formatDesignation>
    </format>
  </objectCharacteristics>

  <relationship>
    <relationshipType>dependency</relationshipType>
    <relationshipSubType>requires</relationshipSubType>
    <relatedObjectIdentifier>
      <relatedObjectIdentifierType>local</relatedObjectIdentifierType>
      <relatedObjectIdentifierValue>DBPTK</relatedObjectIdentifierValue>
    </relatedObjectIdentifier>
    <relatedEnvironmentPurpose>render</relatedEnvironmentPurpose>
  </relationship>
</object>
```

Figure 6 – PREMIS linking from format to software

As can be seen in Figure 6 the nature of the relationship, <relationshipType>[77] is used (value, e.g. 'dependency'); intimately linked to this it is also the indication of a <relationshipSubType>[78], e.g. 'requires'.

In order to identify the Access Software, which is used to render the representation, the <relatedObjectIdentifier> is employed; and the <relatedEnvironmentPurpose> gives us a hint about what the purpose is (here: to 'render').

The Database Preservation Toolkit (DBPTK) also has a relationship, this time to another environmental object, which is a RDBMS. This can be seen in Figure 7:

[77] Controlled vocabulary: http://id.loc.gov/vocabulary/preservation/relationshipType.html
[78] Controlled vocabulary: http://id.loc.gov/vocabulary/preservation/relationshipSubType.html

```
<object xsi:type="intellectualEntity">
  <objectIdentifier>
    <objectIdentifierType>local</objectIdentifierType>
    <objectIdentifierValue>DBPTK</objectIdentifierValue>
  </objectIdentifier>
  <environmentFunction>
    <environmentFunctionType>software</environmentFunctionType>
    <environmentFunctionLevel>1</environmentFunctionLevel>
  </environmentFunction>
  <environmentFunction>
    <environmentFunctionType>software application</environmentFunctionType>
    <environmentFunctionLevel>2</environmentFunctionLevel>
  </environmentFunction>
  <environmentDesignation>
    <environmentName>DBPTK</environmentName>
    <environmentVersion>2.4.1</environmentVersion>
    <environmentDesignationNote>Documentation at github.com/eark-project/software/DBPTK</environmentDesignationNote>
  </environmentDesignation>
  <relationship>
    <relationshipType>dependency</relationshipType>
    <relationshipSubType>requires</relationshipSubType>
    <relatedObjectIdentifier>
      <relatedObjectIdentifierType>local</relatedObjectIdentifierType>
      <relatedObjectIdentifierValue>RDBMS</relatedObjectIdentifierValue>
    </relatedObjectIdentifier>
    <relatedEnvironmentPurpose>render</relatedEnvironmentPurpose>
  </relationship>
</object>
```

Figure 7 – PREMIS link from software to software

4.3.2.2.2 Metadata regarding Access specific Events

The PREMIS Event Entity holds information about actions that Digital Objects[79] are subject to.

From an Access perspective, it is important to record information about the actions that pertain to:

1. **Digital migrations**[80]
    a. an object is migrated from an AIP to a DIP. This will be done by using the semantic unit 2.2 <eventType> with the value “creation”[81];
    b. an object is migrated and re-ingested for preservation reasons. This will be done by using the semantic unit 2.2 <eventType> with the value “migration”;
2. **Digital provenance**[82]. The ability to record the history of custody of an object is important to Authenticity. This is done using the semantic unit 2.6 <linkingAgentIdentifier> and link it to specific

[79] An object composed of a set of bit sequences. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf

[80] Cf. OAIS.

[81] Cf. http://id.loc.gov/search/?q=&q=cs%3Ahttp%3A%2F%2Fid.loc.gov%2Fvocabulary%2Fpreservation%2FeventType

[82] Cf. Documentation of processes in a Digital Object’s life cycle. Digital provenance typically describes Agents responsible for the custody and stewardship of Digital Objects, key Events that occur over the course of the Digital Object’s life cycle, and other information associated with the Digital Object’s creation, management, and preservation. Source PREMIS: http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf

agents. If an IP has had several different custodians over time, this complexity is expressed by using the semantic unit 1.13 <relationship>;

3. **Assembling of data**. If for example several AIPs are assembled to constitute one DIP, this can to be tracked by using the semantic unit 1.13 <relationship>.

Dissemination for end-user purposes will not be registered in PREMIS, but can be registered in the local archives' Data and User Management System. Metadata pertaining to end-user activity is described in section 4.3.3.2 DIP Access

4.3.2.2.1 Metadata regarding rights and roles

Regarding Access Rights Information, PREMIS mainly addresses the handling of intellectual property rights, and supports these using the semantic components 4.1.3 <copyrightInformation>, 4.1.4 <licenseInformation> and 4.1.5 <statuteInformation>. The semantic unit 4.1 <rightsStatement> that governs these components has had its scope widened in PREMIS 3.0 to include the recording of other permissions, for example information pertaining to confidentiality (cf. access dates, i.e. is this IP publicly available in 20 years, in 75 years?)[83]. However, the PREMIS rightsStatement is for internal use, and as such does not touch upon Access issues[84]. That is why E-ARK recommends recording this piece of information in the EAD tag <accessrestrict>, the advantages being that:

1. Access Rights Information can always be found in one place and one place only, namely in the descriptive metadata, which, per default, is the metadata that are displayed in the Access Software (Finding Aids and different viewers)
2. EAD supports the description of potentially very complex hierarchical levels of an IP and can therefore if necessary differentiate access restrictions all the way down to the individual file level.
3. EAD information is very often added upon Ingest and Finding Aids can immediately be populated with this kind of information.

Different roles will be handled by the local applications, i.e. one logs in as e.g. administrator, archivist, end-user.

4.3.2.3 Use of EAD in an E-ARK Dissemination Information Package (DIP)

EAD3[85] version 1.0 is the standard that the E-ARK project implements and uses for descriptive metadata. From the Access perspective, EAD will be used to populate the Finding Aid that the E-ARK project proposes. EAD will thus help the end-user find relevant information in the archive as well as contributing to the user-friendliness of the relevant information when DIPs are consulted by the end-user.

[83] Cf. the semantic component 4.1.6 otherRightsInformation.

[84] Cf. PREMIS3.0 Data Dictionary: "The PREMIS rightsStatement is intended to allow a preservation repository to determine whether it has the right to perform a certain action in an automated fashion, with some documentation of the basis for the assertion.".

[85] The Library of Congress (January 15, 2016). <EAD> - Encoded Archival Description. Official site. Retrieved 4th of April 2016 at:https://www.loc.gov/ead/

EAD is fully described in another deliverable[86], so there's no reason to repeat it here. It is however convenient to briefly describe the common architecture of an EAD-description and highlight the tags that will be available in the E-ARK GUIs, as well as the tags that are needed in one way or another by the Access Software (e.g. <unitid>)[87].

Before listing the requirements and recommendations that have been found particularly important for Access, please note the three following considerations:

Firstly, the Access Rights Information that concerns the end-user has to be available in EAD, not in PREMIS (see above), and <accessrestrict> will be used for this.

Secondly, the hierarchical archival descriptions will be addressed using the component tag (<c>)[88],[89] and to the extent possible, the Access Software will address all hierarchical levels in a description of the Archive's collections, thus providing user friendly information for every consulted element.

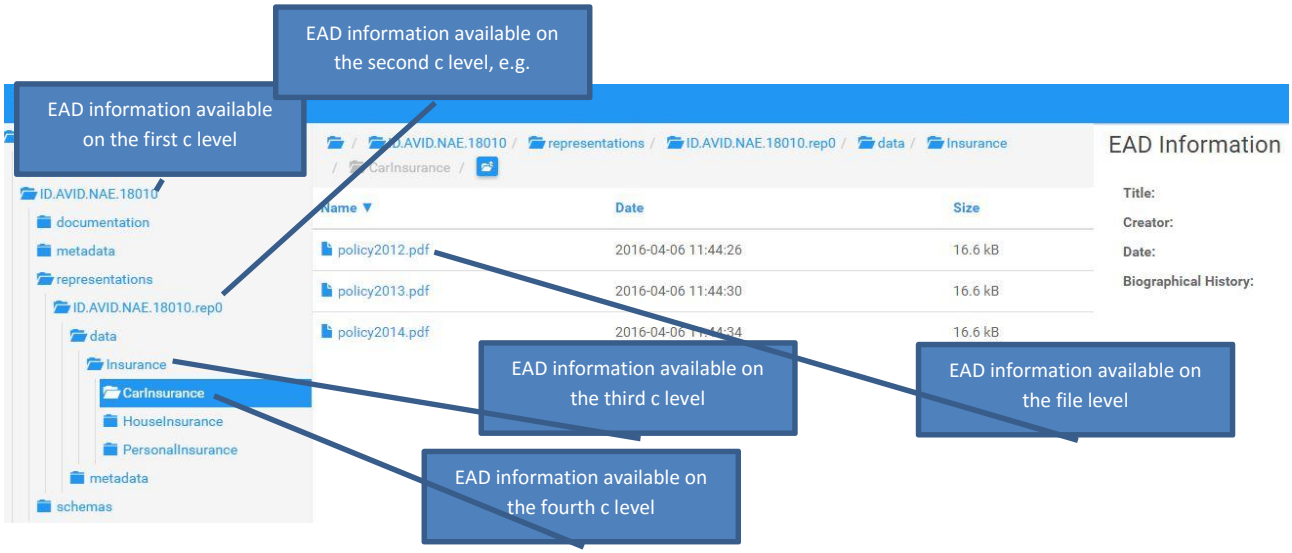

Figure 1 - Illustration of the EAD component tag <c>

[86] D3.3 E-ARK SMURF http://www.eark-project.com/resources/project-deliverables/52-d33smurf

[87] Two element names ("Geographical name", "Biography or History") are not currently described in D3.3 E-ARK SMURF, but are described in the table. They will be introduced in the final SMURF specification.

[88] According to Society of American Archivist a typical Finding Aid has "two or three views of a collection, each of which describes the same body of materials, but at varying levels of detail. The first level describes the entire collection in a very general way [...] The next level might focus on groupings of material within the collection, describing each in more detail than was done at the first level, highlighting more specific material types and additional individuals and subjects represented. This mid-level description may be represented in a Finding Aid by narrative descriptions of series or subseries within the whole. Finally, each file, or possibly each item, may be described."# The EAD format has been created and developed to follow this hierarchical way of describing collections, intellectual units, etc.

[89] At the time of writing it has not been decided yet whether the E-ARK project will adopt unnumbered ot numbered <c> tags.

Thirdly, the <dao> element is particularly important[90], both in order to:

1. Connect searches from the search across IPs and from the search in metadata (Finding Aid);
2. Allow for an appropriate visualisation of both data and EAD metadata inside the E-ARK Access Software GUIs.

The search and visualisation functionality of E-ARK Access Software supports by default the faceted searching and native visualisation of the following EAD elements[91]. The elements below are valid for each descriptive level of the hierarchical description.

Element name	EAD element	Description and usage[92]	Cardinality
ID of the Unit	<unitid>	**<unitid>** may contain any alpha-numeric text string that serves as a unique reference point or control number for the described material, such as a lot number, an accession number, a classification number, or an entry number in a bibliography or catalog. **<unitid>** is primarily a logical designation, which sometimes indirectly provides location information, as in the case of a classification number.	1..1
Title of the Unit	<unittitle>	**<unittitle>** is for recording the title statement, either formal or supplied, of the described materials. The title statement may consist of a word or phrase. **<unittitle>** is used at both the highest unit or **<archdesc>** level (e.g., collection, record group, or fonds) and at all the subordinate **<c>** levels (e.g., subseries, files, items, or other intervening stages within a hierarchical description).	1..1
Date of the Unit, Structured Date of the Unit	<unitdate>, <unitdatestructured>	**<unitdate>** is for indicating the date or dates the described materials were created, issued, copyrighted, broadcast, etc. **<unitdate>** may be in the form of text or numbers, and may consist of a single date, a date range, or a combination of single dates and date ranges; **<unitdatestructured>** provides a machine-processable	1..*

[90] The linking mechanism will be established in the pilots of the E-ARK project.
[91] Note that this can change during the third year of the E-ARK project where the GUIs are piloted. Also, these elements are the basic ones: Each DIP representation format may have different requirements, especially the ERMS format specification *(*SMURF ERMS).
[92] The information captured in this column is taken directly from Society of American Archivists (2015). EAD 3 TAG Library. Retrieved at http://www2.archivists.org/sites/all/files/TagLibrary-VersionEAD3.pdf

Element name	EAD element	Description and usage[92]	Cardinality
		statement of the date or dates the materials described were created, issued, copyrighted, broadcast, etc. **<unitdatestructured>** must contain one of the following child elements: **<datesingle>**, **<daterange>**, or **<dateset>**. **<unitdatestructured>** may contain only one child, therefore **<dateset>** must be used in situations where complex date information needs to be conveyed and requires at least two child elements. A date set may combine two or more **<datesingle>** and **<daterange>** elements.	
Scope and Content	<scopecontent>	**<scopecontent>** contains a narrative statement that summarises the range and topical coverage of the materials. It provides the researcher with the information necessary to evaluate the potential relevance of the materials being described. **<scopecontent>** may include information about the form and arrangement of the materials; dates covered by the materials; significant organizations, individuals, events, places, and subjects represented in the materials; and functions and activities that generated the materials being described. It may also identify strengths of or gaps in the materials.	0..1
Conditions Governing Access	<accessrestrict>	Record in **<accessrestrict>** information about the availability of the described materials, whether due to the nature of the information in the materials being described, the physical condition of the materials, or the location of the materials. Examples include restrictions imposed by the donor, legal statute, repository, or other agency, as well as the need to make an appointment with repository staff. May also indicate that the materials are not restricted;	0..1
Conditions Governing Use	<userestrict>	Use **<userestrict>** for information about any limitations, regulations, or special procedures imposed by a repository, donor, legal statute, or other agency. These conditions may be related to reproduction, publication, or quotation of the described materials after access to the materials has been granted. **<userestrict>** may also be used to indicate the absence of restrictions, such as when intellectual property rights have	0..1

Element name	EAD element	Description and usage[92]	Cardinality
		been dedicated to the public.	
Language of the Material	<langmaterial>	**<langmaterial>** records information about languages and scripts represented in the materials being described. **<langmaterial>** must contain one or more **<language>** or **<languageset>** elements, but cannot contain text.	0..*
Related Material	<relatedmaterial>	**<relatedmaterial>** is used to identify associated materials in the same repository or elsewhere. These materials may be related by sphere of activity, or subject matter.	0..*
Description of Subordinate Components	<dsc>	Use **<dsc>** to wrap subordinate components in the archival hierarchy of the materials being described. Although **<dsc>** may repeat, it is recommended to include only a single **<dsc>** element. Because it is a wrapper element and not an essential part of archival description, **<dsc>** may be deprecated in future versions of EAD. Avoiding multiple **<dsc>** elements within an EAD instance will make future migrations simpler.	0..*
Component	<c>	As a wrapper for a set of elements, **<c>** provides information about the content, context, and extent of a subordinate body of materials. It is always a child or descendant of **<dsc>** and often a child and/or parent of another **<c>**. Each **<c>** identifies a logical section, or level, of the described materials. The physical filing separations between components need not always coincide with the intellectual separations. For example, a **<c>** that designates dramatic works might end in the same box in which the next **<c>** begins with short stories. Also, not every **<c>** directly corresponds to a folder or other physical entity. Some **<c>** elements simply represent a logical point in a hierarchical description.	0..1
Digital Archival Object	<dao>	**<dao>** is a linking element that uses @href to connect to born digital records or digital representations of the described materials. Digital representations may include graphic images, audio or video clips, images of text pages, and electronic transcriptions of text. The objects can be selected examples, or digital surrogates of all the materials in a collection, fonds, or an	1..*

Element name	EAD element	Description and usage[92]	Cardinality
		individual file.	
Description	<abstract>	Description of the entity.	0..1
Geographic Coordinates	<geographiccoordinates>	Use to express a set of geographic coordinates such as latitude, longitude, and altitude representing a point, line, or area on the surface of the earth.	0..*
Subject	<subject>	Indicates a topic reflected in the described materials.	1..1
Geographical name	<geogname>	"An element for identifying the name of a place, natural feature, or political jurisdiction." (p. 197 in TAG Library)	1..1
Biography or History	<bioghist>	"A concise essay or chronology that places the archival materials in context by providing information about their creator(s). Includes significant information about the life of an individual or family, or the administrative history of a corporate body. Use a series of **<p>** elements to capture a narrative history, and/or **<chronlist>** to match dates and date ranges with associated events (and, optionally, places)." (from p. 65 in TAG Library)	1..1
Creator	<origination label="Creator">	An entity primarily for making the content of the resource; an entity primarily responsible for making the resource (comment: Examples of a Creator include a person, an organization, or a service)	0..*
Keywords	<index> <head>"keywords"	Keywords	0..*

Table 9 – EAD DIP Elements

4.3.3 Access related metadata that will not be in the DIP

Not all Access related metadata should be included in the DIP. The dissemination process will depend on and generate other metadata than that inside the DIP. This can for example be metadata that the archives use to administer the DIPs and the dissemination process, or information about who has accessed a DIP and when. Orders are essential in the dissemination process, but not all the information stored within them should be part of the DIP.

Even though not all of these metadata belong inside the DIP, archives may choose to keep them elsewhere, for example for statistical purposes. It is up to each local archive to make policies for what to do with these pieces of information.

4.3.3.1 IP Order

To be able to create and manage requests from an end-user effectively, E-ARK tools make use of a specific order.xml file. The order is a separate XML file and is not to be included in the DIP as it carries the order from the end-user to the archive independently of the IP which is requested. This process happens automatically: The end-user finds what she searches for in the E-ARK search GUIs; presses a button that puts the order in an order basket and generates the order.xml; it is sent to the Order Management Tool (OMT)[93], which is the tool that E-ARK has conceived to make it possible for Archives to process orders.

The elements to be included in the order.xml are specified below. The order.xsd can be viewed here[94]:

Element name	Item name	Element description	Datatype	Occurrence	Mandatory / Optional (M/O)	Notes
Order ID		Each order has a unique ID.	Text	1	M	
Order Title		Title of ordered UD[95]. Human friendly identification.	Text	1	M	
	Ordered Item Reference Code	Reference code of ordered UD. One order can contain multiple units of description (items).	Text	1..n	M	
	Level of description of	Level of description indicates, whether user is ordering entire fonds	Text	0..n	O	

[93] The E-ARK tool that manages orders created in the E-ARK access system.
[94] https://github.com/eark-project/end_user_gui/blob/develop/app/models/schemas/Order.xsd
[95] Unit of Description.

Element name	Item name	Element description	Datatype	Occurrence	Mandatory / Optional (M/O)	Notes
	ordered item	or unit of description at any of its lower levels of arrangement.				
	AIP URI	ID of ordered AIP or ordered computer file. One order can contain data from one or more AIPs or computer files.	Text	1..n	M	
Order Origin		Means of placing order (email, telephone, reading room etc.)	list	0..1	O	
End-User Order Notes		Additional notes for the order added by end-user.	Text	0..1	O	
Archivist Order Notes		Additional notes for the order added by the archivist.	Text	0..1	O	
Order Date		Date, on which end-user issues the order. Can be added manually. Could be legal requirement.	Date	1	M	
Dossier of orders (of the user)		All orders from each user are grouped in a dossier.	Text	0..1	~~O~~	
Order Validation Date		Date, on which order was accepted and confirmed in the archives.	date	1	M	
Order Planned Date		Date, on which end-user wants to access the records.	Date	0..1	O	
Access Date		Date, on which order is expected to be ready	Date	1	M	

Element name	Item name	Element description	Datatype	Occurrence	Mandatory / Optional (M/O)	Notes
		for the end-user.				
Access Date Comments		Any comments relevant for the access date	Text	0..1	O	
End user UniqueID		User's ID number	Text	1	M	The end-user UniqueID gives access to personal user information needed for contact, e.g. the user's e-mail address
Responsible person UniqueID		The archivist who checked the order and finalised it.	Text	1	M	The responsible person's UniqueID gives access to personal user information needed, e.g. the archivist's e-mail address
Order Status		Indicates the status of the order (e.g. 'created'; 'processing'; 'ready')	Text	1	M	
Access Restriction		Taken from <accessrestrict> in EAD	Text	1..n	M	
Internal Note		Notes internal to the Archive and not to be seen by End-Users	Text	1..0	O	

Element name	Item name	Element description	Datatype	Occurrence	Mandatory / Optional (M/O)	Notes
Access End Date		Indicates the date on which the IP is no longer available for the End-User	Date	0..n	O	
Delivery Format		Indicates the format in which the order is delivered to the End-User		1..n	M	

Table 10 – DIP order elements and their descriptions

4.3.3.2 DIP Access

While the user accesses the DIP and its contents, the Order Management Tool (OMT) or any data & user management system may capture metadata pertaining to this activity. These metadata will not be included in the DIP, but in a separate XML file: The access.xml. Archives can collect and store them separately for the purposes of statistical analyses or the keeping of registers of access and use. The elements to be included in the access.xml are specified below:

Element name	Element description	Datatype	Occurrence	Mandatory / Optional (M/O)	Notes
Order UniqueID	Each order has a unique ID.	Text	1	M	
Order Title	Title of ordered UD. Human friendly identification.	Text	1	M	
End user UniqueID	User's ID number	Text	1	M	
Access period start	Starting date of the period when the DIP was available for access.	Timestamp	1	M	
Access period planned end	Proposed end of DIP availability date	Date	1	M	This parameter is used to automatically terminate access to DIP. It is created on DIP creation and can be prolonged

Element name	Element description	Datatype	Occurrence	Mandatory / Optional (M/O)	Notes
					upon request.
Access period end	Actual date when the DIP availability ended.	Timestamp	1	M	This parameter is created on the date when DIP stops being available.
DIP Items list	List of the accessible computer files in the DIP.	Text	0..n	O	The purpose of this element is to provide information on available computer files within DIP or which files from DIP were used to prepare the web service
Accessed Item Reference Code	ID of accessed DIP and accessed computer file.	Text	1..n	M	Metadata is particularly relevant for the access restricted records.
Date of access	Date and time (hh:mm:ss) when end user starts accessing individual Item (DIP or computer file).	Timestamp	1..n	M	Metadata is particularly relevant for the access restricted records
Downloaded Item Reference Code	ID of downloaded DIP or downloaded computer file. In one download action entire DIP or just selected computer files can be downloaded.	Text	1..n	M	Metadata is particularly relevant for the access and/or re-use restricted records
Download Date	Date and time (hh:mm:ss) when end user starts downloading individual Item (DIP or computer	Timestamp	1..n	M	Metadata is particularly relevant for the

Element name	Element description	Datatype	Occurrence	Mandatory / Optional (M/O)	Notes
	file).				access and/or re-use restricted records

Table 11 – DIP Access elements and their descriptions

4.3.4 Access Scenario and E-ARK Access Software for the reference DIP: the End-User Working Area and the DIP Viewer

The DIP representation format specifications are complemented with a description of their respective access scenarios (see below). Similarly, the reference DIP is accompanied with a description of the users' ability to view and examine

1. Their order(s), i.e. via the End-User Working Area, and
2. the DIP as a whole, i.e. via the DIP Viewer.

When the end-user is notified about granting of access, he logs into his own space. This space is the End-User Working Area. This area consists of a simple webpage that gives an overview of the ordered DIPs together with other trivial information, and access to editing personal profile information. When clicking on a requested DIP from the list, the DIP Viewer is loaded and it presents the structure and the contents of the DIP.

The purpose of the DIP Viewer is to enable the end-user to acquire a quick overview of the DIP that has been delivered to her by the archive. However, the DIP Viewer is also a tool which can be employed by the archivist to perform any necessary modifications in order to make the DIP ready for the end-user. For example, the DIP Viewer not only allows for viewing metadata files, but also editing them, if you are logged in with the appropriate rights. Insufficient descriptions in EAD can thus be enhanced; incorrect ones can be rectified, etc.

As described above, the DIP is a physical folder structure that contains content information and associated metadata. The DIP Viewer imports and presents this folder structure and it is possible to navigate it, just like in a regular file system. As mentioned, the DIP Viewer also offers the ability to view and edit files - both metadata and data files. It goes without saying that restrictions can be imposed so that only reading rights are granted. The DIP Viewer also enables search in and across the reference DIP, and depending on the file format viewers that are available in the local implementation of the DIP Viewer, it will also be possible to view and search the file formats contained inside the representation of the DIP. Very importantly, the DIP Viewer also renders EAD information pertaining to the level of description, which is selected (cf. c-levels in EAD, section 4.3.2.3 Use of EAD in an E-ARK Dissemination Information Package (DIP)).

The DIP Viewer can - while under development - be consulted here[96]. Below is a screenshot of the current DIP Viewer:

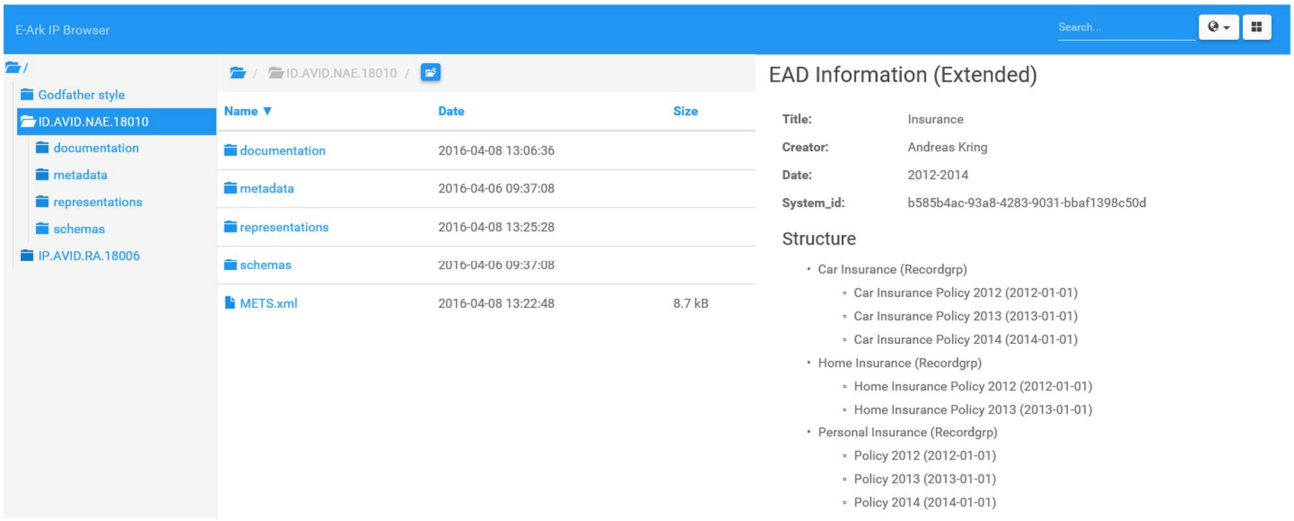

Figure 8 – Screenshot of DIP Viewer

The front-end of the DIP Viewer is the open source web application framework, AngularJS. The back-end is Alfresco, which is why it is possible by default to view a large number of different file formats and, not least, also to render complex ERMS structures.

The ambition is that the DIP Viewer not only is for viewing and navigating the structure of the DIP, but that it is also an SFBS Viewer and an ERMS Viewer, so everything is built into one application, and one presentation. This 3-in-1 application will be relatively lightweight (~3GB) and demand few resources (~4GB RAM). The installation of the program is automated (script-driven) and it can be run off-line, thus varying the use case scenarios.

4.4 Specifications for DIP representation formats and description of pertaining access scenarios

This section describes the DIP representation formats and the access scenarios in which they are rendered.

When made fit for long-term preservation, these content information types are normalised[97] into what the E-ARK project has named 'representation formats'. Representation formats can be in the SIP format, the AIP format and in the DIP format. The DIP representation formats of the E-ARK project are:

Content information types	DIP representation formats	Sections

[96] http://178.62.194.129/ipviewer/

[97] The term is used in two meanings: Firstly, - here – in the sense in which the digital preservation community is employing the word: On Ingest, Content Data Objects are transformed into long-term friendly formats. Secondly, in database normalisation where columns and tables are organised in order to reduce redundancy.

Content information types	DIP representation formats	Sections
ERMS and case files	SMURF ERMS	4.4.1 E-ARK DIP SMURF representation formats for ERMS and SFBS
Simple File-System Based Records	SMURF SFSB	
Databases	SIARD1.0, SIARD2.0, and SIARDDK	4.4.2 E-ARK DIP SIARD representation formats for relational databases
Data warehouse	OLAP	4.4.3 E-ARK DIP OLAP representation format for data warehouse
Geodata	GML, and GeoTIFF and GML Frame	4.4.4 E-ARK DIP GML and GeoTIFF representation formats for vector and raster geodata

Table 12 – Section overview of representation formats

Note that .xml samples for each DIP representation format as well as the associated .xsd files will be created during the pilots. They will all be available by the end of the project (January 2017), and can be found on GitHub[98] as they are being created.

4.4.1 E-ARK DIP SMURF representation formats for ERMS and SFBS

The content information type SMURF has been defined within deliverable 3.3 as the **s**emantically **m**arked **u**p **r**ecord **f**ormat. As such the purpose of this chapter is to define the access use cases, and metadata and tool requirements for data which has been initially submitted as a SMURF SIP and preserved as a SMURF AIP.

When looking at the use cases from an Access and end-user point of view we can differentiate between four different scenarios[99]:

a) **Access to individual computer files or folder structures.** In this scenario the data are managed within an archive as a set of computer files and/or folder structures, for example as archived from a hard drive. In this scenario the data have very little descriptive metadata attached to them. However, we assume that there is a basic archival hierarchy available which in turn is described using the core elements of EAD.
In regard to the access use case the user is able to carry out either a full text search or browse the basic hierarchy. Once the user finds the suitable computer file (by full text search) or aggregation

[98] https://github.com/eark-project
[99] Please note that the scenarios below are not exclusive but should be implemented in parallel in an institution.

(by hierarchical browsing or search within the catalogue) the user can order the DIP. The DIP includes the computer file(s) as well as the basic metadata which the user shall be able to view with an E-ARK viewer.

b) **Access to single records.** In this scenario the records in question originate from an ERMS. The assumption for this scenario is that during ingest the content of the ERMS has been mapped to the hierarchical aggregations of the archival catalogue[100] and therefore the user is able to search for single records within the catalogue. As well, in addition to simple EAD metadata the record is assumed to include additional descriptive metadata originating from the ERMS. In most cases the record would include only a few computer files though in exceptional cases some records might even include tens of different computer files.
In this scenario the user is again able to carry out either a full text search, a catalogue search or browse the archival hierarchy. If the user finds a record of interest through any of these means[101] he can order it as a DIP. The DIP includes the relevant EAD metadata, original records management metadata and the computer files. The user shall be able to view the DIP with an E-ARK viewer.

c) **Access to a case file.** In this scenario the case files originate from an ERMS. The assumption for this scenario is that during ingest the content of the ERMS has been organised according to a case logic and the different case files have been mapped to the hierarchical aggregations of the archival catalogue. As well, both the records in the case file and the case file itself have specific metadata originating from the source ERMS.
In this scenario the user is able to carry out either a full text search, a catalogue search or browse the archival hierarchy. If the user finds a case file of interest through any of these means he can order it as a DIP. The DIP includes EAD metadata, original records management metadata and the computer files. The user will be able to view the full DIP with an E-ARK viewer.

d) **Access to an ERMS.** In this scenario the content originates from an ERMS. The assumption is that the ingest process included ERMS data (e.g. classification scheme, records, metadata) as a whole and the integrity of the whole transfer is also maintained within the archival data management and preservation layers. It is also worth to note, that subsequent ingests from the same ERMS are possible to be treated either as additions to the same AIP or as a new AIP – the E-ARK Common Specification of Information Packages, reference DIP specification and the current DIP representation format specification do not pose any restrictions as such, and institutions are able to follow local archival policies.

In this scenario the user is able to carry out either a full text search, a catalogue search or browse the archival hierarchy. As well, there might be a dedicated search or browsing capability for

[100] Cf. Finding Aid in the Glossary.
[101] As an example, the user might find a computer file using full-text search. As the computer file is part of a record the user has the possibility to order the full record instead of accessing only the single file (which would lead back to scenario a).

"transfers". If the user finds a computer file, record or any aggregation which he wants to access he has the possibility to "order the whole transfer" and not only view the aggregation unit itself (which would fall under scenarios a) – c)). In response the archive prepares a large DIP from one or several AIPs which include the original classification scheme, records, metadata and additional elements defined in the SMURF profile. The user can access the DIP with a dedicated viewer which allows the user to browse and search using the original classification and view the records and computer files in a "close to original" environment. Ideally the full-ERMS DIP could be also exported and accessed in an end-user's own ERMS platform.

Below we will look further into the specific needs arising from these scenarios. Please note that as scenarios a) and b) are very similar in regard to access and tool requirements these have been joined in the next section 4.4.1.1 Access to single records, computer files or folder structures.

4.4.1.1 Access to single records, computer files or folder structures

4.4.1.1.1 Specific requirements for the sub-format

The "single records DIP" scenario is the simplest among all the DIP access scenarios and sub-formats. As such there are no specific requirements additional to the ones mentioned in the Common Specification and for the DIP.

4.4.1.1.2 Sub-format data and structure

The single records DIP:

- **Must** include only one representation of the data
- **Must** follow the simple folder structure as defined in the E-ARK Common Specification for Information Packages, and the reference DIP format (cf. section 4.3.1 DIP Data Model and Physical Folder Structure).

As such we expect that the DIP structure:

- **Must** include one root METS file;
- **Must** include one root *metadata* folder which includes a sub-folder *descriptive* and might include a sub-folder *preservation;*
- **Must** include one *representations* folder which includes exactly one sub-folder with the name or ID of the representation;
- **Must** include exactly one *Data* folder within the folder of the representation for all ordered computer files.
- In addition the package **could** include additional folders for *documentation* (for example on how to install a specific piece of software needed for a computer file), schemas or XSLT files (might be needed for user-friendly rendering of XML data or metadata).

In addition we recommend that the computer files within the *Data* folder **should** be further split into sub-folders according to either the archival hierarchy or the original order. For example, if the DIP in question is

"Personal letters of Prime Minister John Smith" which are split in the archival hierarchy annually, we recommend including into the *Data* folder:

- A top-level folder called "Personal letters of …".
- Sub-folders for "2000", "2001", "2002" etc.

Another example would be the ordering of a single record originating from an ERMS. In this case we recommend adding a folder with the title of the record (for example: Circular no 25 from 24.03.2001) and include all computer files directly in this folder.

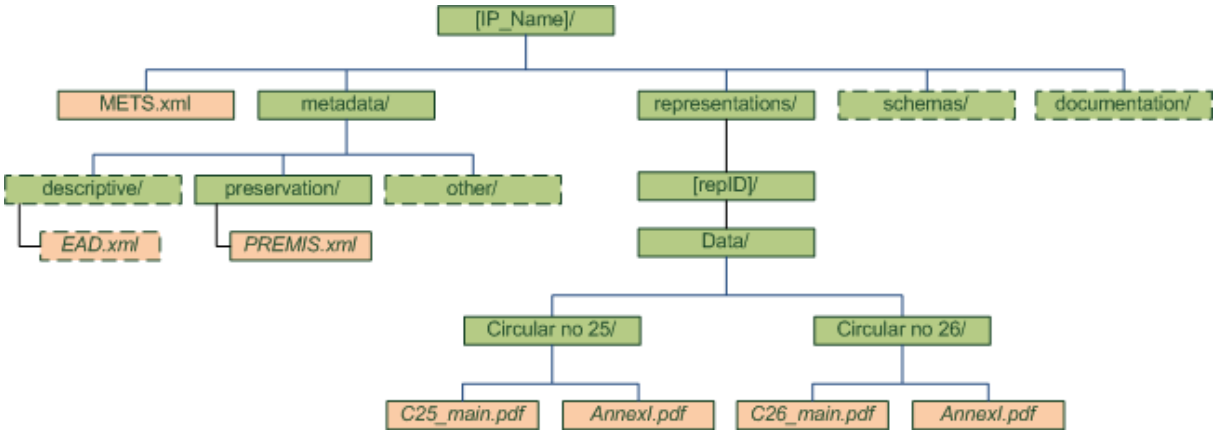

Figure 9 – Example of the DIP structure including two single records

This is, however, not a mandatory requirement but a recommendation in order to increase the human-readability of the whole package.

As well, if this recommendation is followed then, to reduce confusion and make the browsing of the package easier, we recommend including only the lowest relevant levels of folder hierarchy. To continue with the first example, you should not include a folder for the fonds level "Personal fonds of Prime Minister John Smith" and any other high-level aggregations into the DIP's folder structure.

4.4.1.1.3 Sub-format metadata

As mentioned above we expect that the single records DIP will include very little descriptive metadata in EAD format for scenario a) above, and might include some additional records-level ERMS originated metadata for scenario b).

Therefore metadata for the single records DIP should, to a large extent, follow the requirements set down in the E-ARK Common Specification and the reference DIP specification above (cf. section 4.3 The reference DIP).

Root METS file

In regard to the root METS file the only additional requirement is that the attribute <mets TYPE> must use the value of "DIP:SMURF:Single_Record".

Example: <mets TYPE= “DIP:SMURF:Single_Record”>.

However, as a special note we would like to stress the importance of the Common Specification requirement to have all data files described in the METS file section (element <fileSec>) and to have the internal structure of the *Data* folder described within the mandatory structural map instance (element <structMap TYPE=”physical” LABEL=”E-ARK structural map”>).

Descriptive metadata

The package must include a simple EAD file which uses the mandatory elements as defined for the general DIP format.

In short, EAD metadata **must** use the tags set out in the E-ARK DIP Pilot Specification D5.3, among others the following mandatory elements need to be used:

- The title for each descriptive unit within the package (element *did/unittitle*);
- The creation date of the record or descriptive unit (one of the elements *did/unitdate* or *did/unitdatestructured*);
- The relative path to the file or folder within the \representations\[repID]\Data folder of the DIP (element *did/dao* with the *href* attribute referring to the relative location of the file, other attributes used according to the official EAD3 standard). If the record includes multiple computer files the additional use of the element *did/daoset* is needed[102].

The use of all other elements is optional.

Example:

```
…
<c02 level="record">
   <did>
      <unittitle>Circular no 25</unittitle>
      <unitdate type="single" normal="20010324">March 24, 2001</unitdate>
      …
      <daoset label="Computer Files" coverage="whole">

         <dao daotype="borndigital" linktitle="Circular no 25 body"
href="/representations/ID0023dfc33/Circular_no_25/C25_main.pdf">
         </dao>

         <dao daotype="borndigital" linktitle="Circular no 25 Annex I"
href="/representations/ID0023dfc33/Circular_no_25/AnnexI.pdf">
         </dao>

      </daoset>
   </did>
</c02>
…
```

[102] Like mentioned before, the final linking method is yet to be established.

For scenario b) above the package is also expected to include further metadata taken from the extended EAD set as described in Appendix 1 of the E-ARK SMURF Profile (D3.3). The **mandatory** elements which need to be included:

- archival aggregation level(s) to which the records belong (using the @level attribute on the *archdesc, c*, *c01* etc element);
- and archival history of the record (using the *custodhist* element);
- if applicable, EAD metadata must also include information about any access restrictions to the records (using the *accessrestrict* element).

Preservation metadata

The single records DIP should include a PREMIS metadata file which includes information about preservation events applicable to the records as well as all information pertaining to the rendering environment necessary for end users to view and use the records. However, the PREMIS file should follow the overall requirements as set down above (cf. section 4.3.2.2 Use of PREMIS in an E-ARK Dissemination Information Package (DIP)) and there are no further restrictions necessary.

However, to increase the human-readability of the DIP we recommend putting all PREMIS metadata into one single metadata file (like *premis.xml*) and also embedding any external technical metadata in it. This recommendation is however not mandatory and is mainly valid for smaller DIPs which include a small number of data objects.

4.4.1.1.4 Access tools

It should be possible to open and browse the single records DIP with the help of the E-ARK DIP Viewer software. A basic description of the software is delivered in section 4.3.4 Access Scenario and E-ARK Access Software for the reference DIP: the End-User Working Area and the DIP Viewer. The detailed requirements for the tool originating from this sub-format are:

- Ability to view the overall package metadata in a user friendly rendering
- Default presentation of the archival hierarchy as a navigable tree
- Alternate presentation of the DIP folder structure as a navigable tree
- Ability to view the metadata for each aggregation level in a user friendly way
- Clear highlighting of access restrictions if relevant
- Ability to view the names of computer files within the appropriate aggregation level
- Ability to view the technical metadata of the computer files
- Ability to extract the computer files for opening in external applications (for example Adobe Acrobat)
- Clear highlighting of relevant rendering information if the file in question is not expected to be in a “typical format” and the user is not expected to have relevant software available.

Further, the tool must be implemented as both stand-alone software (used for downloaded DIPs) and as a portal component (for online access).

4.4.1.2 Access to a case file

4.4.1.2.1 Specific requirements for the sub-format

The “case file DIP” in its main aspects follows the requirements set down in the E-ARK Common Specification for Information Packages and the generic DIP format.

However, there are some additional requirements for the case file DIP:

- The DIP **must** include relevant information about the integrity of cases and their components;
- It must be possible to access the case file within the DIP in the overall context of the archival hierarchy and the source system.

4.4.1.2.2 Sub-format data and structure

The folder structure of the case file DIP must follow the same requirements as for the single records DIP (see previous chapter).

Additionally we recommend that the *Data* folder of the DIP includes additional sub-folders with the titles or reference numbers of the case files. If the records inside the case file consist of more than one computer file, the case file folder should also include sub-folders with the title of the record.

Please note that this is not a mandatory requirement but a recommendation to improve the human-readability of the DIP.

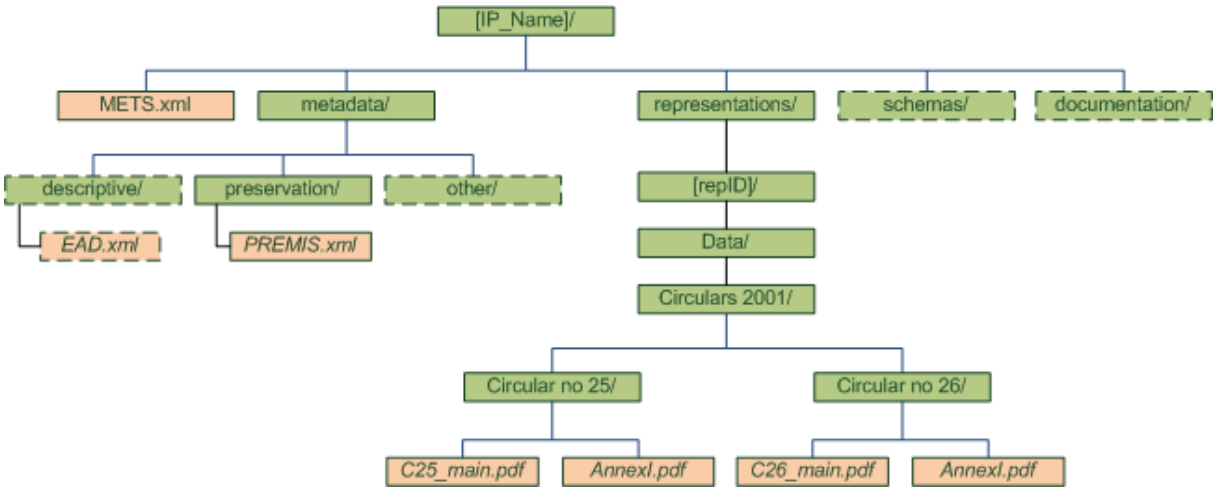

Figure 10 – Recommended folder structure for the case file scenario.

4.4.1.2.3 Sub-format metadata

In general the case file DIP is expected to include:

- detailed information about the case file;
- information about the context of the case file in the original production (creation and use) environment;
- the EAD description of the case file, its contents (if available) and its archival context;

- the root METS metadata and (if required by the user) preservation metadata in PREMIS format.

Root METS file

In regard to the root METS file the only additional requirement is that the attribute <mets TYPE> must use the value “DIP:SMURF:Case_file”.

Example: <mets TYPE=“DIP:SMURF:Case_file”>.

Descriptive metadata

As mentioned above descriptive metadata within the case file DIP should allow the user to understand both the archival context as well as the original context of the case file.

We expect that the package includes, therefore, EAD metadata which details the location of the case file within the archival hierarchy and provides archival metadata. The requirements for EAD metadata are the same as described above for the general DIP format.

In addition, the extended EAD file **must** contain additional metadata which includes:

- The title for each descriptive unit within the package (element *did/unittitle*);
- The creation date of the record or descriptive unit (one of the elements *did/unitdate* or *did/unitdatestructured*);
- The relative path to the file or folder within the \representations\[repID]\Data folder of the DIP (element *did/dao* with the *href* attribute referring to the relative location of the file, other attributes used according to the official EAD3 standard). If the record includes multiple computer files the additional use of the element *did/daoset* is needed;
- an aggregation type of “file” (using the @level or @otherlevel attribute at the appropriate c-level);,
- the original classification schema in full or in part (using the *fileplan* element);
- full ERMS originated metadata about the case file (embedded into EAD using the *odd* element at the appropriate c-level);
- full ERMS originated metadata about the records within the case file (embedded into EAD using the *odd* element at the appropriate c-level);
- the internal structure of the case file (presented by the hierarchical use of the *c*, *c01*, etc element);
- relevant metadata about the actors and events as exported from the source ERMS and recorded in the extended EAD (sub-elements of the EAD element *odd* as described in D3.3 SMURF specification, for example information about signatures, opening and closing of the file, etc.);
- if applicable, EAD metadata must also include information about any access restrictions to the records (using the *accessrestrict* element).

Preservation metadata

The case file DIP should include a PREMIS metadata file including information about all preservation events applicable to the records as well as all information pertaining to the rendering environment necessary for

end users to view and use the records. However, the PREMIS file should also follow the overall requirements as set down above and there are no further restrictions of specifications necessary.

4.4.1.2.4 Access tools

It should be possible to view the case file with the E-ARK DIP Viewer. All the requirements mentioned above for the single records DIP apply. In addition, we expect some further requirements to be fulfilled:

- ability to view the case file in its original context (ability to browse the original classification)
- ability to view/browse the internal structure of the case file;
- ability to view the detailed case file metadata.

4.4.1.3 Access to an ERMS

4.4.1.3.1 Specific requirements for the sub-format

The full-ERMS DIP is in main aspects following the requirements set down in the E-ARK Common Specification for Information Packages and the generic DIP format.

However, specific Access related requirements for the full-ERMS DIP are as follows:

- The DIP **must** include the original structure and metadata as present during the transfer and ingest of the package;
- The DIP **should** include sufficient information to recreate a near original browsing and access experience for the user;
- The DIP **must** include relevant information about the integrity of ERMS data (e.g. classification scheme, cases, individual records and their components);
- The ERMS data (e.g. classification scheme, cases, individual records and their components) within the DIP **could** be accessed in an overall context of the archival hierarchy and the source system (scenario a), b), c));

4.4.1.3.2 Sub-format data and structure

The folder structure of the ERMS DIP has to follow the same requirements as for the case file scenario. However, in regard to the internal structure of the *Data* folder the ERMS scenario **might** alternatively follow the original ERMS export structure in case it has been used within the according AIP(s).

4.4.1.3.3 Sub-format metadata

In regard to the root METS file the only additional requirement is that the attribute <mets TYPE> must use the value “DIP:SMURF:Full_ERMS”.

Example: <mets TYPE= “DIP:SMURF:Full_ERMS”>.

Descriptive metadata should be available according to the extended EAD metadata as described in the D3.3 SMURF profile. As a mandatory requirement the package **must** include metadata which details the location of aggregations (e.g. the case files, referenced using the <dao> or <daoset> element) within the archival hierarchy. All other EAD elements shall be used as with the reference DIP specification and essentially need

to provide sufficient archival metadata about all aggregations (described as EAD c-levels) within the DIP to allow the end-user to identify, select, and access a single computer file, record and/or case file.

The ERMS DIP should include a PREMIS metadata file which includes information about preservation events and detailed access restrictions if relevant for the user. Metadata about the restrictions may be related to entire DIP or to individual case file, record or computer file. However, the PREMIS file should follow the overall requirements as set down above and there are no further restrictions of specifications necessary.

Further, PREMIS metadata should indicate the designated tools (environments) which can be used to access the full ERMS DIP as described in section 4.3.2.2.1 Metadata regarding Representations and Access Software.

4.4.1.3.4 Access tools

Since the full ERMS DIP shares most of characteristics with other SMURF DIP sub-formats all the requirements mentioned above for the single records DIP and case file DIP apply.

In addition we expect the following scenario specific requirements related to the *Data* folder of the Full ERMS DIP to be fulfilled:

- ability to view/browse the classification schema, preferably presented in hierarchal structure;
- ability to view the metadata of the whole classification schema;
- ability to view/browse the classes and their metadata;
- search over the full classification schema with the use of class metadata as filters;
- ability to view/browse the aggregations (e.g. case file) in original context (possibility to browse the original classification);
- ability to view detailed aggregations metadata;
- search over classes with the use of aggregation metadata as filters;
- ability to view/browse the internal structure of the archived aggregation (e.g. case file) with the possibility to browse contained records;
- ability to view the detailed record metadata;
- search over aggregations with the use of records metadata as filters;
- ability to view/browse the archived record with the possibility to browse contained attachments (imbedded/related computer files);
- ability to access content of contained attachments (embedded/related computer files).

4.4.2 E-ARK DIP SIARD representation formats for relational databases

The DIP is created from a representation within an AIP that contains a relational database in SIARD format.

4.4.2.1 AIP with more than one representation

In the case where the AIP contains more than one representation only one representation at a time can be chosen as basis for the DIP[103].

A representation may contain a SIARD archive file with a modified[104] version of the database in the form of de-normalization and/or supplemental views in order to make it transparent.

The difference between the representations must be indicated by specific PREMIS metadata.

Different representations must be treated as different (representation) object entities belonging to the same intellectual (object) entity. The relationship between the representations is expressed by the semantic unit (PREMIS) "relationship". This semantic unit includes "relationshipType" and "relationshipSubType" as semantic components. In regard to the recommended use of PREMIS, both components' values should be taken from a controlled vocabulary.

4.4.2.1.1 More than one database in a SIARD archive file

In the unlikely, but possible case, that the SIARD archive file itself contains more than one database. Only one database at a time can be used for the DIP.

4.4.2.2 SIARD versions and SIARD format structure

The SIARD version can be SIARD 1.0, SIARDDK or SIARD 2.0[105], the latter being recommended by E-ARK for use with relational databases.

The SIARD format consists of the SIARD archive file (named [database name].siard) and possibly associated files outside the SIARD archive file (not applicable for SIARD 1.0), representing LOBs from the database. These associated files are referenced from the SIARD table files inside the SIARD archive file, and are listed in the IP's METS file (or several METS files in case of segmented IPs due to size of the LOBs). The SIARD archive file and the associated LOBs files outside the SIARD archive file are named a SIARD database hereafter.

4.4.2.3 Folder structure

The folder structure of a DIP containing a relational database in SIARD format is fully compliant with the folder structure of the E-ARK Information Packages described in section 4.1 of the Introduction to the Common Specification for Information Packages in the E-ARK Project.

4.4.2.4 Data

The primary data of the relational database (rdb) to be delivered to the user is stored in a SIARD database. The actual content of the SIARD database (db) within an E-ARK DIP may be

1. all or a subset of the SIARD db from the AIP;

[103] In the unusual case where the end-user requests access to more than one representation of an AIP it will be delivered as separate DIPs. In the final DIP specification we may offer more than one representation in a DIP.
[104] The assumption being that a database is always archived in accordance with its original data model.
[105] Created by the Swiss Federal Archives in cooperation with E-ARK

2. all or a subset of the SIARD db from the AIP supplemented with certain views[106] for better understanding and use of the database;
3. a de-normalised version of the SIARD db from the AIP or from a subset of it (i.e. case 1 or 2).

4.4.2.5 Metadata

In addition to the standard E-ARK DIP and SIARD metadata inside the SIARD archive file the following information should be stored:

1. Metadata on whether the SIARD db in the DIP is generated from the original database in the AIP or a modified version. If it was generated from a modification, then there should also be metadata about the modification. In this phase of the implementation of the project we recommend providing this information as a description within (PREMIS) <environmentDesignationNote>.

2. Metadata (Representation Information) about recommended rendering tools.
    a. Metadata about recommended E-ARK tools using <environmentFunction> and <environmentDesignation> (see section 4.3.2.2.1 Metadata regarding Representations and Access Software above).
    b. A longer description like metadata about recommended rendering tools, scenarios. <environmentDesignationNote> (see section 4.3.2.2.1 Metadata regarding Representations and Access Software above).

4.4.2.6 Access scenario: Relational databases in SIARD format

The first step is the search of the IP(s) which contain the data that meet the end-user's requirement. One may find one or more relevant AIP(s). Then the following questions have to be taken into consideration:

1. Are there any access restrictions regarding the occurrence of sensitive data, copyright or other legal regulations?
2. Which parts of the IP(s) are needed by the user? The archivist has to make a decision based on the specific request of the user as to whether the whole data content of the IP(s) will be delivered to the user, or if it is sufficient to deliver only a part of the IP(s).
3. How many DIPs have to be created to fulfil the user's requirements?
4. How will the data content and associated metadata and necessary documentation of the DIP be rendered to the user? An appropriate tool has to be chosen to make the data usable by the user.

[106] In database theory, a view is the result set of a stored query on the data, which the database users can query just as they would in a persistent database collection object. This pre-established query command is kept in the database dictionary. Unlike ordinary base tables in a relational database, a view does not form part of the physical schema: as a result set, it is a virtual table computed or collated dynamically from data in the database when access to that view is requested. Changes applied to the data in a relevant underlying table are reflected in the data shown in subsequent invocations of the view. In some NoSQL databases, views are the only way to query data. Source Wikipedia https://en.wikipedia.org/wiki/View_(SQL).

The archivist may have considerable flexibility when looking for an appropriate solution for these issues, but the solution may also be highly dependent on the following circumstances:

1. How is the rdb in the AIP(s) archived?
    a. Is it archived as a simple SIARD db generated from the production database with or without documentation?
    b. Or is it archived as a modified version of the original database for better understanding?
2. The software tools and IT infrastructure available to the archivist.
3. The technical skills (or qualifications) of the archivist.
4. The inner regulation or guidelines of the archive regarding how much effort may be spent on the task of generating a DIP.

The access scenarios may appear quite different when taking into account the above mentioned issues.

The simplest scenario is when the whole data content (SIARD db) of an AIP will be delivered to the user and the db can be properly rendered by a user friendly tool. For that purpose the E-ARK DB Viewer might be sufficient. This scenario may represent the easiest situation for the archive, but not necessarily for the user. The AIP may contain several representations of the db, and the user may need access to more than just one.

In most cases the production of an appropriate DIP for the user is rather complicated.

1. After having found the relevant AIP(s) the archivist has to decide in terms of the relevant access restrictions and user requirements from which part of the AIP(s) the DIP(s) will be generated. This involves:
    a. Selecting the relevant database
    b. Selecting the relevant records of the table(s);
    c. Selecting the relevant columns of the table(s);
    d. Changing/removing sensitive data if needed;
    e. Choosing the appropriate tool to make these modifications. For some of these a tool like the E-ARK DB Viewer would be sufficient but in most cases the SIARD db, stored in the AIP, has to be extracted from the AIP and imported with the E-ARK DBPTK into an RDBMS. Here the modifications have to be made in the GUI or more likely by SQL scripts. After that a new SIARD db may be exported from the RDBMS with the DBPTK.
2. In some cases, or for certain purposes, modifications have to be made on the structure of the data model to facilitate understandability or to allow for sophisticated analysis. These modifications can be done by SQL commands, scripts or by specialised tools after importing the SIARD db into a RDBMS. The modified database may be rendered by
    a. a standard tool like DB Viewer after generating a new SIARD db;
    b. or by a special tool connected directly to the RDBMS (using standard applications or specially developed applications).

3. In some cases it may need a special newly developed rendering tool (e.g. an Oracle APEX application for particular reporting).

Access scenarios:

1. **For the user using the DB Viewer without structural modifications**

    a. The archivist takes the SIARD db from the AIP.
    b. The archivist checks the access restrictions of the data stored in the AIP.
    c. If needed, the archivist selects the data meeting the user's requirements and data accessibility rules, and copes with sensitive data. This will be performed using SQL commands after loading the SIARD db into a RDBMS.
    d. The archivist exports the modified content with DBPTK into SIARD 2.0 format and packages it as a DIP using a DIP creator tool.
    e. The archivist delivers the DIP to the end-user.
    f. The end-user imports the SIARD db into the DB Viewer using the DBPTK.
    g. The end-user uses the DB Viewer to browse, search and analyse the database.

2. **For the user using the DB Viewer with structural modifications**

    a. The archivist modifies the data model by adding views, and possibly de-normalises the database to make the database transparent to the end-user.
    b. The archivist creates a new SIARD db and DIP using DBPTK. Specific metadata will be added to the new DIP package about the new presentation and modification of the database.
    c. The archivist delivers the DIP to the end-user.
    d. The end-user imports the SIARD db into the DB Viewer using the DBPTK.
    e. The end-user uses the DB Viewer to browse, search and analyse the database.

    (Steps d and e can be performed both in the archive institution and in the end-user's own environment).

3. **For the experienced user using RDBMS and specific tool(s)**. **Analysing the DIP using RDBMS and specific tool(s)**

    a. The archivist modifies the data model by adding views, possibly de-normalises the database to make the database easier to understand (transparent) to the end-user.
    b. The archivist provides access to the end-user to connect to the RDBMS.
    c. The archivist provides specific tool(s) to search and analyse data.

4.4.3 E-ARK DIP OLAP representation format for data warehouse

This section is about Online Analytical Processing (OLAP) of data from relational databases. Therefore it will refer to the previous section about relational databases.

4.4.3.1 Data warehouse

The typical way of creating and using OLAP objects requires a data warehouse, for which there are different kinds of data sources: relational databases (rdb), spreadsheets, statistical file formats (SPSS, SAS, etc.), data in XML formats, etc. All the data from these data sources will be extracted and imported into a staging area and from there the data will be transformed and loaded into a particular kind of third normal form rdb and into a "dimensional model", known as a star schema[107]. This is what we call a data warehouse.

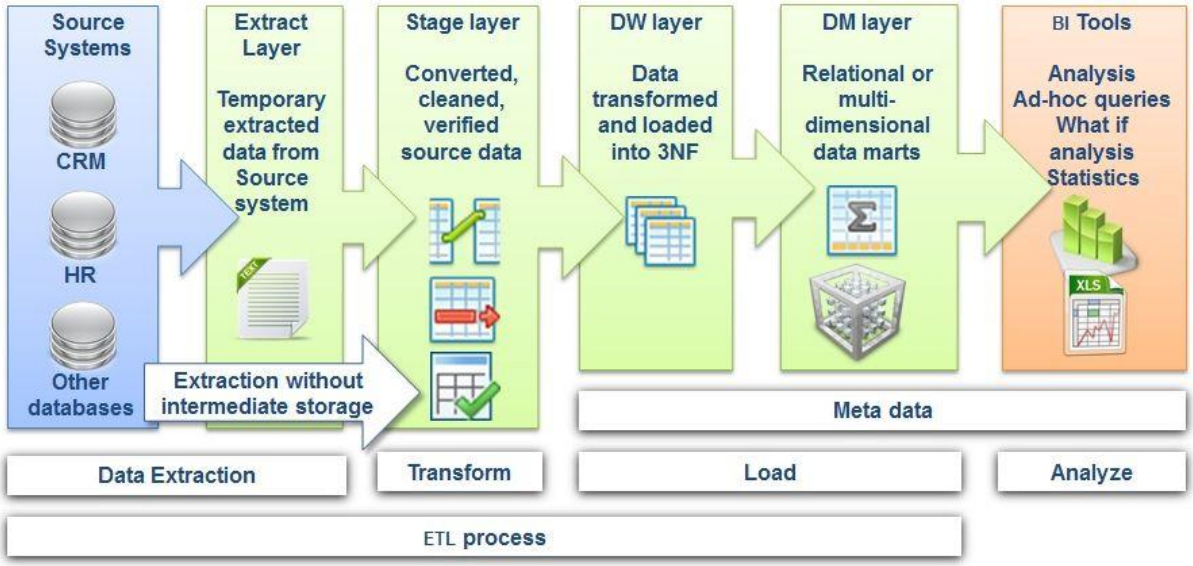

Figure 11 - Typical flow from source to BI

4.4.3.2 Data mart

For particular purposes, such as running specific analytical processes on the data, one or more data marts will be developed. These are usually based on the star schema and have a denormalised data model, but data marts do not necessarily contain all data that are stored in the data warehouse and the specific star schema model is highly dependent on the purpose of the required analysis[108].

4.4.3.3 OLAP Cube

An OLAP cube is a multi-dimensional dataset, typically from a data mart. The definition and creation of an OLAP Cube is extremely vendor specific and there is no unified exchange format for it. Therefore the vendor specific export formats of OLAP Cube definitions or of OLAP Cubes (i.e. with data) are not interchangeable between the different OLAP applications.

---

[107] See deliverable D4.3 available at http://www.eark-project.com/resources/project-deliverables/53-d43earkaipspec-1 for a full description of data mining, data warehousing, data marts, OLAP, star schemas, de-normalisation, MultiDimensional DBMSs (MDDBMSs) etc.

[108] It is possible to create a top-down data warehouse which then populates smaller data marts. This is the DW lifecycle set out by W.H. (Bill) Inmon. Conversely, Ralph Kimball's bottom-up approach advocates creating data marts first, then carefully joining them to form a DW. See http://www.computerweekly.com/tip/Inmon-vs-Kimball-Which-approach-is-suitable-for-your-data-warehouse for a full discussion of such issues.

4.4.3.4 Data warehouse stored as rdb in SIARD format in the DIP

E-ARK has decided that everything that will be done with data after the data warehouse stage will be treated as the rendering of the dataset/database, i.e. the denormalised data will be exported into the SIARD format, and all specific information about that will be stored in the documentation folder in the DIP.

The following two subcases will be taken in account:

1. Specific data warehouses will be provided as rdb in SIARD format in the DIPs (to allow for better understanding and easier use in a simple RDBMS)
2. Specific data warehouses will be provided as rdb in SIARD format in the DIPs with sufficient information to generate and analyse OLAP Cubes.

4.4.3.5 Folder structure

The folder structure of a DIP containing relational databases in SIARD format is fully compliant with the folder structure of E-ARK Information Packages described in section 4.1 of the Introduction to the Common Specification *for Information Packages in* the E-ARK Project.

In the case where the IP contains more than one representation of the database, the difference between the representations must be indicated in PREMIS.

Different representations should be treated as different (representation) object entities belonging to the same intellectual (object) entity. The relationship between the representations can be expressed by the semantic unit (PREMIS) “relationship”. This semantic unit includes “relationshipType” and “relationshipSubType” as semantic components. In regard to the recommendation to use PREMIS, both components’ values should be taken from a controlled vocabulary.

The documentation folder should include all relevant OLAP specific data which may be:

1. A textual description of the definition of the OLAP objects.
2. Vendor specific exports (files) of the OLAP cube.
3. Vendor specific documentations to particular software tools for creating and analysing OLAP data.
4. Software tools for creating and analysing OLAP data (always together with sufficient documentation).

Primary data of the rdb to be delivered to the user are stored in the form of a SIARD db. The actual content of a SIARD db within an E-ARK DIP will be a subset of the content of the SIARD db (from a certain representation) of the AIP.

4.4.3.6 Metadata

In addition to the standard E-ARK DIP and SIARD 2.0 metadata, the following information should be stored:

1. Metadata on whether the actual SIARD db is generated from a modified schema or from the original database schema. If it was generated from a modified one, then there should also be

metadata about the modification. In this phase of the implementation of the project we recommend providing this information as a description within the <environmentDesignationNote>.

2. Metadata about recommended rendering tools.
    a. Metadata about recommended E-ARK tools using <environmentFunction> and <environmentDesignation> (see section 4.3.2.2.1 Metadata regarding Representations and Access Software above);
    b. A longer description about
        i. the creation of star schema (if needed) and OLAP cube(s);
        ii. recommended rendering tools, scenarios. <environmentDesignationNote> (see section 4.3.2.2.1 Metadata regarding Representations and Access Software above).

The PREMIS standard provides an opportunity to store information about the applications, environments and other circumstances by and in which a digital object was created. A detailed plan about which semantic units should be (and how) used will be worked out during the pilot projects.

4.4.3.7 Access scenario: Data warehouse and OLAP cubes

A data warehouse (without OLAP objects) can be rendered as a SIARD db by the E-ARK DB Viewer. It can also be loaded into an RDBMS using the DBPTK and accessed by a vendor specific application (such as Oracle BI) or by an in-house developed application.

Accessing and rendering OLAP objects is quite complicated. First, the SIARD db, which contains the data, has to be loaded into a RDBMS. After that the following scenarios may happen:

1. A particular vendor specific OLAP definition file is stored within the DIP (e.g. Oracle dimensional object definitions can be exported/saved either in an XML template or in an EIF format), and the needed software infrastructure is also available, then:
    a. The definition file has to be imported into the particular RDBMS. In case of Oracle this can be done:
        i. by running a PLSQL package, and the definition file (name) will be a parameter of the package,
        ii. or by applying the Oracle Analytic Workspace Manager.
    b. A vendor specific tool has to be applied to perform the possible/required analytical processes on the OLAP cube. In case of Oracle, this specific tool is the Oracle BI.
2. The DIP contains only descriptions and documentation about how to generate (define) the needed OLAP objects (in the Documentation folder). These descriptions can be highly vendor specific, but they can be general, not requiring specific software.
    a. An appropriate software infrastructure has to be chosen and installed and connected to the RDBMS, in which the OLAP objects can be created and the required analytical processes can be performed. This software infrastructure may differ from the one where the data were originally produced.

b. The OLAP objects have to be created. This step is very vendor specific. All vendors provide their own tool for this task. This step consists of two sub-steps:
    i. Creating the star schema;
    ii. Defining the OLAP cube(s).
c. After the definition of the OLAP objects, a specific tool is needed for the OLAP analysis. The previous two steps determine which analytical tools can connect to the database and interpret the particular OLAP definitions.

Access scenarios:

1. **For the user using the DB Viewer**

    a. The archivist takes the SIARD db from the AIP.
    b. The archivist checks the access restrictions of the data stored in the AIP.
    c. If needed, the archivist selects the data regarding the user requirements and data accessibility, and changes sensitive data. This will be performed by SQL commands after loading the SIARD db into a live RDBMS.
    d. The archivist exports the modified content with DBPTK into SIARD2 format and packages it into a DIP.
    e. The archivist delivers the DIP to the end-user.
    f. The end-user extracts the SIARD db from the DIP.
    g. The end-user imports the SIARD db into the DB Viewer using the DBPTK.
    h. The end-user uses the DB Viewer to browse, search and analyse the database (data warehouse).

2. **For the experienced user using a special, vendor specific tool (without OLAP object)**

    a. The end-user or archivist loads the SIARD db into a live RDBMS.
    b. The end-user uses the special, vendor specific tool (e.g. Oracle BI) to browse, search and analyse the database (data warehouse).

    (these steps can be performed both in the archives or in the end-user's own environment/equipment).

1. **For the experienced user using a special, vendor specific tool (with OLAP object)**

    e. The archivist checks the description or definition of the OLAP object(s) stored in the AIP's documentation folder. If needed, the archivist modifies it.
    f. The archivist packages the (modified) definition of the OLAP object(s) stored in the AIP's documentation folder into the DIP.
    g. The archivist delivers the DIP to the end-user.
    h. The end-user or archivist extracts the SIARD db and the documentation from the DIP

i. The end-user or archivist loads the SIARD db into a live RDBMS with the DPTK.
j. The end-user or archivist uses a special, vendor specific tool (e.g. Oracle BI, Oracle Analytic Workspace Manager) to create OLAP cubes based on the definition in the documentation folder. In case of Oracle RDBMS the definition can be stored as an XML file and the OLAP Cube can be created by importing it into a properly installed ORACLE environment.
k. The end-user uses the special, vendor specific tool (e.g. Oracle BI, Oracle Analytic Workspace Manager) to browse, search and analyse the database (data mart).

(steps h-k can be performed both in the archives or in the end-user's own environment/equipment. When using archive's infrastructure f-i steps may be skipped).

4.4.4 E-ARK DIP GML and GeoTIFF representation formats for vector and raster geodata

This section is about Geospatial data (in short: geodata) and how they are represented in DIP format.

Geodata is a combination of the graphical representation of objects in space and their descriptions or attributes. Increasingly, geospatial formats include geospatially focused datasets or databases that contain primary information about a geographic location. In addition, ancillary and supplementary data – that can be either included or derived using spatial analysis – are considered necessary for rendering, interpretation and re-use of the data.

The basis of geodata is generally either vector or raster graphics data which can be stored as a set of files or as a database.

Vector data represents spatial objects as points, lines or polygons or, if complex, a combination of them. Descriptive or derived attributes are stored in tables (one row per spatial object).

Raster data represents spatial objects as cells in a matrix. Those cells can contain different types of values from binary to decimal numbers.

But for proper interpretation of geodata we also need some other elements that go beyond coordinates and attribute tables. The main elements are georeferencing, geoprocessing and visualization[109].

4.4.4.1 Folder structure

Within the IP geodata will be structured according to the representation-based model as shown in Figure 2 – Folder structure of the geodata DIP below:

[109] Described in greater detail in the chapter 4.3 of the official deliverable D.4.3.

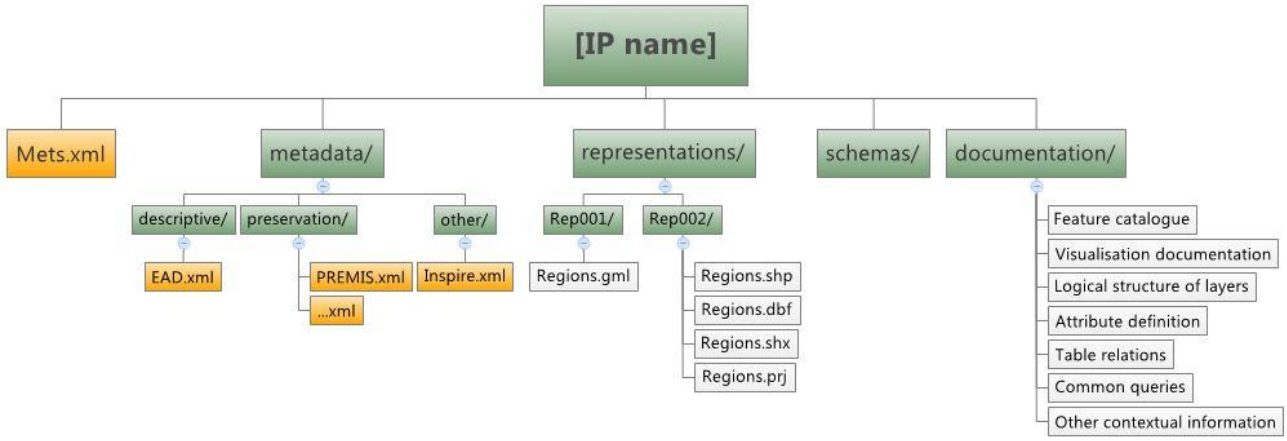

Figure 2 – Folder structure of the geodata DIP

4.4.4.2 Representations directory – the data folder

The representations directory contains at least one representation of geodata in a long term preservation format (GML for vector and GeoTIFF for Raster) and all information that is needed to properly render the information. Information needed for proper rendering of geodata can contain the elements described in the following sections.

4.4.4.2.1 Graphical information

Graphical information can be in vector or raster format.

Geodata in a vector format is organised in vector datasets which contain only one type of geometry (point, line, polygon …) per dataset. A dataset also contains a set of features (representations of real world objects) that are of the same topic or project. For instance, a roads dataset has all the roads in one dataset.

Vector data can be stored in different formats, such as SHP[110], KML[111], DXF[112], GML[113], etc.; the GML format as defined by the ISO19136:2007 standard was chosen as the long term preservation format.

It is also possible to store a representation of the data in the original format if that format is still commonly used and well documented, as is the case with the ESRI[114] Shapefile (SHP) format. This is widely accepted and supported by most GIS software today.

Geodata in a raster format is stored in a "geo-enabled" raster file. Some common formats for raster geodata are: GeoTIFF[115], JPEG2000[116], BIM[117], GRID[118], ASCII GRID[119], TIFF[120], etc.; for long term preservation format

[110] http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf
[111] http://www.opengeospatial.org/standards/kml
[112] http://www.autodesk.com/techpubs/autocad/acad2000/dxf
[113] http://www.opengeospatial.org/standards/gml
[114] http://www.esri.com
[115] http://www.remotesensing.org/geotiff/spec/geotiffhome.html

we chose GeoTIFF or ordinary TIFF with a GML bounding box. Both GeoTIFF and GML bounding box need to contain the mandatory georeferencing information.

Attribute information in GML is already a part of the GML itself.

Attribute information in an ESRI Shapefile is stored in the mandatory *.dbf[121] file. Attribute definitions are required in the documentation for interpretation purposes.

Attribute information in a Raster File can be stored in the value of the pixel itself. Attribute definitions are required if pixel value is not a grayscale value of an image.

4.4.4.2.2 Georeferencing information - Coordinate Reference System (CRS)

CRS tells us how to locate objects in geodata on the earth's surface. Elements of the spatial reference system are projection, geodetic datum, and unit of measure. All the elements can be defined by an EPSG code[122] .

Georeferencing information in GML is a mandatory part of the file itself and it is embedded in the geodata file itself.

```
<gml:boundedBy>
   <gml:Envelope srsName="urn:x-ogc:def:crs:EPSG:6.6:4326">
       <gml:lowerCorner>50.23 9.23</gml:lowerCorner>
       <gml:upperCorner>50.31 9.27</gml:upperCorner>
   </gml:Envelope>
</gml:boundedBy>
```

The attribute "`srsName`" holds the value of the coordinate reference system code.

Georeferencing information in ESRI Shapefile (shp)

An ESRI shapefile needs a <shapefilename>.prj file in order to be properly georeferenced. A Prj file is a txt file, containing a definition of the coordinate reference system and all of its elements.

[116] http://jpeg.org/jpeg2000
[117] http://www.buildingsmart.org
[118] ESRI Grid Format, http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=About_the_ESRI_Grid_format (Informal specification published by ESRI)
[119] http://desktop.arcgis.com/de/desktop/latest/manage-data/raster-and-images/esri-ascii-raster-format.htm
[120] http://partners.adobe.com/public/developer/tiff
[121] http://www.dbase.com/Knowledgebase/INT/db7_file_fmt.htm
[122] http://www.epsg.org

```
PROJCS["NAD_1983_UTM_Zone_10N",GEOGCS["GCS_North_American_1983",
DATUM["D_North_American_1983",SPHEROID["GRS_1980",6378137,298.257222101]],
PRIMEM["Greenwich",0],UNIT["Degree",0.0174532925199433]],
PROJECTION["Transverse_Mercator"],PARAMETER["False_Easting",500000.0],PARAMETER["False_Northing",
0.0],PARAMETER["Central_Meridian",-123.0],PARAMETER["Scale_Factor",0.9996],
PARAMETER["Latitude_of_Origin",0.0],UNIT["Meter",1.0]]
```

Georeferencing information in GeoTIFF

GeoTIFF contains the CRS information in its header. A GeoTIFF raster needs to be validated for the existence of CRS metadata within its header. Otherwise CRS needs to be added as for an ordinary TIFF format.

Georeferencing information in ordinary TIFF with GML bounding box

If geodata comes in an ordinary TIFF files it should be accompanied by an additional "world" file. For example a D240143.tif file would be accompanied by a D240143.tfw file. That is a txt file, containing information about the coordinates and size of the first top left pixel.

```
0.42333
0.0
0.0
-0.42333
394250.00
```

CRS information is then provided within the GML vector file that represents the location of the raster file in space and accompanies an ordinary TIFF.

4.4.4.2.3 Visualization information

If geodata needs visualization information in order to be properly interpreted, we also require this information. A large number of GIS software enable an export of the symbology definitions to a special file, however they may be in a proprietary XML structure (like the *.qml in QGIS, *.tab in MapInfo…) or even in a binary format (*.lyr in ArcGIS). The only OGC standard for symbology is the OGC Styled Layer Description XML format (sld files). If the Producer[123] cannot provide the Archive with SLD files, these can be recreated from the description which is provided in the documentation in QGIS. Raster files can have a colour map associated with the pixel value.

The SLD standard is used for rendering geodata in OGC web services and is therefore an appropriate input for easier DIP creation.

[123] The role played by those persons or client systems that provide the information to be preserved. This can include other OAIS's or internal OAIS persons or systems. Source OAIS: http://public.ccsds.org/publications/archive/650x0m2.pdf

```
<StyledLayerDescriptor xmlns="http://www.opengis.net/sld"
       xmlns:ogc="http://www.opengis.net/ogc"
       xmlns:xlink="http://www.w3.org/1999/xlink"
       xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
       version="1.0.0"
       xsi:schemaLocation="http://www.opengis.net/sld StyledLayerDescriptor.xsd">
   <NamedLayer>
       <Name>Simple Point</Name>
       <UserStyle>
           <Title>SLD Cook Book: Simple Point</Title>
           <FeatureTypeStyle>
               <Rule>
                   <PointSymbolizer>
                       <Graphic>
                           <Mark>
                               <WellKnownName>circle</WellKnownName>
                               <Fill>
                                   <CssParameter name="fill">#FF0000</CssParameter>
                               </Fill>
                           </Mark>
                           <Size>6</Size>
                       </Graphic>
                   </PointSymbolizer>
               </Rule>
           </FeatureTypeStyle>
       </UserStyle>
   </NamedLayer>
</StyledLayerDescriptor>
```

Listing 1: Example of an SLD file

4.4.4.2.4 Management of Large geospatial datasets

In case of extremely large datasets (more than 1 GB), we can experience difficulties in the validation process of such an XML file or with file manipulation. That is when datasets need to be split into two or more parts for technical reasons.

In case of vector data, we divide the GML file according to the record ID in the table. For instance if the dataset contains 4,000,000 records and is 1.8 GB large, we can split it to 2 GML files, each containing consequential 2,000,000 records (for example the Dataset001.gml from 1 – 2,000,000, and in Dataset002.gml from 2,000,001 –4,000,000). This action can be performed using standard GIS tools (QGIS).

When creating a DIPu we need to merge the individual parts into one dataset using standard GIS tools (QGIS).

These actions needs to be documented in PREMIS metadata.

4.4.4.3 Documentation directory

In the folder »Documentation« information is stored which will not be already included in the »Metadata« or »Data« folders. It will help archivists and end-users to understand the data-set in a wider social context and will provide a better understanding of the meaning, use and structure of the spatial data. The goal is to provide the end-user with all the necessary information for a proper understanding and interpretation of the geodata data set.

4.4.4.3.1 Feature Catalogue

The feature catalogue represents a logical structure of attributes. It provides a better understanding of the meaning, use and structure of the spatial data and provides a unified classification of spatial data in feature types (classes). Feature types are distinguished by their attributes (properties), by importance and by the relations between them.

4.4.4.3.2 Visualisation and cartographic representation

Data visualisation provides an illustration and representation of spatial data. The catalogue of cartographic symbols is a collection of agreed cartographic symbols, which are used during the process of visualising spatial data sets to display objects in space. Cartographic symbols are shown in the legend, which explains their meaning.

For certain spatial data the visualisation is already made by the Producer or owner of a spatial data set in the form of (geo-located) raster images or paper maps. In these cases, it is reasonable to archive the visualisation. For each spatial data set it is possible to produce any number of different visualisations with the appropriate software in the archive.

4.4.4.3.3 The logical structure of layers and attribute definition

The logical structure of layers shows the organisation of the data layers at the level of logical tables contained in the database or in a connection of unstructured objects and their attributes organised in a GIS application. It also contains the descriptions of the attributes of each data layer and the code values for each attribute. This information is similar to information contained in Feature Catalogue, but it is described in a different manner. We can choose what to archive in the appraisal process.

4.4.4.3.4 Table relations

In the case of a complex system of tables, the documentation directory should contain diagrams of relations between tables in a database or within a GIS project in order to enable the reconstruction of queries and provide greater understanding of the usage of tables.

4.4.4.3.5 Common queries

A list and a description of the most common queries provide additional information about how the data set was used in production environment from the end user's point of view. The main goal is to enable the re-creation of the original functionality of tools, which enables users to get information in the form of common queries and common reports, similarly to how they were used in the original production environment.

4.4.4.3.6 Other contextual documentation

This chapter combines all the documents (links) to the relevant documentation describing the lineage and provenance of the spatial data set. The list of documentation includes: user manuals, related practices in EU and worldwide, methodological rules, scientific articles, publications, etc.

4.4.4.4 DIP Metadata Directory

Besides the standard archival metadata like, METS, PREMIS and EAD geodata is often accompanied by geodata specific metadata. We expect this data to be in a standardised structure based on standards EN ISO 19115[124] or on the EC INSPIRE directive[125]. This data could in the future be accessed by the actual GIS tools available at the time. So we expect an Inspire.xml file or a number of them named after their geodata counterparts.

DIP specific elements of metadata for geodata will be created in PREMIS if the original archival record is altered using geoprocessing tools in order to fulfil the requirements of the users.

4.4.4.5 Access and search scenarios

4.4.4.5.1 Geodata specific search

End-users can browse and search for geodata within archival Finding Aids. Beside standard E-ARK search tools (full-text search, catalogue search) geo-search is identified as the only geodata specific search. Geo-search requires a separate tool with a spatial index[126] that can be a base for executing geo-search functions. It allows the end-user to search for geodata contained or intersecting a boundary or within a time window. Users can limit the area they are interested in or they can limit the search results in selected filters, respectively. Geoweb service allows end-users to set different filters to refine their search results, e. g. language; geodata formats (vector, raster, etc.).

All searches result in the end-user ordering one or more IP packages and then accessing the geodata DIP_u. Since geodata requires a special geodata viewer[127] some Access specific circumstances emerge.

4.4.4.5.2 Access to geodata

Archives can offer two different forms of access to geodata: unstructured files; and geodata rendered via a web service. The manner in which geodata is accessed depends on the user's ability to work with raw geodata sets and their knowledge of geodata tools (QGIS, etc.).

[124] EN ISO 19115 defined at http://www.iso.org/iso/catalogue_detail?csnumber=26020

[125] Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE) was published in the official Journal on the 25th April 2007. The INSPIRE Directive entered into force on the 15th May 2007.

[126] Peripleo is one of the tools that we can use for Geosearch - https://github.com/pelagios/peripleo

[127] Geodata can be previewed within standalone GIS tools like QGIS or it can be served as an OGC service with a tool like Geoserver and then previewed within a web browser.

4.4.4.5.2.1 Unstructured files

Advanced end-users such as geodata experts can work with raw data using QGIS. They can work with QGIS in the reading room using the archive's infrastructure or order a copy of a specific set of data to use within their own tools outside the reading room, according to archive's policy.

4.4.4.5.2.2 Web service

Serving geodata through OGC web services[128], though technically more demanding, enables a broader access to geodata and enforces greater control over how data is accessed and manipulated; it can also prevent reuse. It brings geodata to those without specific knowledge of working with geodata. Access can be made possible via a Web or mobile application (login or some sort of authentication is needed) or externally (free web access for everyone), depending on the archival policy. The archive can also set up its own GeoServer, which is accessed only within the archival network.

A web service allows viewing layers of geodata - permanent DIPs that have no access restriction. Depending on archival policy, the DIP_u can be prepared for a certain end-user, based on the end-user's search results and order. Such access would require log in for viewing the order. The end-use can view his/her order from the reading room or from anywhere using the internet, again depending on the archival policy. It is also possible to recreate maps from multiple layers of geodata and add query access according to documentation.

4.4.4.5.2.2.1 Edited and customised view in QGIS

The QGIS viewer enables the end-user (archivist or the user in the reading room) manipulation with geodata. Access can again be full or restricted, which is negotiated in the process step of the initial order. The same rules (depending on archival legislation) as for any other data type apply also for geodata.

1. The user with full access (archivist – the keeper of the archived records and user in the reading room with allowance) can edit, manipulate and view geodata DIP_u in whatever way they see fit. Within QGIS they can perform several types of action, like simplification, selection of elements, transformation. With access to documentation, they can recreate the visualization and can recreate GIS projects.
2. If the geodata has to be anonymised an archivist or employee with knowledge of geodata manipulation modifies the DIP_0. The DIP_0 is then transformed into a DIP_u and is made ready for the end-user for reading room use or reproduced for outside use.

The end-user with restricted access can work with a geodata DIP_u which has been modified by an archivist or employee with geodata knowledge. (S)he can manipulate data within QGIS in the same way as users with full access, only the data-sets are different.

4.4.4.5.2.2.2 Reproduction of geodata

The end-user can order a reproduction of set of geodata. They can order an electronic copy or create maps in QGIS, which can be printed.

[128] Serving Geodata through OGC services can be done using a variety of opensource or proprietary tools like Geoserver, Mapserver, ArcGIS Server… (https://en.wikipedia.org/wiki/Web_Map_Service).

5 Glossary

Access.xml	The xml file that captures access metadata pertaining to end-user access activity. These metadata will not be included in the DIP, but in a separate XML file: The access.xml. Archives can collect and store them separately for the purposes of statistical analyses or the keeping of registers of access and use.
Access Aid	A software program or document that allows Consumers to locate, analyse, order or retrieve information from an OAIS. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf
Access Functional Entity	The OAIS functional entity that contains the services and functions which make the archival information holdings and related services visible to Consumers. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf
Access Rights Information	The information that identifies the access restrictions pertaining to the content information, including the legal framework, licensing terms, and access control. It contains the access and distribution conditions stated within the Submission Agreement, related to both preservation (by the OAIS) and final usage (by the Consumer). It also includes the specifications for the application of rights enforcement measures. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf
Access scenarios	Access scenario is used as a term to describe the environment, the DIP and the Access Software which altogether are used to render content information and associated metadata.
Access Software	A type of software that presents part of or all of the information content of an Information Object in forms understandable to humans or systems. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf
Archival Information Package	An Archival Information Package, consisting of the content information and the associated Preservation Description Information (PDI), which is preserved within an OAIS. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf
Archival Catalogue	See Finding Aid.
Archival record	Materials created or received by a person, family, or organization, public or private, in the conduct of their affairs that are preserved because of the enduring value contained in the information they contain or as evidence of the functions and responsibilities of their creator. Source Society of American Archivists: http://www2.archivists.org/glossary/terms/a/archival-records#.VyB5VXqd9iN
Authenticity	The degree to which a person (or system) regards an object as what it is purported to be. Authenticity is judged on the basis of evidence. Source OAIS

	http://public.ccsds.org/publications/archive/650x0m2.pdf
Common Specification	The common IP specification for E-ARK IPs conceived to constitute a common basis for the E-ARK SIP, AIP and DIP Specifications.
Compound Object	A Digital Object composed of multiple Files: for example, a Web Page composed of text and image Files.
Consumer	The role played by those persons or client systems, which interact with OAIS services to find preserved information of interest and to access that information in detail. This can include other OAIS's, as well as internal OAIS persons or systems. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf In E-ARK "Consumer" is an umbrella term that designates all users of archival holdings, thus both internal users, cf. archivists, and external users, cf. end-user.
Content Data Object	The Data Object that together with associated Representation Information comprises the Content Information. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf
Content Information	A set of information that is the original target of preservation or that includes part or all of that information. It is an Information Object composed of its Content Data Object and its Representation Information. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf
Content Information Type	The data types for which format specifications have been created, cf. Electronic Management Systems (ERMS), Simple File-Based System Records (SFBS), databases, and geo-data.
Database	A database is an organised collection of data. It is the collection of schemas, tables, queries, reports, views and other objects. Source: Wikipedia: https://en.wikipedia.org/wiki/Database
Database Preservation ToolKit (DBPTK)	The Database Preservation Tool Kit is a piece of software which, from an Access perspective, enables the loading of a SIARD file into an RDBMS http://keeps.github.io/db-preservation-toolkit/. It is developed by KEEP SOLUTIONS which is a partner of the E-ARK project http://www.keep.pt/en
Data warehouse	In computing, a data warehouse (DW or DWH), also known as an enterprise data warehouse (EDW), is a system used for reporting and data analysis, and is considered as a core component of Business Intelligence [1] environment. DWs are central repositories of integrated data from one or more disparate sources. They store current and historical data and are used for creating analytical reports for knowledge workers throughout the enterprise. Examples of reports could range from annual and quarterly comparisons and trends to detailed daily sales analysis.
DB Viewer	GUI conceived by the E-ARK project to view and analyse databases.

Descriptive metadata	Also named Descriptive Information in OAIS: The set of information, consisting primarily of Package Descriptions, which is provided to Data Management to support the finding, ordering, and retrieving of OAIS information holdings by Consumers. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf The standard that E-ARK recommends for descriptive metadata is EAD.
Digital Object	An object composed of a set of bit sequences. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf
Digital Provenance	Documentation of processes in a Digital Object's life cycle. Digital provenance typically describes Agents responsible for the custody and stewardship of Digital Objects, key Events that occur over the course of the Digital Object's life cycle, and other information associated with the Digital Object's creation, management, and preservation. Source PREMIS: http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf
DIP_0	A provisional Dissemination Information Package directly derived from one or more AIPs, which may or may not be ready for use, according to the user's order and access rights.
DIP_p	A permanent Dissemination Information Package, available to be accessed indefinitely by users due to frequent requests for the same data. The DIPP can be available on-line.
DIP_u	A Dissemination Information Package, ready to be accessed, and previously checked against user's order and access rights.
DIP reference format	Refers to the E-ARK container format which is conceived to store the content information and its associated metadata.
DIP Representation Formats	The DIP representation formats are content specific implementations of the DIP reference format and offer examples of content information type specific scenarios.
DIP Status	The E-ARK DIP can have three statuses: See DIP_0, DIP_u and DIP_p.
Dissemination Information Package (DIP)	Dissemination Information Package: an Information Package, derived from one or more AIPs, and sent by archives to the Consumer in response to a request to the OAIS. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf
EAD	Encoded Archival Description. A non-proprietary de facto standard for the encoding of Finding Aids for use in a networked (online) environment. Finding Aids are inventories, indexes, or guides that are created by archival and manuscript repositories to provide information about specific collections. While the Finding Aids may vary somewhat in style, their common purpose is to provide detailed description of the content and intellectual organization of collections of archival materials. EAD allows the standardization of collection information in Finding Aids within and across repositories. http://www.loc.gov/ead/eadabout.html

Electronic Records Management System (ERMS)	Electronic Records Management System is a type of content management system and refers to the combined technologies of document management and records management systems as an integrated system.
End-User	The end-user designates an external user who seeks content information in archival holdings.
ERMS Viewer	GUI conceived by the E-ARK project to view ERMS systems.
Exchange	Refers to the DIP as an exchange format, and as such it is essential that it is possible to transfer DIPs, for example between a repository and various Access environments.
Finding Aid	A type of Access Aid that allows a user to search for and identify Information Packages of interest. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf
Geodata	Geodata is information about geographic locations that is stored in a format that can be used with a geographic information system (GIS). Geodata can be stored in a database, geodatabase, shapefile, coverage, raster image, or even a dbf table or Microsoft Excel spreadsheet.
GeoTIFF	GeoTIFF is a public domain metadata standard which allows georeferencing information to be embedded within a TIFF file. The potential additional information includes map projection, coordinate systems, ellipsoids, datums, and everything else necessary to establish the exact spatial reference for the file.
GML	The Geography Mark-up Language: the XML grammar defined by the Open Geospatial Consortium (OGC) to express geographical features. GML serves as a modelling language for geographic systems as well as an open interchange format for geographic transactions on the Internet.
Graphical user interface (GUI)	A Graphical user interface (GUI) is a graphical interface to a program on a computer. It takes advantage of the computer's graphics capabilities to make the program easier to use.
Information Object	A Data Object together with its Representation Information. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf
Information Package	A logical container composed of optional content information and optional associated Preservation Description Information. Associated with this Information Package is Packaging Information used to delimit and identify the content information and Package Description information used to facilitate searches for the content information. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf
Intellectual Entity	A set of content that is considered a single intellectual unit for purposes of management and description: for example, a particular book, map, photograph, or database. An Intellectual Entity can include other Intellectual Entities; for example, a

	Web site can include a Web page; a Web page can include an image. An Intellectual Entity may have one or more digital representations. Source PREMIS http://www.digitizationguidelines.gov/term.php?term=intellectualentity
METS	The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language of the World Wide Web Consortium. The standard is maintained in the Network Development and MARC Standards Office of the Library of Congress, and is being developed as an initiative of the Digital Library Federation. Source http://www.loc.gov/standards/mets/
MultiDimensional DBMS	A MultiDimensional DBMS is a particular kind of RDBMS that is specifically geared towards OLAP (in fact MDDBMS is often used co-terminously with OLAP).
Normalisation	The term is used in two meanings: Firstly, in the sense in which the digital preservation community is employing the word: On Ingest, Content Data Objects are transformed into long-term friendly formats. Secondly, in database normalisation where columns and tables are organised in order to reduce redundancy.
OAIS	The Open Archival Information System is an archive (and a standard: ISO 14721:2003), consisting of an organization of people and systems that has accepted the responsibility to preserve information and make it available for a Designated Community. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf
OLAP	In computing, online analytical processing, or OLAP, is an approach to answering multi-dimensional analytical (MDA) queries swiftly. OLAP is part of the broader category of business intelligence, which also encompasses relational database, report writing and data mining. Typical applications of OLAP include business reporting for sales, marketing, management reporting, business process management (BPM), budgeting and forecasting, financial reporting and similar areas, with new applications coming up, such as agriculture. Source Wikipedia https://en.wikipedia.org/wiki/Online_analytical_processing
OLAP Cube	OLAP is an acronym for online analytical processing. An OLAP cube is an array of data understood in terms of its 0 or more dimensions. OLAP is a computer-based technique for analysing business data in the search for business intelligence.
Order Management Tool (OMT)	The E-ARK tool that manages orders created in the E-ARK Access system.
order.xml	The xml-file that specifies an order in the E-ARK Access system.
Packaging Information	The information that is used to bind and identify the components of an Information Package. For example, it may be the ISO 9660 volume and directory information used on a CD-ROM to provide the content of several files containing content information and

	Preservation Description Information. Source: OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf
PREMIS	The PREMIS Data Dictionary for Preservation Metadata is the international standard for metadata to support the preservation of Digital Objects and ensure their long-term usability. Developed by an international team of experts, PREMIS is implemented in digital preservation projects around the world, and support for PREMIS is incorporated into a number of commercial and open-source digital preservation tools and systems. The PREMIS Editorial Committee coordinates revisions and implementation of the standard, which consists of the Data Dictionary, an XML schema, and supporting documentation. Source: http://www.loc.gov/standards/premis/
Preservation metadata	Preservation metadata is an essential component of most digital preservation strategies. As an increasing proportion of the world's information output shifts from analog to digital form, it is necessary to develop new strategies to preserve this information for the long-term. Preservation metadata is information that supports and documents the digital preservation process. Preservation metadata is sometimes considered a subset of technical or administrative metadata. Source https://en.wikipedia.org/wiki/Preservation_metadata The standard that E-ARK recommends for preservation metadata is PREMIS.
Producer	The role played by those persons or client systems that provide the information to be preserved. This can include other OAIS's or internal OAIS persons or systems. Source OAIS: http://public.ccsds.org/publications/archive/650x0m2.pdf
QGIS	A Free and Open Source Geographic Information System. http://www.qgis.org/en/site/
Record	Any 'information created, received and maintained as evidence and information by an organisation or person, in pursuance of legal obligations or in the transaction of business' (ISO 15489-1:2001, 3.15). In MoReq2010®, a record may be further characterised as follows. • It has an extensible set of metadata that describe it. • It has one or more components that represent its content. • It is classified with a business classification. • It has a disposal schedule that describes explicitly if, how and when it will be disposed of or destroyed. • It belongs to an aggregation of records. • Access to it is controlled and limited to authorised users. • Its destruction may be prevented by a disposal hold. • It may be exported to another MCRS while retaining all of the characteristics listed above. [MoReq 2010, v 1.1]
Relational Database Management	A relational database management system (RDBMS) is a computer software application that interacts with the user, other applications, and the database itself to capture and analyse data. A general-purpose RDBMS is designed to allow the definition, creation,

System	querying, update, and administration of databases.
Representation	The set of files, including structural metadata, needed for a complete and reasonable rendering of an Intellectual Entity. For example, a journal article may be complete in one PDF file; this single file constitutes the representation. Another journal article may consist of one SGML file and two image files; these three files constitute the representation. A third article may be represented by one TIFF image for each of 12 pages plus an XML file of structural metadata showing the order of the pages; these 13 files constitute the representation. Source PREMIS: http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf , p.8
Representation Information	Representation Information is metadata that that transforms a Digital Object into an Information Object and thereby making it understandable by a human being. It consists of Semantic and Structure Information. Source OAIS: http://public.ccsds.org/publications/archive/650x0m2.pdf
Semantically marked up records formats (SMURF)	The SMURF is an IP format for ERMS systems and SFSB (simple file-system based records) conceived by the E-ARK project.
SFSB Viewer	GUI conceived by the E-ARK project to view Single File-Based System Records.
Simple File-Based System Records (SFSB)	Simple file-system based records: records that contain simple file-system based folders or files, including those originating from content and data management systems, such as SharePoint, that are not based on true file systems. They address the submission of computer files or folders from the file Producers rather than from an ERMS. They require manual enrichment with additional descriptive metadata
SIARD	IP format for databases. Currently there exist three versions: SIARD1.0, SIARDDK and SIARD2.0.
SIARD 1.0	SIARD1.0 is the original SIARD format developed by SFA. Available at: http://www.ech.ch/vechweb/page?p=dossier&documentNumber=eCH-0165&documentVersion=1.0
SIARD 2.0	SIARD2.0 is developed in E-ARK in collaboration with the Swiss Federal Archives (SFA), and is based on the original SIARD format developed by SFA. Available at: http://www.eark-project.com/resources/specificationdocs/32-specification-for-siard-format-v20
SIARDDK	SIARDDK is a format used in Denmark since 2010 and is a slight deviation from SIARD1.0.
Simple File-System Based Records (SFSB)	Simple file-system based records (SFSB) are records that contain simple file-system based folders or files, including those originating from content and data management systems, such as SharePoint, that are not based on true file systems. They address the submission of computer files or folders from the file Producers rather than from an

	ERMS. They require manual enrichment with additional descriptive metadata.
Structural metadata	Structural metadata describes the physical and/or logical structure of digital resources; it expresses the intellectual boundaries of complex objects and can be used to describe relationships between an object's component parts. Structural metadata is commonly used to facilitate navigation and presentation of complex items by defining structural characteristics such as pagination and sequence. And, like METS, can be used to aggregate related metadata. Source http://www.library.illinois.edu/dcc/bestpractices/chapter_11_structuralmetadata.html The standard that E-ARK recommends for structural metadata is METS
Submission Information Package (SIP)	An Information Package that is delivered by the Producer to the OAIS for use in the construction or update of one or more AIPs and/or the associated Descriptive Information. Source OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf
Views (SQL)	In database theory, a view is the result set of a stored query on the data, which the database users can query just as they would in a persistent database collection object. This pre-established query command is kept in the database dictionary. Unlike ordinary base tables in a relational database, a view does not form part of the physical schema: as a result set, it is a virtual table computed or collated dynamically from data in the database when access to that view is requested. Changes applied to the data in a relevant underlying table are reflected in the data shown in subsequent invocations of the view. In some NoSQL databases, views are the only way to query data. Source Wikipedia https://en.wikipedia.org/wiki/View_(SQL)

Table 13 - Glossary

6 References

Common Specification, draft http://www.eark-project.com/resources/specificationdocs/50-draftcommonspec-1

D2.1 General pilot model and use case definition http://www.eark-project.com/resources/project-deliverables/5-d21-e-ark-general-pilot-model-and-use-case-definition

D2.2 Legal Issues Report: European Cultural Preservation in a Changing Legislative Landscape http://www.eark-project.com/resources/project-deliverables/33-d22-legal-issues-report-european-cultural-preservation-in-a-changing-legislative-landscape

D2.3 Detailed Pilots Specification http://www.eark-project.com/resources/project-deliverables/60-23pilotsspec

D3.1 Report on available best practices http://www.eark-project.com/resources/project-deliverables/6-d31-e-ark-report-on-available-best-practices

D3.3 E-ARK SIP Pilot Specification http://www.eark-project.com/resources/project-deliverables/51-d33pilotspec

D3.3 E-ARK SMURF http://www.eark-project.com/resources/project-deliverables/52-d33smurf

D4.3 E-ARK AIP Specification http://www.eark-project.com/resources/project-deliverables/53-d43earkaipspec-1

D5.1 GAP report between requirements for access and current access solutions http://www.eark-project.com/resources/project-deliverables/3-d51-e-ark-gap-report

D5.2 E-ARK DIP Draft Specification http://www.eark-project.com/resources/project-deliverables/31-d52

D6.1 Faceted Query Interface and API http://www.eark-project.com/resources/project-deliverables/34-d61-faceted-query-interface-and-api

D6.2 Integrated Platform Reference Implementation http://www.eark-project.com/resources/project-deliverables/54-d62intplatformref-1

Describing and Preserving Digital Object Environments, New Review of Information Networking, 2013 Dappert, A., Peyraud, S., Delve, J., Chou, C.,ISSN 1361-4576, 106-173

DIP, ERMS and SFSB Viewer http://178.62.194.129/ipviewer/

dm-file-ingest https://github.com/eark-project/dm-file-ingest

EAD3 https://www.loc.gov/ead/

E-ARK Web https://earkdev.ait.ac.at:8443/cas/login?service

ESSArch Preservation Platform (EPP) http://epp.essarch.org/

Hadoop https://hadoop.apache.org/

Lily http://www.lilyproject.org

OAIS, Space data and information transfer systems -- Open archival information system (OAIS) -- Reference model, ISO 14721:2012 http://public.ccsds.org/publications/archive/650x0m2.pdf

PREMIS 3.0 https://www.loc.gov/standards/premis/v3/

QGIS http://www.qgis.org/en/site/

Redmine https://e-ark-redmine.magenta-aps.dk/

Repository of Authentic Digital Objects (RODA) http://www.roda-community.org/

SIARD2.0 http://www.eark-project.com/resources/specificationdocs/32-specification-for-siard-format-v20