Towards a Common Approach for Access to Digital Archival Records in Europe

Alex Thirifays
Danish National Archives
Rigsdagsgården 9
1218 København K
+45 41 71 72 47
alt@sa.dk

Kathrine Hougaard Edsen Johansen
Danish National Archives
Rigsdagsgården 9
1218 København K
+45 41 71 72 20
khej@sa.dk

ABSTRACT

This paper describes how the E-ARK project (European Archival Records and Knowledge Preservation) aims to develop an overarching methodology for curating digital assets. This methodology must address business needs and operational issues, proposing a technical wall-to-wall reference implementation for the core OAIS flow – Ingest, Archival Storage and Access.

The focal point of the article is the Access part of the OAIS flow. The paper first lays out the access vision of the E-ARK project, and secondly describes the method employed to enable information processing and to pin-point the functional and non-functional requirements. These requirements will allow the E-ARK project to create a standardized format for the Dissemination Information Package (DIP), and to develop the access tools that will process this format. The paper then proceeds to describe the DIP format before detailing what the access solution will look like, which tools will be developed and, not least, why the E-ARK Access system will be used and work.

General Terms

Frameworks for digital preservation, Preservation workflows, Infrastructure opportunities

Keywords

Access management, DIP format, E-ARK-project, digital archives.

1. INTRODUCTION

European National Archives have partnered up with vendors of digital preservation services, research institutions and interest groups to develop common ground for OAIS compliant [3] digital archiving. Common tools and common Information Packages (IP) are at the heart of this project. The differences in legislation and practices in place across Europe mean that workflows and processes cannot meaningfully be aligned. Using the same tools and IP formats will, however, lead to new possibilities for collaboration across boarders and national differences. Among the benefits of closer collaboration on core parts of digital archiving is the possibility of more cost effective use of resources that uniting of efforts can lead to.

An OAIS compliant end-to-end methodology for digital archiving with common formats for OAIS Information Packages and tools will be the outcome of the project. The objective is to provide a single, scalable, computational and robust approach capable of meeting the needs of diverse organisations, public and private, large and small, with capacity of supporting from simple to complex content data types. Common formats for SIPs, AIPs and DIPs are being developed along with supporting tools. The methodology will cover multiple content data types of which databases, ERMS data[1], geo-data, and multidimensional data sets (OLAP cubes[2]) for data mining are the most prominent.

The ambition is that the methodology becomes an international standard allowing archival institutions to hand-pick single components, e.g. access presentation tools, or plug-in the whole reference implementation and be ready to do OAIS compliant digital archiving.

The E-ARK initiative is a 3-year multinational research project that runs from 1st of February 2014 to 31st of January 2017. It is co-funded by the European Commission under its ICT Policy Support Programme (PSP) within its Competitiveness and Innovation Framework Programme (CIP). The consortium holds 16 partners from 11 countries and collaborates with archives in Sweden and Switzerland.

The project is divided into 8 Work Packages, and the Danish National Archives is leading the Work Package on Access.

More information about the project is available from the website at www.eark-project.eu.

2. A COMMOM APPROACH TO ACCESS

2.1 Filling a Gap

The ultimate goal of preservation actions must be to ensure that access and reuse is possible. However, experience related to providing access to born-digital archival material is still limited. A study carried out in 2014 early in the E-ARK project [1]

[1] Electronic Records Management System (ERMS)

[2] OLAP is an acronym for online analytical processing. It is a technique originating from Business Intelligence needs and is used for fast and in-depth analysis of large data sets that are logically arranged in so-called snowflake schemas.

confirmed this view. The study further revealed extensive gaps between the requirements for access and existing services for access [2].

Few archives provide access to databases, geo-data, ERMS data, OLAP cubes, and other complex born-digital materials. However, many archives have expressed the need to provide access to one or more of these complex content data types now or in the near future. When comparing user needs with existing access solutions the gaps are also obvious. Existing access solutions are not very user-friendly and generally only meet users' needs poorly. The most significant gaps in relation to users' needs are:

- Lack of functionalities in access tools to support intended use of data
- Lack of comprehensive and qualitative metadata in finding aids makes it difficult to find data of interest
- Lack of flexible and modern access services
- Lack of interoperability between access components

Throughout the project E-ARK will aim to develop solutions that will help bridge the identified gaps and improve possibilities for access to digital archival materials.

2.2 Powerful Access Tools

The vision in E-ARK is to create components for access and re-use that bridge the most significant gaps. Tools for providing access to born-digital archival records will be developed and a special focus will be on complex content data types. User-friendliness and flexibility of the tools are top priorities. The same goes for ensuring that tools are easy to use and will allow consumers to access and use material for their intended purposes. Behind the tools lies a robust, common DIP format that enables efficient access via the developed tools.

2.3 For the Benefit of Archives and Users

Archivists and end-users alike will benefit from a closer collaboration and use of common tools and formats across borders. For archives an obvious benefit is that much needed access tools for complex content data types will be available, but among the possible benefits are also efficiency and cost effective use of resources. The extensive focus on user-friendliness and the determination to create tools that meet the present needs for access services will bridge some of the largest gaps between existing solutions and users' needs. Users will benefit from modern digital archival access solutions developed with user requirements in mind. Further the implementation of a common DIP format across national archives introduces new perspectives for cross-border research in archival material.

3. METHOD

The approach adopted to identify requirements for a common DIP format and access tools was formalized in a requirements specification template designed to be used by all work packages dealing with the creation of IPs and development of tools (work package on Ingest, Archival Storage, Access and Services and Integration). The information processing approach was double and consisted in a bottom-up and a top-down identification of requirements.

3.1 Bottom-Up Approach

The bottom-up approach entailed detailed analyses of requirements from essentially three different sources: User; tools; and metadata standards.

3.1.1 User Needs for Digital Archival Access Solutions

First step was to study the existing landscape of access to digital archival materials, identify user needs for access services, and then find the gaps between existing solutions and user needs. This was done in spring 2014 where a survey and a series of qualitative follow-up interviews were carried out with a broad range of stakeholders [2].

The results of the gap analysis form an important foundation for the onward work and will be referenced continuously to ensure that what is developed is something which is in demand and, equally as important, that it will meet the quality goals of users.

3.1.2 Requirements for Pilot Sites and Access Tools in General

The E-ARK project plans to operationalize the tools at specific pilot sites prior to their release. It was therefore relevant for multiple purposes to examine the requirements of each pilot site and to include them directly in the requirements specification. The access pilot sites include: KEEP SOLUTIONS, the Estonian Business Archives, the National Archives of Estonia, the National Archives of Hungary, the National Archives of Slovenia, and most probably also the Danish National Archives. The foci of these pilot sites are different and the identified requirements therefore reflect this and also cover different services, legislations, and data types, such as business records and databases, CMIS access to single records, access to Moreq compliant Electronic Document and Records Management Systems (ERMS), and access to geo-data.

3.1.3 The Metadata Elements for the DIP

In order to identify which metadata elements were needed in the common DIP format, a number of existing metadata standards were chosen for examination: METS[3], PREMIS[4], apeEAC-CPF[5], GML (INSPIRE)[6], SIARD[7], BagIt[8], Dublin Core[9], and EAD[10].

To allow for comparison of the standards, it was agreed that the analysis of the standards should not be at the level of individual metadata elements, but rather in terms of semantic metadata categories. There are a number of possible ways in which metadata elements can be categorised and there is no single "right" way of doing this. The following categories were chosen to proceed with the analysis: Provenance, Context, Discovery, Relations, Rights, Reference Information, Preservation, Integrity, Storage, and Datatype.

---

[3] METS (Metadata Encoding & Transmission Standard) http://www.loc.gov/standards/mets/

[4] PREMIS (Preservation Metadata: Implementation Strategies) http://www.loc.gov/standards/premis/

[5] apeEAC-CPF (Encoded Archival Context for Corporate Bodies, Persons, and Families) http://eac.staatsbibliothek-berlin.de/

[6] GML(INSPIRE) http://wiki.osgeo.org/wiki/INSPIRE

[7] SIARD (Software Independent Archiving of Relational Databases)http://www.bar.admin.ch/dienstleistungen/00823/01911/index.html?lang=en

[8] BagIt https://tools.ietf.org/html/draft-kunze-bagit10

[9] Dublin Core http://dublincore.org/

[10] EAD (Encodes Archival Description) http://www.loc.gov/ead/

In order to produce a more detailed impression of the coverage of each standard, the number of metadata elements belonging to each of the categories was recorded. This quantification made it possible to highlight potentially interesting differences between the standards which could subsequently be further investigated by drilling down into the metadata elements of the standards themselves. In addition, investigators were asked to provide a brief description of how they counted the elements, the nature of the standard itself and any other comments relevant to comparing the standards.

These thorough and detailed examinations of user needs, requirements for tools, and metadata standards were the first part of the adopted method for specifying the E-ARK DIP format and the associated access tools.

The second part of the job was to employ a top-down approach to complete the identification of requirements.

3.2 Top-Down Approach

The top-down approach consisted in making a comprehensive breakdown of the activities that make up the access flow; identify use cases and attach acceptance criteria and constraints to these; as well as identifying functional requirements.

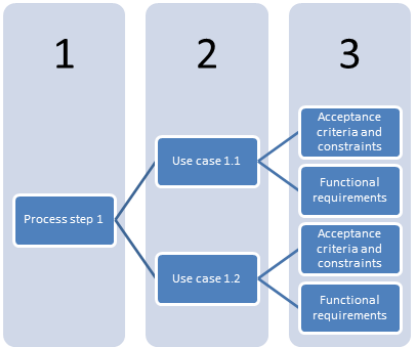

Figure 1. Decomposition of the Access flow

In the following sections, this top-down method will be described.

3.2.1 High-level Illustration and Process Step Description

Creating high-level illustrations and descriptions of the generic process steps for the whole access flow has contributed to reaching a common understanding between users (archivists) and developers. It has also defined the scope of the access activities that need to be underpinned by tools developed in E-ARK. Furthermore this work has been used to create a first platform for discussion, enabling the identification use cases.

Both high-level illustrations and descriptions are based on the E-ARK General Model, which sets up a common conceptual framework for the entirety of digital archival activities.

The overall access flow consists of 4 main steps:

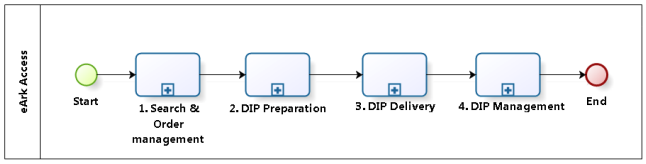

Figure 2. High-level access flow illustration

Each step is broken down in several sub-processes, an example of which is the DIP Delivery:

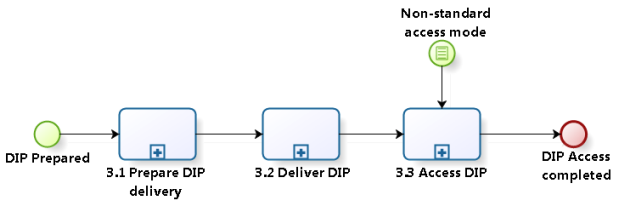

Figure 3. DIP Delivery

And each of these sub-processes is equally decomposed one more time, but those illustrations are too detailed to be inserted into this article.

In addition to the illustrations and textual descriptions of each process step, considerations were provided regarding the *product* context, which details in which environment do the products need to function and what the relationships are to native products and systems; the *assumptions* that represent the lowest common denominators that the system needs to heed, for example equipment availability, user expertise, and legal requirements.*;* and the *dependencies* that take into account relationships to other components and formats adopted or developed by the E-ARK project.

3.2.2 Identification of Use Cases

The identification and the description of the generic process steps enabled the creation of use cases, which have

1. served as communication platforms between archivists and developers and thus been used to facilitate the creation of an agile development environment where iterations rectify potential misconceptions;
2. informed the requirements of the access tools and the DIP format.

The use cases that were identified are as follows:

- Search in descriptive metadata and data
- Create initial order
- Validate order
- Check DIP availability and suitability
- Create DIP from AIP
- Modify DIP
    - Modify DIP for databases
    - Modify DIP for records
    - Modify DIP for GIS content
- Negotiate / prepare delivery method
- Provide access rights
- Notify end user
- Search in Database & ERMS
    - Search in database with Google functionality
    - Search with existing search forms
    - Search with SQL / DBMS functionality
    - Search with combination of google search and SQL / DBMS functionality
- Search in single records
- Search in GIS data
- Analyze with OLAP
- Deletion or maintenance of DIP
- Non procedural (generic) tasks
    - Update roles and users
    - Update access restrictions
    - Check logs
    - Log in

The use-cases are subject to change, and should be, throughout the whole agile development process.

3.2.3 Acceptance Criteria, Constraints and Functional Requirements

The use cases were enriched with acceptance criteria and constraints that in essence define quality goals (how will the product satisfy the user?) and boundaries (how is the product limited by external circumstances?) of the access services.

From each use case it was also possible to derive the functional requirements that were then matched to the functional requirements identified by the bottom-up approach described previously.

The bottom-up and the top-down approach have thus been used as complementary information processing methods and were adopted to secure a comprehensive understanding of the field of research at hand; and that all requirements were taken into account.

4. THE EARK DIP

4.1 A Common Specification for Information Packages

A set of common principles for all Information Packages in E-ARK have been developed to ensure consistency and coherence across IP formats. This framework called the Common Specification [4] outlines the structure of IPs, defines common metadata profiles, naming conventions and other matters that need identical handling across IP formats. The Common Specification makes up the core specification, but is amended and enriched for ingest, archival storage and access purposes in SIP, AIP and DIP formats respectively.

Information Packages are wrapped and described by a METS file. A specific E-ARK METS profile has been developed that defines core set of mandatory metadata and optional metadata. Widely used metadata standards, e.g. PREMIS and, are incorporated and used for their specific purposes (i.e. respectively preservation metadata and encodings for the finding aid).

4.2 The DIP Format

The purpose of the E-ARK DIP format is to create a format which is as standardized as possible and which observes technical, legal, user and other identified requirements. It is primarily an exchange format used for access purposes. The DIP is built on the principles from the Common Specification, but is extended to fit the specific purpose of access.

4.2.2 The DIP data model

The IP data model defined in the Common Specification is depicted below, and replicated in the DIP:

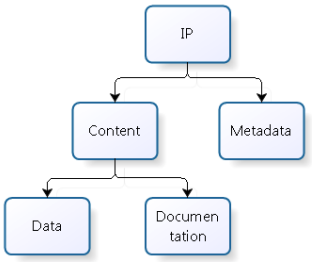

Figure 4. Common data model for Information Packages in E-ARK

As can be seen from the data model the DIP has a logical spilt between metadata and content, and content is further split into data and data-documentation.

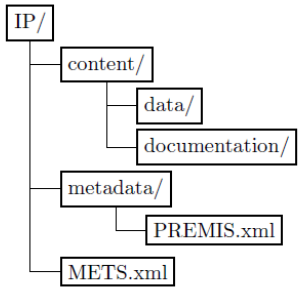

Figure 5. Common structure for Information Packages in E-ARK

The logical split is reflected in the Information Package structure (figure 5) where Information Packages are split into content and metadata at the root level of the package. The top level document of the IP is a METS document which describes the structure and encapsulates different types of metadata about the digital objects and inter-related metadata entities of the package. A common E-ARK METS profile is used across all E-ARK Information Packages. It has a basic set of mandatory elements that are common across all three types of IPs, but it can also be used to mark up metadata specific to SIPs, AIPs and DIPs respectively.

The E-ARK DIP format uses PREMIS to capture information targeted at supporting the digital preservation process. This means for example that the E-ARK PREMIS profile will capture preservation events pertaining to migrations. The conversion of a file into another format (named a 'representation' in E-ARK) will thus be documented in PREMIS. Another example is that E-ARK will make it optional to use PREMIS to capture relationships between an intellectual object or a representation to the documentation documents that are relevant for them. E-ARK will most probably adopt the new PREMIS 3.0[11].

I addition to the METS and PREMIS the DIP may include other metadata. This can be both structured metadata, including for example administrative-, preservation- and descriptive metadata in XML-format, and unstructured metadata, which could be scanned documents such as a user manual giving instructions about how the archived system was used when in production, classification schemas or filing plans, etc. It is up to each archive to decide what additional metadata and metadata files are included in the DIP. The E-ARK pilots will showcase different uses of the DIP format where it is adapted to local needs, specific content data types, and implementations.

The way "Content" is structured and documented in the DIP is not elaborated in this article because it has not been finalized at the time of writing. Detailed content type specifications will be developed for each of the content types in scope of EARK. The content type specifications will build on existing work e.g. SIARD2.0[12] will be used for databases and whole IT-systems.

[11] http://www.loc.gov/standards/premis/v3/index.html

[12] http://www.eark-project.com/news/29-siardfeedback

4.2.3 Metadata in the DIP

The majority of metadata in the DIP will be inherited from the SIP and the AIP and build on existing metadata files. In addition the DIP will contain DIP specific metadata, which for example could be information about which tool will be used to display the DIP in, i.e. "Rendering information" or supplementary authenticity information generated in the AIP-DIP process. Rights metadata are naturally also an important part of the DIP specific metadata, even though these can be embedded in the SIP and AIP profiles if needed. .

The METS file in the DIP will use the profile developed for the Common Specification, but as a DIP specific version of the profile holding DIP specific metadata elements. This means that if the "IP type" is set to "DIP" in the METS file, the elements in the file will accordingly be the ones relevant for access.

4.2.4 DIP Specific Metadata

The DIP specific metadata that are added in the DIP are divided into the following categories:

- Access rights
- User roles and permissions
- Rendering information, specifying for example that geo-data is to be rendered by QGIS[13]
- DIP status (there are three: One for when the DIP has been created (DIP_0); one when it has been prepared for the user (DIP_u); one for when it has been assigned to the DIP permanent storage (DIP_p))
- Dissemination notes and metadata enrichments made in the dissemination process
- Supplementary authenticity metadata (needed if for example multiple AIPs make up one DIP or if adjustments have been made to the DIP)

For each of the above categories specific metadata elements have been identified from existing standards and examinations of user needs. The metadata elements are described in-depth together with information about their datatypes, occurrence and whether or not they are mandatory.

4.2.5 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 those 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. The purposes of them can be multiple, one of which is statistical.

These metadata types do not belong in the DIP, and it is up to each local archive to decide whether to keep them and where they should keep them (e.g. in their data management system).

4.2.6 Data Formats in the DIP

The formats of the data content types are not going to change whether they reside in the SIP, AIP or DIP.

As already mentioned they have not yet been entirely specified. The data content types that the E-ARK project will focus on are:

1. Single records, e.g. from ERMS (e.g. PDF, TIFF)
2. Databases (in SIARD 2.0 format)
3. Geo-data (in GML format)
4. Datasets for data mining (in OLAP cubes)

When the DIP has been created and exported to a staging area, it needs to be rendered to a viewer. If it's a database, it could for example to be loaded into a Database Management System (DBMS) and displayed in a Graphical User Interface (GUI), which is put on top of this DBMS.

5. ACCESS TOOLS

The access tools that will be developed represent all the components necessary to establish a fully functioning digital OAIS archive. This is the so-called "reference implementation" and will consist of open source code, which is downloadable from the source code management platform, GitHub[14]. The whole reference implementation can be downloaded and implemented, or just the desired components.

If an archive decides to download and install a single component, the open source code as well as a series of textual guidelines will facilitate the installation process. However, integration code will be necessary to develop as things like storage adapters are not within the scope of the E-ARK project.

Even though the intention is to offer a comprehensive digital archival solution, it is not all components that will be developed by the E-ARK project; actually, most won't, since the project essentially will build on existing resources: Where solid open source components exist, these will be integrated into the reference implementation; and where there's no urgent need for a component in the archival community, because every archive already has it, only basic functionality will be developed – this goes for example for the Finding Aid or the archival catalogue component.

As depicted in *Figure 2 - High-level access flow illustration* the E-ARK access system consists of four high-level processes for which four main software components will be developed. These will be made up of a yet undefined number of software modules.

The first one is the Search and Order Management component. This component consists of two main modules. The first one, Search and Select Information Objects, allows a consumer (cf. OAIS) to make searches in both data (the AIPs are loaded into a distributed storage (HDFS[15] and indexed in the Lily[16] repository) and metadata, using an open source Finding Aid yet to be decided upon. If an information object is not directly accessible (e.g. for reasons of access restrictions), the Manage Order module allows the archivist to validate or reject the order.

In case green light is given, the DIP Preparation component provides the appropriate DIP, either by fetching it in the DIP storage, or by creating it from parts of an AIP, one AIP or several AIPs. Since the IP formats are fashioned by the E-ARK project, it is necessary also to develop an AIP-DIP module that can access and transform the AIP(s) into a DIP. Another module in the DIP

Preparation component allows for modifying the DIP, if this is needed for reasons of for example anonymization of sensitive data.

[13] QGIS http://www.qgis.org/en/site/

[14] GitHub https://github.com/eark-project

[15] HDFS http://hadoop.apache.org/

[16] Lily http://www.lilyproject.org/lily/index.html

Once the DIP is prepared, it is sent into the DIP Delivery component. The DIP Delivery component notifies the consumer and choses the appropriate delivery scenario. Delivery scenarios depend on the nature of the content of the DIPs – the so-called data content types. The delivery scenario of for example a database can be executed in two ways: The priority will be to load the content of the database (which is wrapped in a SIARD 2.0 file) into a Database Management System (DBMS), which has an E-ARK built GUI sitting on top of it. This access scenario will require that users possess knowledge about SQL-queries in order to perform searches, but these searches will be powerful. If E-ARK resources allow for it, the second scenario for delivering a DIP database will be a NoSQL solution: The delivery module retrieves the database structure and data and sends them to an index engine (e.g. the de facto standard Lucene[17]).This allows for a far more user-friendly 'Google' like search, where the technical user requirements are minimal. The downside of this NoSQL scenario is however that the "Google" search method does not at all achieve the same level of pertinence of the results as queries made in a running DBMS do.

The last module – the DIP Management module – closes the process by sending the DIP to the DIP storage or deleting it.

The components, modules and storage areas used in the access system described above could look something like what is outlined below but not everything is settled at the time of writing.

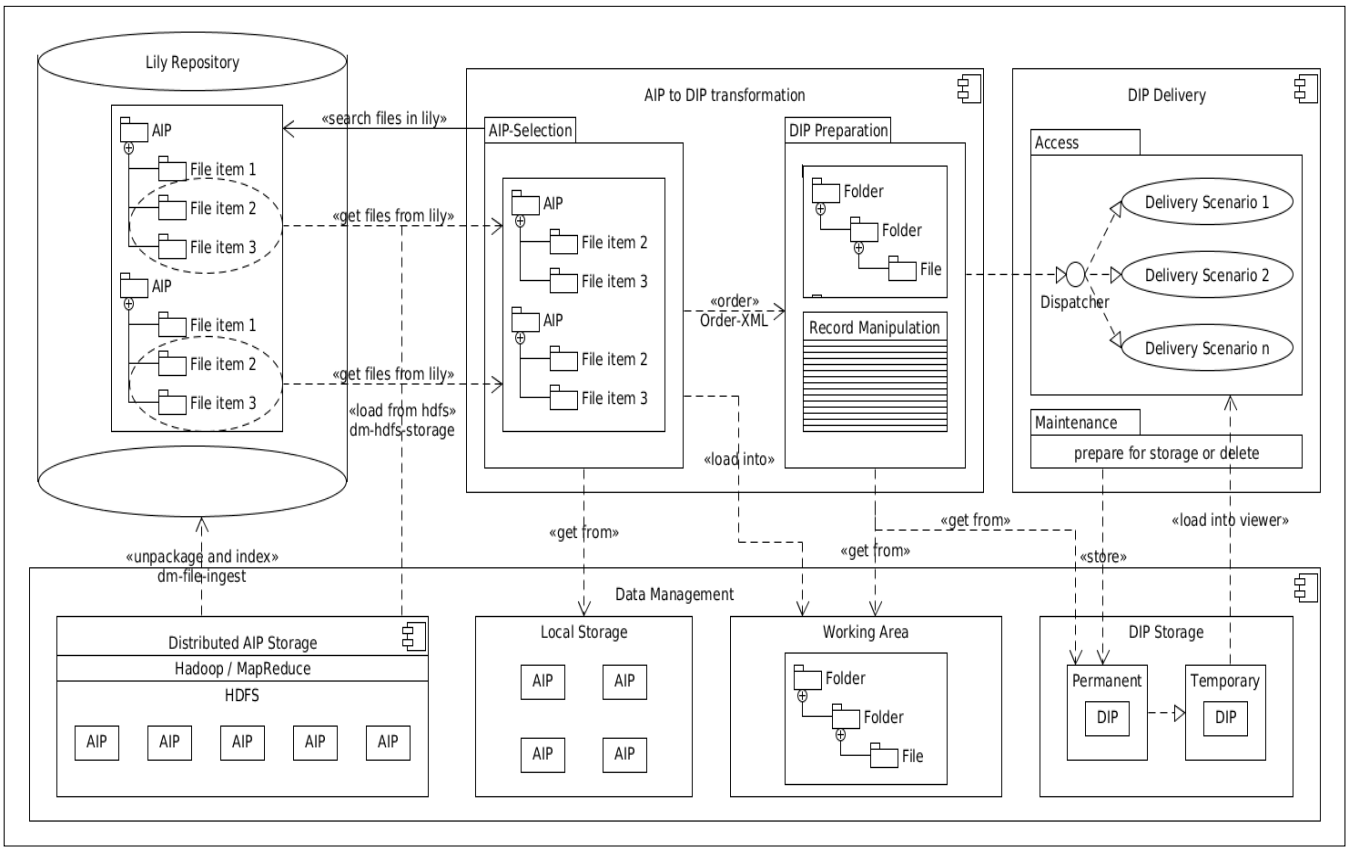

Figure 6. Overview over access components and modules

[17] Lucene https://lucene.apache.org/

6. PROOF OF CONCEPT: THE PILOTS

Ensuring that outcomes of the E-ARK have practical use and meet the needs of stakeholders and potential takers is crucial. To this end extensive piloting will be carried out.

A reference implementation of all components comprising an end-to-end solution will be hosted at the Austrian Institute of Technology (AIT). This will showcase the E-ARK end-to-end methodology for digital archiving.

In addition to the reference implementation seven pilot sites will test and implement E-ARK components. As can be seen from table 1, five pilots will focus on testing the access components and one will incorporate parts of the access components.

A real life local implementation of an end-to-end solution will be tested at Estonian Business Archives (Pilot 4).As an institution not directly involved in the E-ARK project the Estonian Business Archives will test products from a point of view of an institution wishing to implement a complete solution. Since this is a local implementation it will not necessarily include all E-ARK components but just the ones found suitable to meet the need in that particular situation.

Equally important to the full end-to-end pilot is the piloting of single E-ARK components. These will test the ability of E-ARK components to be implemented into existing environments. Testing the 'plug-and-play' aspect is vital because a major part of the archives in scope of E-ARK outcomes will already have a digital archiving solution in place and only wish to implement a subset of specific components.

Four pilots will test access components in combination with existing digital archiving environments. The focus of each pilot is different and different access components will be tested.

National Archives of Estonia will provide seamless access to public records (Pilot 3). National Archives of Slovenia will provide access to spatial data (Pilot 5). KEEP SOLUTIONS will load database into a DBMS for access purposes (Pilot 6). National Archives of Hungary to provide access to databases (Pilot 7). The Danish National Archives (Pilot 1) will test database access components in parallel with the pilot.

The extensive piloting and testing in multiple countries, technical environments and archival practices will ensure that E-ARK outcomes will in fact be scalable, robust and capable of meeting the needs of diverse organisations.

Table 1. Data Management and Access pilots of the E-ARK project

Full-scale Pilot		Data Management	Access
Pilot 1	SIP creation of relational databases (Danish National Archives)		
Pilot 3	Ingest from government agencies (National Archives of Estonia)		
Pilot 4	Business archives (National Archives of Estonia, Estonian Business Archives)		
Pilot 5	Preservation and access to records with geo-data (National Archives of Slovenia)		
Pilot 6	Seamless integration between a live document management system and a long-term digital archiving and preservation service (KEEP SOLUTIONS)		
Pilot 7	Access to databases (National Archives of Hungary)		

Focus of the pilot

Elements also used/tried within the pilot

The pilots will start in late 2015 and continue throughout 2016 and thus run for about a third of the project time.

7. NEXT STEPS AND SUSTAINABILITY

At the time of writing the E-ARK DIP *draft* format was about to be handed over to the European Commission as an official deliverable of the project.

The next steps consist in establishing an environment where developers and archivists iteratively can enhance the DIP format and the requirements specification for the tools to be developed.

The last two tasks of the Work Package on Access will focus on respectively the development of an AIP-DIP transformation component; and the development of Search, Access, and Display Interfaces. The tools will handle single records, geo-data, databases, ERMS, and showcase data mining possibilities via OLAP cubes.

The E-ARK project believes that there is an interest for take-up of the tools that will be developed.

First of all, there's a flagrant need for handling databases and EDRM systems in a standardized way, let alone at all. Every country has increasingly digitized administrations, which all use both. However, only a few national archives ingest databases on a scale that can keep the pace up with the public authorities' use of these systems, and even fewer give access to them. There's thus a growing need for ERMS and database archiving in Europe. Secondly, there is also a growing understanding of the fact that the only way of giving value to archival records is to make them accessible: Dark Archives are in nobody's interest, and especially researchers and the public authorities themselves seek more and more frequently access to these records. And if national and local archives can display increased use of the records they hold, for example facilitated via E-ARK methodology and tools, increased funding cannot be too far away.

Thirdly, the E-ARK quantitative and qualitative interviews showed that user friendly tools as well as tools that help exploiting the information that lies within the IP's, are much sought for. If E-ARK can fashion tools that respond to modern users' expectations, there's a good chance that they will be endorsed internationally.

The fourth consideration that will help increase take-up of E-ARK methodology is the ambition that the common IP format which is developed and operationalized by the E-ARK tools actually becomes the de facto standard of international archiving. Not only will it help the exchange of information packages and standardize

the search for them and within them, but it will also reduce the number of tools needed in the archival community, and thus their development and maintenance cost.

Backing up the fourth consideration is a fifth one, which is all about that the fact that the tools are open source, and available on GitHub from the reference implementation. They will be accompanied by guidelines, but of course technical IT knowhow is indispensable for installing them.

Lastly, the pilot sites will integrate and use the tools, proving their worth; the tools will actually be running out there, before the E-ARK project ends, and be ready for direct implementation in an archive near you by the first months of 2017.

8. CONCLUSION

The E-ARK consortium is in the process of developing requirements specifications for the SIP, AIP and DIP formats as well as for the tools that will process these formats. These are based on thorough examinations of user needs, of the existing landscape of digital archival solutions in Europe, and of a series of other requirements that are relevant for the development of the E-ARK methodology, e.g. legislative and tools' requirements. The DIP format that results from these investigations is also to a very high degree based on existing standards, such as METS and PREMIS. In order to do a reality check regarding both the format and the tools, the E-ARK project envisages validation at pilot sites, which will prove the concepts. Measures for sustainability and up-take of products of the E-ARK tools and formats include an open source reference implementation holding independent software modules that can be downloaded and plugged in.

9. ACKNOWLEDGMENTS

The authors wish to thank Sven Schlarb, Austrian Institute of Technology (AIT) and the rest of the E-ARK group for their contributions.

10. REFERENCES

[1] Billenness, C. S. G., Anderson, D. and Johansen, K. H. E. 2014: E-ARK project – best practice survey results on archiving of digital material *In the proceeding of iPRES 2014 (Melbourne, Australia (6-10 October), 337-339* https://www.nla.gov.au/sites/default/files/ipres2014-proceedings-version_1.pdf

[2] Johansen, K. H. E., Thirifays, A., Randmäe, P., Škoflanec, J., Delve, J. and Anderson, D. 2014: Gap Report Between Requirements for Access and Current Access Solutions http://eark-project.com/resources/project-deliverables/3-d51-e-ark-gap-report/file

[3] Reference model for an Open Archival Information System (OAIS), ISO 14721:2012: https://www.iso.org/obp/ui/#iso:std:iso:14721:ed-2:v1:en:sec:3

[4] Common Specifications for IP's in the project E-ARK [This is an internal E-ARK project document; however it will be made available upon request. Contact one of the authors of this article to make a request.]

[5] Bergin, M. B., 2013: Sabbatical Report: Summary of Survey Results on Digital Preservation Practices at 148 Institutions, University of Massachusetts, Amherest: http://works.bepress.com/cgi/viewcontent.cgi?article=1012&context=meghan_banach

[6] Kristmar, K. V. 2012: Common challenges, different strategies", EBNA, 29 May 2012, Copenhagen http://www.sa.dk/content/us/about_us/danish_national_archives/25th_european_board_of_national_archivists_conference/presentations

[7] Ruusalepp, R. & Dobreva, M. 2012: Digital Preservation Services: State of the Art Analysis, Digital Cultural Heritage Network, DC-NET: http://www.dc-net.org/index.php?en/201/publications