[redacted]

DELIVERABLE

EKA Collections Migration Plan

Document Author(s)

Name	**Role**
Anssi Jääskeläinen	XAMK Research Manager
Karin Oolu	EKA Records Manager
Mart Klanberg	EKA Systems Architect
Andres Uueni	EKA Researcher
Stephen Mackey	PIQL

Document Approver(s)

Name	**Role**
Radoslav Jakub	HaDEA Project Adviser

Digitalia Research Centre for Digital Information	[redacted]
DISEC	piql

1. An overview

The following plan describes EKA collections (CMS, ERMS) migration to an E-ARK compatible SIP format. This work is a part of a wider objective of Activity 2 of the project. The main goal of the task is to automate SIP creation based on exports from existing digital archives, CRM, CMS, or ERMS. The activity will be performed with a combination of modern web technologies and the backend scripting language, Python.
A detailed description of sample export packages can be found in M2.2: CMS/CRM/ERMS Exports Report (Dec 2021). The current document "EKA Migration Plan" describes the results of fully completed migrations of three EKA export packages as follows:

1	Pilot: **EKA ERMS Webdesktop (WD) Public Document Registry (PDR)**
	Public Document Registry (PDR) contains all of the corporate management records of EKA except those of correspondence (incoming, outgoing). PDR was chosen for the pilot export package because not all sensitive data is not being properly handled and published in the main ERMS. PDR records are published and hence accessible according to the Estonian legislation and Public Information Act. The first objective of export was to see if and how the export criteria will work as the software and content of WD are hosted by a private company, Webware and a similar export has not been done before.
2	**EKA CMS Digiteek**
	Digiteek contains a web-based digital collection of graduation works/theses created since 1914. Digiteek stores image files and links each object to key information. Each artifact/object has its own unique "object lifecycle map", where information related to the objects is held and the history of the condition of the object is preserved. All data is stored on a central server: from the processing of status reports and the periodic inventory of artifacts to the graphical mapping of highly detailed research and conservation work. The main goal of this export is to prepare the collection for transferring to the new software.
3	**EKA ERMS Webdesktop records with archival value (AV)**
	According to Archives Act of Estonia, records with archival value at EKA must be transferred to the National Archives of Estonia (NAE) for long-term preservation. The list of classified information with assigned archival value was provided by the Appraisal Act by NAE (17.09.2015) and captures all the information generated in the process of the activities of the institution. At EKA, documents with archival value are usually located in ERMS together with metadata. The goal of this export is to prepare the collection for transfer to NAE, also because a similar export has not been done before.

Necessary evidence of the migration process will be preserved in the form of full conversion logs produced by the Onelick SIP Creator tool.

2. Architecture

AS-IS: WD (Webdesktop) export file, Digiteek export file
TO-BE: E-ARK SIP, with the most important metadata embedded into METS and dc.xml files. The rest of the original metadata is handled as payload files and can be placed in a separate directory inside the SIP package.

2.1 AS-IS status of EKA collections

(1). ERMS Webdesktop (WD) Public Document Registry (PDR) pilot export package:

The pilot export package includes the records/items with
1. Metadata (only from the fields that have been set to be public).
2. Files (only from the fields that have been set to be public).
3. Relations between published records.

The export files are presented in two formats:
1. XML files, which will include metadata, relations between the records, and references to the files connected with the records.
2. Document files attached to those records.

It was decided that the ERMS Webdesktop PDR pilot export as received from the vendor should be restructured. There is at least one XML file per ERMS object, some containing also nested copies of child objects. The attachment (document) file names were extracted from the XML and document files were renamed to their original form and grouped by object ID into subfolders (see below the before and after structure).

The pilot export package was first organized into a folder structure where one folder (XML/) under the topic root folder contained all XML files and the other folder (files/) contained all document files belonging to all of the objects of that type (topic).

The original directory structure looks like this:

- types_export/
    - 57835/
        - unit_58119/

        - object_type_name/
            - XML/
                - XXX.xml
            - files/
                - YYY

In this structure:

- 57835 is EKA as WD-s client (an organization, EKA)
- unit_58119 is an organizational unit (in EKA-s case an actual organization)
- object_type_name refers to the data-type e.g. noukogu_otsused (decisions/resolutions of the council).
- XXX is the object ID (positive integer)
- YYY is the file ID (positive integer) as referenced in the XML

Furthermore, folders outside the topic root folder might have files related to this certain topic. These files were linked to each other via XML metadata. The pilot export contained the files without the file names and endings. Files were just a series of numbers with IDs, names, and other metadata described inside the XML metadata file. It was discussed among project partners that it will be simpler to modify the structure with a shell script, than waiting for a new export from the vendor. EKA created a simple shell script that transforms the original structure into a more manageable format. The new structure is organized so that the topic root folder contains the metadata XML files and all the payload files related to each XML file are inside similarly named subfolders with their names and extensions.

The new structure of the ERMS export file looks as follows:

- renamed_1/
    - object_type_name/
        - XXX.xml
        - XXX/
            - real_file_name_1.ext
            - real_file_name_N.ext

(2). CMS EKA Digiteek export package:

EKA Digiteek (Metfond collection) export package is structured as follows:

- metfond_deref/
    - category_object.sql
    - category_object.xml
    - description.sql
    - description.xml
    - event_objects.xml
    - Metfond.xml (Main metadata file which will be used to capture content into mets.xml and dc.xml files)
    - object_event_1.sql
    - object_event_images.sql
    - object_event_instructors.sql
    - object_event_pub_authors.sql
    - object_events.sql
    - object_events.tsv
    - object_events.xml
    - AAA/
        - object_event_1.tsv
        - object_event_images/
            - BBB/
                - CCC/
                    - filename.ext
        - object_event_images.tsv
        - object_event_pub_authors.tsv
        - title_image/
            - filename.ext

The top-level directory name “metfond_deref/” refers to the fact that symbolic links to media have been de-referenced - i.e. the export contains all the full-quality media files. For processing efficiency, the original files are normally kept separate and referenced in the export structure by symbolic links. De-referencing is only done for deliverable exports.

metfond.xml contains the package-level metadata - a description of the collection and a few details about the contents.
The SQL files are included for reference and the TSV files contain all deeper-level metadata until we know more precisely what exactly might have restrictions for publication or value for archival. Due to the number of objects, determining the copyright/publishing rights issues is not feasibly done by examining the public web view of each object, which might be incorrect at the moment.

In the structure above AAA is the main object/event ID, CCC is the media file ID and BBB is only created to match the physical layout of the file location in the CMS web server, for reference and easier restoration of data from the archive if such need arises.

(3). ERMS EKA Webdesktop records with archival value (AV) export package:

The third export package was extracted from ERMS Webdesktop by sub-contractor Webware according to the sample file and schema provided by NAE (Appendix 1).
Similarly to the first export package (PDR), the third export package includes the records/items with metadata, files, and relations between published records.

The export files are presented in two formats:
1. XML files, which include metadata, relations between the records, and references to the files connected with the records.
2. Document files attached to those records.

Similarly to the first PDR export package the third export package was organized in a folder structure where one folder (XML/) under the topic root folder contained all XML files and the other folder (files/) contained all document files belonging to all of the objects of that type (topic). The biggest problem was the files were presented without their endings. File names in export were “failiID” and in metainfo “file name/ID”. Even though the filename and the ending could be found from the associated metadata the script of OneClick SIP Creator couldn’t handle such situations correctly. The reason is that it is not forbidden to upload files with exactly the same name to the ERMS system, but exporting files with exactly similar names was not possible. Therefore, the new export package was provided by the sub-contractor with added file extensions, replacing “failiID” to “failiID.extension” by adding xlsx, txt, pdf, jpg etc.

2.2 TO-BE status of EKA collections

All export packages will be migrated to E-ARK SIP.

3. Identification of the data format, location, and sensitivity

1	Pilot: **EKA ERMS Webdesktop (WD) Public Document Registry (PDR)**
	○ **Data format:** there are XML files with all the metadata of a record and document/attachment files for those records in a directory structure described above. ○ **Location:** the Webdesktop ERMS is hosted by Webware OÜ and all exports of the contents are done case-by-case based on negotiations. There’s no way for a client to do an export using existing functionality at the moment. ○ **Sensitivity:** there was no sensitive information in the export, as it was based on the PDR, which is freely accessible.
2	**EKA CMS Digiteek**
	○ **Data format**: all the data is kept in a PostgreSQL database and filesystem of a web server. The public and administration UI-s are served by a legacy Perl application. ○ **Location**: the CMS is self-hosted by EKA in a legacy CMS. We have full access to the application's source code, database, and media files. The proprietary source code is not documented but it is possible to get information and limited support from the developer (Trinidad Wiseman LLC), e.g. for a server upgrade. ○ **Sensitivity**: Sensitive information was avoided by utilizing similar SQL queries that are used to render the public view.
3	**EKA ERMS Webdesktop records with archival value (AV)**
	o **Data format**: there are XML files with all the metadata of a record and document/attachment files for those records in a directory structure described above and in Appendix 1. o **Location**: the Webdesktop ERMS is hosted by Webware OÜ and all exports of the contents are done case-by-case based on negotiations. o **Sensitivity**: this export package includes sensitive information, restrictions seen in the Appendix 2.

4. The size and scope of the migration

1	Pilot: **EKA ERMS Webdesktop (WD) Public Document Registry (PDR)**
	○ **The scope** of the pilot export was all records except correspondence, marked as public and having no restrictions on publication. ○ **The export size** was moderate (i.e. not considerable - 351 records with 649 document attachments in total).
2	**EKA CMS Digiteek**
	○ **The scope** of this initial export is one collection from the Digiteek CMS called Metfond. ○ **The size** of a full export of Metfond with only the original (full-quality) media files is approximately 27GB and contains 2521 artworks or projects with 7080 media files in total. It contains the most valuable works from the years 1921 to 2020.
3	**EKA ERMS Webdesktop records with archival value (AV)**

		o **The scope** of this export consists of all records appraised as AV to be transferred to NAE for LT preservation. The detailed content of the package is seen in Appendix 2 with restrictions information. o **The size** of the export package of AV records is 1,93 GB (8048 Files, 2 829 Folders)

5. Backup

1	Pilot: **EKA ERMS Webdesktop (WD) Public Document Registry (PDR)**	
		○ We received a zip file from the vendor and transformed it as described above. We have no direct access to the ERMS system to do extractions and rely on the vendor to provide us with packages that we can then transform as we see fit.
2	**EKA CMS Digiteek**	
		○ It was done by first making an offline copy of all the media files and then running read-only (SELECT) SQL queries in the online database. There was no risk of damaging the original data, plus we have nightly backups of the CMS.
3	**EKA ERMS Webdesktop records with archival value (AV)**	
		○ We received a zip file from the vendor and transformed it as described above. We have no direct access to the ERMS system to do extractions and rely on the vendor to provide us with packages that we can then transform as we see fit.

6. Team roles, responsibilities, and migration tool

- EKA ERMS Webdesktop is a proprietary system and the export files are provided by local company Webware, which owns and operates this system. EKA performed a minor transformation of an export package provided by Webware.
- EKA CMS Digiteek is also a proprietary system but it is self-hosted and EKA has access to the database (PostgreSQL) and source code (mostly Perl). The export was done by EKA since the company that developed Digiteek considers it a legacy system and won't offer any support.
- Migration tool, AKA the Oneclick SIP creator, is built by using the Python Anaconda environment with few suitable extension packages and with a simple PHP8.1 powered WebUI. The SIP creator is currently fully dockerized and operational, however, the repository will be kept private until the final release of the tool, by the end of this year. After the release, the SIP creator can easily be deployed by using the included Docker script.

7. Execution of the data migration/Migration process

- Pre-migration activities
  - The SIP creator accepts all uploaded files and folders as they are but if requested, a virus check is conducted before further processing. By default, this check is enabled in a Docker-based installation of the SIP creator.
- Migration activities
  - At present the implemented SIP generator ingests any number of files, folders, and combinations and creates a number of SIP packages based on the input data. Everything that is uploaded via web UI at the same time ends up inside the same SIP package. The SIP creator process follows the following workflow.
    - Ensure that the upload is completed
    - Check if the uploaded folder contains file/files that might contain exported metadata (name(s) of the metadata files can be defined in config.ini file included in GitLab download)
      - If metadata file(s) are found, suitable metadata for the SIP package DC.xml and METS.xml fields are collected
    - Basic SIP directory structure is created
    - Payload files are moved into the SIP structure and original metadata files are moved into the originals folder (from a SIP point of view these metadata files are also handled as payload files)
    - Representation level METS.xml file is generated
    - Package level METS.xml and DC.xml files are generated
    - Folder structure is backed into a zip file named according to SIP UUID
    - Generated package is validated against the commons-IP version 2.3
    - If the validation is successful, the validation report, conversion log, and the SIP file are packed into a zip file
      - In case the validation is unsuccessful the final package is still created but with a prefix INVALID_
    - Created final package is returned to the web UI

8. Testing/validation

Generated SIP packages are automatically validated against the newest E-ARK SIP specification using the latest version of the commons-IP tool (https://github.com/keeps/commons-ip). At the time of writing the latest version is 2.3. All generated packages have been successfully validated.

Migrated content will be verified by multiple persons to ensure that the integrity of the data is OK and that everything is migrated successfully. Furthermore, a comprehensive audit of the process and its results is conducted by the project personnel based on the migration logs.

9. Calendar

The migration of PDR content (export 1) will be done annually for EKA's own use and backup security reasons.

It is estimated that EKA will migrate CMS Digiteek collection (export 2) regularly to E-ARK archival format until the creation (or purchase) of an E-ARK-compatible repository solution in the future. EKA Digiteek collection will be migrated to E-ARK archival format annually as the collection is replenished with new theses at the end of each academic year.

Migration of EKA ERMS content to E-ARK archival format can be done annually only after the regular control of series integrity and correctness of the previous year's documents which is performed each year in March. Hence, the schedule of EKA ERMS migration follows the principle of the DCC Curation Lifecycle Model of transferring documents from the active usage/curation phase to the passive, preservation phase and will take place annually, in April.

In the future, the migration of EKA ERMS content will be easier to perform as the customer-oriented solution of records selection and extraction from ERMS is planned to be implemented at the beginning of 2023. The solution will eliminate the problematic ordering of special development work, and the customers of WD can mark appropriate records and extract the content more easily by themselves.
The migration of records with archival value (export 3) depends on the future requirements of NAE, also the frequency of transfer will depend on future developments of a direct interface between data providers and NAE. The first testing during this project has been successful and the cooperation within these developments is ongoing.

10. Risks

The biggest risk of migrating EKA ERMS collection to E-ARK format is related to the problem that EKA cannot export the data out from ERMS WD and depends on the service agreement between software provider Webware. All possible reasons for increasing risks (personnel, business, or technical issues, also cyber-attacks), must be settled in the agreement. Moreover, the risk of customer-oriented selection and extraction solution development is also high, meaning that EKA will probably continue to order a data export each year.

Risks related to Digiteek collection migration depend mostly on human mistakes, system bugs, cyber-attacks, and major disasters of nature or society (war).

Migration risks are related to other risks of digital preservation and will be covered in more detail in the M2.19 EKA collections preservation plan

Appendix 1

1. Sample file by NAE – a separate file attached to this document
2. NAE-API Schema – a separate file attached to this document

Appendix 2

List of ERMS Webdesktop AV records with restrictions and location data:

Series ID	Series Indicator	Series title	Value	Retention Schedule	Restriction	Restriction base	Data types	Type ID	Core type
58541	1-2	Nõukogu koosolekute protokollid	AV	permanent	internal use	AvTS § 35 lg 2 p 1	Nõukogu koosolekute protokollid	#2261385	Protocols
							Nõukogu koosolekute protokollid (KUNI 31.12.2019)	#62239	Protocols
58544	1-3	Nõukogu määrused KUNI 31.12.2019	AV	permanent	N/A		Nõukogu määrused KUNI 31.12.2019	#58464	Regulations/Decisions
58547	1-4	Nõukogu otsused KUNI 31.12.2019	AV	permanent	N/A		Nõukogu otsused KUNI 31.12.2019	#58475	Regulations/Decisions
58944	1-5	Kuratooriumi koosolekute protokollid KUNI 31.12.2019	AV	permanent	internal use	AvTS § 35 lg 2 p 2	Kuratooriumi koosolekute protokollid (KUNI 31.12.2019)	#62168	Protocols
58946	1-6	Kvaliteedikomisjoni protokollid, otsused ja aruanded (Õppeteenuste n	AV	permanent	Not published		Kvaliteedikomisjoni protokollid	#62195	Protocols
							Saabunud kiri	#57913	Correspondence
						N/A	Väljasaadetav kiri	#57941	Correspondence
58948	1-7	Valitsuse koosoleku protokollid	AV	permanent	internal use	AvTS §35 lg 1 p 12,17	Juhtimiskoosoleku protokollid	#2258691	Protocols
58950	1-8	Rektori käskkirjad	AV	permanent	N/A		Rektori käskkirjad	#58570	Directives
61653	1-13	Nõupidamiste protokollid ja otsused (sh ajutised komisjonid, töögrupp	AV	permanent	internal use	AvTS § 35 lg 1 p 17 AvTS § 35 lg 1 p 12, AvTS § 35 lg 2 p 3	Ajutiste komisjonide jm nõupidamiste protokollid	#62277	Protocols
61656	1-14	Teadusnõukogu koosolekute protokollid ja otsused	AV	permanent	internal use	AvTS § 35 lg 1 p 12, AvTS § 35 lg 2 p 3	Teadusnõukogu protokollid ja otsused	# 62291	Protocols
64228	1-17	Kirjavahetus valitsusasutustega (Riigikantselei, Vabariigi Valitsus, min	AV	permanent	N/A	AvTS § 35 lg 2 p 2, AvTS § 35 lg 2 p 3, AvTS § 35 lg 1 p 12, AvTS § 35 lg 1 p 16, AvTS § 35 lg 1 p 17	Saabunud kiri	#57913	Correspondence
							Väljasaadetav kiri	#57941	Correspondence
3445592	1-19	AV - (KUNI 31.12.2018) Koostöö- ning konsortsiumlepingud teiste kõr	AV	permanent	internal use	AvTS § 35 lg 1 p 17	Lepingud (koostöö, üld, laen, vara, IT jm)	#62827	Agreements
119689	1-21	Siseauditiga seotud dokumentatsioon	AV	permanent	internal use	AvTS § 35 lg 1 p 12	Saabunud kiri	#57913	Correspondence
							Väljasaadetav kiri	#57941	Correspondence
1581907	1-22	Koostöö- ja konsortsiumlepingud teiste kõrgkoolide ja teadusasutuste	AV	permanent	Not published		Lepingud (koostöö, üld, laen, vara, IT jm)	#62827	Agreements
2230652	1-23	Nõukogu määrused	AV	permanent	N/A		Nõukogu määrused	#2224560	Regulations/Decisions
2230658	1-24	Nõukogu otsused	AV	permanent	N/A		Nõukogu otsused	#2227440	Regulations/Decisions
2230727	1-25	Senati määrused	AV	permanent	N/A		Senati määrused	#2230700	Regulations/Decisions
2230739	1-26	Senati otsused	AV	permanent	N/A		Senati otsused	#2230744	Regulations/Decisions
2450447	1-27	Senati koosolekute protokollid	AV	permanent	internal use	AvTS § 35 lg 2 p 1	Senati koosolekute protokollid	#2261916	Protocols
64223	2.1-12	Rektori valimiskomisjoni protokollid	AV	permanent	internal use	AvTS § 35 lg 1 p 12	Rektori valimiskomisjonide ja -kogude protokollid	#62344	Protocols
64225	2.1-17	Kirjavahetus rektori valimistega seotud küsimustes	AV	permanent	internal use	AvTS § 35 lg 1 p 12	Saabunud kiri	#57913	Correspondence
							Väljasaadetav kiri	#57941	Correspondence
684168	4-10	Protokollid (ehitus ja -hankeprotokollid)	AV	permanent	internal use	AvTS § 35 lg 1 p 9 AvTS § 35 lg 1 p 17	Ehituskoosoleku protokollid	#684134	Protocols
							Hankeprotokoll	#2307825	Protocols
61619	6.1-1	Vastuvõtukomisjoni protokollid, erialakomisjonide protokollid, eksamit	AV	permanent	internal use	AvTS § 35 lg 1 p 12	Vastuvõtukomisjoni protokollid	#114345	Protocols
61805	6.1-7	Teaduskonna nõukogu koosolekute protokollid	AV	permanent	internal use	AvTS §35 lg 1 p 12, 17	Teaduskonna nõukogu koosolekute protokollid	#62790	Protocols
61815	6.1-12	Lõputööde kaitsmise komisjonide koosolekute protokollid (BA, MA, D	AV	permanent	internal use	AvTS § 35 lg 1 p 12	Lõputööde kaitsmise komisjonide koosolekute protokoll	#62440	Protocols
61834	6.2-1	Doktorikooli juhatuse koosoleku protokollid ja otsused	AV	permanent	internal use	AvTS § 35 lg 2 p 3	Doktorikooli juhatuse koosoleku protokollid ja otsused	#62477	Protocols
61838	6.2-3	Eelretsenseerimise komisjoni koosolekute protokollid	AV	permanent	internal use	AvTS § 35 lg 2 p 3	Eelretsenseerimise komisjoni koosolekute protokollid	#269916	Protocols
408459	6.2-6	Doktoritööde kaitsmisnõukogude protokollid	AV	permanent	internal use	AvTS § 35 lg 2 p 3	Muud doktorikooli protokollid	#62305	Protocols
61846	7-2	Koosolekute protokollid	AV	permanent	internal use	AvTS §35 lg 1 p 12, 17	Täiendõppe koosolekute protokollid	#62421	Protocols
61863	8-7	Kirjavahetus asutuste, ettevõtete ja eraisikutega ARENDUSTEGEVUS	AV	permanent	internal use	AvTS § 35 lg 1 p 17	Saabunud kiri	#57913	Correspondence
							Väljasaadetav kiri	#57941	Correspondence
1651653	8-10	Teadusgrantide aruanded	AV	permanent	Not published		Projektide/grantide aruanded	#2280582	Reports
61869	9.1-1	Kommunikatsiooniga seotud dokumentatsioon (sh suhtlemine meediag	AV	permanent	N/A		Saabunud kiri	#57913	Correspondence
							Väljasaadetav kiri	#57941	Correspondence