The E-ARK project:
An EU-funded archival cooperation
... Just finished

Arne-Kristian Groven,
Senior Advisor,
The National Archives of Norway,
NorDig,
Stockholm, 25.-26.10 2017
This presentation aims at

• Giving a brief introduction to the E-ARK project
• Presenting our pilot, performed at the National Archives of Norway, both goals and results
• Showing the (main) tools used in the pilot from an end-user perspective, using a few screenshots.
• Discussing, briefly, both project and pilot
Welcome to the E-ARK Project

(European Archival Records and Knowledge Preservation)

E-ARK is a multinational big data research project that aims to improve the methods and technologies of digital archiving, in order to achieve consistency on a Europe-wide scale.

Tackling a range of problems associated with independent record-keeping technologies, systems and practices, E-ARK aims to impact the development of internationally accessible archives through: the provision of technical specifications and tools, the development of open source software, and a programme of international knowledge exchange.

E-ARK is an innovative and collaborative research project that brings together experts from across Europe to improve the quality and accessibility of digital archives. The project is funded by the European Commission and is led by a consortium of leading research institutions.

The project’s goals are to:

1. Develop a framework for the creation of high-quality digital archives that is consistent with international standards.
2. Create open-source software tools that can be used by researchers and institutions to improve the accessibility and usability of digital archives.
3. Improve the training of researchers and archivists in the latest techniques and tools for digital archiving.

The E-ARK project is an important step towards creating a more accessible and sustainable future for digital archives in Europe and beyond.
Partners

University of Brighton (E-ARK Project Co-Ordinator)

AIT Austrian Institute of Technology GmbH

Archives of the Republic of Slovenia - Arhiv Republike Slovenije

The DLM Forum Foundation

Statens Arkiver: The Danish National Archives

The Digital Preservation Coalition

Universität zu Köln: The University of Cologne

Instituto Superior Técnico - Lisbon Technical University

The National Archives of Hungary

Rahvusarhiiv: The National Archives of Estonia

Arkivverket: National Archival Services of Norway

ES Solutions (Sweden)

Magenta (Denmark)

KEEP Solutions LDA (Portugal)

Agência para a Modernização Administrativa IP - Agency for Public Services Reform, IP (Portugal)

Ministerio de Hacienda y Administraciones Pública - The Ministry of Finances and Public Administrations (Spain)

University of Portsmouth

(Plus the Swiss Federal Archives)
The E-ARK Objectives

• In co-operation with commercial systems providers, E-ARK will create and pilot a pan-European methodology for electronic document archiving.

• [E-ARK will be] synthesising existing national and international best practices, that will keep records and databases authentic and usable over time.

• E-ARK will demonstrate the potential benefits [for public... ] by providing simple, efficient access to the workflows for the three main activities of an archive - acquiring, preserving and enabling re-use of information.
Technologies used in our pilot

• The ESSArch software, which is open source software available from:
  – ESSArch Tools Producer, ETP, (v0.93.1) http://etp.essarch.org/
  – ESSArch Tools Archive, ETA (v0.93.1) http://eta.essarch.org/
  – ESSArch Preservation Platform, EPP, (v2.7.3) http://epp.essarch.org/

• The vendor is ES Solutions AB.
Initial pilot goals

• To evaluate ESSArch tools on three real world production data sets, containing a fairly substantial number of digital records
• To specifically look into the ability to get a job done with a satisfactory result
• Also to provide quantitative measurements around time consumption
• Preferably using the E-ARK package definitions
Deviations/challenges

• When these pilot goals were defined at the beginning of the project, the direction of the E-ARK work regarding OAIS package structure implementations was still not known to us.

• We hoped it would imply minor changes compared to OAIS package structure we had defined at the National Archives of Norway a few years earlier.

• But the resulting E-ARK package structure represented more substantial differences than anticipated.
Options available for us

• Either to apply the E-ARK package definitions on (constructed) data sets on test servers, or
• Evaluate the E-ARK tools on the production data set, but with another OAIS package structure
• Or both
How we addressed these challenges

• We ran the three original scenarios on real world data sets, using tools developed in E-ARK, but not on the E-ARK package structure.

• In addition we ran (in December 2016 to January 2017) additional test scenarios on both real and constructed data sets in a test environment, using the E-ARK package structure.

• By doing it this way we could also evaluate the tools ability to handle different OAIS package structures.
Screenshots:
ESSArch Tools Producer

• Preparation of the digital material, including submission agreements, specifying profile details etc.

• SIP creation

• SIP submission to ESSArch Tools Archive
### ESSArch Tools for Producer

#### Create SIP Submit SIP

**Prepare IP Create SIP**

<table>
<thead>
<tr>
<th>Label</th>
<th>Responsible</th>
<th>Date</th>
<th>State</th>
<th>Events</th>
<th>Status</th>
<th>Delete</th>
</tr>
</thead>
<tbody>
<tr>
<td>test06</td>
<td>admin</td>
<td>2017-01-25 19:28:26</td>
<td>Created</td>
<td>Events</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>test03</td>
<td>admin</td>
<td>2016-12-02 21:04:03</td>
<td>Creating</td>
<td>Events</td>
<td>61%</td>
<td></td>
</tr>
<tr>
<td>test union</td>
<td>admin</td>
<td>2016-12-02 20:41:34</td>
<td>Creating</td>
<td>Events</td>
<td>75%</td>
<td></td>
</tr>
<tr>
<td>test02</td>
<td>admin</td>
<td>2016-11-18 11:43:32</td>
<td>Prepared</td>
<td>Events</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

**Event detail**

<table>
<thead>
<tr>
<th>ID</th>
<th>Event detail</th>
<th>Event time</th>
</tr>
</thead>
<tbody>
<tr>
<td>311f1769-86df-40ef-ad9f-c80e77ac32924</td>
<td>Update IP status of test06</td>
<td>2017-01-27 10:14:46</td>
</tr>
<tr>
<td>022e37b-7612-431a-bc0f-4517ba38a7</td>
<td>Create TAR file /ESSArch/data/algorithm/eb/022e37b-7612-431a-bc0f-4517ba38a7.tar from /ESSArch/data/algorithm/eb/022e37b-7612-431a-bc0f-4517ba38a7</td>
<td>2017-01-27 10:14:45</td>
</tr>
<tr>
<td>7cb9256c-085f-94f9-36b2-5e5752eae22</td>
<td>Validate XML file /ESSArch/data/algorithm/eb/022e37b-7612-431a-bc0f-4517ba38a7.mets.xml</td>
<td>2017-01-27 10:14:45</td>
</tr>
<tr>
<td>952b8740-a8e0-4c8b-bd9e-5728e3f1c4b8</td>
<td>Validate checksum for /ESSArch/data/algorithm/eb/022e37b-7612-431a-bc0f-4517ba38a7.mets.xml against 2a800bcb70280c622b33c5797f5865a03</td>
<td>2017-01-27 10:14:45</td>
</tr>
<tr>
<td>c66a9e7-0669-4141-81c5-979208c40e9d</td>
<td>Validate XML file /ESSArch/data/algorithm/eb/022e37b-7612-431a-bc0f-4517ba38a7.mets.xml against 2a800bcb70280c622b33c5797f5865a03</td>
<td>2017-01-27 10:14:45</td>
</tr>
<tr>
<td>d809a737-1b63-4ce5-bdb1-040463332694</td>
<td>Validate logical representation in /ESSArch/data/algorithm/eb/022e37b-7612-431a-bc0f-4517ba38a7.mets.xml against physical representation at /ESSArch/data/algorithm/eb/022e37b-7612-431a-bc0f-4517ba38a7.mets.xml</td>
<td>2017-01-27 10:14:45</td>
</tr>
<tr>
<td>db716e64-d2e4-4728-b8ba-85debf70ee2a</td>
<td>Validate checksum for /ESSArch/data/algorithm/eb/022e37b-7612-431a-bc0f-4517ba38a7.mets.xml against 2a800bcb70280c622b33c5797f5865a03</td>
<td>2017-01-27 10:14:45</td>
</tr>
<tr>
<td>f4d794b7-ab9d-4f9a-ba78-0c272c505d5f</td>
<td>Create ZIP file /ESSArch/data/algorithm/eb/022e37b-7612-431a-bc0f-4517ba38a7.mets.xml</td>
<td>2017-01-27 10:14:45</td>
</tr>
</tbody>
</table>
### Package dependencies

<table>
<thead>
<tr>
<th>Archival Institution</th>
<th>Archivist organization</th>
<th>Archival type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Riksarkivet</td>
<td>Skatteverket</td>
<td>document</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Archival location</th>
<th>Archive Policy</th>
<th>Container format</th>
</tr>
</thead>
<tbody>
<tr>
<td>sweden-stockholm-nacka</td>
<td>archive policy 1</td>
<td>TAR</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Container format compression</th>
<th>Checksum algorithm</th>
<th>Submission reception validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>MD5</td>
<td>yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Submission reception exception handling</th>
<th>Submission reception receipt confirmation</th>
<th>Submission risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>none</td>
<td>none</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Submission mitigation</th>
<th>Information package file</th>
<th>Submission information package file</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>ip.xml</td>
<td>sip.xml</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Archival information package file</th>
<th>Dissemination Information package file</th>
<th>Submit description file</th>
</tr>
</thead>
<tbody>
<tr>
<td>aip.xml</td>
<td>dip.xml</td>
<td>info.xml</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Content type specification file</th>
<th>Archival description file</th>
<th>Authority information file</th>
</tr>
</thead>
<tbody>
<tr>
<td>erms.xml</td>
<td>ead.xml</td>
<td>eac_cpf.xml</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Preservation description file</th>
<th>IP event description file</th>
<th>Mime types definition file</th>
</tr>
</thead>
<tbody>
<tr>
<td>premis.xml</td>
<td>ipevents.xml</td>
<td>mime types</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Preservation organization receiver email</th>
<th>Preservation organization receiver url</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:receiver@archive.xxx">receiver@archive.xxx</a></td>
<td>https://eta-demo.essarch.org.reta_reti</td>
</tr>
</tbody>
</table>

### Included profiles

- event
- submit_description
- transfer_project
- sip

### Validator choices
File and folders, the package structures created
File and folders, the package structures created
Screenshots:
ESSArch Tools Archive

• Receives SIPs created in ESSArch Tools Producer
• Transfer SIPs into ESSArch Preservation Platform

Arne-Kristian Groven, 2017
Screenshots:
ESSArch Preservation Platform

• Receives and validates the SIPs
• Creates the AIPs
  – but no automated file conversion (migration)!
• Storage management
• Checks in AIPs (and SIPs)
• Checks out AIPs upon requests
• Media management and monitoring
EPP (ESSArch Preservation Platform) 2.8.1 - Status at 2017-01-27 11:40

Processes:
- SIPReceiver: 1
- SIPValidateAIS: 1
- SIPValidateApproval: 1
- SIPValidateFormat: 1
- AIAPreprocessor: 1
- AIAPreprocessor: 1
- AIAPreprocessor: 1
- AIPPreprocessor: 1
- TLD: 1
- db_sync_ais: 1
- ESSlogging: 1
- AccessEngine: 1
- FTPServer: 1

Proc logfile alarms:
None

Storage media overview
Archive tapes online:
None
Add new tapes starting with:
None
Write tapes:
None
Empty tapes:
None
Error tapes:
None
Ongoing requests
None

No uncompleted requests
Select which information package to checkin from reception

<table>
<thead>
<tr>
<th>IDENTIFICATION</th>
<th>ARCHIVIST ORGANIZATION</th>
<th>LABEL</th>
<th>CREATE DATE</th>
<th>START DATE</th>
<th>END DATE</th>
<th>TYPE</th>
<th>MEDIA</th>
<th>STATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>f37654a-670d-4fd2-a8cf-b1b885ee7506</td>
<td>the archivist</td>
<td>Test (2000 - 2012) - 1234</td>
<td>2016-11-18T11:04:46.011000+01:00</td>
<td>2016-11-10</td>
<td>2016-12-20</td>
<td>SIP</td>
<td>EFT</td>
<td>Reception</td>
</tr>
<tr>
<td>i4a34i50-0806-481f-b0c5-abb3082226eba</td>
<td>the archivist</td>
<td>test 04</td>
<td>2017-01-24T14:12:25.168000+01:00</td>
<td>2016-11-10</td>
<td>2016-12-20</td>
<td>SIP</td>
<td>EFT</td>
<td>Reception</td>
</tr>
<tr>
<td>e32ceeb7-0345-4c85-8f15-6b8e4ab78e30</td>
<td>the archivist</td>
<td>test06</td>
<td>2017-01-27T10:14:45.851000+01:00</td>
<td>2016-11-10</td>
<td>2016-12-20</td>
<td>SIP</td>
<td>EFT</td>
<td>Reception</td>
</tr>
</tbody>
</table>

Showing 1 to 3 of 3 entries
Request to check in IP from reception

Result:
Category: controlarea
Label: Check in from Reception
User: admin
Request purpose: sjekk inn
Status: OK

Summary:
Found logfile: /ESSArch/data/gate/reception/e32cf6b7-0345-4c85-8f15-6b8e4ab78e30_ipevents.xml for package: /ESSArch/data/gate/reception/e32cf6b7-0345-4c85-8f15-6b8e4ab78e30.tar
Copy /ESSArch/data/gate/reception/e32cf6b7-0345-4c85-8f15-6b8e4ab78e30.tar to package content directory: /ESSArch/data/epc/control/044a758a-001b-4c65-b777-fb5cf9b9426d/e32cf6b7-0345-4c85-8f15-6b8e4ab78e30/content
Success to get METS agents.altRecords and label from e32cf6b7-0345-4c85-8f15-6b8e4ab78e30.xml
Add altRecordID ['POLICYID', 'u'c46cfff1675f5db368e6a6df3579de1'] to metsfile
Add altRecordID ['INFORMATIONCLASS', 'u'i'] to metsfile
Add altRecordID ['DELIVERYTYPE', 'u'NA'] to metsfile
Create new PREMIS: /ESSArch/data/epc/control/044a758a-001b-4c65-b777-fb5cf9b9426d/e32cf6b7-0345-4c85-8f15-6b8e4ab78e30/metadata/premis.xml
Add new entry to D8 for AIC_UUID: 044a758a-001b-4c65-b777-fb5cf9b9426d
Add new entry to D8 for ID_UUID: e32cf6b7-0345-4c85-8f15-6b8e4ab78e30
Add new entry to D8 for eventIdentifierValue: 28854a80-7822-4087-b70f-151d03a63b02
Add new entry to D8 for eventIdentifierValue: 29779aa2-0301-4d8a-8866-82ed0db72d6d
Add new entry to D8 for eventIdentifierValue: dbd70cc0-4ac0-4de5-b767-367cd2b59237
Add new entry to D8 for eventIdentifierValue: d50e601c-4301-4d79-a322-240e7f35d3d
Add new entry to D8 for eventIdentifierValue: b6d6f99b-954c-45ba-9f43-c73c475f6f68
Add new entry to D8 for eventIdentifierValue: 286d9941-5ed8-4103-842b-c63280d099ea
Add new entry to D8 for eventIdentifierValue: 432624a4-c786-8b86-2b233dc75d7e2
Add new entry to D8 for eventIdentifierValue: 073d0539-3a86-4a85-90fd-9a34a03c349a
Add new entry to D8 for eventIdentifierValue: d11ed8b8-de22-425a-b1aa-1f21109f72a0
Add new entry to D8 for eventIdentifierValue: 0da8d77a-9c70-4768-9408-d53852fa87d7
Add new entry to D8 for eventIdentifierValue: 0718514a-75d1-4582-9c4c-9e404b1e63ca
Add new entry to D8 for eventIdentifierValue: fe9b1091-ef02-4c8a-a793-30139563f79c
Data set characterization: Scenario 1

- Data used in this pilot originated from an archival producer called Direktoratet for utviklingssamarbeid (Norad). Norad is the Norwegian Agency for Development Cooperation. The period of data creation was 2003-04-01 to 2005-12-31.

- The data is the standardized output from an EDRMS system, generated according to the national NOARK-4 standard. The output consisting of a generated Noark-4 XML file containing (meta-) data descriptions of the records created in the EDRMS, in addition to the documents (records) represented as (mostly) PDF/A files. The overall size is 20 GB.

Arne-Kristian Groven, 2017
Data set characterization: Scenario 2

• Data used in this pilot originated from an archival producer called Folketrygdfondet, which is a professional investment manager whose main task is to manage the Government Pension Fund Norway on behalf of the Ministry of Finance. The period of data creation was 2005-01-01 to 2014-03-31.

• The data is the standardized output from an EDRMS system, generated automatically according to the national NOARK-5 standard. The output consisting of a generated Noark-5 XML file containing (meta-) data descriptions of the records created in the EDRMS, in addition to the 8194 documents (records) represented as (mostly) PDF/A files. The overall size is 5 GB.
Data set characterization: Scenario 3

• Data used in this pilot originated from an archival producer called Direktoratet for Naturforvaltning, a no longer existing directorate related to environmental administration. The period of data creation was 1985-01-01 to 1999-12-30.

• The data set is a register, originating from an old database made in the 1980s. The data is represented as a set of 10 CSV files, containing 338,500 registrations. The size is 105 MB.
Data set characterization: Real-world additional scenario

• Data used in this pilot originated from a private archival producer, consisting of 12,252 element, 10,135 files in 2,116 directories. The overall size of the content is 2.1 GB.
Summary of pilot results/experiences

• The tools we evaluated were able to produce satisfactory results (output).

• The time consumption on the automated parts of EPP, ESSArch Production Platform, was 1 hour 50 minutes for the largest of the original scenario
  – time consumption was not measured on the additional scenarios
A few words about software risks...

• In general, (negative) risiks associated with software acquisitions, are high...
• If the software has few users
• If few people have deep knowledge about and are able to maintain the software
Software risks continued

• Additionally, for long-term preservation and access software, where data need to undergo many format transformations over time
  – Proprietary software will make quality assurance difficult due to lack of transparency, while
  – Free and open source software will reduce these negative risks, through transparency (open code/open community)

• This is the main reason for many leading archival institutions worldwide to use free and open source software as a principle

Arne-Kristian Groven, 2017
Concluding remarks about the ESSArch products...

- The ESSArch tools coming out of the E-ARK project is of better quality than the tools brought into the project three years ago.
- The software has become open source software, allowing any technically skilled person to gain knowledge about the software/source code.
- Hence, the challenge for ESSArch is to attract enough external software developers and users (user organisations).

Arne-Kristian Groven, 2017
Concluding remarks about the ESSArch products

• In order to do so at least one improvement has to be done, as soon as possible: Providing (good) user documentation!

• Additionally, to integrate, one way or another, with market leading products in order to form an holistic approach based on OAIS
Concluding remarks about E-ARK

• The field of digital long-term preservation and access is about to move from its early phases into a higher level of maturity
• E-ARK and other similar projects play an important role in this, by bringing together many stakeholders and approaches, producing fewer approaches and standardisation as an output
• This was appreciated by the EU appointed reviewers
• The European Commission's Year 3 Review of the E-ARK Project has rated it as "Excellent" and have described it in their initial feedback as a "A European Showcase Project"
• But follow-up projects are needed!

Arne-Kristian Groven, 2017