#### European Archival Records and Knowledge Preservation

#earkproject www.ea

www.eark-project.eu

@EARKProject

# E-ARK format for storage and longterm preservation and the integrated prototype

Meeting of the Member States Expert Group on Digitisation and Digital Preservation



13.10.2015, Luxembourg





# Part I: E-ARK format for storage and long-term preservation







#### WP4 overview



#### WP4 deliverables



#### WP4 main work areas

The AIP format specification is the basis The for the SIP-AIP implementation, nevertheless the development is a two-way dynamic process. SIP-AIP Conversion

Component Implementation

The SIP-AIP Conversion component creates an AIP that complies with the AIP format specification.

EARK-AIP format specification







#### **E-ARK Information Package**

- Directory structure
  - Data and Metadata separated
  - Structure that allows to distinguish descriptive and digital provenance metadata
- METS as the structural metadata standard
- PREMIS as the preservation metadata standard







### AIP format specification

- Hierarchical structure of the information package
  - aligned with a high level specification that is valid across package types (S/A/D)IP
- The AIP format specification consists of
  - a proposed structure that allows storing a SIP which was submitted.
  - a proposed structure that allows adding new representations during ingest or after the AIP was archived.
- Directory tree as well as the use of metadata standards describing structure and preservation data of the information package
  - METS as the structural metadata standard
  - PREMIS as the preservation metadata standard







## Representations

#### • PREMIS definition

"The set of files, including structural metadata, needed for a complete and reasonable rendition of an Intellectual Entity." \*

- Representations can be provided as part of the SIP already
- Creating representations happens either during ingest or after the AIP was archived.
- A new Representation is not necessarily the result of a file format migration, it can also be a set of instructions how to create an emulation environment to render a set of files.

\* Introduction and Supporting Materials from PREMIS Data Dictionary, p. 7.







# Let's start with the SIP ...









































































































XML Schemas to validate XML instances















# ... and now to the AIP ...









It is stored in a separate branch to clearly separate it from representations that are created during ingest or post-ingest.









(Rep-001 was migrated during ingest into Rep-001.2)

































AIP METS

























Representation METS files and

**Descriptive Metadata** 









































Other metadata folders possible at the various levels. For example, PREMIS rights can be placed at the AIP level whereas PREMIS events would be recorded at the representation level. "Documentation" folder as a sibling to "Metadata" at each level (depends on whether the documentation relates to a specific representation or to all representations)















# ... but hold on, what's the point of all this splitting of METS files?









In this example we were dealing with three image files ...

















#### In this case we must be able to separate representations ....





















# Part II: E-ARK integrated prototype







### Goal of the Integrated Prototype

- Downloadable package
  - Software components from WP4, WP5, and WP6
  - Packaging; Search in/across packages; Access
- Usable in combination with existing systems
  - Extends existing commercial and proprietary archiving products (loosely coupled)
  - Provides concrete integration hooks for ESSArch Preservation
     Platform supporting (loose and tight coupling)
- Based on Scalable Technologies
  - Deployed for small/medium workloads on single host
  - Ability to scale out (by design) for large volumes







### **Processing Information Packages**

- Information Packages are processed using automated workflows during ingest and access
  - Implemented as scripts
  - for Information Package creation, validation, storage, ...
- Workflow execution is delegated to a backend processing infrastructure
  - Makes task processing configurable, portable, distributed, asynchronous, …
- End-users can manage workflows using the graphical E-ARK Web Portal
  - Browse, configure, execute, and track workflows
  - create AIP, store AIP, ...







### **Content Repository and Indexing**

- An infrastructure that can be used to extend existing ingest tools, archiving and preservation platforms.
  - Provides a staging area for storing Information Packages, e.g. in addition to tape-based archival storage.
  - Provides support for extracting data and metadata from contained items, which are indexed and stored in a content repository.
  - Provides interfaces that supports searching in and across
     Information Packages on a per-item basis.
  - Provides interfaces that support **random access** on a per-item basis.
- Technically based on broadly used technologies for large-scale content processing
  - Hadoop, Lucene, SoIR, Tika, Lily, HBase







## **ESSArch Preservation Platform**

- Production environment that supports traditional archiving preservation processes (pre-ingest, ingest, different storage methods, ...).
  - Provides a backend for digital libraries and archive information systems
  - Provides the E-ARK reference implementation supporting E-ARK specifications and software components
- Integration I: Using Prototype Workflows inside EPP
  - Direct compatibility based on aligned technology stack.
  - Python (workflow language), Celery (task processing), same SQLbased data model.
- Integration II: Using custom Storage Method in EPP
  - Enables EPP to stage Information Packages to E-ARK content repository and index infrastructure.







#### Why "Integrated" Prototype?

localho	st Ha	dooj	o Ma	p/Redu	ice Adr	ninistra	tion		~	( )n			Quick Links	Auch	Request-Handler (d)	## + ()
State: RUNING Started: In Sep Version: 2.0.0- Compiled: Web	G p 04 08:49:18 mr1-cdh4.7.0 i May 28 10:2 babarana	CEST 20 Unknow 6101 PDT	15 n 2014 by j	erkins from Un	known				h					Dashboard	- common q	This
Churtor E				in is 243	MB/990	40								# Core Admin	ittella.	_
Cluster 3	Running	, (ne	ap su	e 15 342	Occupied	(in the second	Bernard	Bergmand	Mar Tank	Reduce Test		Rechtered	Embedant	🗿 Jive Properties	14	- <1
Map Tasks	Reduce Tasks	51	bnissio	ns Nodes	Map Slots	Reduce Sile	s Map Slots	Reduce Slots	Capacity	Capacity	Tasks,Node	Nodes	Nodes	E Thread Dump		
	0	3			1	0	0	0	2	2	4.00		0	4 1645	sort	
														* goudace	stat. Ines	
Scheduli	ng Infor	matic	on											L carkt	0 10	
Queue Name	e State 1	schedu	ing info	rmation										# /mg		-
06974	[id-ning]	44	_											(L) Ourry		
liker (Jobid, P	vierby, User	, Name		the user field of	d There is all free									Estatu		
Running	Jobs													Ta Nephonine	et.	
jobi	id .	Priorit	y User	Name	Ma	Ma	Maps	Reduce %	Reduce	Reduce	s job	scheduling	Diagnostic	T sealers	C indent	-
-				batchindext	uld 100.000	6 .	n Compete	0.00%		Compres	-			E prostat	C debugtuery	
Complete	ed Jobs			Jub		*			-					Colorgan	C dismax	
			-				Marra	Destroy M	Deathers	. Bertana	n Inte	-had day	Manage 11		CN.	
jobi	id	Priorit	y User	Name	Com	plete Tot	d Complete	d Complete	Total	Complet	ed in	ormation	into		C spatial	
job_20150904	40549,0001	NORM	& niner	BatchindexB jsb	ulid 100.009	1	1	100.00%	•	0	NA.		NA		C spelkheck	
ke 20150904		NORM	L caber	Betchindext	uld 100.009	6 1	1	100.00%		0	NA		NA		Execute Query	

Solr	rselect	
Cashboard	- common	This XML file does not appear to have any style information associated with it. The document tree is shown below
Logging	meta	
# Core Admin		Construction of the second
A JINS Properties		- <response></response>
II Thread Durne	4	<int name="status">0</int>
		<int name="QTime">5</int>
+ news	sort	- <lst name="perans"></lst>
a another		<str name="wt">xmi</str>
* postel	stat, tows	clists
L carki	10   10	
# (Trg		- <result name="response" numfound="4" start="0"></result>
Downy		+ <doc></doc>
E Literar	đ	+ <doc></doc>
a contra		= stir name="ath">
	wt.	DNA_AVID.SA.18001.01_141104/DNA_AVID.SA.18001.01_141104/Metadata
	ini .	/IP_CS_mets.xsd
T contra	O indent	
E prostat	debugbuery	<str name="contentType">application/xml</str> <long name="size">148099</long>
A Placet of States	C dismax	- <str name="content"> METS_Materials_Encoding and Transmission Standard METS is intended to provide a</str>
all Contractions	ediamax	standardized XMI format for transmission of romniev dinital library chierts between
	CN.	systems. As such, it can be seen as filing a role similar to that defined for the
	C facet	Submission Information Package (SIP), Archival Information Package (AIP) and
	C special	Dissemination Information Package (DIP) in the Reference Model for an Open Archive
	C spelkheck	Information System. The root element <mets> establishes the container for the</mets>
	Concession of the local division of the loca	information being stored and/or transmitted by the standard, metsType. Complex Typ
	Execute Query	to HETS Sectors A HETS bounded, designs of seven possible subjects and y sectors.









#### Why "Integrated" Prototype?

State: RUNIN State: FUNIN		acob	1110	/Reduct	Aunn	naciae			1	100					S	olr
Version: 2.0.0 Compiled: We	mr1-cdh4.7.0. i May 28 10:21	Unknown 5131 POT 2	014 by je	kins from Unknow					lit	Tala		1			• Di	bisodrite
identifier: 203	509040849										-				10 Lo	0300
Cluster 9	ummary	/ (Hea	p Siz	e is 342 MB	3/889 MB)										# CO	JIE AGRIN
Running Map Tasks	Renting	Sub	Total	Nodes 0	p Slots Re	occupied duce Slots	Reserved Map Slots	Reserved Reduce Slots	Map Task Capacity	Reduce Task Capacity	Avg. Tasks,Node	Diacklisted Nodes	Decluded Nodes		2.04 2.75	vead Durg
1	0	3		1 1	0		0	0	2	2	4.00		0		4.74	ews .
															* #	orders .
Scheduli	ng Infori	matio	n												Lee	inter-
Queue Nam	e State S	ichedulin	ig inter	nation												
default	running N	49.													P	Overy
Filter (Jobid,	riority, User	Name)													5	
Dample 'seene	109-3209 will fil	for by 'and	R' only in I	he uper field and '32	IP is all fields									. 8		
Running	lobs													<u> </u>	12	
jot	id .	Priority	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce	Reduce	s job 5 rd infe	icheduling irmation	biagnostic Info		T	
job_2015090	0949_0003	NORMAL	rainer	Batchindextfulid ISD	100.00%	1	1	0.00%	0	0	Pail.		NA			
Complet	ed Jobs														6	
jot	id	Priority	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce	Reduce	s jeb S ed inte	icheduling prmation	Clagnostic Info			
job_2015090	40849_0001	NORMAL	rainer	BatchindexBuild Isb	100.00%	1	1	100.00%	0	0	NA		NA			

E-ARK

Apacha	Request-mandler (qt)	## Http://kicalited.it/Millinationals1.public/http://withaid.
Solr	nelect	
· Castlinant	- common	This XML file does not appear to have any style information associated with it. The document
- Construction	4	tree is shown below.
Logging	meta	
# Core Admin		Security and the second s
A Java Properties	Fa	- <ist "response="" a="" leader"="" name=""></ist>
Thread Dump		<int name="status">0</int>
		<int name="QTime">5</int>
o mean	sort	- <lst name="perams"></lst>
-		<str name="wt">xml</str>
+ footex	stati, tows	<str name="q">mets</str>
L carkt		- Clarks
Million .	1	- <result name="response" numfound="4" start="0"></result>
Dame		+ <doc></doc>
and the second s	0	+ <doc></doc>
And a		- <doc></doc>
0.000	-	- 550 hame= pon A
12 Replication	101 .	/IP_CS_mets.xsd
1.00000	Directoret	
Tel De Marrie	O debugbuery	<str name="contentType">application/xml</str>
prowset.		<iong name="mrs">140099</iong>
A Players / State	C dismax	METS: Metadata Encoding and Transmission Standard. METS is intended to provide a
48 Calaireant	- ediamar	standardized XML format for transmission of complex digital library objects between
	C N	systems. As such, it can be seen as filling a role similar to that defined for the
	T facet	Submission Information Package (SIP), Archival Information Package (AIP) and
	C south	Dissemination Information Package (DIP) in the Reference Model for an Open Archival
	C melitheck	Information System. The root element <mets> establishes the container for the</mets>
	a sport age	information being stored and/or transmitted by the standard, metsType. Complex Type
	Execute Query	metsHdr (METS document header), dmdSec (descriptive metsdata section), amdSec







#### Why "Integrated" Prototype?

Iocalhost Hadoop Map/Reduce Administration	Anste Wegestersteinder (g) #192.0623400.082300.08530000000000
Skated hr Sig M Re-19 II CS2 2155 Vender: 22.6-11 cdA J. Lichoon Complete Verlay 28 102 cd1 POT 2314 by jerkins from Unknown	Comparing Control of the second
Cluster Summary (Heap Size is 342 MB/889 MB)	S (or siden     S (or sid
Running         Didd Subsection         Total Subsection         Total Reserved         Occupied Reserved         Reserved Mig Stats         Messarved Reserved         Messarved Mig Stats         Reserved Reserved Reserved         Messarved Reserved         Reserved Reserved         Messarv	If Treed Dump <init faitur's="" names'=""> c/init &gt; c/init</init>
Scheduling Information	* godici van tree <pre>Sut name='u'&gt;nut</pre>
Duose Name State Scheduling Internation	■ = - <reall name="reports" numfound="4" start="0"> ↓ </reall>
Harr gene, Henrie yn de Anno () o yn de sae feit ad 100 m a fleets Renning jobs	- <
jobd         Prisrby         User         Name         High N Complete         Map Total         Complete Complete         Roduce Solution         Roduce Notice         Roduce Solution         Roduce Information         Information	Torusso         Order
Completed Jobs	Ampenume Gome     General
jobid přísníky User Name Maje N Mag Maje Reduce 10 fordat Complete Complete Complete Complete Notal No	Submission Information Peckage (SIP), Archivel Information Peckage (SIP), Archivel Information Peckage (SIP), Archivel Information Peckage (SIP) in the Reference Model for an Open Archive Information System. The root element = metas- establishes the container for the
PR_2015060484_000 NDFNA_ mine Path/modBuld 100.00% 1 1 100.00% 0 NA NA	information being stored and/or transmitted by the standard, metType: Complex 7 Exocutivary for METS Sections A METS document consists of seven possible subsidiary sections: metbid/f /METS document headeri, danside (description); and de





# **SIP-AIP Conversion Component**

- Close to ESS' product ESSArch EPP (<u>http://www.essarch.org</u>) with regards of the technology stack used
  - Python programming language
  - Django-WebUI-Development framework
  - Celery distributed task queue
- Implemented as tasks which can be parallelized on a cluster
  - Modular: Tasks can be combined in different workflows
  - Extensible: Tasks can be added by using a template that is easy to understand (task implementation interface)
- Results will be published as deliverable D4.4 "Final version of SIP-AIP conversion component".







# SIP to AIP Conversion

#### Customizable conversion workflows



#### Characteristics

#### • Modular:

Individual tasks can be used to build workflow variants, e.g. to build a package type specific workflow.

#### • Extensible: Specific tasks can be added at any point in the workflow

#### • Scalable:

Scalability by parallel execution; tasks can be executed as isolated processes (e.g. independent from a database)







# Task interface (DefaultTask)

#### class SIPDeliveryValidation(DefaultTask):

```
accept input from = [SIPtoAIPReset. name , SIPPackaging. name ]
def run task(self, task context):
                                                                                   ——Task configuration
    tl = task context.task logger
    deliveries = get_deliveries(task_context.path, task_context.task_logger)
    if len(deliveries) == 0:
         tl.adderr("No delivery found
                                             working directory")
         task context.task status = 1
    else:
         for delivery in deliveries:
             tar_file = deliveries[delivery]['tar_file']
delivery_file = deliveries[delivery]['delivery_xml']
             tl.addinfo("Package file: %s" % delivery_file)
tl.addinfo("Delivery XML file: %s" % delivery_file)
schema_file = os.path.join(task_context.path, 'schem
                                                                         as/IP CS mets.xsd')
             tl.addinfo("Schema file: %s" % schema file)
             sdv = DeliveryValidation()
             validation result = sdv.validate delivery(task context.path, delivery file, schema file, tar file)
             tl.log = tl.log + validation result.log
             tl.err = tl.err + validation result.err
                                                                      size/checksum): %s" % validation result.valid)
             tl.addinfo("I
             if not validation result.valid:
                  tl.adderr("Delivery invalid:
                                                    %s" % delivery)
```







# Information package status

- Information package status is defined by:
  - Last executed task
  - Success/failure of last task execution
- Task execution control
  - Accepted inputs defined for each task
  - Example: SIPRestructuring task
    - accept\_input\_from = [SIPExtraction]
    - $\rightarrow$  To restructure the SIP it needs to be extracted













REPUBLIKA SLOVENIJA MINISTRSTVO ZA KULTURO ARHIV REPUBLIKE SLOVENIJE



The Danish National Archives





THE E-ARK PROJECT IS CO-FUNDED BY THE EUROPEAN COMMISSION UNDER THE ICT-PSP PROGRAMME

#### www.eark-project.eu





# ama



# MAGENTA<sup>aps</sup>



I THE NATIONAL ARCHIVES OF NORWAY



1111

R A H V U S A R H I I V THE NATIONAL ARCHIVES OF ESTONIA





