The Use of Big Data Techniques for Digital Archiving

Sven Schlarb, Austrian Institute of Technology

Tuesday 15th March 2016, Cambridge

OUTLINE

- E-ARK Project Overview
- Technical Background
- Integrated Prototype
- Data Mining Use Cases

Project Overview

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

www.eark-project.eu

Advisory Boards

Archival

- Archives of Emilia-Romagna, Italy
- Directorate-General of the Book, of Archives & of Libraries, Portugal
- EC Archives & Records Management
- EC Historical Archives
- German Federal Archives
- National Archives of Bulgaria
- National Archives of Finland
- National Archives of France
- National Archives of Sweden
- National Archives of the Netherlands
- Polish Data Archive
- Queensland State Archives
- Swiss Federal Archives
- UK National Archives
- UK Parliamentary Archives

Commercial Technial

- Arkivum
- ARMA Europe
- DigitalForever
- Discovery Garden
- Microsoft Research
- Open Preservation Foundation
- Open Text Initiative
- Preservica
- Versity

Data Providers

- Danish Agency for Digitisation
- Estonian Ministry of Economic Affairs & Communication
- Estonian Unemployment Insurance Fund
- James Lappin, RM Consultant

E-ARK

ICTPSP
ICT POLICY SUPPORT PROGRAMME

Project mission

- Improve access to the archived records of European Archives
- Create guidelines and recommended practices
- Cover relational databases, record management systems, and geographical data
- Create open source implementation evaluated in several pilots

Outcomes

Standardisation of available best-practices

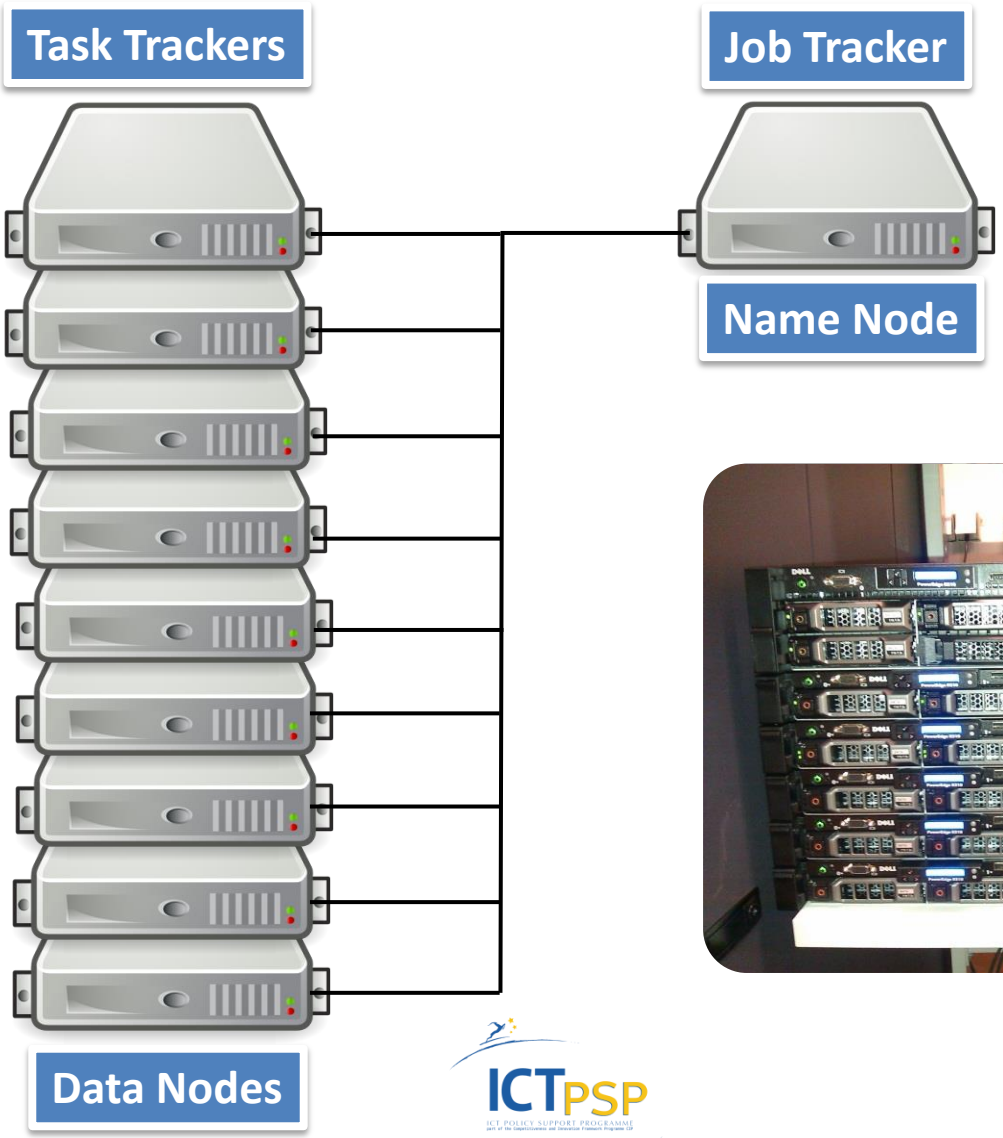
- Common terminology (Knowledge Center)
- SIP, AIP and DIP format specifications
- Pre-ingest, ingest and access workflows

Open source tools

- Scalable, modular, and reusable implementation of specifications
- Individual deployments (Pilots) and an integrated reference implementation

Technical Background

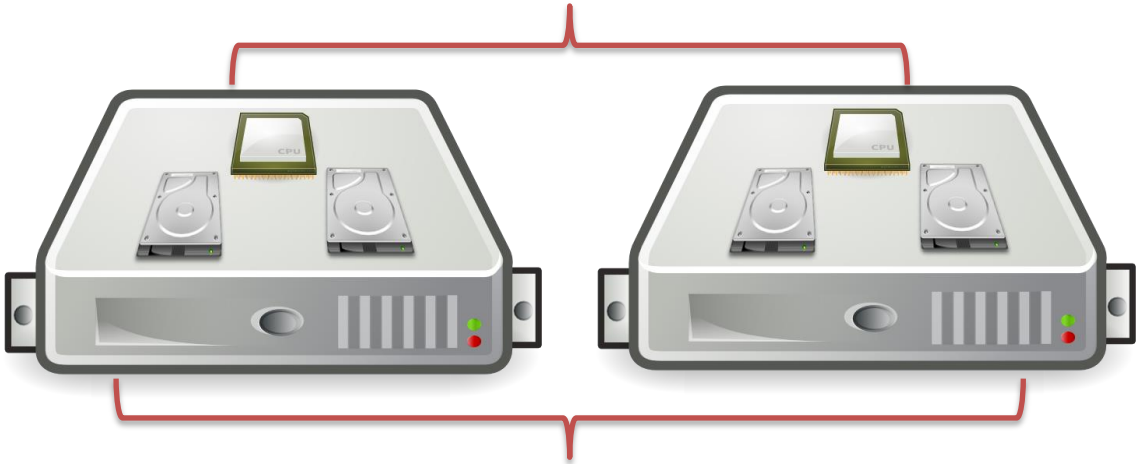
Hadoop Cluster



Hadoop = MapReduce + HDFS

Distributed processing (MapReduce)

example: 2 x Quad-Core-CPUs:
10 Map (Parallelisierung)
4 Reduce (Aggregation)

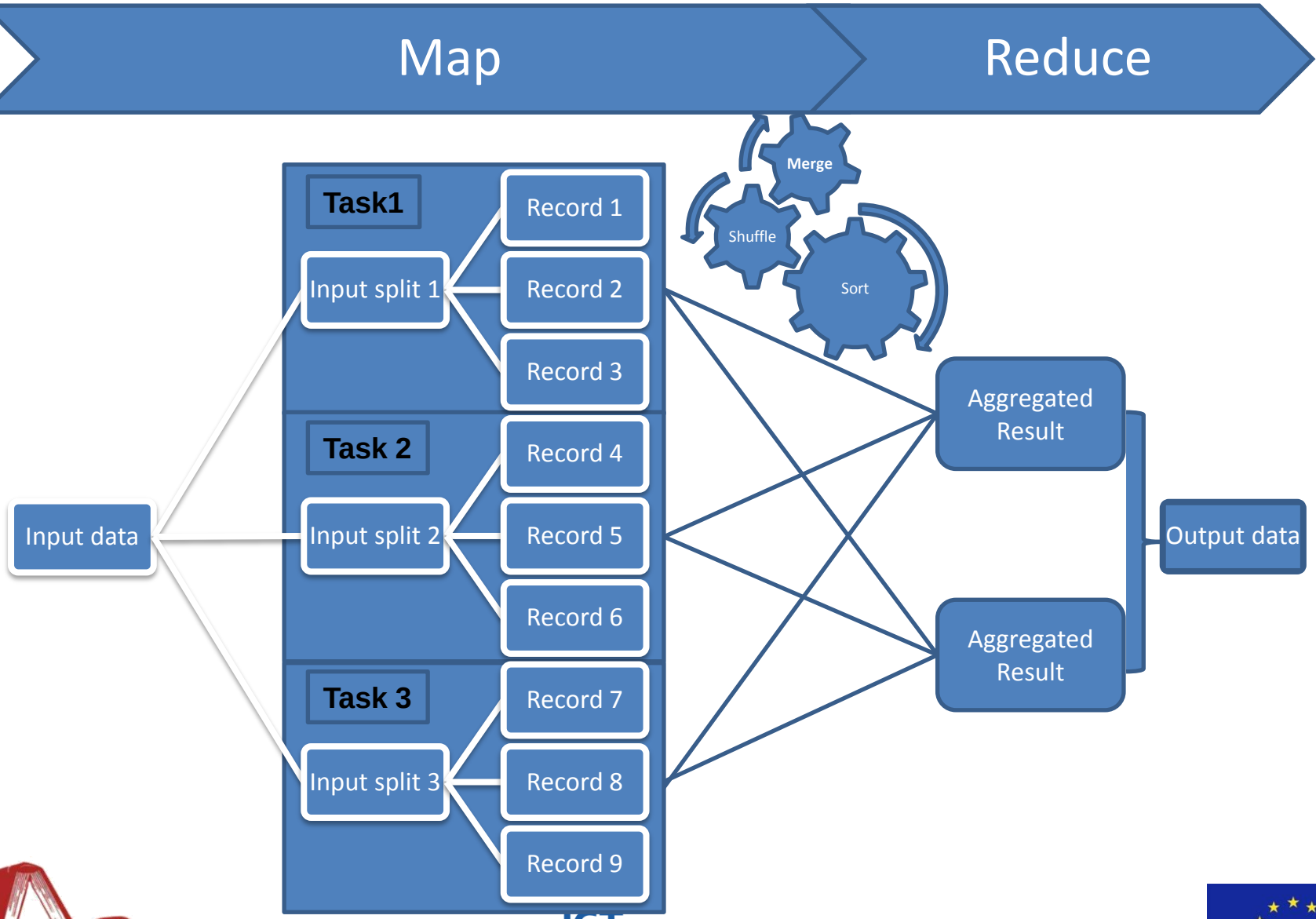


example: 4 x 1 TB Hard-Disks (replication factor 3):
ca. 1,33 TB

Distributed Storage (HDFS)

HADOOP

Map/Reduce in a nutshell



E-ARK

ICTPSP
ICT POLICY SUPPORT PROGRAMME

E-ARK Integrated Prototype Architecture & Implementation

Base technology stack

lily

APACHE HBASE

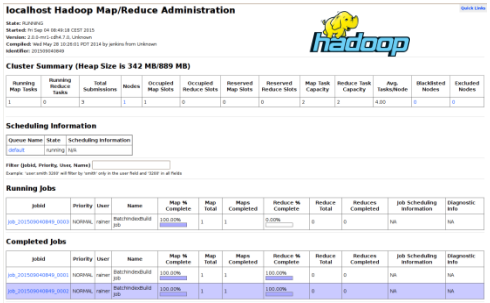
Apache Solr

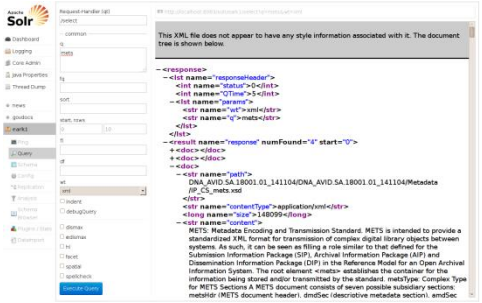
hadoop

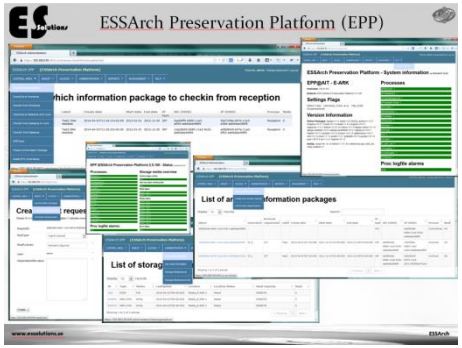
E-ARK

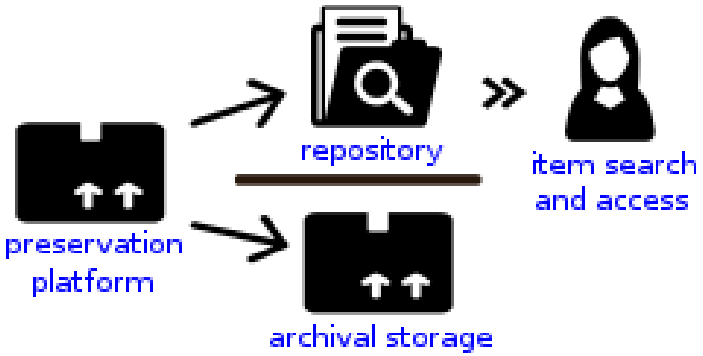
ICTPSP

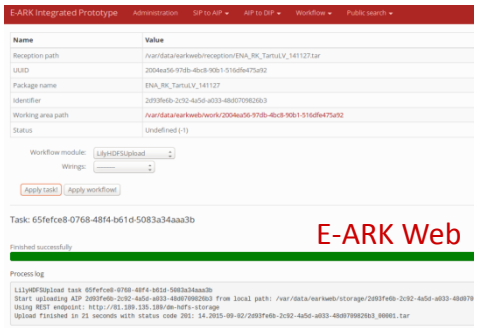
“Integrated” Prototype?

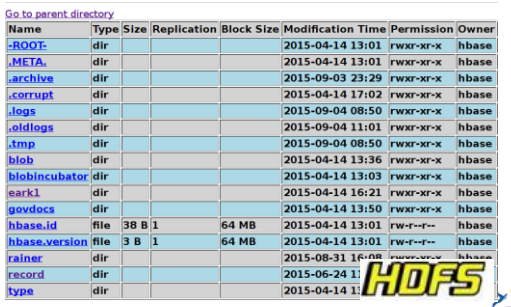




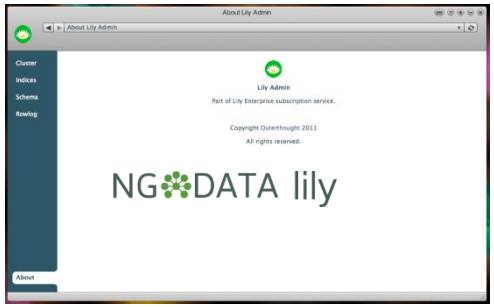




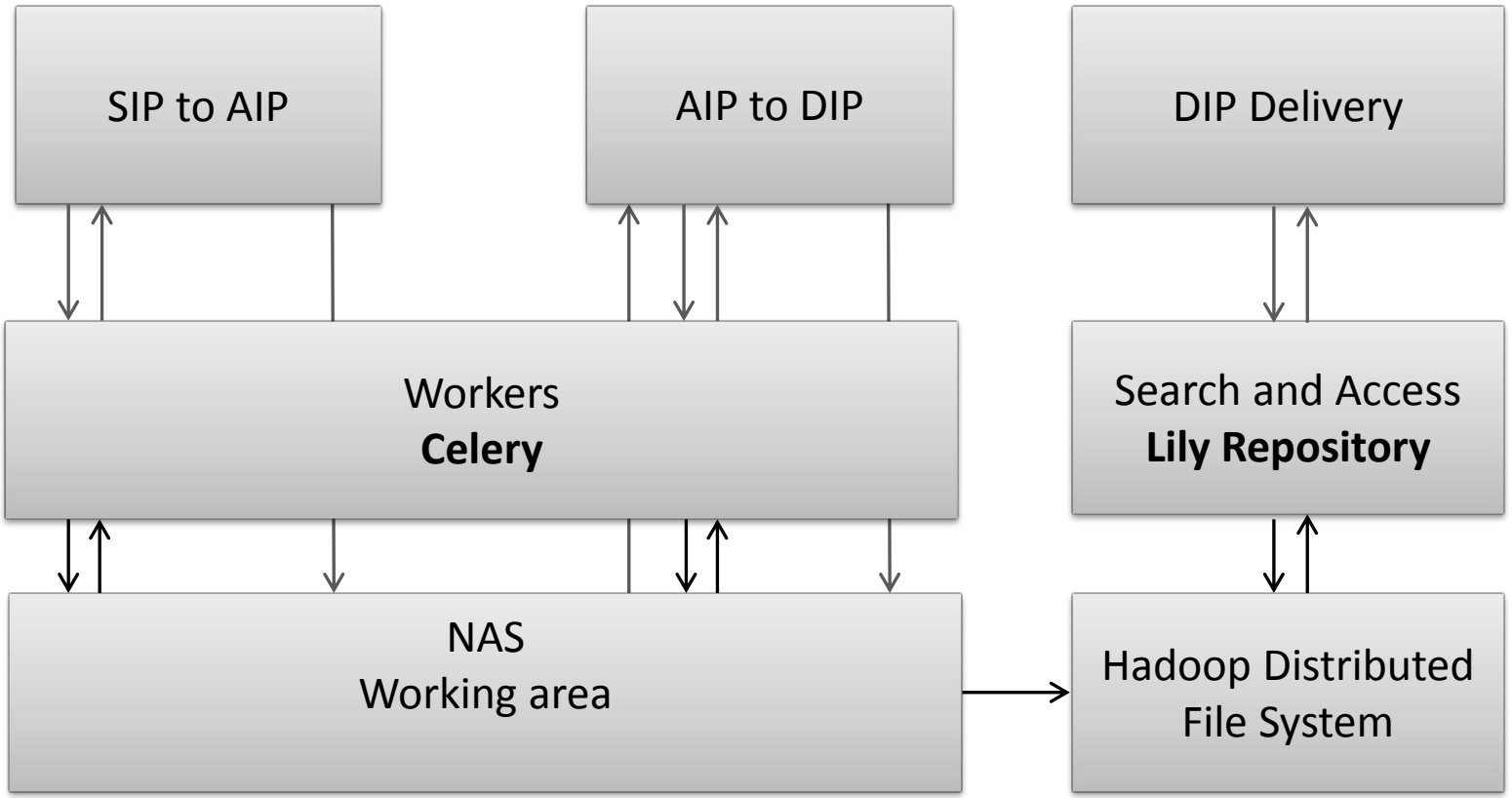




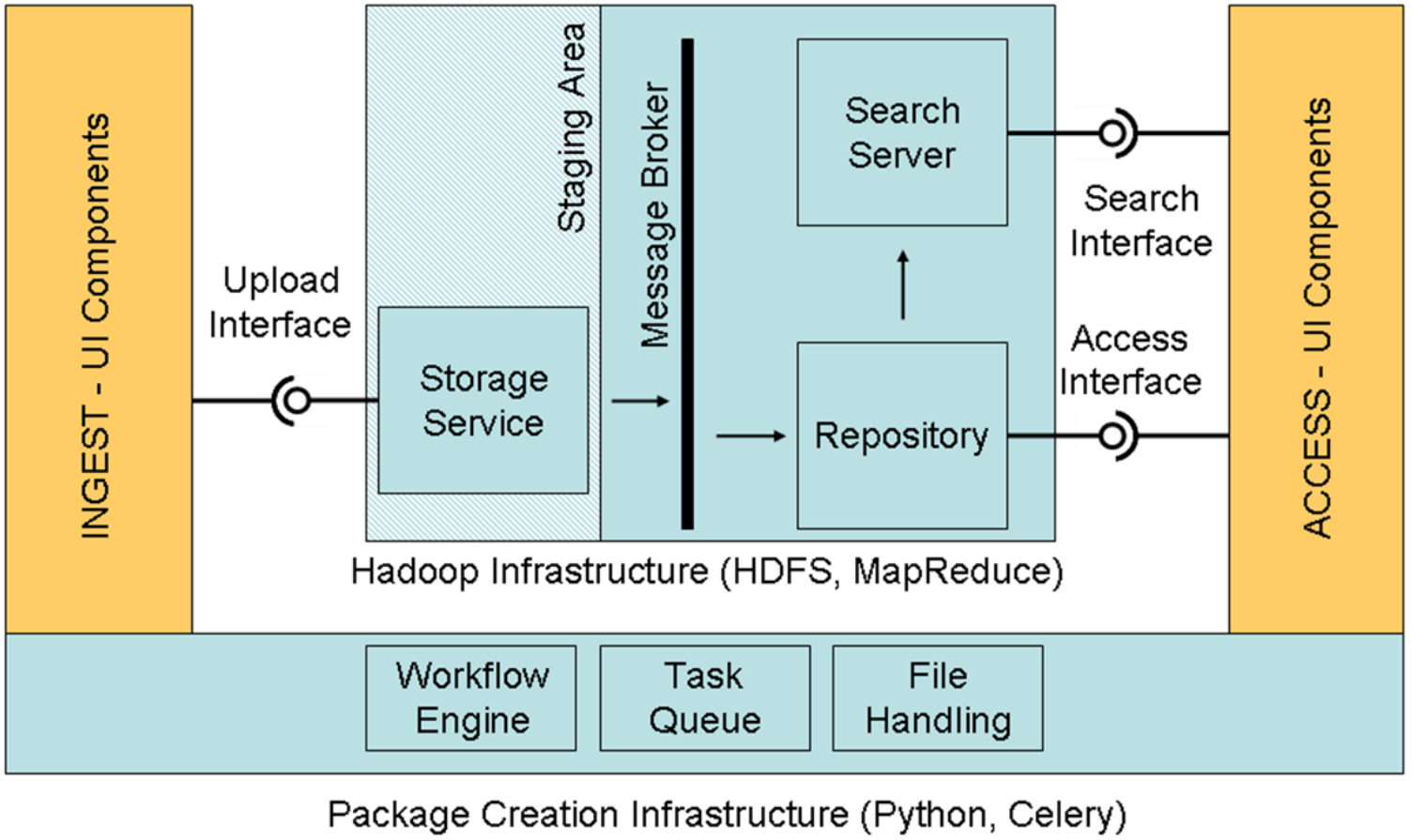
Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner
-ROOT-	dir				2015-04-14 13:01	rwxr-xr-x	hbase
.META.	dir				2015-04-14 13:01	rwxr-xr-x	hbase
.archive	dir				2015-09-03 23:29	rwxr-xr-x	hbase
.corrupt	dir				2015-04-14 17:02	rwxr-xr-x	hbase
.logs	dir				2015-09-04 08:50	rwxr-xr-x	hbase
.oldlogs	dir				2015-09-04 11:01	rwxr-xr-x	hbase
.tmp	dir				2015-09-04 08:50	rwxr-xr-x	hbase
blob	dir				2015-04-14 13:36	rwxr-xr-x	hbase
blobincubator	dir				2015-04-14 13:03	rwxr-xr-x	hbase
eark1	dir				2015-04-14 16:21	rwxr-xr-x	hbase
govdocs	dir				2015-04-14 13:50	rwxr-xr-x	hbase
hbase.id	file	38 B	1	64 MB	2015-04-14 13:01	rw-r--r--	hbase
hbase.version	file	3 B	1	64 MB	2015-04-14 13:01	rw-r--r--	hbase
rainer	dir				2015-08-31 16:08	rwxr-xr-x	hbase
record	dir				2015-06-24 11		
type	dir				2015-04-14 13		

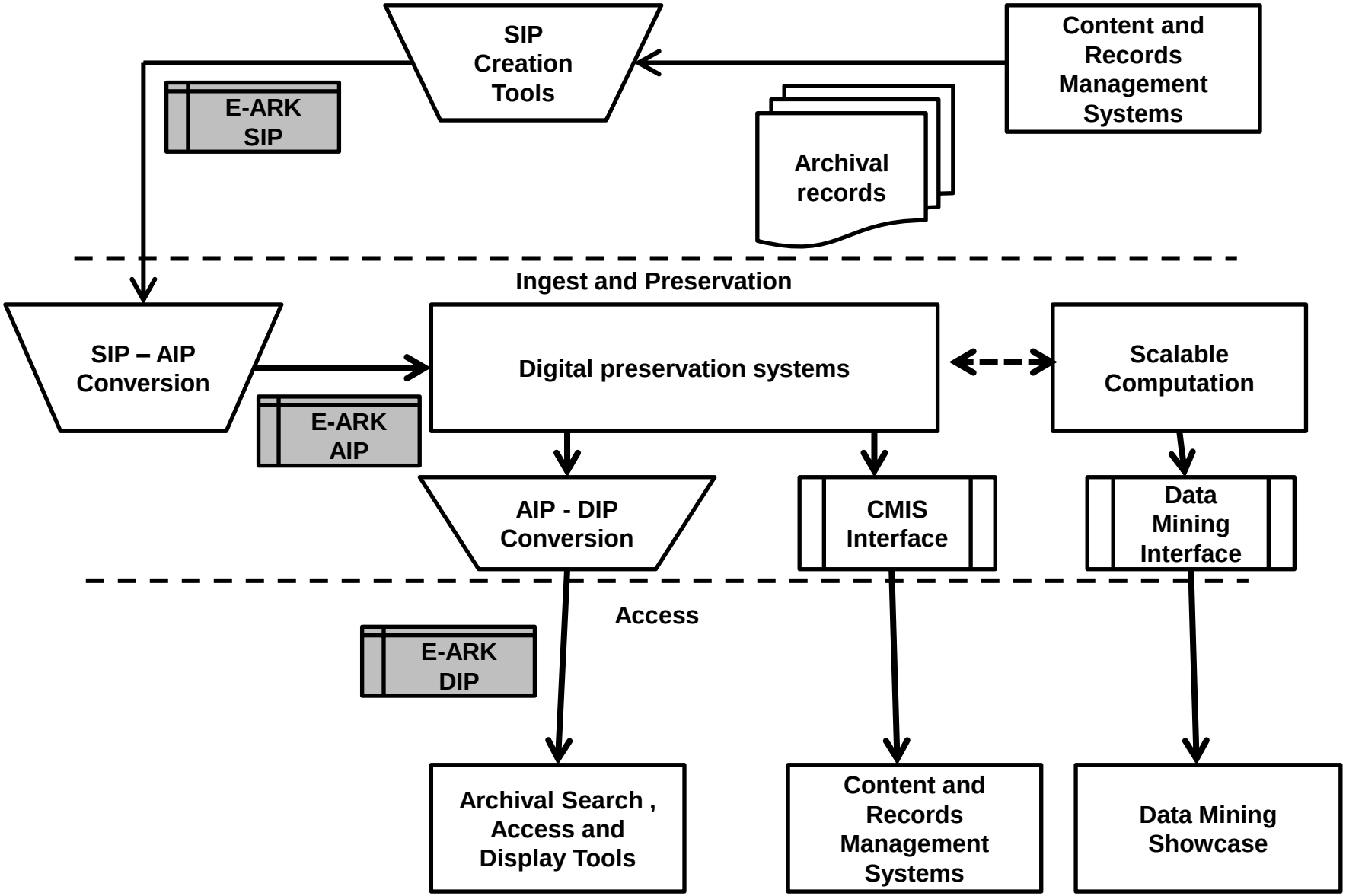


Information Package processing & Access Repository



Access Repository - Interfaces




SIP
Creation
Tools
Content and
Records
Management
Systems
E-ARK
SIP
Archival
records
Ingest and Preservation
SIP – AIP
Conversion
Digital preservation systems
Scalable
Computation
E-ARK
AIP
AIP - DIP
Conversion
CMIS
Interface
Data
Mining
Interface
Access
E-ARK
DIP
Archival Search ,
Access and
Display Tools
Content and
Records
Management
Systems
Data Mining
Showcase

SIP Creator

Package name: **CAMBRIDGE.001**

Process ID: **d5dd80d6-4673-4de4-80da-977b7d7c0e35**

Child-package options

docs + ▾

data

Durchsuchen... Keine Datei ausgewählt.

documentation

Durchsuchen... Keine Datei ausgewählt.

schemas

Durchsuchen... Keine Datei ausgewählt.

metadata
- descriptive

Durchsuchen... Keine Datei ausgewählt.

schemas
- premis-v2-2.xsd
- IP.xsd

Durchsuchen... Keine Datei ausgewählt.

Proceed Delete

Help

In the E-ARK SIP, each representation - as a set of files needed to render an intellectual entity - is stored in a separate directory under the "representations" directory.

It is required to give a name to the representation which will be used as the name of the directory where the actual data, additional documentation, and schemas can be uploaded to.

To create a new representation, enter the name (at least 4 characters long) in the editable select box and click the "plus" symbol which will enable the upload area of the new representation.

To switch between existing representations choose the representation from the select box ('caret' symbol next to the "plus" symbol).

If the upload area of the representation is loaded, files can be uploaded by clicking on 'Browse ...' and selecting a file from the local file system.

Package a set of files using the tar format to upload a collection of files which are automatically extracted in the upload directory.

Hover your mouse over the user interface widgets to get more information about the individiual elements.

SIP to AIP conversion

The AIP – as defined in the OAIS reference model – is an information package used to transmit and/or store archival objects within a digital repository. An AIP contains both, structural and descriptive metadata about the content, as well as the actual content itself.

The SIP to AIP conversion is a set of tasks that can be performed to convert an E-ARK SIP to an E-ARK AIP which both must comply with structural and metadata requirements defined by the E-ARK project.

SIP to AIP task/workflow execution

Name	Value
Process ID	5c6f5563-7665-4719-a2b6-4356ea033c1d
Package name	SLOV.GEO.ALL
Package Identifier	47b1e5a2-50bc-4aa1-8f09-6b438e815420
Working area path	/var/data/earkweb/work/5c6f5563-7665-4719-a2b6-4356ea033c1d
Last task	SIPRestructuring
Last change	10.12.2015 17:17:03
Process status	Success (0)

Task/Workflow execution

Tasks:
- SIPtoAIPReset
- SIPDeliveryValidation
- IdentifierAssignment
- SIPExtraction

Hold down "Control", or "Command" on a Mac, to select more than one.

Run selected tasks!

Process log

Error log

back

contained word: search term

package: package name

content type:
csv
doc
html
pdf
txt
xhtml
xls
xml

sort: relevance

Submit

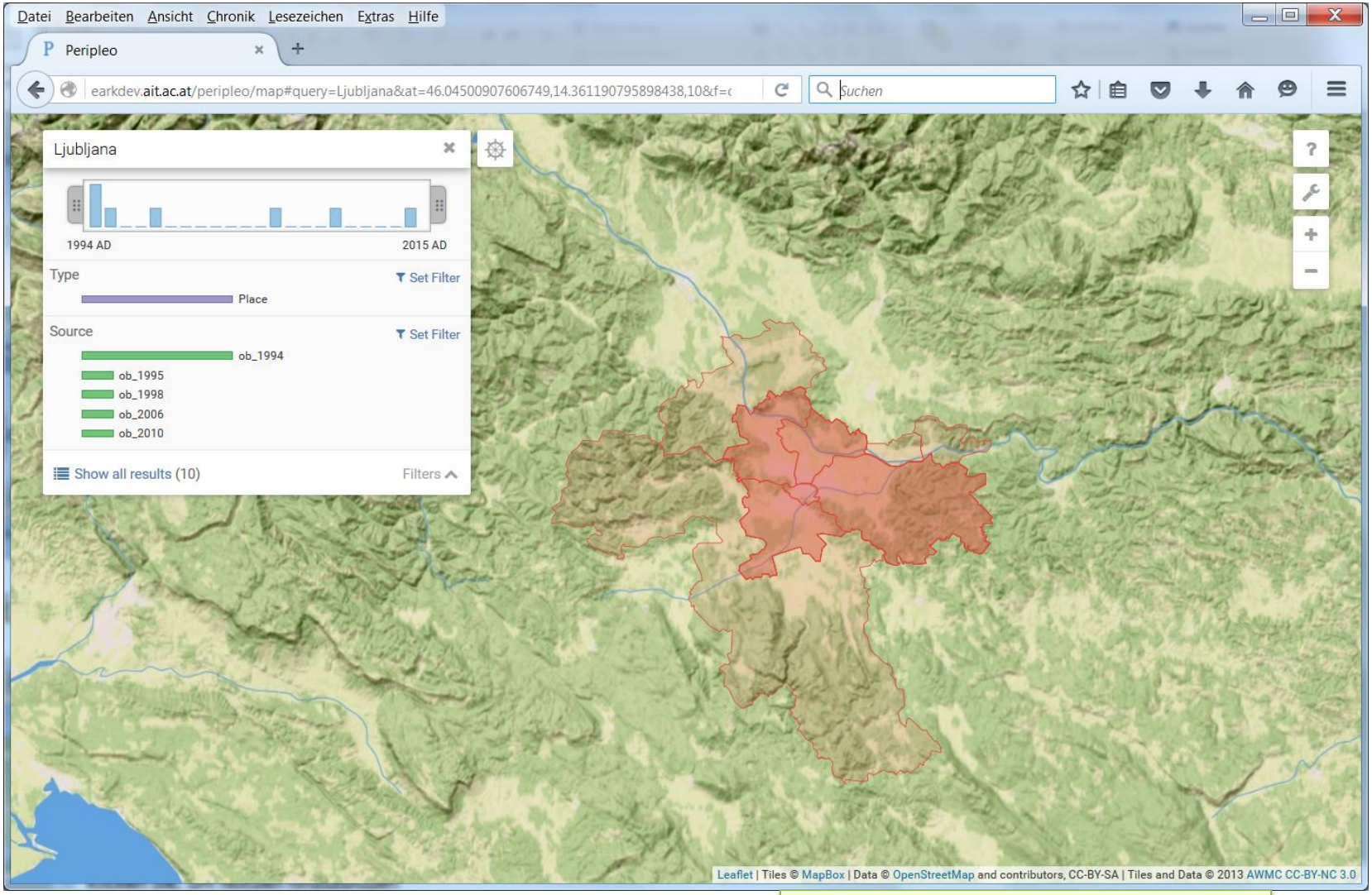
400 results found

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

1e232ce6-9177-401e-8d83-4eeb28dc680b/representations/rep-002/data/Charlemagne.pdf	7 MB
1e232ce6-9177-401e-8d83-4eeb28dc680b/submission/representations/rep-001/data/Charlemagne.pdf	3 MB
1e232ce6-9177-401e-8d83-4eeb28dc680b/submission/representations/rep-001/schemas/schema.txt	66 B
1e232ce6-9177-401e-8d83-4eeb28dc680b/submission/schemas/IP.xsd	142 kB
1e232ce6-9177-401e-8d83-4eeb28dc680b/submission/schemas/mets_1_11.xsd	129 kB
1e232ce6-9177-401e-8d83-4eeb28dc680b/submission/schemas/premis-v2-2.xsd	63 kB
1e232ce6-9177-401e-8d83-4eeb28dc680b/submission/schemas/xlink.xsd	3 kB
1e232ce6-9177-401e-8d83-4eeb28dc680b/submission/state.xml	201 B
1f12e1e4-bad6-486b-b523-4206bcecc352/METS.xml	6 kB
1f12e1e4-bad6-486b-b523-4206bcecc352/metadata/earkweb.log	4 kB
1f12e1e4-bad6-486b-b523-4206bcecc352/representations/rep-002/METS.xml	1 kB
1f12e1e4-bad6-486b-b523-4206bcecc352/submission/metadata/earkweb.log	332 B
1f12e1e4-bad6-486b-b523-4206bcecc352/submission/representations/rep-001/METS.xml	4 kB
1e232ce6-9177-401e-8d83-4eeb28dc680b/submission/representations/rep-001/data/bike.gif	561 kB
1f12e1e4-bad6-486b-b523-4206bcecc352/submission/schemas/IP.xsd	142 kB
1f12e1e4-bad6-486b-b523-4206bcecc352/submission/representations/rep-001/data/bike.gif	561 kB
1f12e1e4-bad6-486b-b523-4206bcecc352/submission/schemas/mets_1_11.xsd	129 kB
1f12e1e4-bad6-486b-b523-4206bcecc352/submission/schemas/premis-v2-2.xsd	63 kB
1f12e1e4-bad6-486b-b523-4206bcecc352/submission/state.xml	201 B
1f12e1e4-bad6-486b-b523-4206bcecc352/submission/schemas/xlink.xsd	3 kB

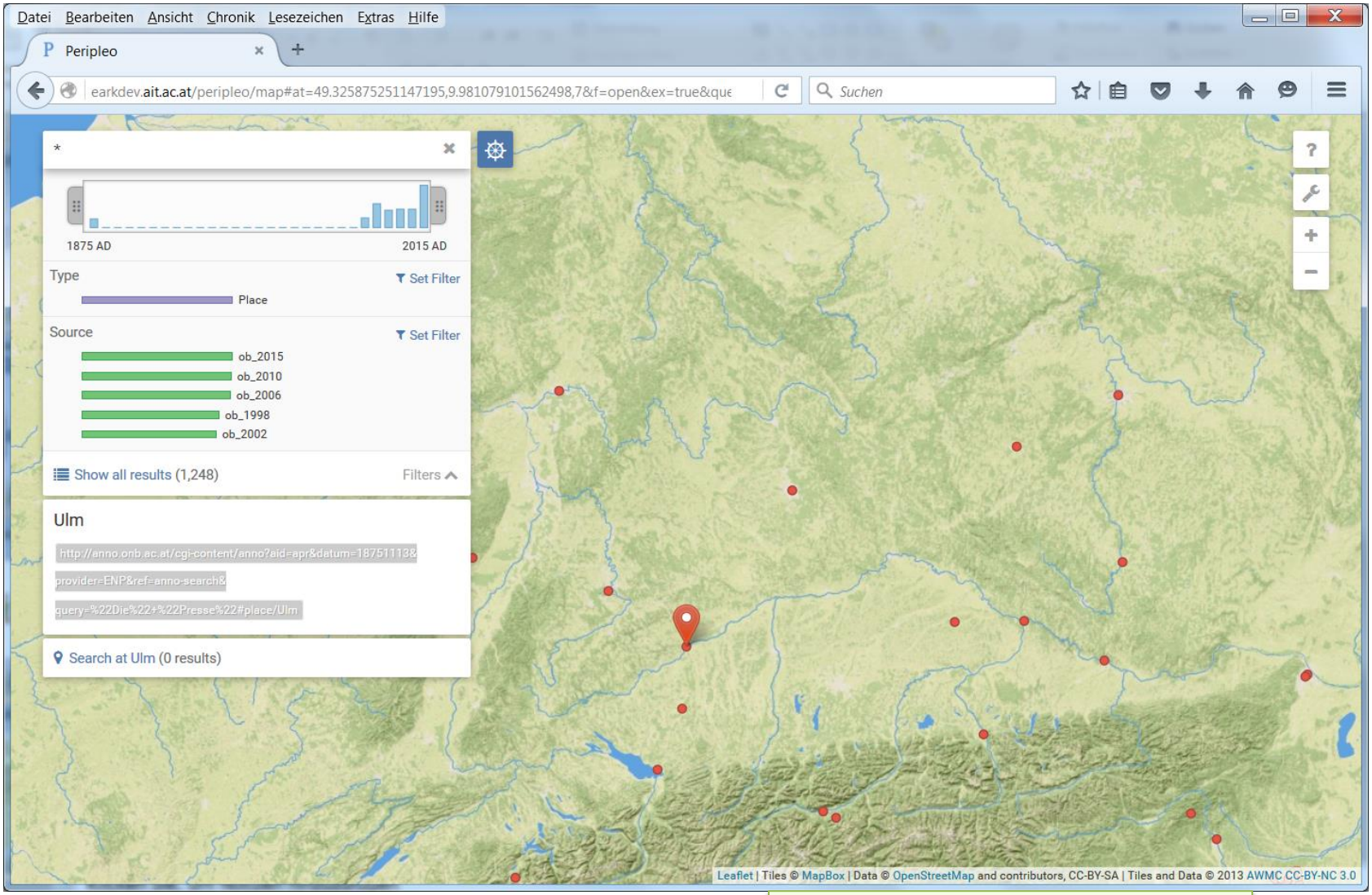
E-ARK Data Mining

Geographical/timeline search



Peripleo - PELAGIOS Project

Geographical/timeline search



Peripleo - PELAGIOS Project

Text mining: Text classification

Training

- Train classifier using annotated text corpus
- SVM – based on statistical features

Classification

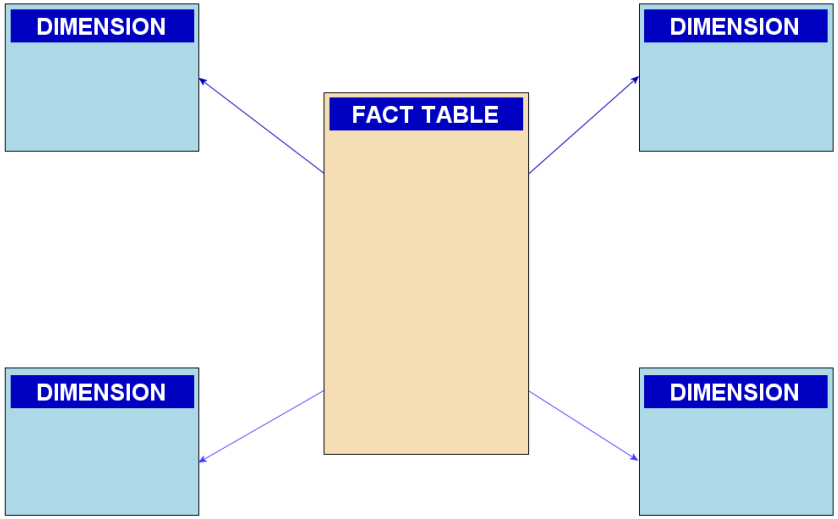
- Scan for texts during ingest (or run MR after)
- Text category estimation

Search

- Add category as a searcheable field to Lily index
- Full-text search using Lily‘s SolR search interface

OLAP (Online Analytical Processing)

- Database archiving and re-use (SIARD2)
- Normalization - OLAP/Oracle Data Warehouse



Thank you!

- http://www.eark-project.eu
- https://github.com/eark-project