CSC

The National Digital Library

TTA
TUTKIMUKSEN TIETOAINEISTOT

Data Archiving and Data Stewardship

Pirjo-Leena Forsström, Heikki Helin, Kimmo Koivunen, Juha Lehtonen, Kuisma Lehtone
CHEP 2013

Index

- Background and motivation
- What is preservation?
- For whom and what?
- How is it done?
- Pilots
- In conclusion

Source: wikipedia PD
Image resources

Digitalization of research and cultural processes

- Typical:
  - Growing volume of data and sources
  - Complexity of data processing
  - data is dynamic
  - High demand of data
  - Complicated interaction between users and data
- Most important challenges:
  - Managing and processing exponentially growing datasets
  - Significant acceleration in analysis cycle
  - Combining data sources

Source: wikipedia PD
Image resources

Long-term preservation of research and cultural heritage data

- Preservation of digital information is the core of research and cultural organization's activity.
- At this time, there has been no controlled and functional way of handling digital information in the long term.
- Long-term preservation of digital data means the reliable preservation of digital information for several decades or even hundreds of years.
- Equipment, software, and file formats will become outdated, but despite this the information must be preserved in understandable form.

Digital processes break easily

- Short-period funding
- Software lifecycle: code, interfaces, formats…
- Dependent on expert knowledge
- Thin documentation and metadata

Source: : wikipedia PD Image resources

Future Aim:

Research and cultural data routinely deposited in well-documented form, regularly and easily consulted and analyzed, and openly accesible while suitably protected and reliably preserved.

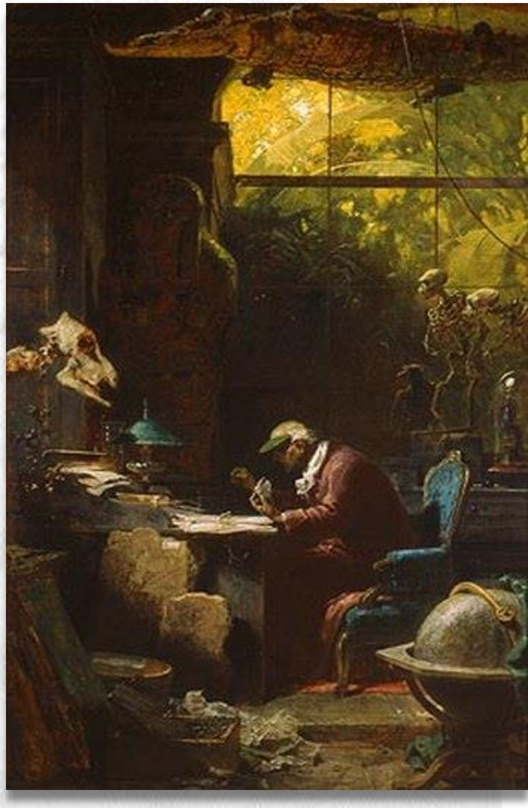
Needs PERSISTENCE

- Coherent organizational framework?
  - Ownership
  - Curation
- Flexible technical architecture:
  - Standard open protocols and interfaces
  - Flexible user access, analysis and visualization of data
  - Address issues of autenthication, authorization, security
  - Supports workflows

Persistence

Holy grail of preservation & information management more generally

- What does persistence mean?
- How long it persists?
- What persists?
- What is “guaranteed” to be accessible?



Source: : wikipedia PD Image resources

Digital Curation

- Curation : The activity of, managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose.
- Archiving : A curation activity which ensures that data is properly selected, stored, can be accessed and that its logical and physical integrity is maintained over time, including security and authenticity.
- Preservation : An activity within archiving in which specific items of data are maintained over time so that they can still be accessed and understood through changes in technology.
- Digital curation : looking after and somehow "adding value" to digital data, ensuring its current and future usefulness. This probably implies creating some new data from the existing, in order to make the latter more useful and "fit for purpose".

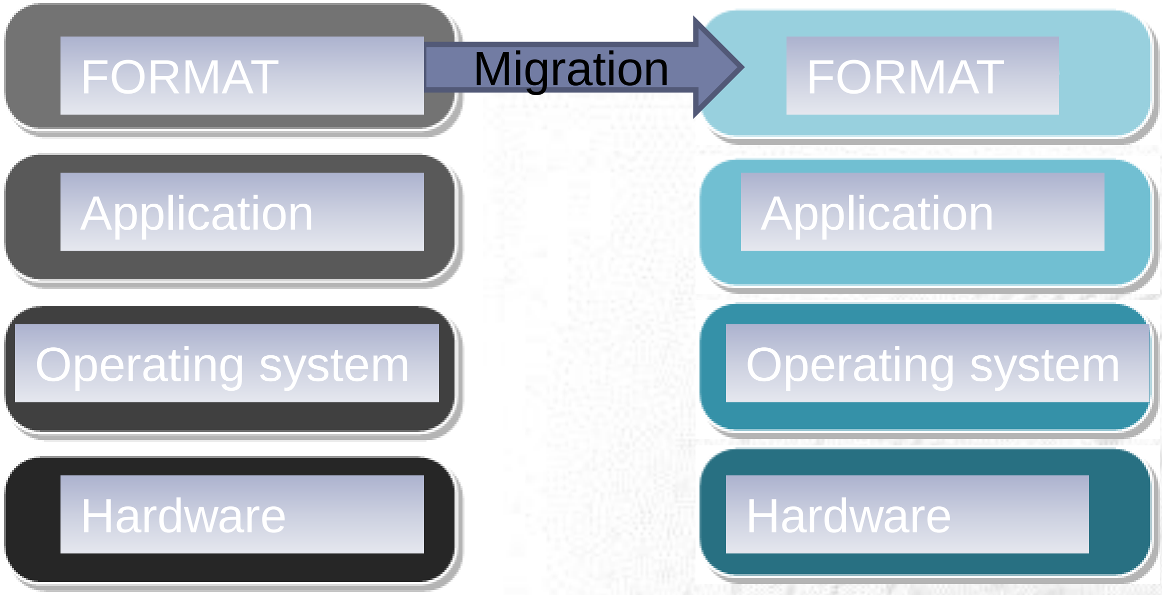
Data Collections

- Research Collections: Authors individual investigators and investigator teams. Maintained to serve immediate group participants during the project lifetime. May not conform to any data standards.
- Resource Collections : Authored by a community of investigators (specific domain of science or engineering) with community-level standards. Lifetime from mid- to long term.
- Reference Collections : Authored by and serve large segments of science and engineering community. Using well-established and comprehensive standards.

Preservation methods

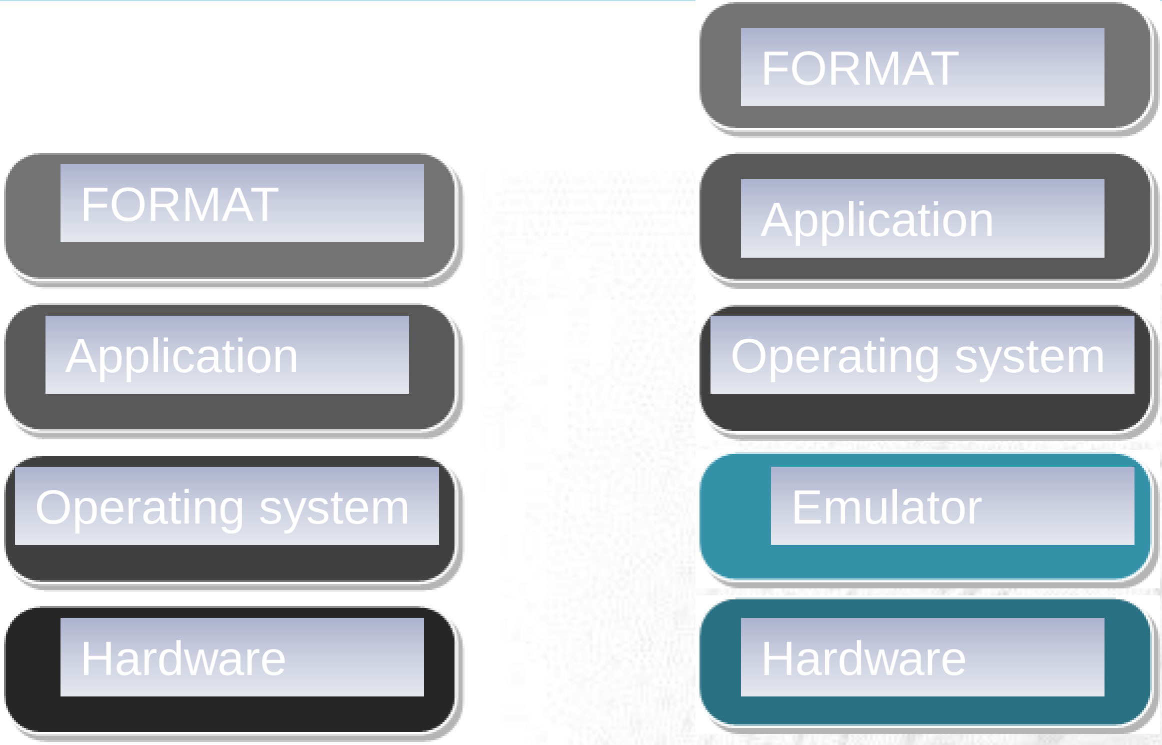
- Preserving the original look-and-feel
  - Emulation
    - Development of emulators to new platforms etc.
    - Active testing and technology watch
- Preserving the content
  - Migration
    - Format development watch (format libraries)
    - Development of transformation processes, testing, implementation, monitoring
    - Preparation for recoveries
- Preserving the bits
  - Integrity
    - File validation and monitoring
    - Management of copies
    - Both objects and metadata

Migration



- Migration enables the utilization of digital objects in new ICT environment
- Special care needed to preserve information content: planning, testing and validation with care

Emulation



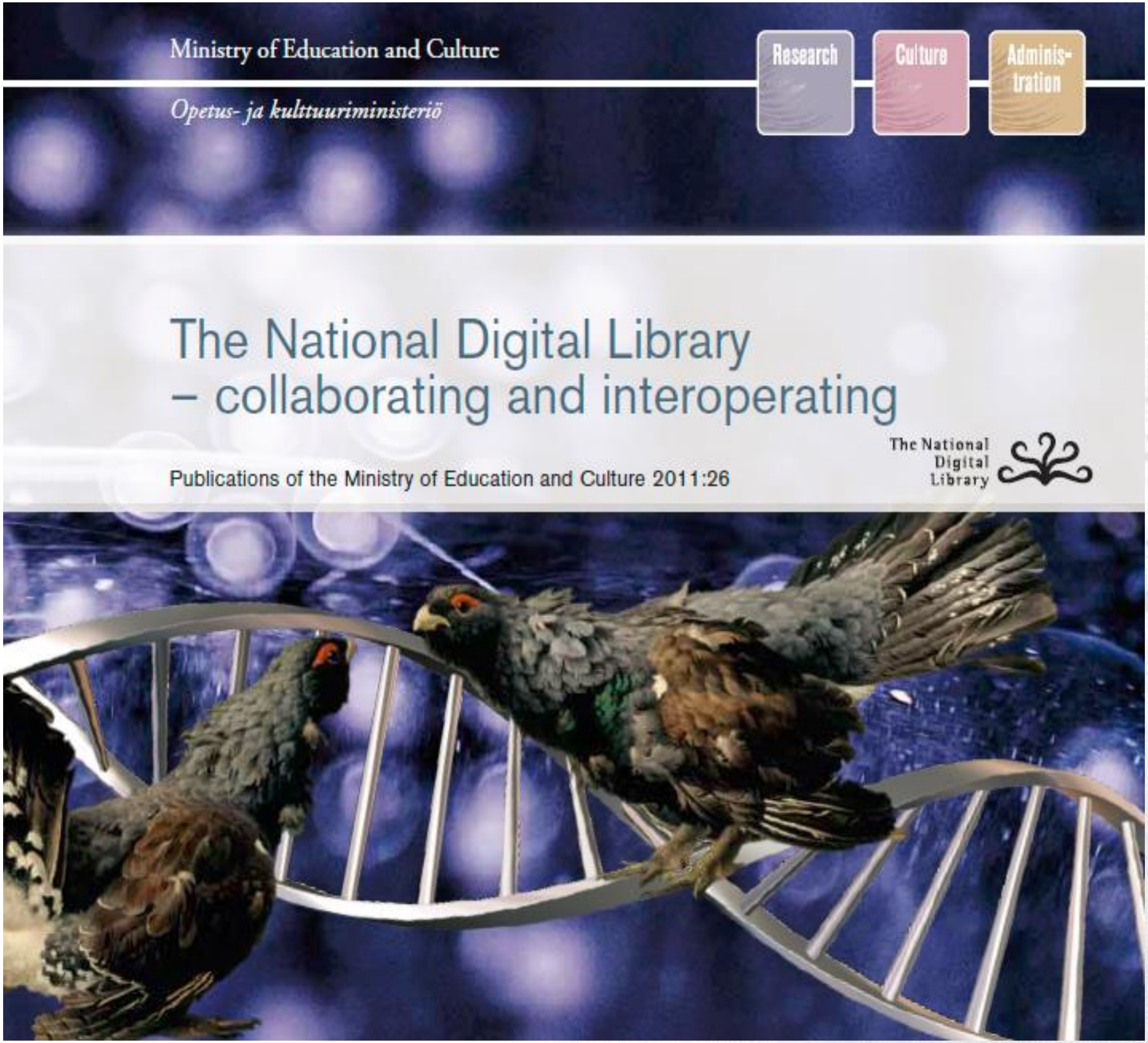
FORMAT
Application
Operating system
Hardware

FORMAT
Application
Operating system
Emulator
Hardware

- Emulation enables the use of old solution on new hardware environment
- Emulation has to solve how the information can be used in context of new data production (copy-paste)

Preservation solution has to manage

- Authencity
- Integrity
- Technology change
- Risk management
- Preservation metadata management
- Scalability of the solution



Ministry of Education and Culture

Opetus- ja kulttuuriministeriö

Research

Culture

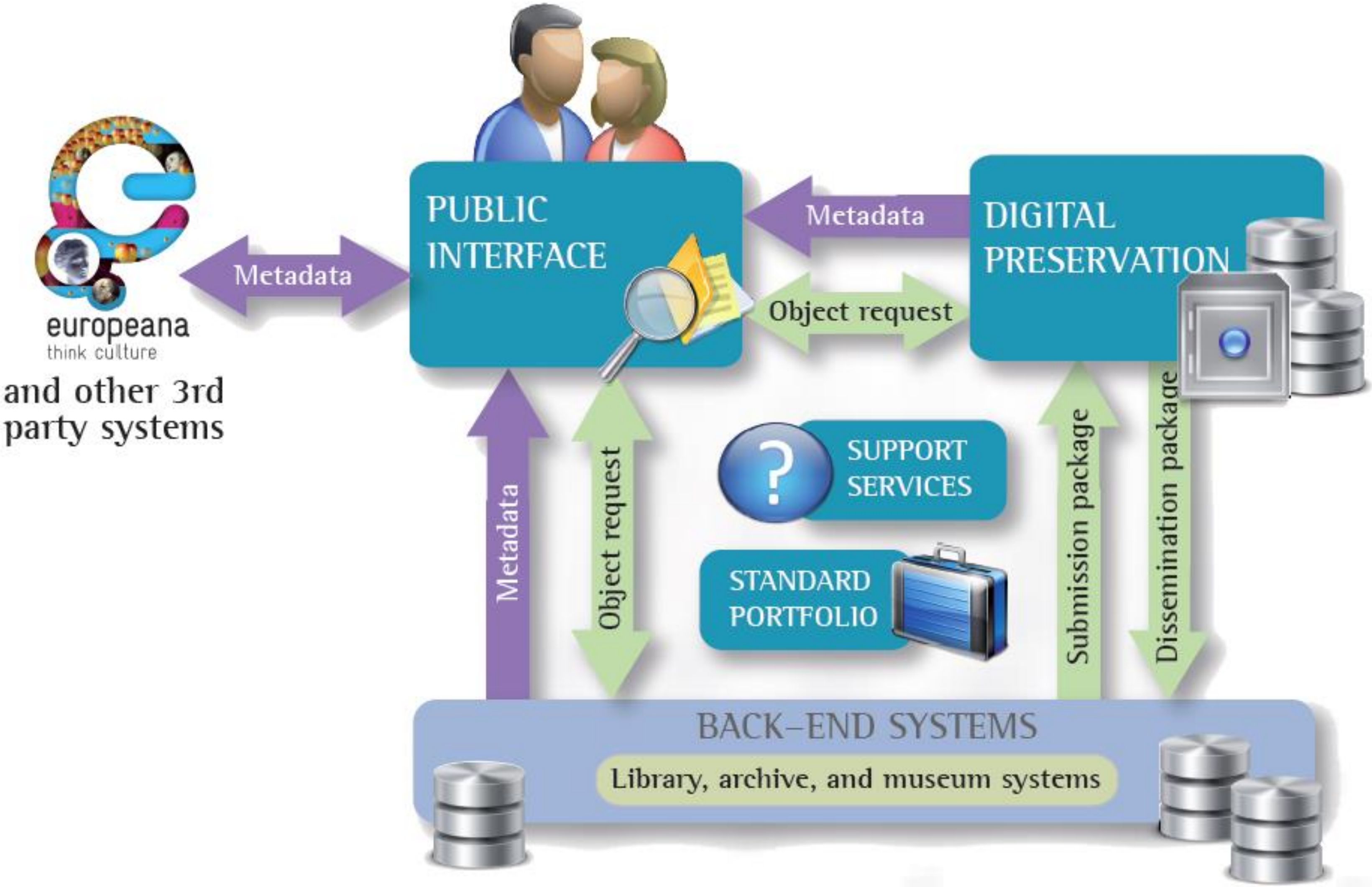
Administration

The National Digital Library – collaborating and interoperating

Publications of the Ministry of Education and Culture 2011:26

The National Digital Library

National Digital Library



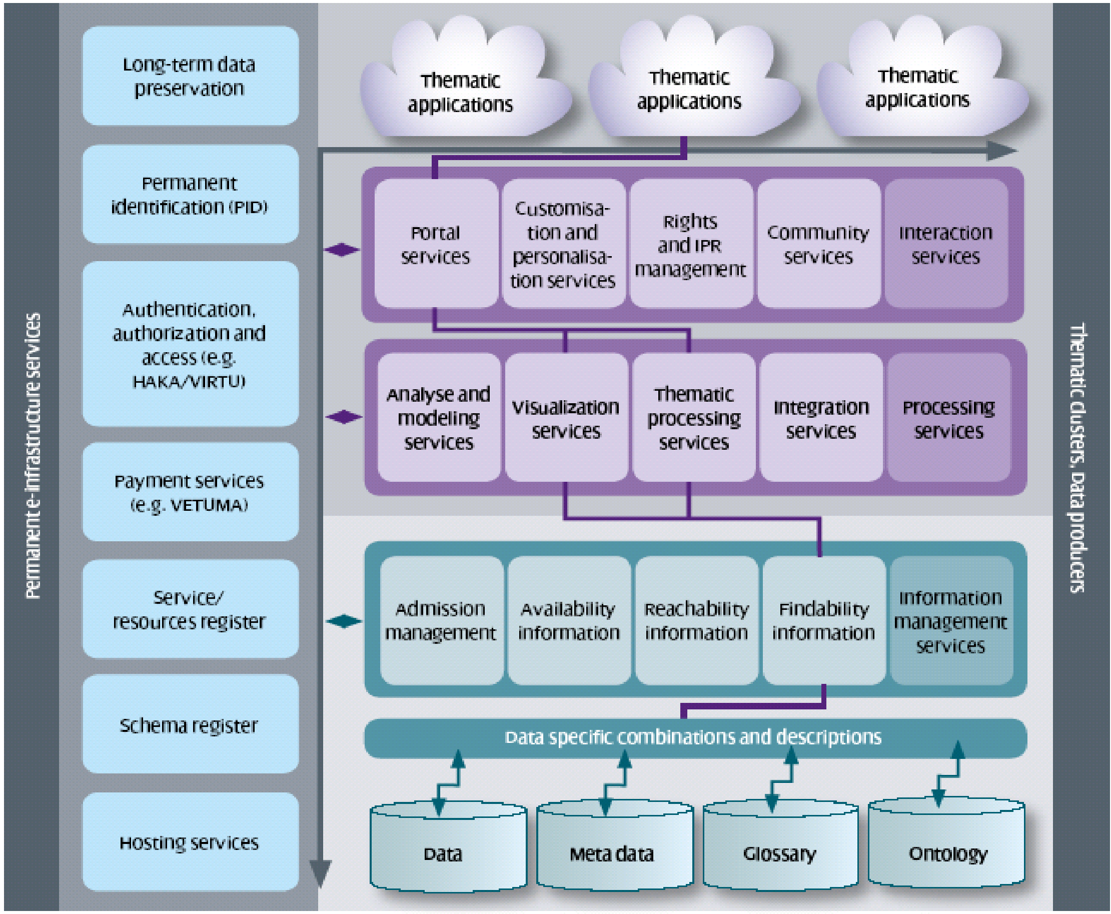
TTA

TUTKIMUKSEN TIETOAINEISTOT

- Ministry of Education: national collaboration and roadmap for research Data
- National Information Infrastructure services for research
- TTA provides 2012-2013
  - Storage solution IDA
  - Metadata catalogue KATA
- Long Term Preservation 2015 –
  - Pilots starting 2014

Research Information infrastructure

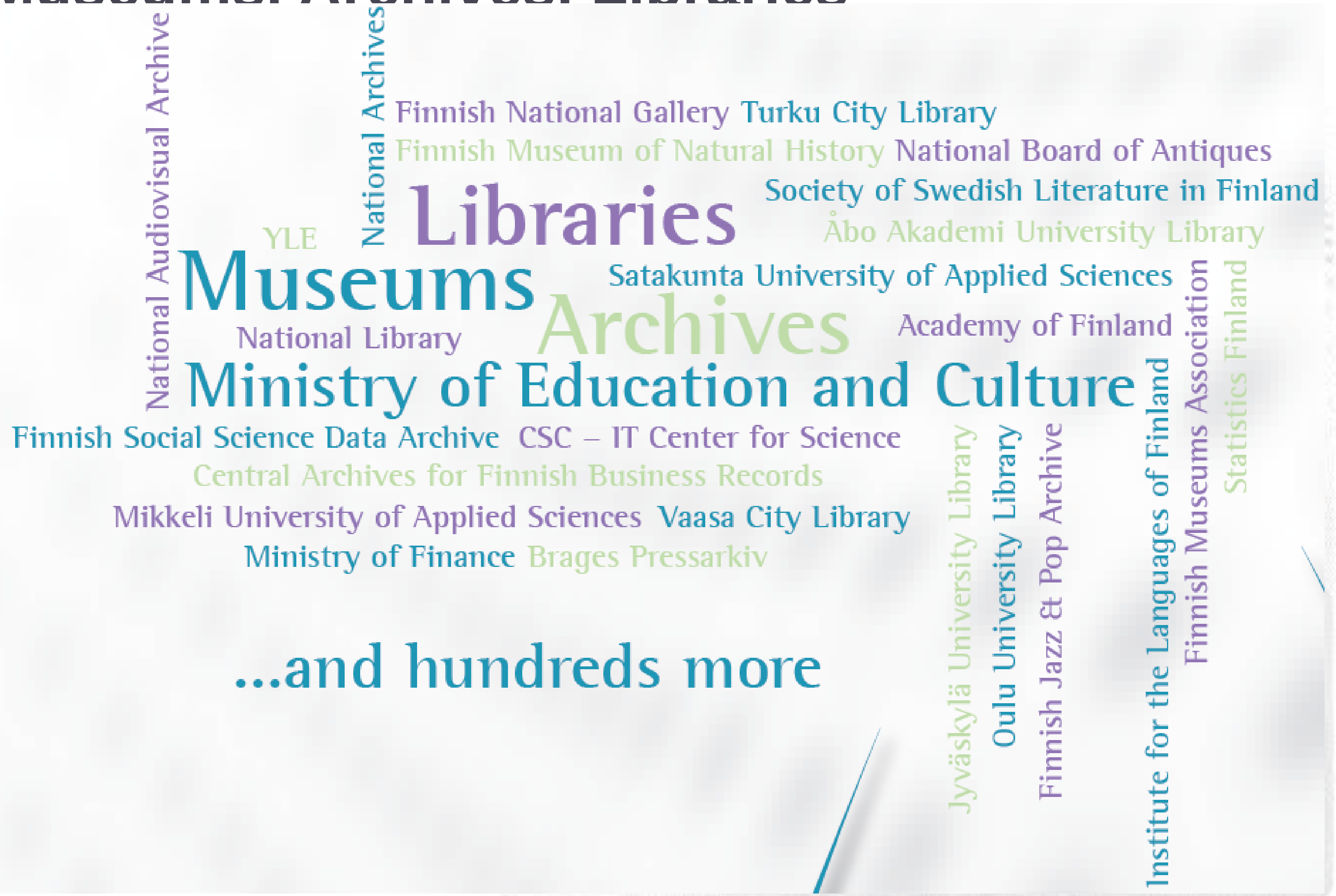
- Embeddedness
- Transparency
- Reach of scope
- Links with conventions of practice.
- Embodiment of standards
- Build on an open platform: Infrastructure does not grow de novo; it wrestles with the “inertia of the installed base” and inherits strengths and limitations from that base.



Permanent e-infrastructure services
Long-term data preservation
Permanent identification (PID)
Authentication, authorization and access (e.g. HAKA/VIRTU)
Payment services (e.g. VETUMA)
Service/ resources register
Schema register
Hosting services
Thematic applications
Thematic applications
Thematic applications
Portal services
Customisation and personalisation services
Rights and IPR management
Community services
Interaction services
Analyse and modeling services
Visualization services
Thematic processing services
Integration services
Processing services
Admission management
Availability information
Reachability information
Findability information
Information management services
Data specific combinations and descriptions
Data
Meta data
Glossary
Ontology
Thematic clusters, Data producers

Who? Research organizations

Museums, Archives, Libraries



National Audiovisual Archive

National Archives

Finnish National Gallery

Turku City Library

Finnish Museum of Natural History

National Board of Antiques

Libraries

Society of Swedish Literature in Finland

YLE

Åbo Akademi University Library

Museums

Satakunta University of Applied Sciences

Archives

National Library

Academy of Finland

Ministry of Education and Culture

Finnish Social Science Data Archive

CSC – IT Center for Science

Central Archives for Finnish Business Records

Mikkeli University of Applied Sciences

Vaasa City Library

Ministry of Finance

Brages Pressarkiv

...and hundreds more

Jyväskylä University Library

Oulu University Library

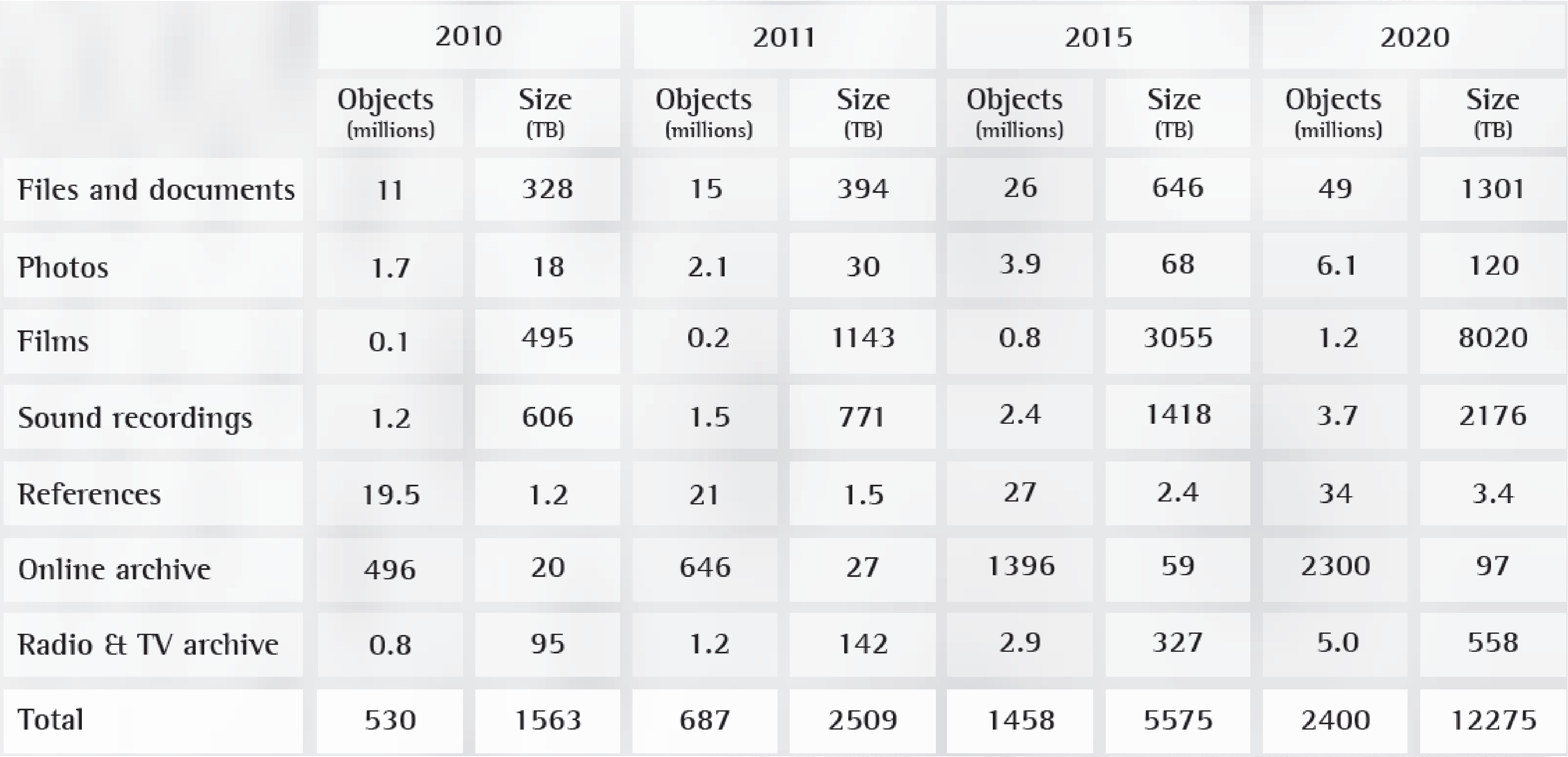
Finnish Jazz & Pop Archive

Institute for the Languages of Finland

Finnish Museums Association

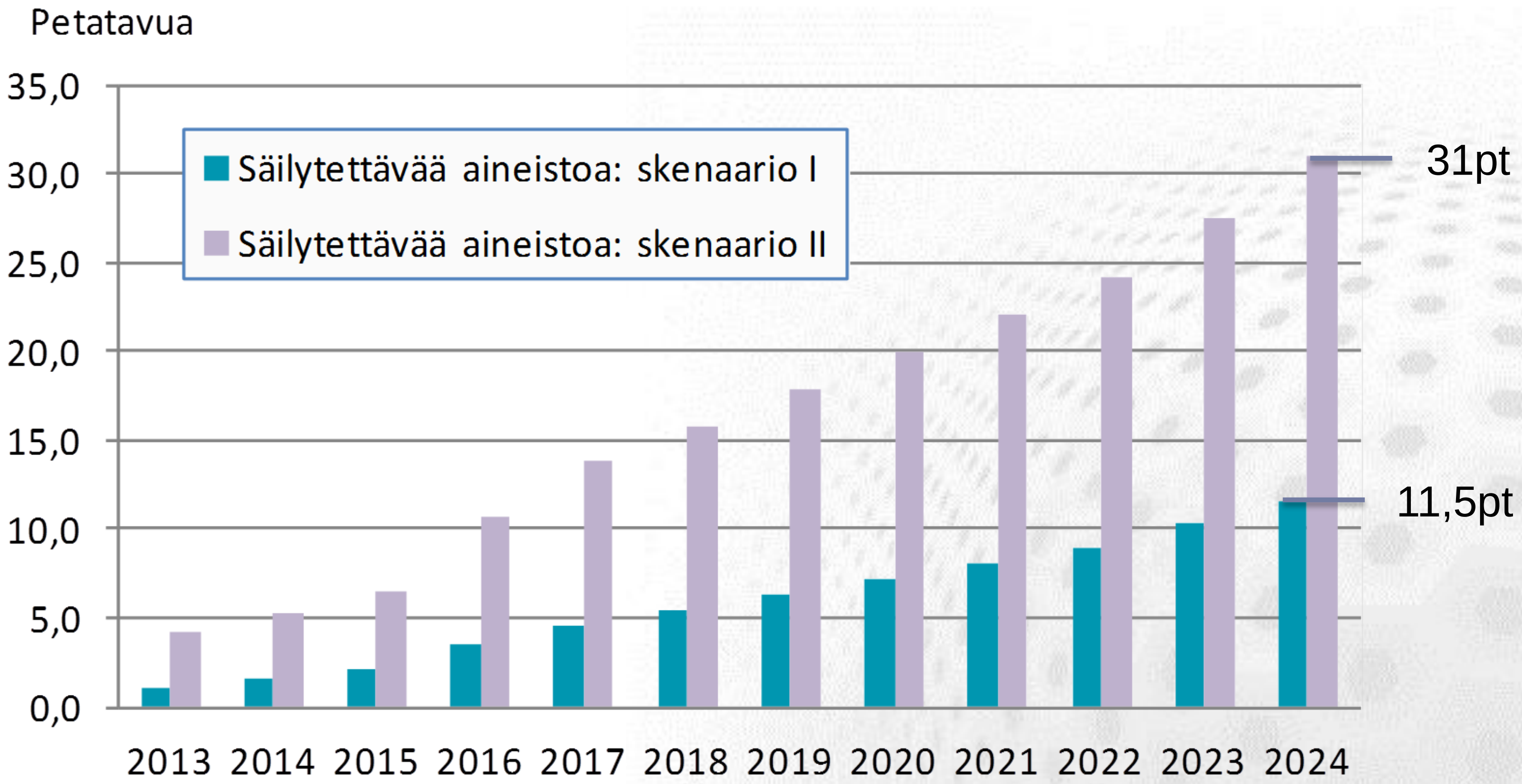
Statistics Finland

What? - Data Volumes



	2010		2011		2015		2020	
	Objects (millions)	Size (TB)	Objects (millions)	Size (TB)	Objects (millions)	Size (TB)	Objects (millions)	Size (TB)
Files and documents	11	328	15	394	26	646	49	1301
Photos	1.7	18	2.1	30	3.9	68	6.1	120
Films	0.1	495	0.2	1143	0.8	3055	1.2	8020
Sound recordings	1.2	606	1.5	771	2.4	1418	3.7	2176
References	19.5	1.2	21	1.5	27	2.4	34	3.4
Online archive	496	20	646	27	1396	59	2300	97
Radio & TV archive	0.8	95	1.2	142	2.9	327	5.0	558
Total	530	1563	687	2509	1458	5575	2400	12275

Research Data volumes, first quess

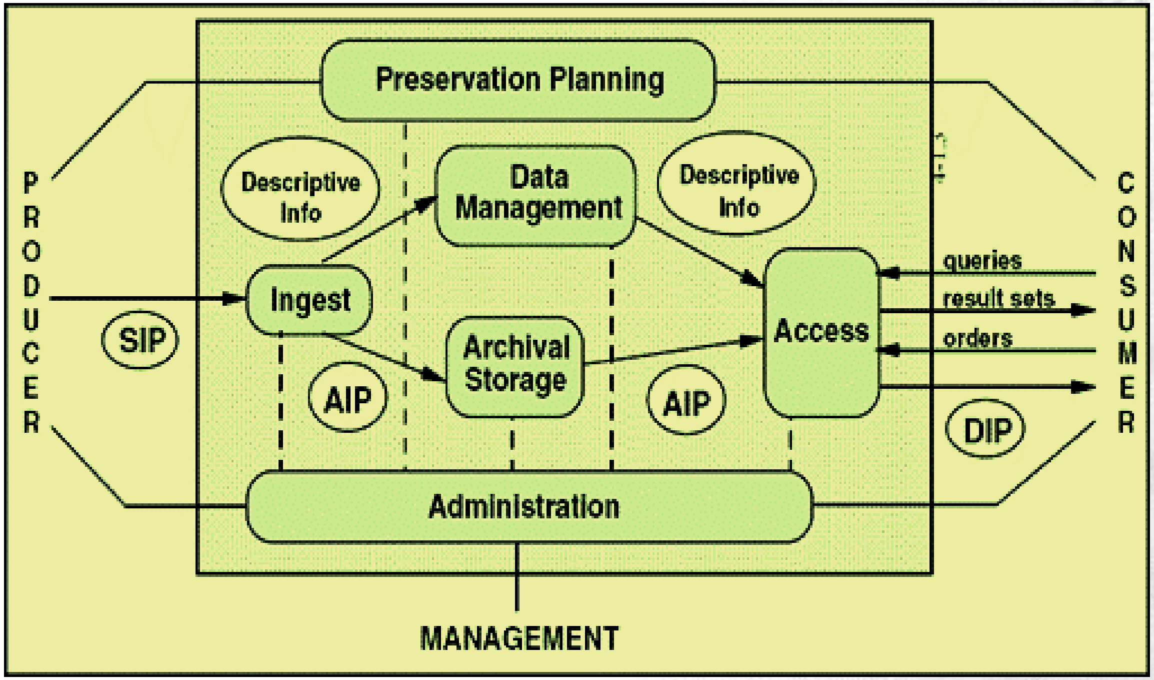


Digital Preservation Services (1/3)

- Digital preservation system will be built according to Open Archival Information System (OAIS) reference model
- Preparation and ingest services
  - Metadata specification
  - Preservation plan preparation
  - Submission information package (SIP) packaging service
  - SIP ingest and validation
  - Processing of acceptable file formats (for transfer) to recommended file formats

Digital Preservation Solution

- the ISO OAIS Reference Model for *an* OAIS. This reference model is defined by recommendation CCSDS 650.0-B-1 of the Consultative Committee for Space Data Systems;[1] this text is identical to ISO 14721:2003.



Source: Long-Term Preservation of Digital Documents. 2006. doi:10.1007/978-3-540-33640-2. ISBN 978-3-540-33639-6.
Public Domain.

Digital Preservation Services (2/3)

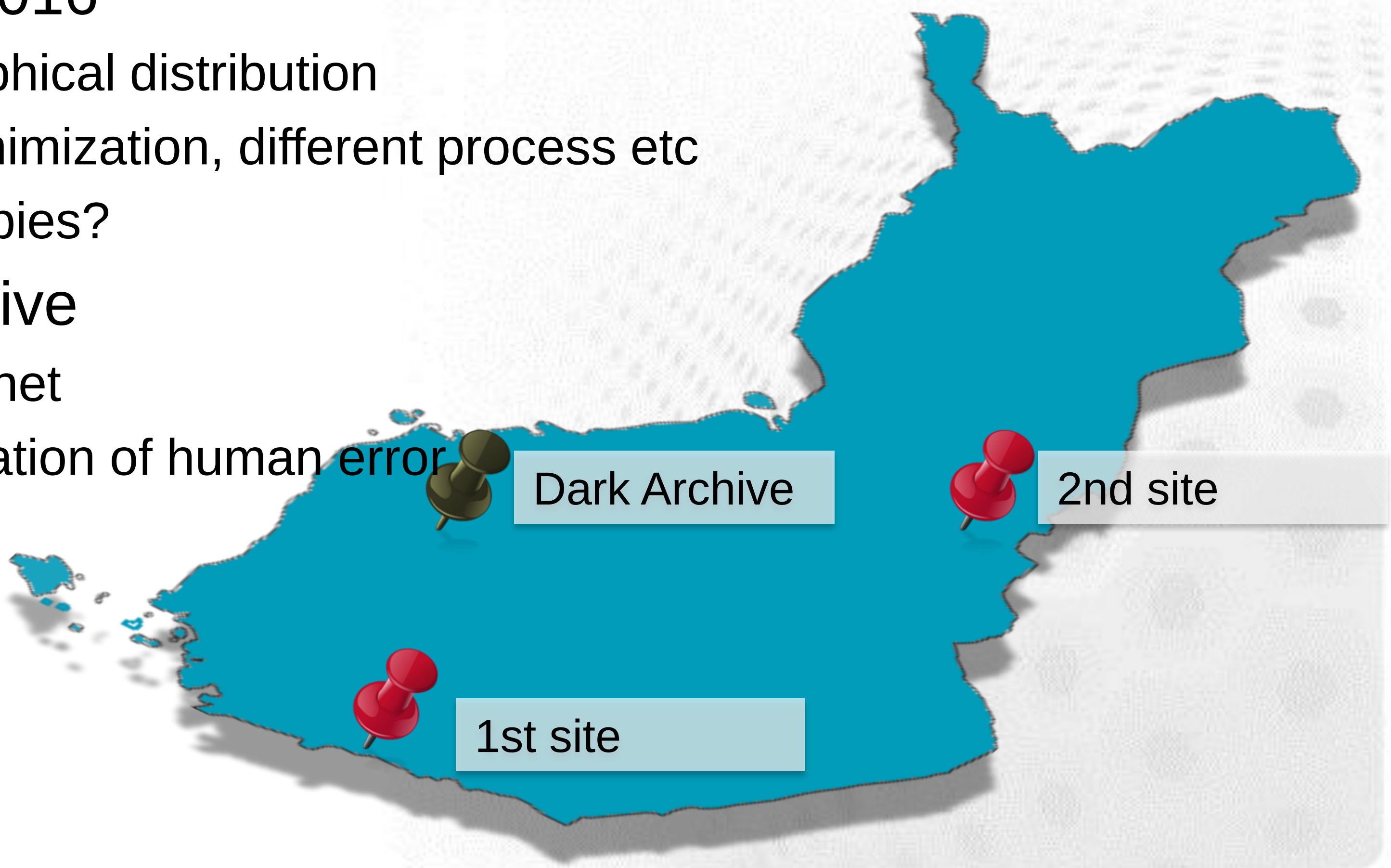
- Preservation services
  - Archival information package (AIP) from SIP
  - Development and monitoring of preservation methods and environment
  - Preservation actions: integrity monitoring, refreshment, replication, migration
  - Geographical distribution
    - E.g. Espoo and Kajaani – 550 km distance between
    - http://goo.gl/maps/J5XkX
- Digital information search functions
  - Dissemination information package (DIP)

Digital Preservation Services (3/3)

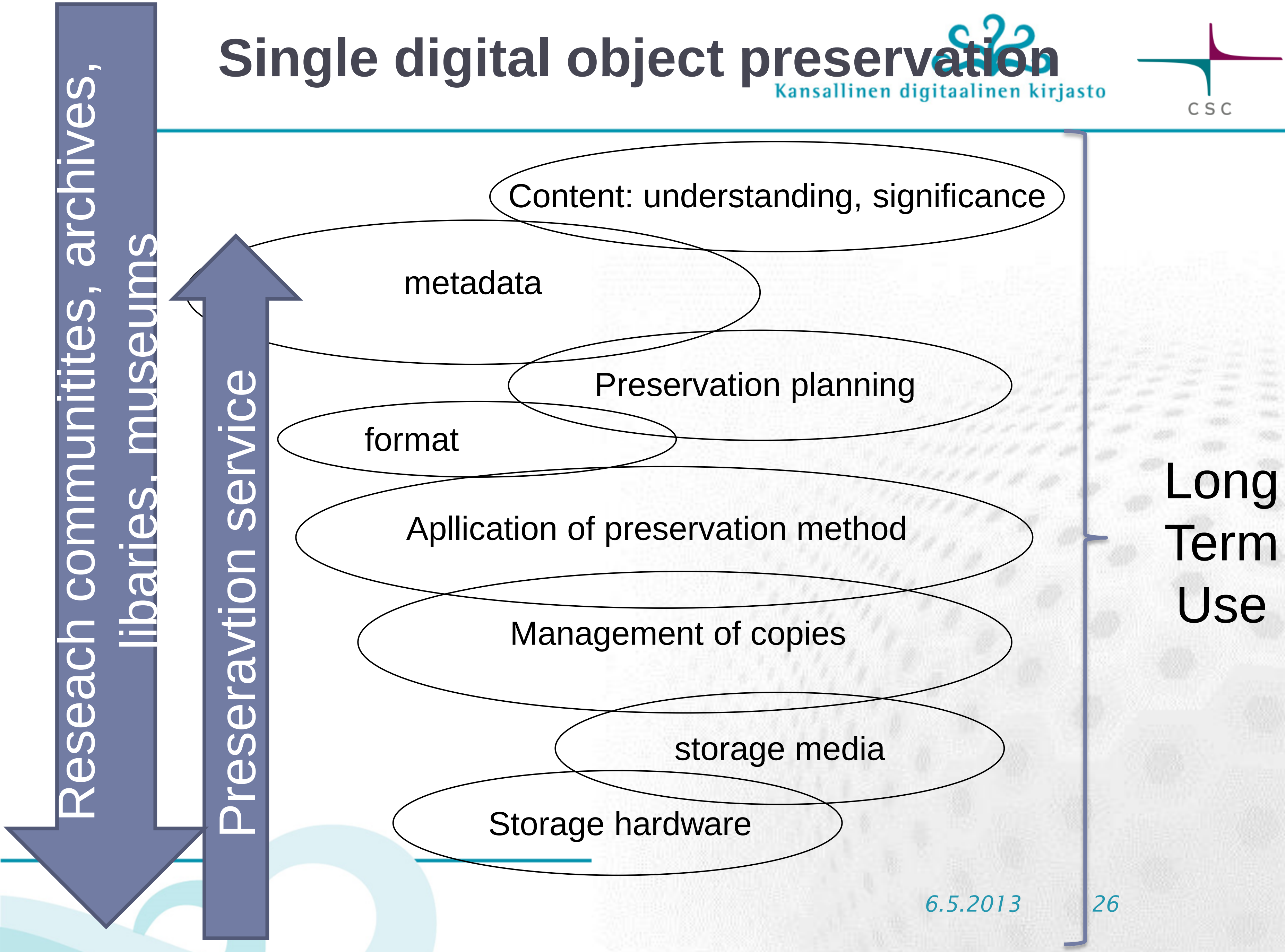
- Digital information management services
  - – Metadata updates
  - – Digital object updates
  - – Removal of digital objects
  - – Preservation plan updates
- Advisory and support services
  - – Usage support of the services and the digital preservation system
  - – Administrative support
  - – Training and information services

Roadmap

- 1st site: Ingest and bit level preservation service starts december 2013
  - 3 copies in 3 different media
- 2nd site 2016
  - Geographical distribution
  - Risk minimization, different process etc
  - Less copies?
- Dark Archive
  - No internet
  - Minimization of human error



Single digital object preservation



Reseach communitites, archives, libaries, museums

Preseravtion service

Content: understanding, significance

metadata

Preservation planning

format

Apllication of preservation method

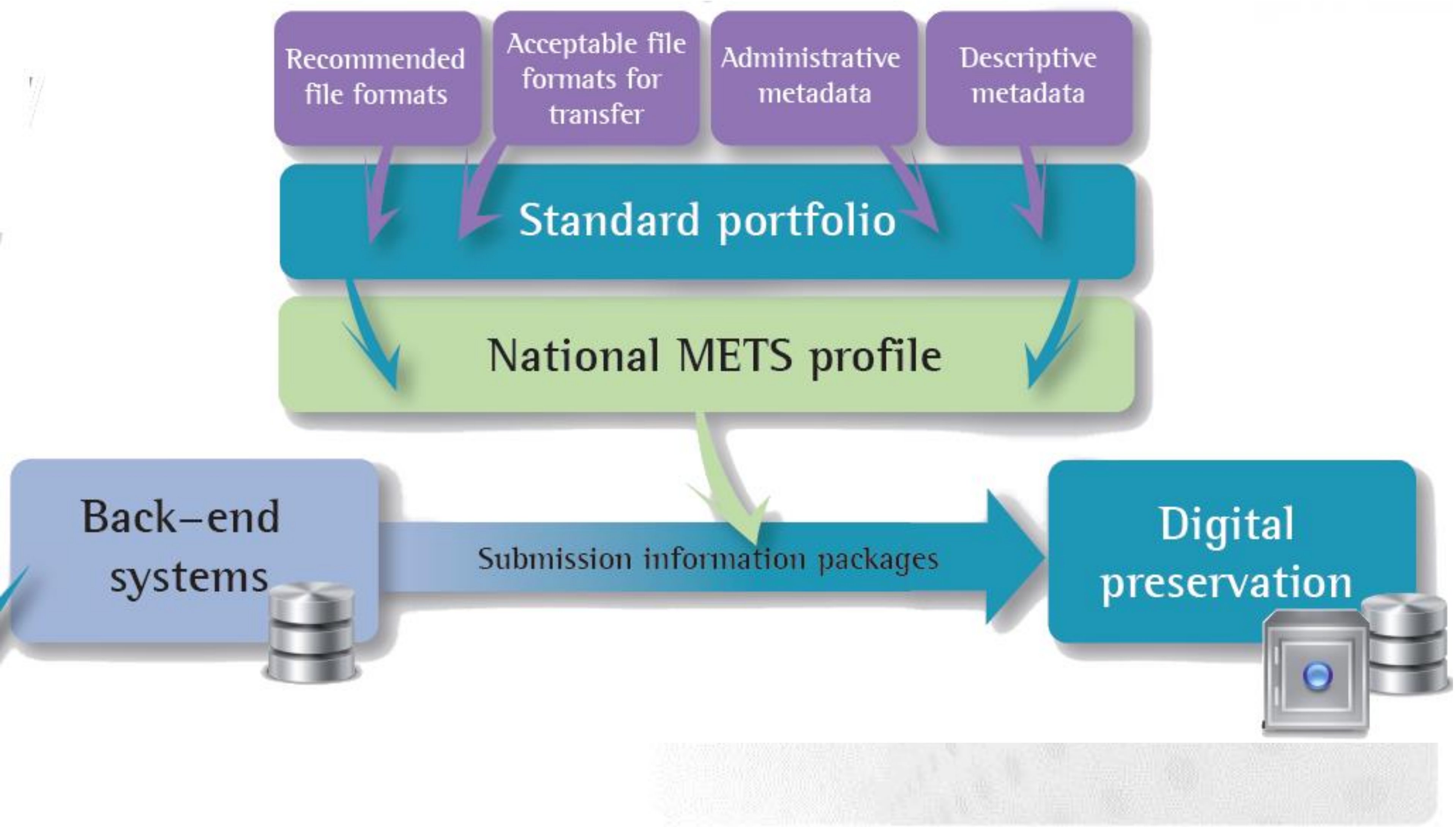
Management of copies

storage media

Storage hardware

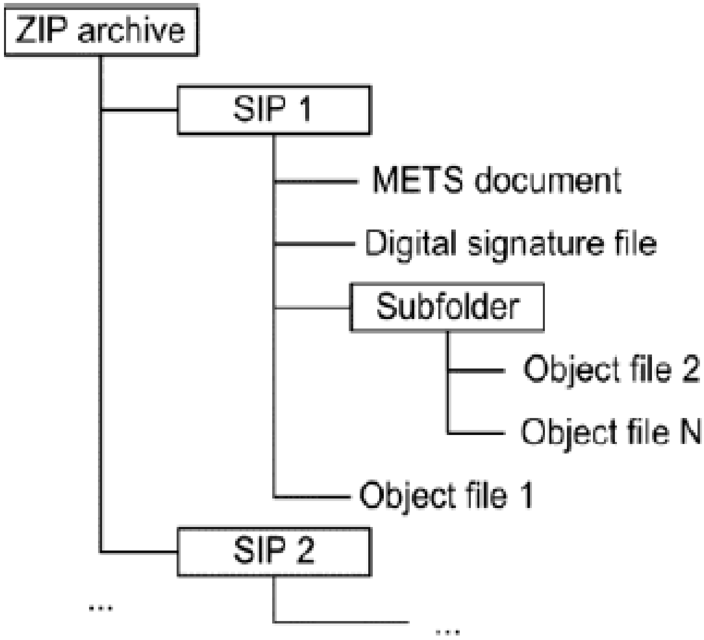
Long Term Use

How? - Nationally Unified Structure



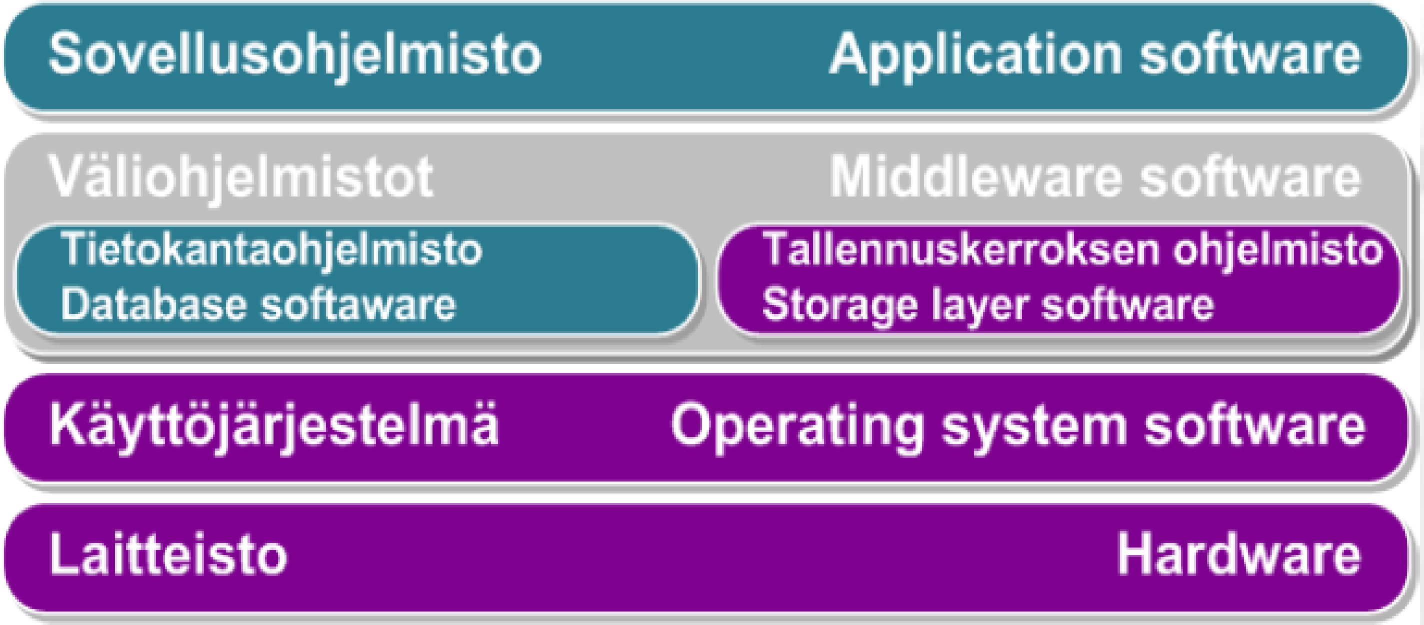
Digital Preservation Solution

- Specifications defined for preparing and creating unified Submission Information Packages (SIPs) with a redifed METS schema
- A closed set of acceptable file formats
- From the acceptable file formats, some are recommended formats, some acceptable for transfer



Technical architecture

- Consists of several layers
- These layers are described in
  - Application architecture
  - Infrastructure architecture



Sovellusohjelmisto Application software

Väliohjelmistot Middleware software

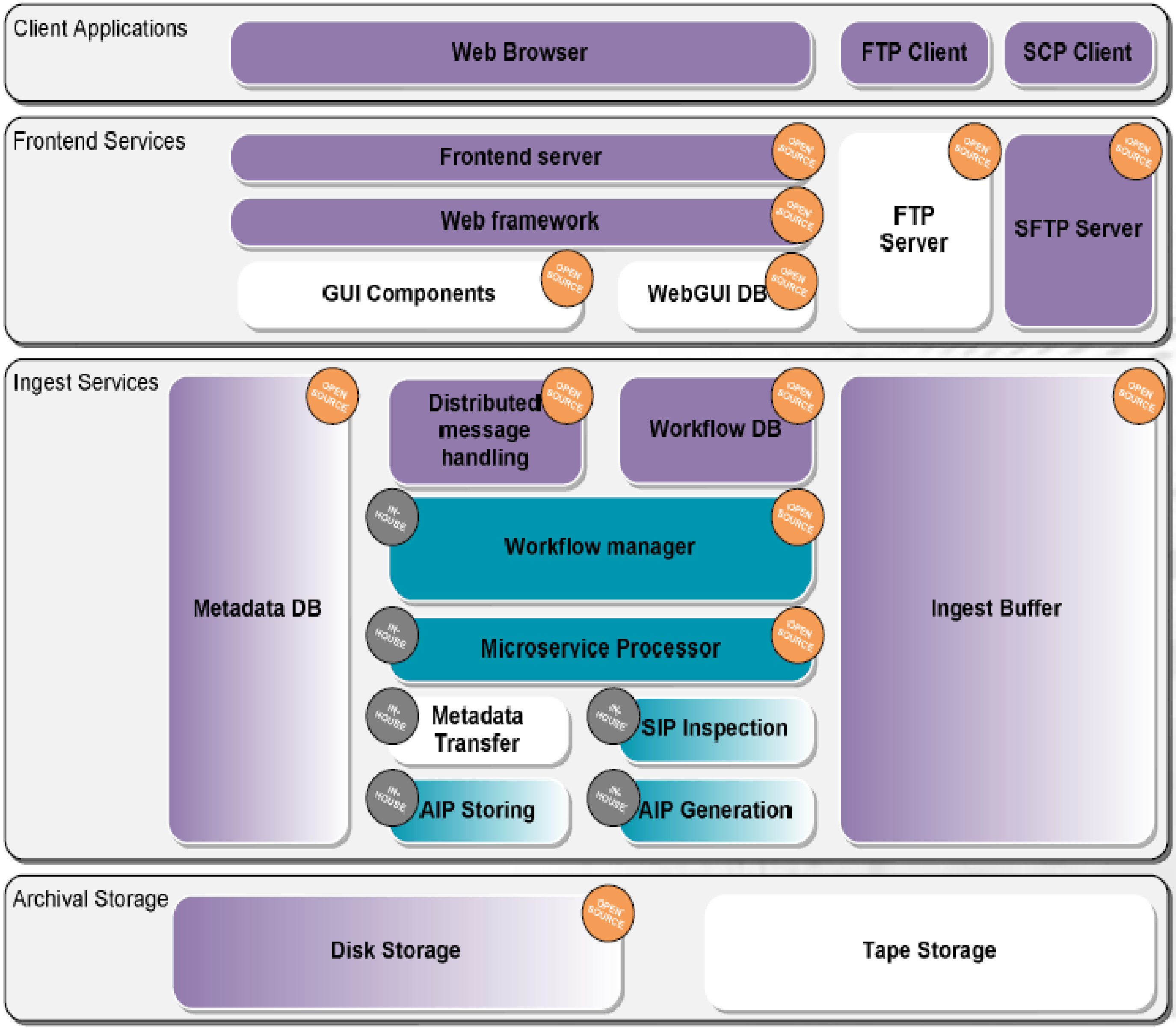
Tietokantaohjelmisto
Database softaware

Tallennuskerroksen ohjelmisto
Storage layer software

Käyttöjärjestelmä Operating system software

Laitteisto Hardware

Ingest



Client Applications

Web Browser

FTP Client

SCP Client

Frontend Services

Frontend server

Web framework

GUI Components

WebGUI DB

FTP Server

SFTP Server

Ingest Services

Metadata DB

Distributed message handling

Workflow DB

Workflow manager

Microservice Processor

Metadata Transfer

SIP Inspection

AIP Storing

AIP Generation

Ingest Buffer

Archival Storage

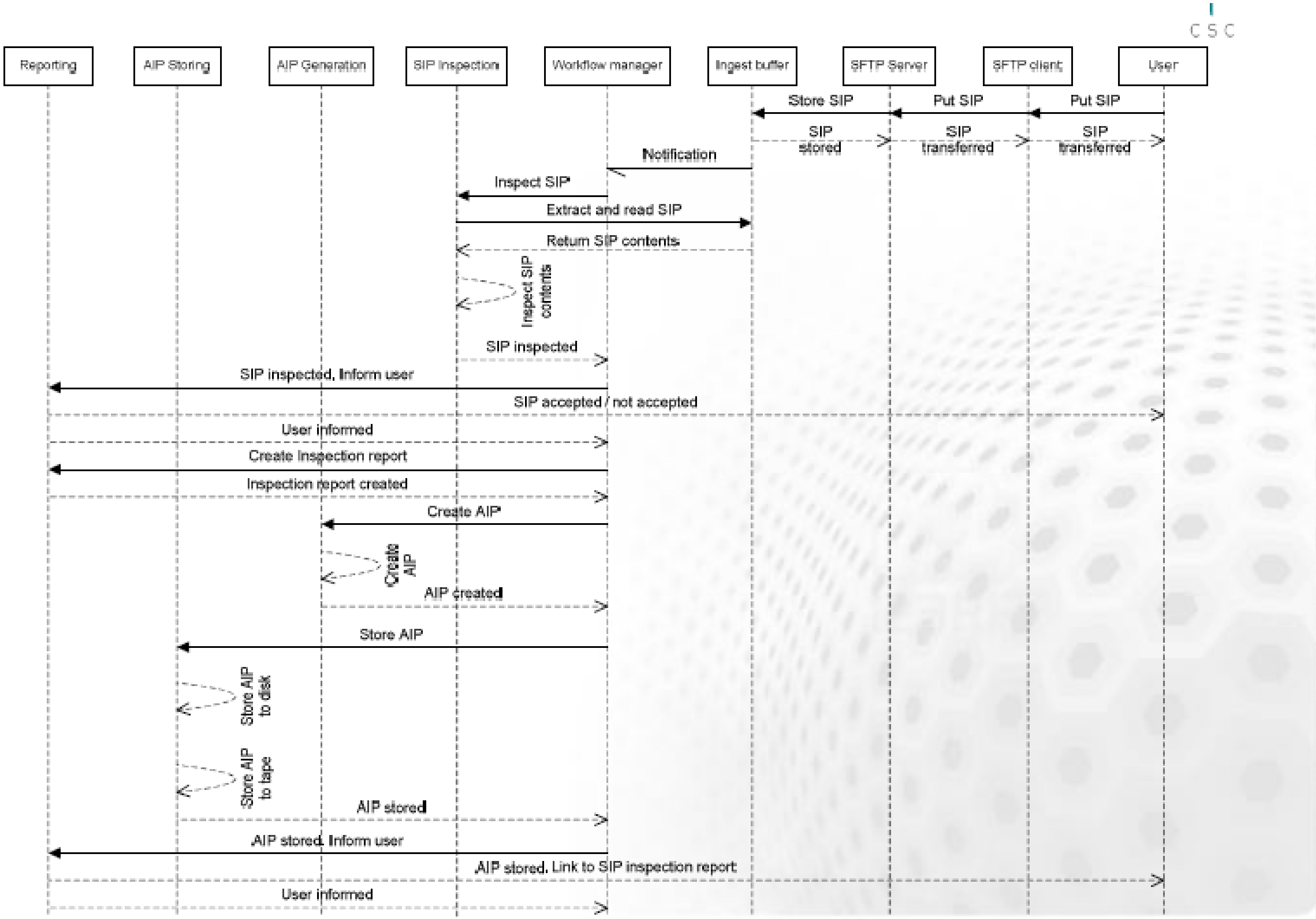
Disk Storage

Tape Storage

Ingest

- We utilize 24 different Open Source components
  - Format checks: 11 components (JHOVE1, JHOVE2, FITS, Epubcheck, Apache ODF Toolkit, Officetron, FLAC, Pngcheck, warc-tools, Ms Office binary File Format Validator, MP3val)
- Missing parts done in-house
- LOT of integration work (technology watch, testing, report handling etc.)

Ingest messages



Reporting

AIP Storing

AIP Generation

SIP Inspection

Workflow manager

Ingest buffer

SFTP Server

SFTP client

User

Store SIP

Put SIP

Put SIP

SIP stored

SIP transferred

SIP transferred

Notification

Inspect SIP

Extract and read SIP

Return SIP contents

Inspect SIP contents

SIP inspected

SIP inspected, Inform user

SIP accepted / not accepted

User informed

Create Inspection report

Inspection report created

Create AIP

Create AIP

AIP created

Store AIP

Store AIP to disk

Store AIP to tape

AIP stored

AIP stored. Inform user

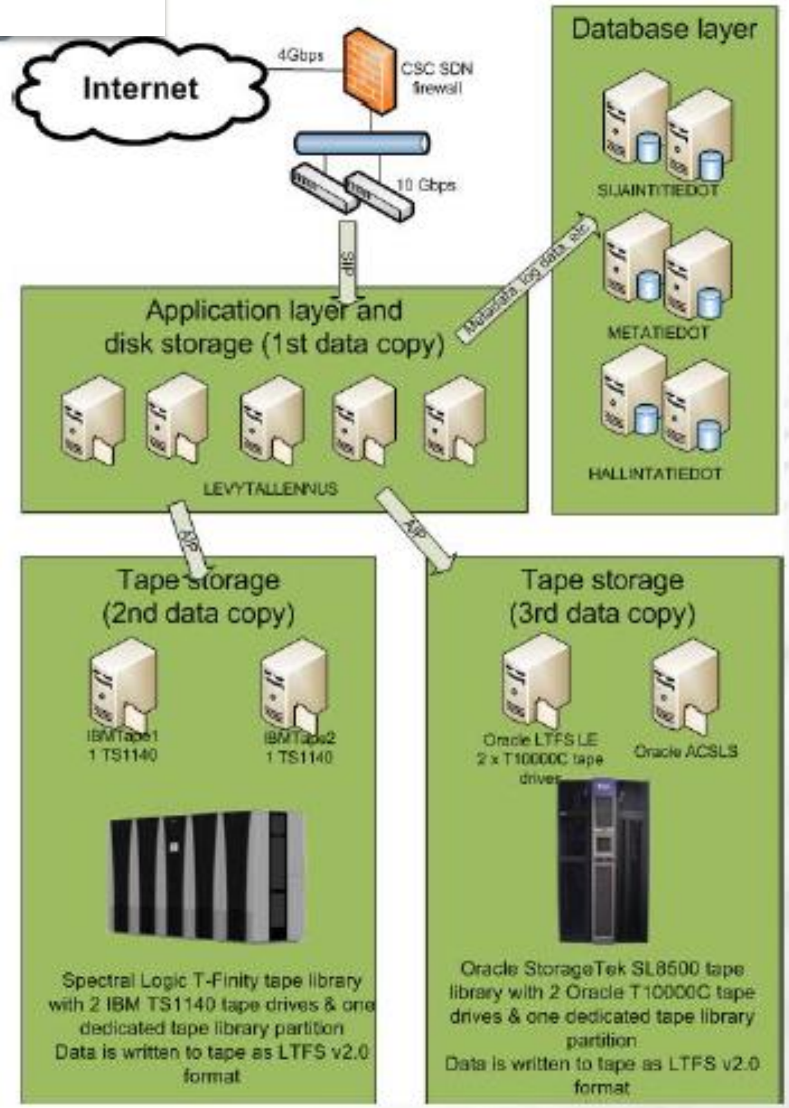
AIP stored. Link to SIP inspection report

User informed

Hardware infrastructure

- Separate database layer
- 3 copies on 3 different media
- Distributed storage group of storage nodes, linked together via tcp/ip & storage software running on operating system

=> no enterprise storage solution



Internet

4Gbps

CSC SDN firewall

10 Gbps

SIP

Metadata, log data, etc

Application layer and disk storage (1st data copy)

LEVYTALLENNUS

Database layer

SIJAINTITIEDOT

METATIEDOT

HALLINTATIEDOT

AIP

AIP

Tape storage (2nd data copy)

IBMTape1
1 TS1140

IBMTape2
1 TS1140

Spectral Logic T-Finity tape library with 2 IBM TS1140 tape drives & one dedicated tape library partition
Data is written to tape as LTFS v2.0 format

Tape storage (3rd data copy)

Oracle LTFS LE
2 x T10000C tape drives

Oracle ACSLS

Oracle StorageTek SL8500 tape library with 2 Oracle T10000C tape drives & one dedicated tape library partition
Data is written to tape as LTFS v2.0 format

In practise

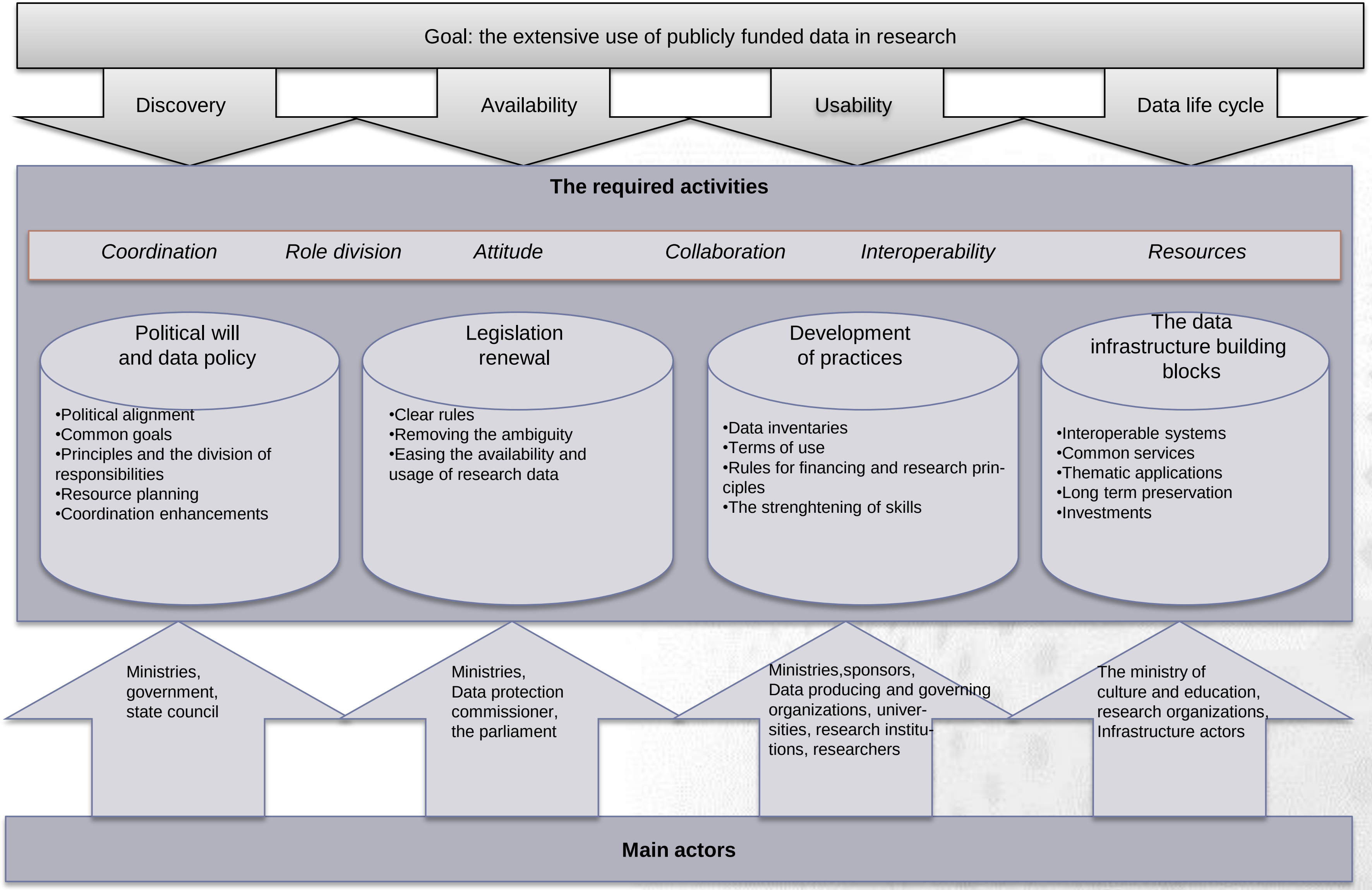
Finnish common digital preservation system:

- Highlights the ownership and roles
- Needs actions starting from policy level and similarly from operational level
- draws practices of partner organizations closer to each other
- reduces the costs and fragmented nature of the ICT systems
- intensifies cooperation

However:

- Common specification (profile) will be most likely updated several times in the future.
  - requires a lot of discussion and collaboration.

Actions on many levels



Goal: the extensive use of publicly funded data in research

Discovery

Availability

Usability

Data life cycle

The required activities

Coordination *Role division* *Attitude* *Collaboration* *Interoperability* *Resources*

Political will
and data policy

•Political alignment
•Common goals
•Principles and the division of responsibilities
•Resource planning
•Coordination enhancements

Legislation
renewal

•Clear rules
•Removing the ambiguity
•Easing the availability and usage of research data

Development
of practices

•Data inventaries
•Terms of use
•Rules for financing and research principles
•The strenghtening of skills

The data
infrastructure building
blocks

•Interoperable systems
•Common services
•Thematic applications
•Long term preservation
•Investments

Ministries,
government,
state council

Ministries,
Data protection
commissioner,
the parliament

Ministries,sponsors,
Data producing and governing
organizations, universities, research institutions, researchers

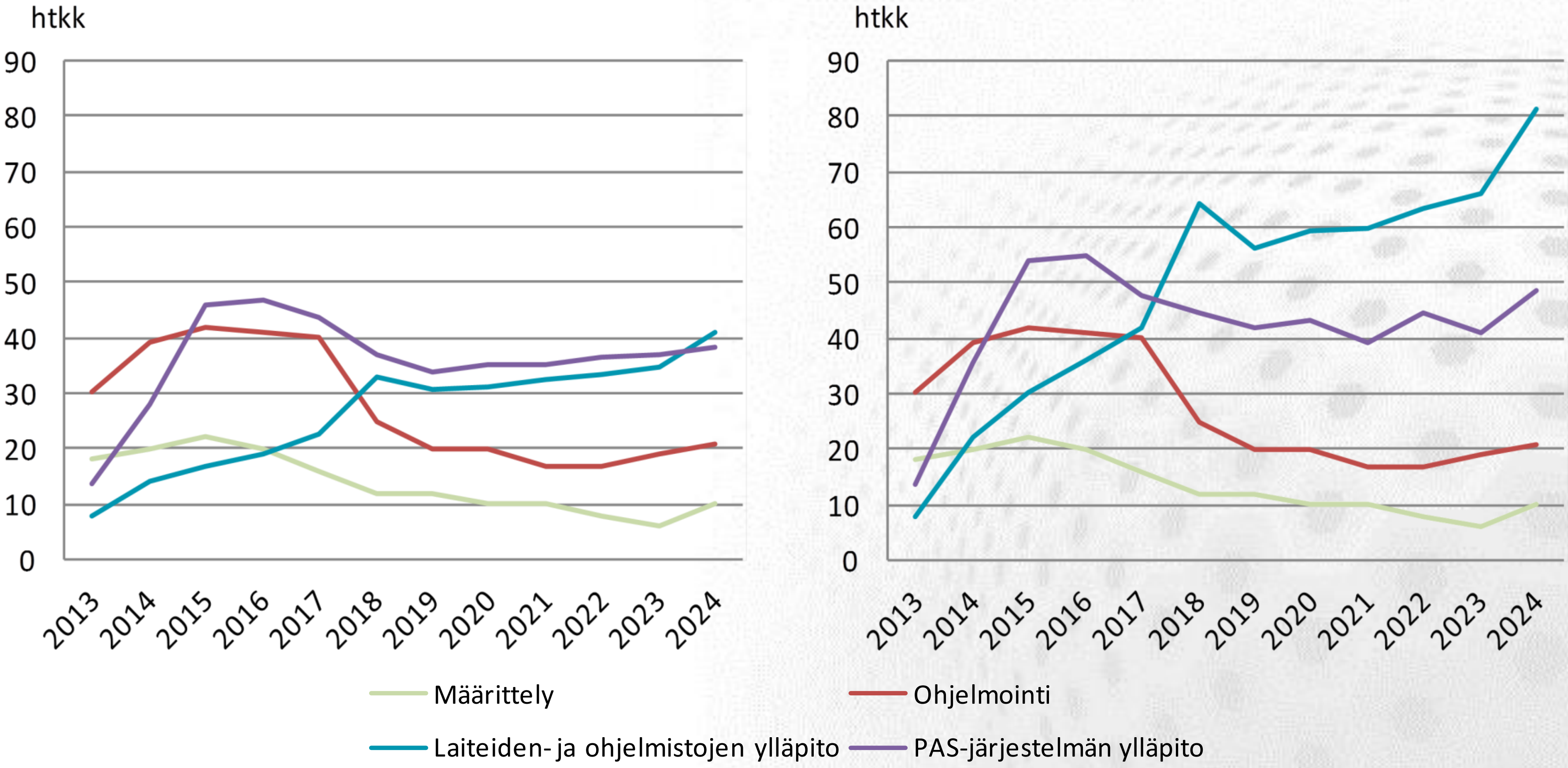
The ministry of
culture and education,
research organizations,
Infrastructure actors

Main actors

Workload distribution

Preservation of cultural heritage

Preservation of scientific data



Pilots

- To understand the preparedness of partner organization
- Ten pilots in total, three libraries, five archives, two museums

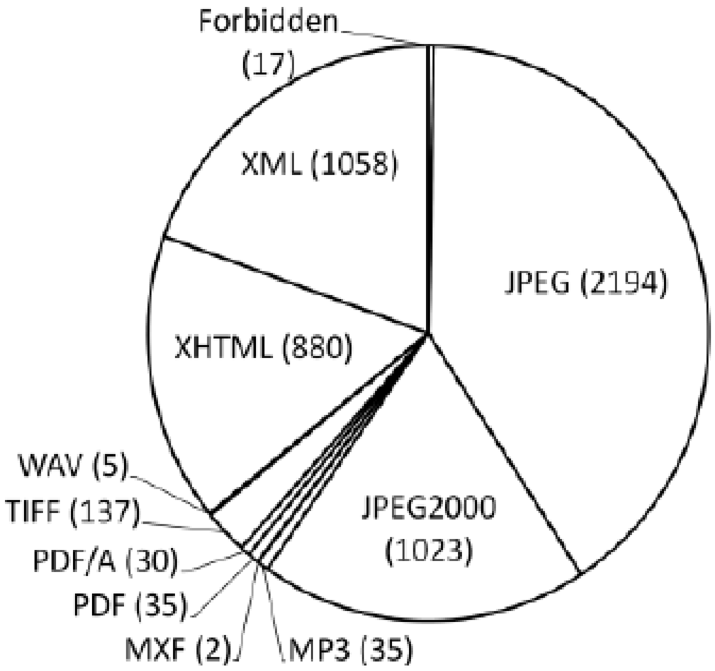


Figure 5. Received number of different file formats.

Source: On preparedness of memory Organizations for Ingesting Data. J.Lehtonen, H.Helin, K.Koivunen, K.Lehtonen, iPRES2013.

Overall packaging results

A grade was given fot each SIP in each validation step

0. The part is missing or does not follow the specifications

1. The part includes sever errors or a large number of minor mistakes

2. The part is flawless or includes only a few minor mistakes

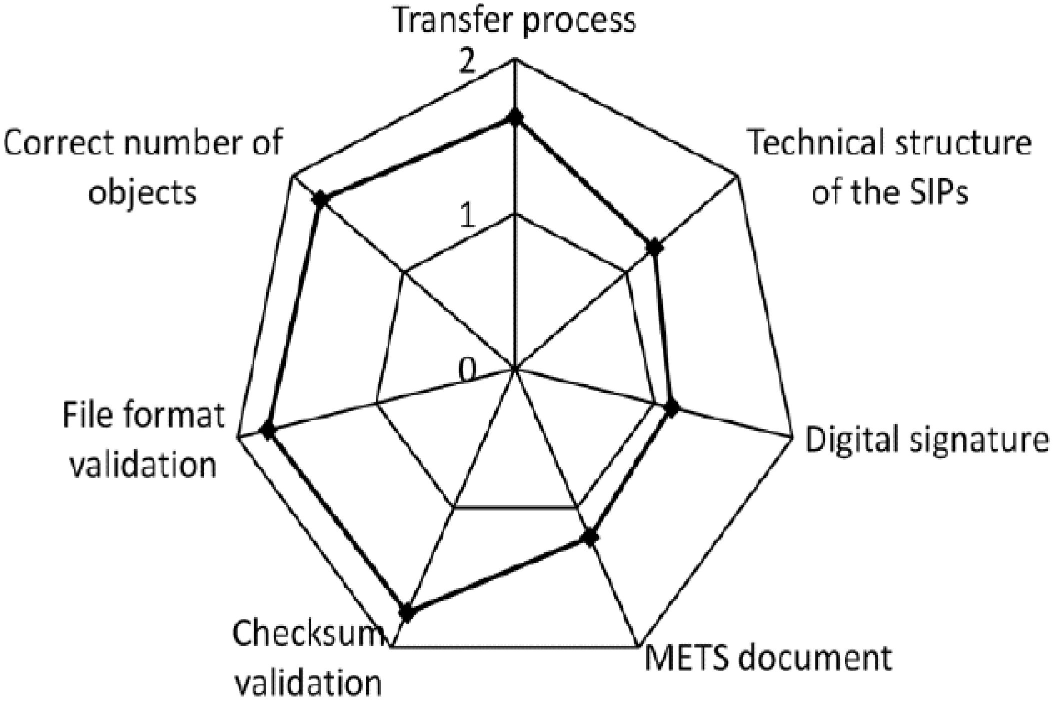


Figure 2. Average grades of the SIPs in validation steps.

Source: On preparedness of memory Organizations for Ingesting Data. J.Lehtonen, H.Helin, K.Koivunen, K.Lehtonen, iPRES2013.

Metadata results

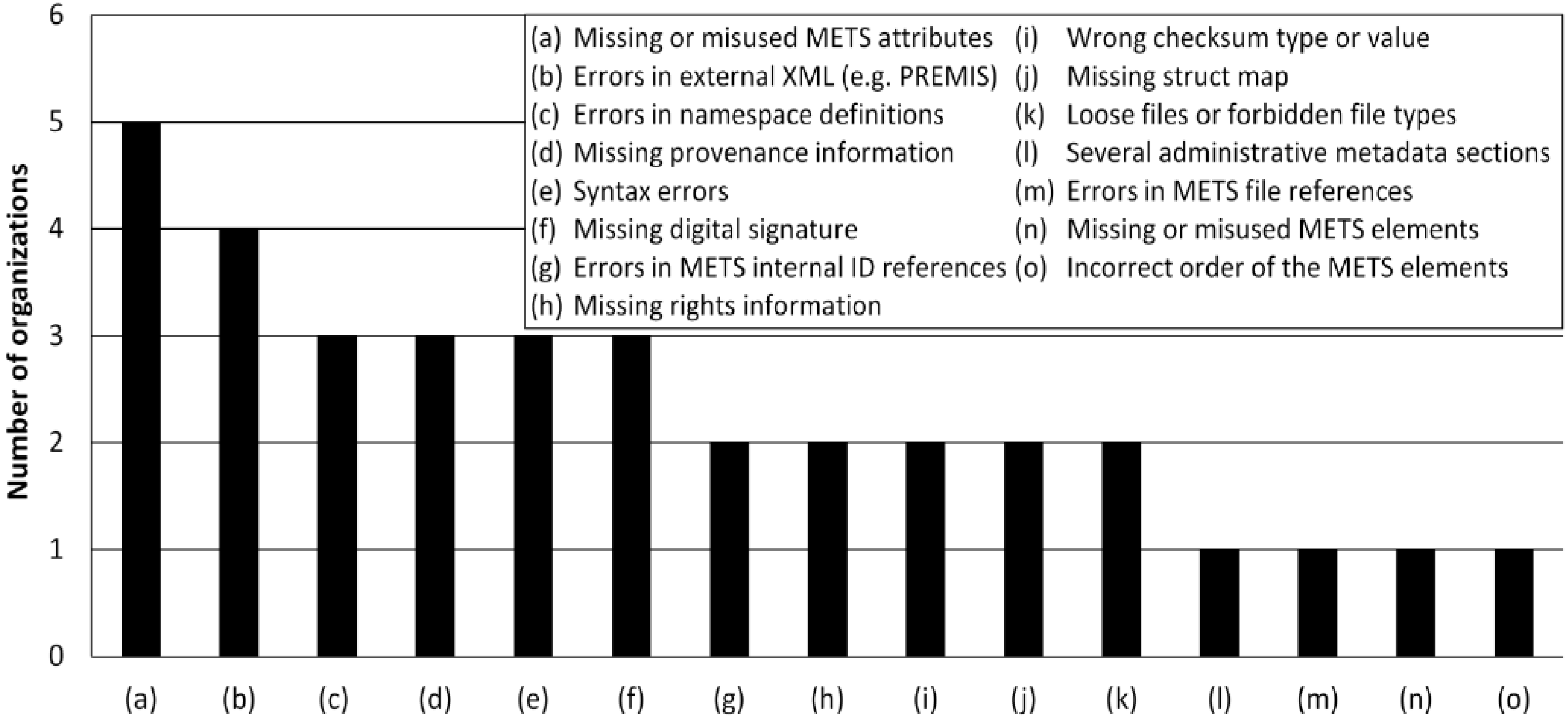


Figure 3. Flaws related to the received METS documents.

Source: On preparedness of memory Organizations for Ingesting Data. J.Lehtonen, H.Helin, K.Koivunen, K.Lehtonen, iPRES2013.

File format results

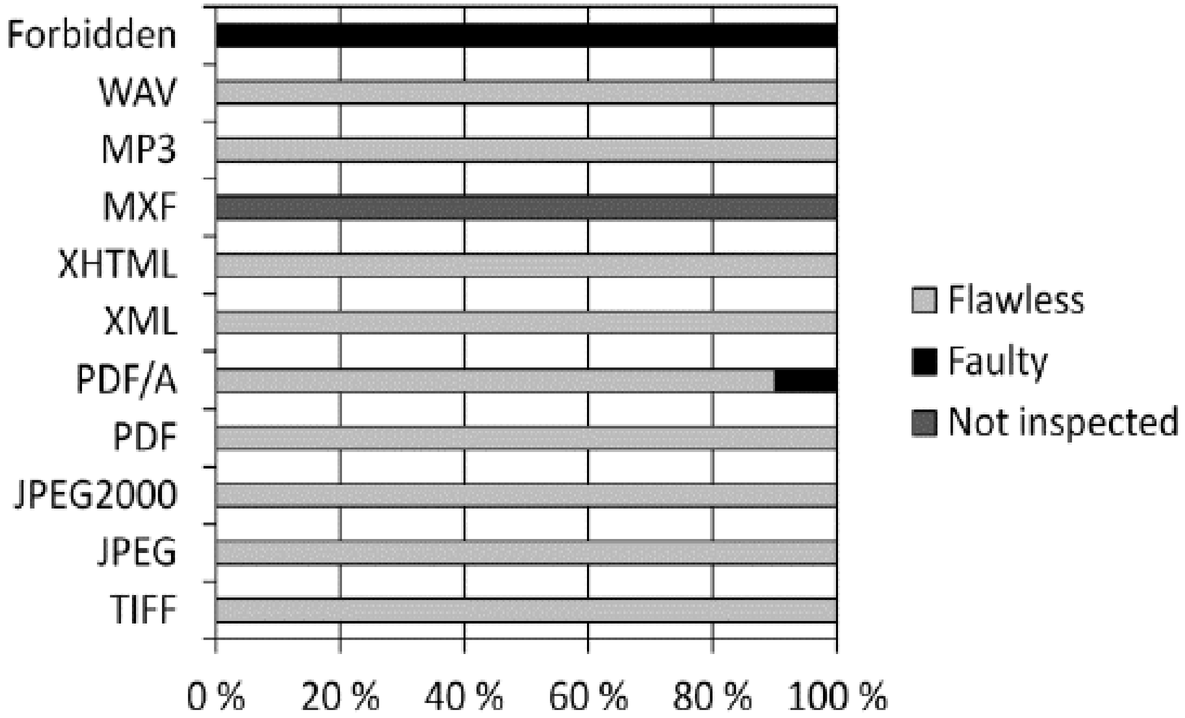


Figure 4. File format validation results.

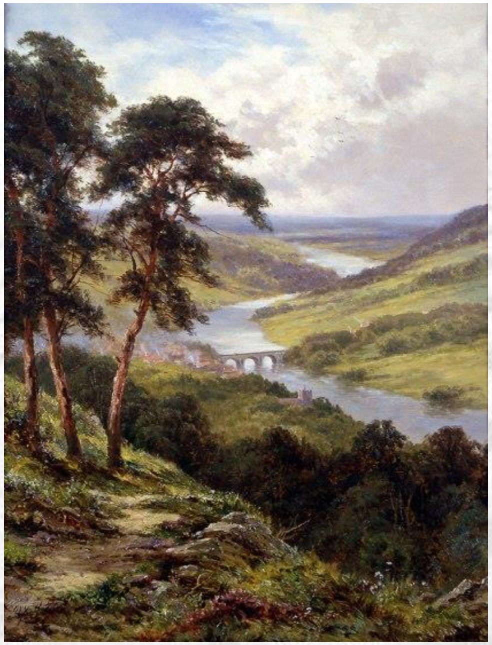
Source: On preparedness of memory Organizations for Ingesting Data. J.Lehtonen, H.Helin, K.Koivunen, K.Lehtonen, iPRES2013.

Never the same again

"πάντα χωρεῖ καὶ οὐδὲν μένει" καὶ "δὶς ἐς τὸν αὐτὸν ποταμὸν οὐκ ἂν ἐμβαίης"
Panta chōrei kai ouden menei kai dis es ton auton potamon ouk an embaies
"Everything changes and nothing remains still ... and ... you cannot step twice into the same stream"[37
Heraclitos

Preservation is an open end problem

- Automatize as much as possible
- Save time: thorough ingest process
- Remember the learning curve
- Accept lifecycles: not everything has to be stored for ever



Source: wikipedia PD
Image resources

Conclusions

- If you want your digital data to survive, start today!
- Equally important:
  - – Ingest
  - – Ownership/stewardship
  - – Preservation planning
  - – Preservation solution
- Clear definition of roles and organization
- Collaboration!
- Exit-strategy

Source: wikipedia PD
Image resources

Thank you!

- http://www.kdk.fi/ http://www.tdata.fi
- *Pirjo-leena.forsstrom@csc.fi*

The National Digital Library

TTA

TUTKIMUKSEN TIETOAINEISTOT