Digital Preservation in Finland

Juha Lehtonen
CSC – IT Center for Science Ltd.
12.9.2014

CC BY NC SA

Mission

CSC, as part of the Finnish national research structure, develops and offers high-quality information technology services.

Vision 2015

- CSC – Pioneer in the Sustainable Development of ICT Services

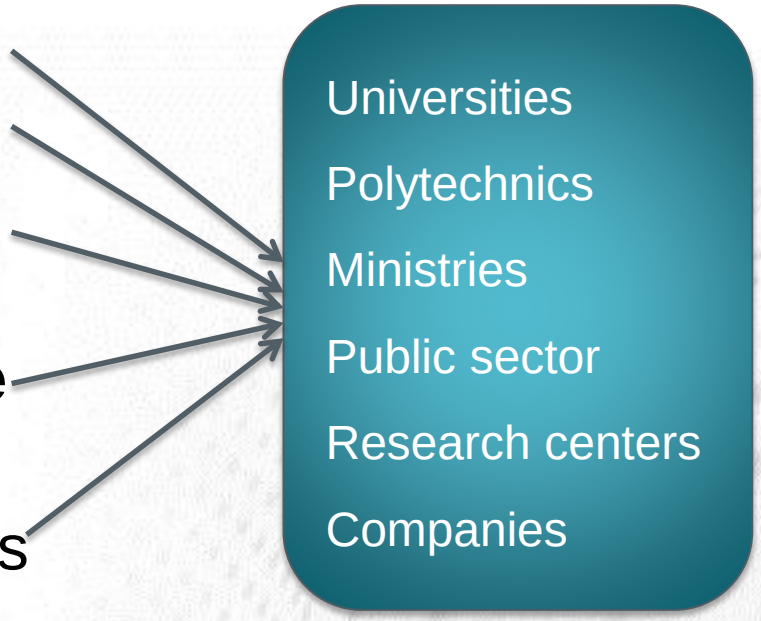
CSC at a Glance

- CSC – IT Center for Science Ltd.
- Owned by Ministry of Education and Culture of Finland
- Operates on a non-profit principle

- Short history:
  - Founded in 1971 as a technical support unit for Univac 1108
  - Connected Finland to the Internet in 1988
  - Reorganized as a company, CSC – Scientific Computing Ltd. in 1993
  - Facilities in Espoo, close to Otaniemi campus (of 15,000 students and 16,000 technology professionals) and Kajaani
  - Staff 260
  - Turnover 2013: 31,2 million euros

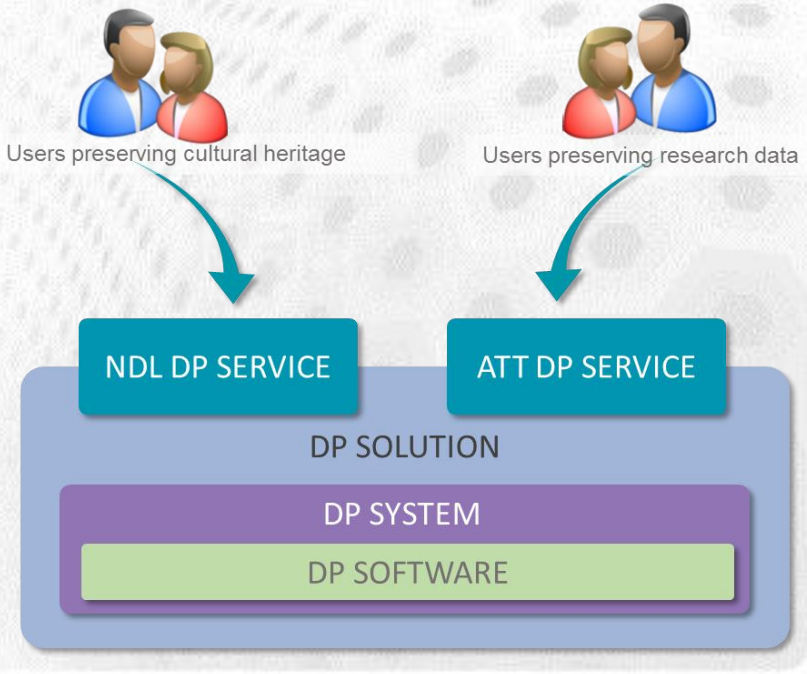
CSC's Services

- Funet Services
- Computing Services
- Application Services
- **Data Services for Science and Culture**
- Information Management Services



Terminology

- DP system
  - Hardware and software for implementing the digital preservation
- DP solution
  - DP system and the organization maintaining and organizing the digital preservation
- DP service
  - Part of the DP solution visible for users
    - NDL DP service for cultural heritage (in production)
    - ATT DP service for research data (being planned)



National Digital Library of Finland

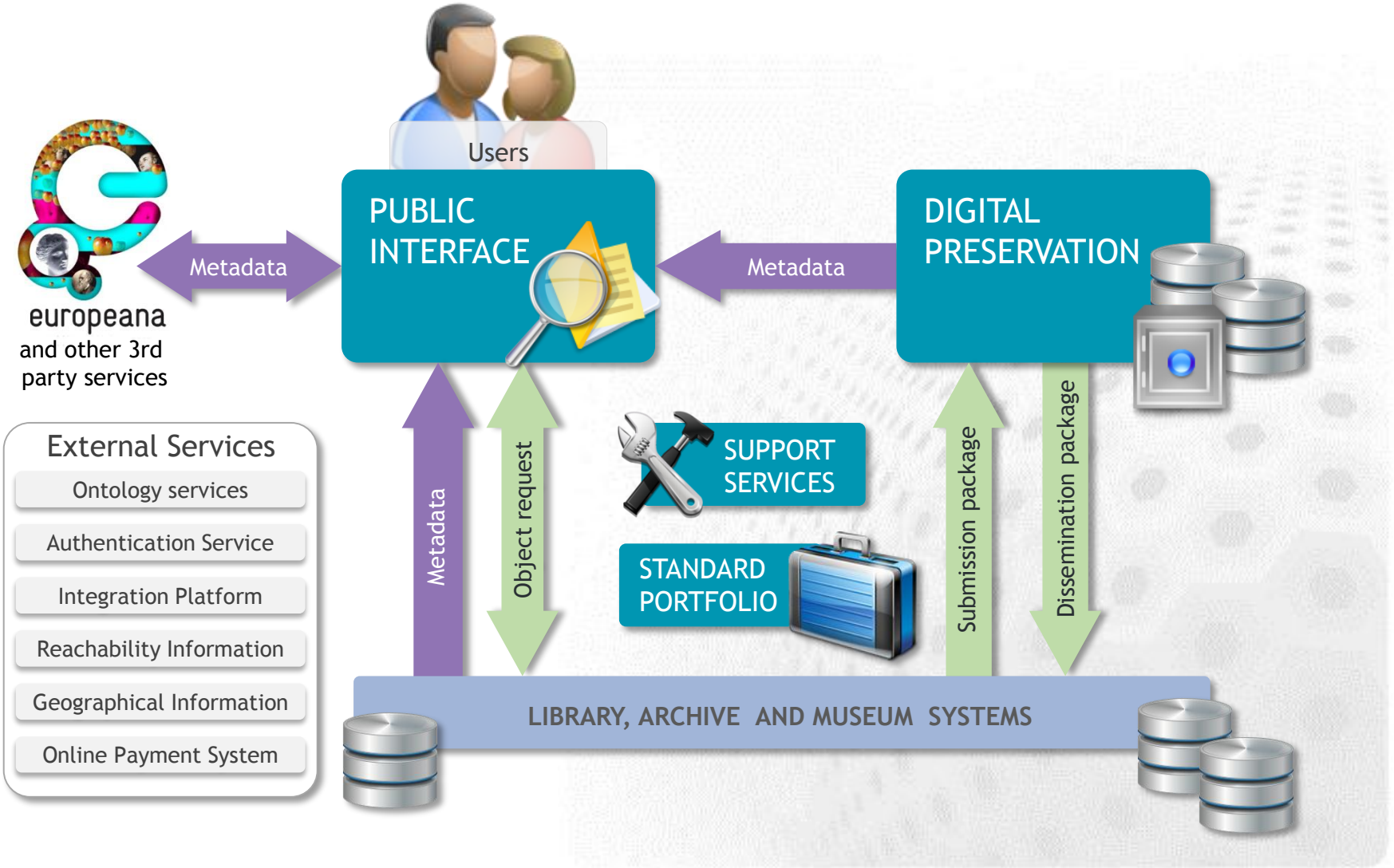
- The Ministry of Education and Culture launched the National Digital Library project (NDL) in 2008.
- The objective of the project is to make the digital data repositories of archives, libraries and museums available to the public now and in the future.
- The project additionally promotes interoperability of processes and IT systems in Finnish memory organizations
- The NDL project includes:
  - Common user interface Finna for the information resources of libraries, archives and museums.
  - Digitisation of the most essential cultural heritage materials of libraries, archives and museums.
  - Development of a digital preservation solution for digital cultural heritage.
  - National Digital Library works as an aggregator for the European Digital Library Europeana.

The National Digital Library

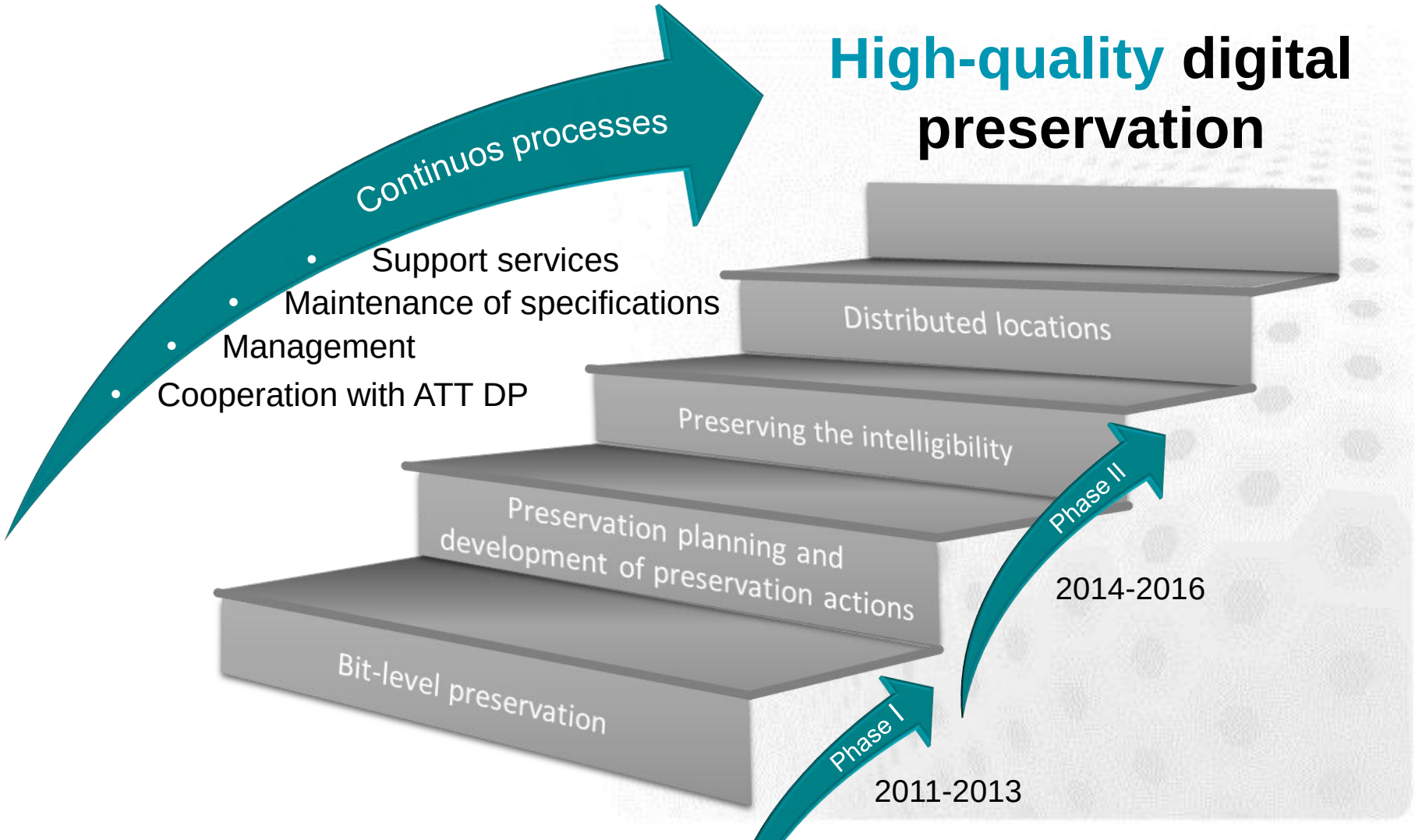
Digital Preservation in NDL

- The Digital preservation (DP) is a system of services which is offered for partner organisations:
  - under the administration of the Ministry of Education and Culture
  - that preserve cultural heritage
- These organisations will transfer the materials intended for long-term or permanent preservation to the NDL's DP service
- Even in the DP system, the ownership of materials will remain with the organisations which stored them.
- Bit level preservation started 2014
- The aim is to have full functionality of the DP system in use by 2016
- Current capacity of the DP system is 0,5 PB which is stored on 3 different media types (disk and 2 tapes)

Enterprise Architecture for NDL



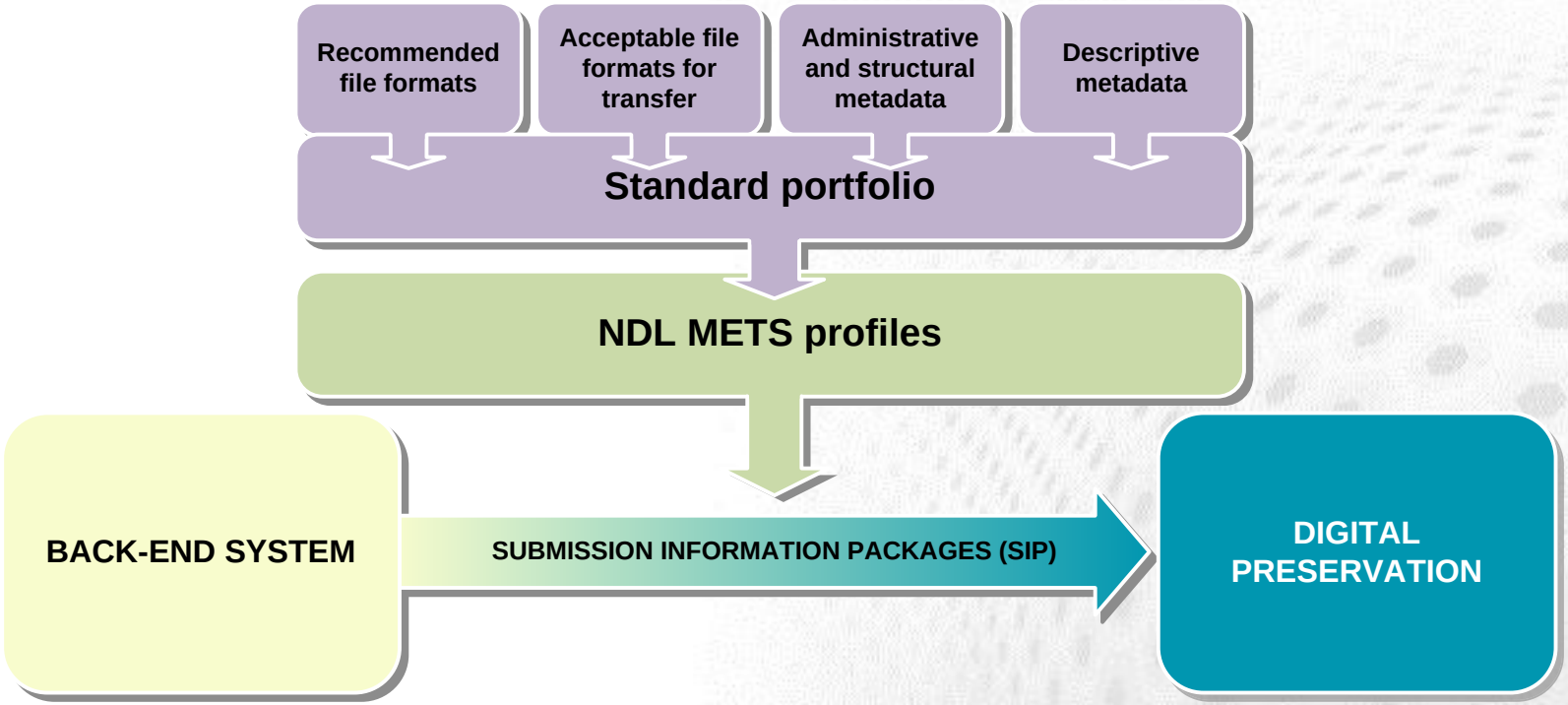
Development of NDL DP System



Standard Portfolio

- Early adoption of common standards
- Description of standards with justifications
  - Metadata, interfaces, …
  - File formats, …
- Enables provision of consistent services, combining data, and developing metadata
  - Semantic commensurable

NDL DP Specifications

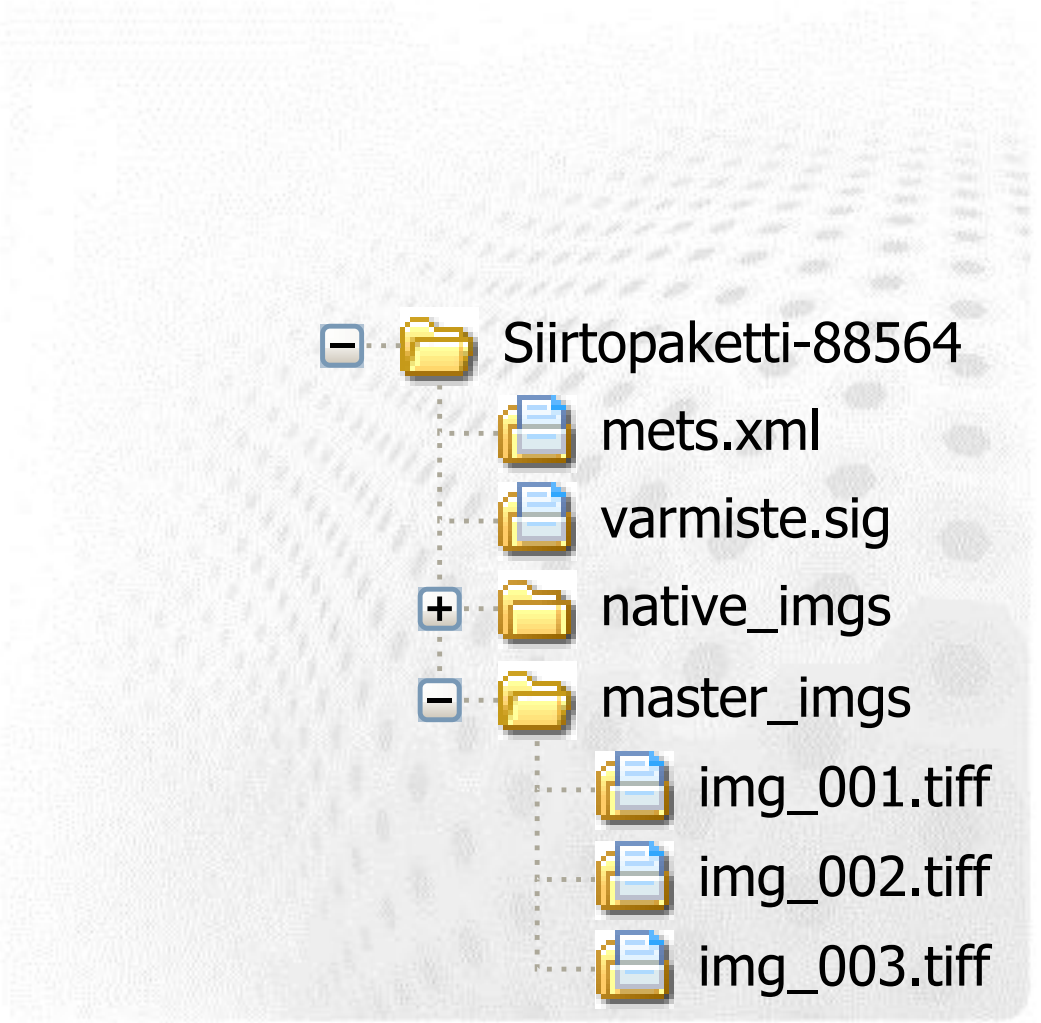


Some Metadata Standards

- METS
- Descriptive metadata:
  - MARC21, FINMARC, DC, MODS, EAC-CPF, EAD, LIDO, VRA Core, Film identification (EN15744), DDI
  - Additionally, other formats may be used:
    - but only additionally and with proper schema
- Technical metadata:
  - PREMIS:OBJECT, MIX, textMD, audioMD, videoMD
- Provenance metadata:
  - PREMIS:EVENT, PREMIS:AGENT
- Rights metadata:
  - Limitations for DP system: PREMIS:RIGHTS
  - Copyright metadata: Currently agreed case by case

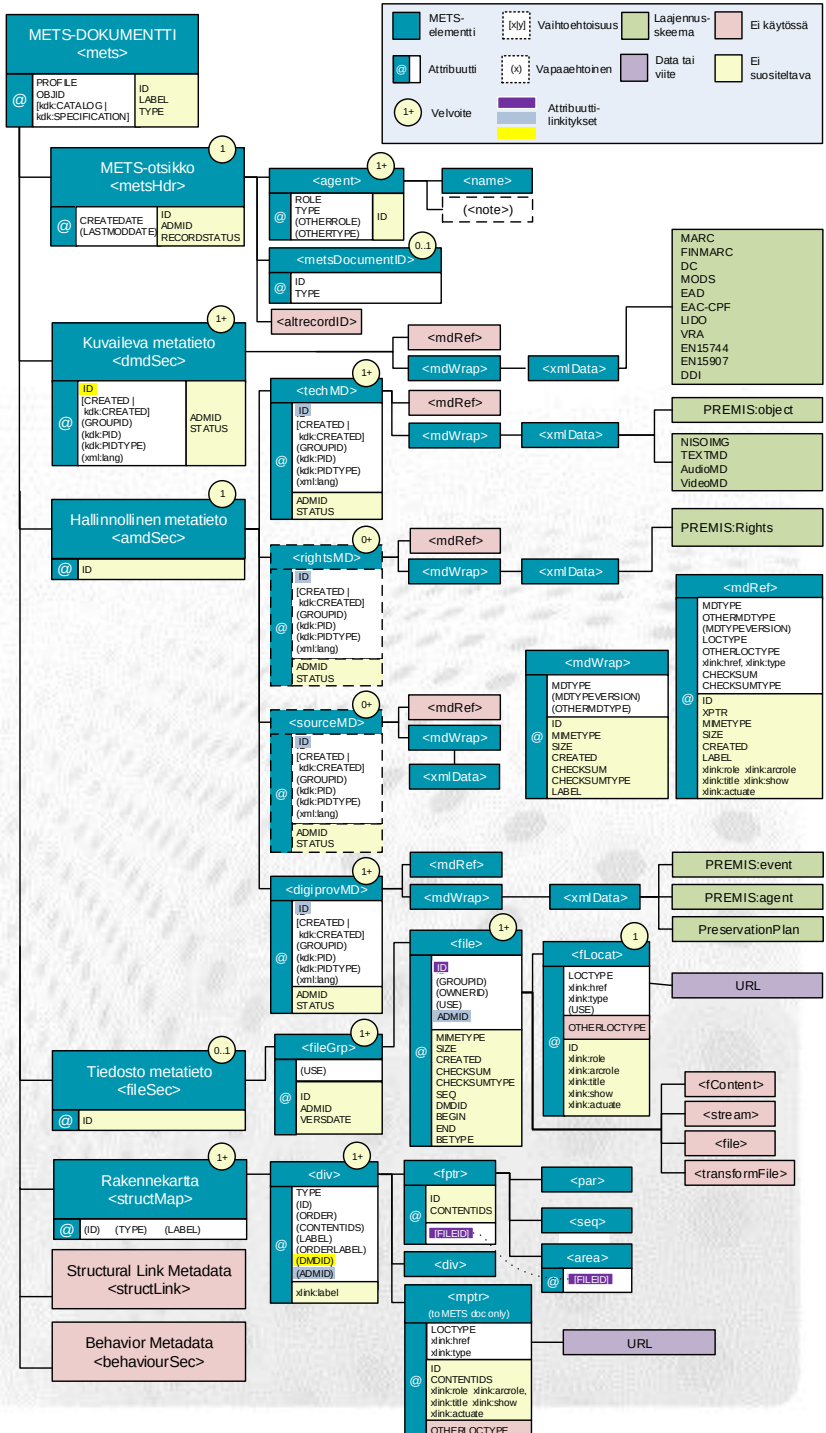
Submission Information Package

- METS document
- Digital signature
- Digital objects



Mandatory Metadata

- Descriptive metadata
- Provenance metadata
  - premis:event
- File format and version
  - premis:format
- Fixity
  - premis:fixity
- Some other technical metadata for some file formats
- Struct map
  - mets:structMap



Mandatory Metadata

- ID for the SIP
  - @OBJID
- IDs for digital objects
  - premis:objectIdentifier
- Creation (and last modification) time of the SIP
  - @CREATEDATE, @LASTMODDATE
- Creation time for metadata sections and files
  - Metadata: @CREATED, @ndl:CREATED
    - NDL attribute extension for EDTF time format
  - Files: premis:dateCreatedByApplication
- Metadata type of wrapped metadata
  - @MDTYPE
- Some other fields

Recommended Metadata

- Technical metadata
  - Partly mandatory
- Rights metadata
- Event history
- Persistent IDs for the metadata sections
  - NDL attribute extension, e.g. techMD@ndl:PID, techMD@ndl:PIDTYPE
- Some other fields

Some Restrictions in NDL METS

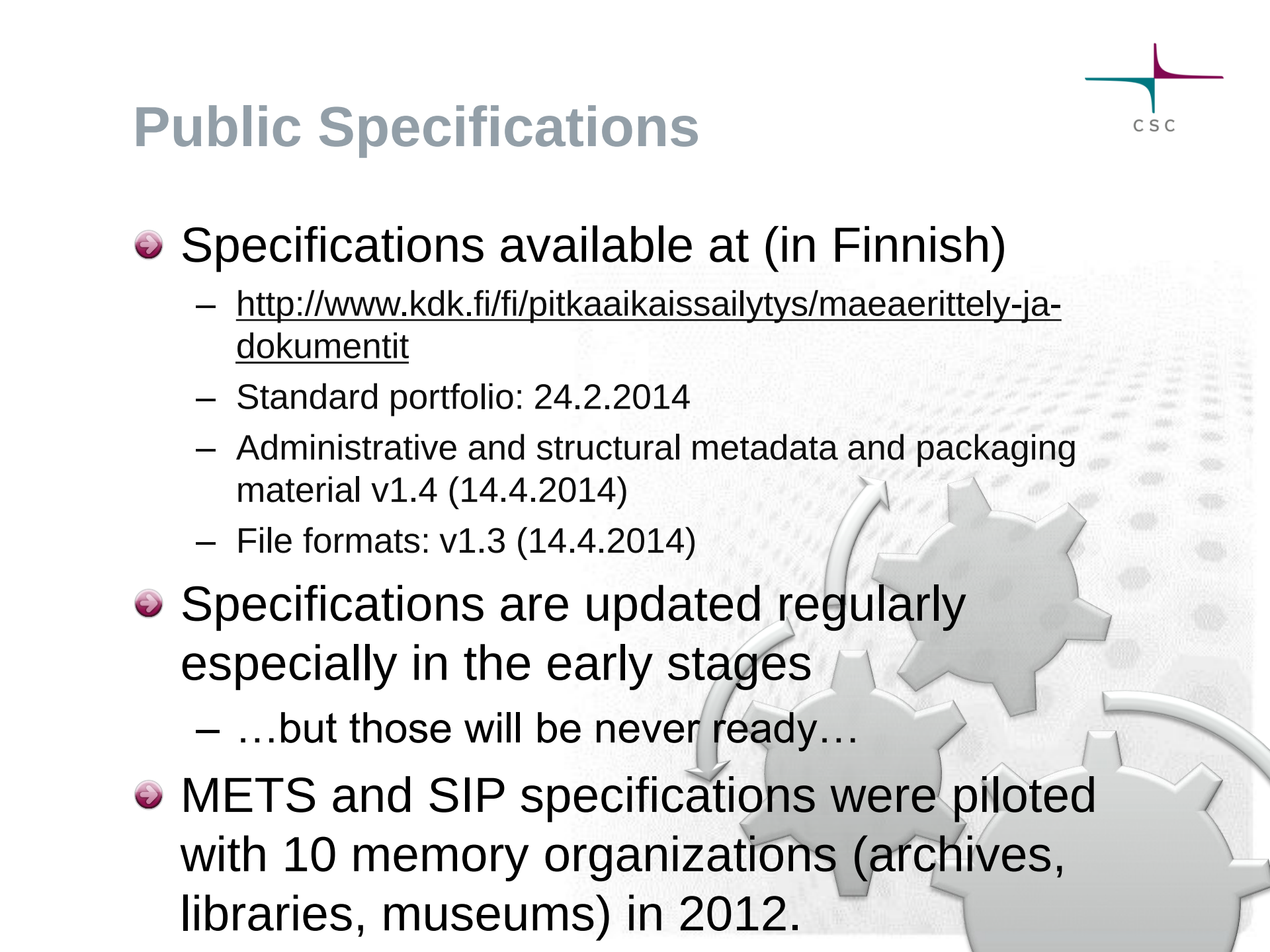
- Only one admSec section allowed
- No structLink section allowed
- No behaviourSec section allowed
- No embedded binary data allowed
- No digital object references outside the SIP allowed (everything needs to be in the SIP)
- No mdRef (metadata references outside the METS file) allowed, except mandatory for the preservation plan
- Various controlled vocabularies given

NDL METS Validation

- XML validation
  - Various metadata formats
  - NDL Schema Catalog
    - Forces to use local copies of schema
    - XML Catalogs v. 1.1, OASIS Standard, 2005
      - http://www.oasis-open.org
- Schematron rules for more complex validation
  - ISO/IEC 19757-3:2006
  - Validates via XSLT conversion
  - e.g. ”For a file, which format is image/tiff, validate that there exists a technical metadata section done with MIX, and that the file and the metadata section are linked together in METS.”

Public Specifications

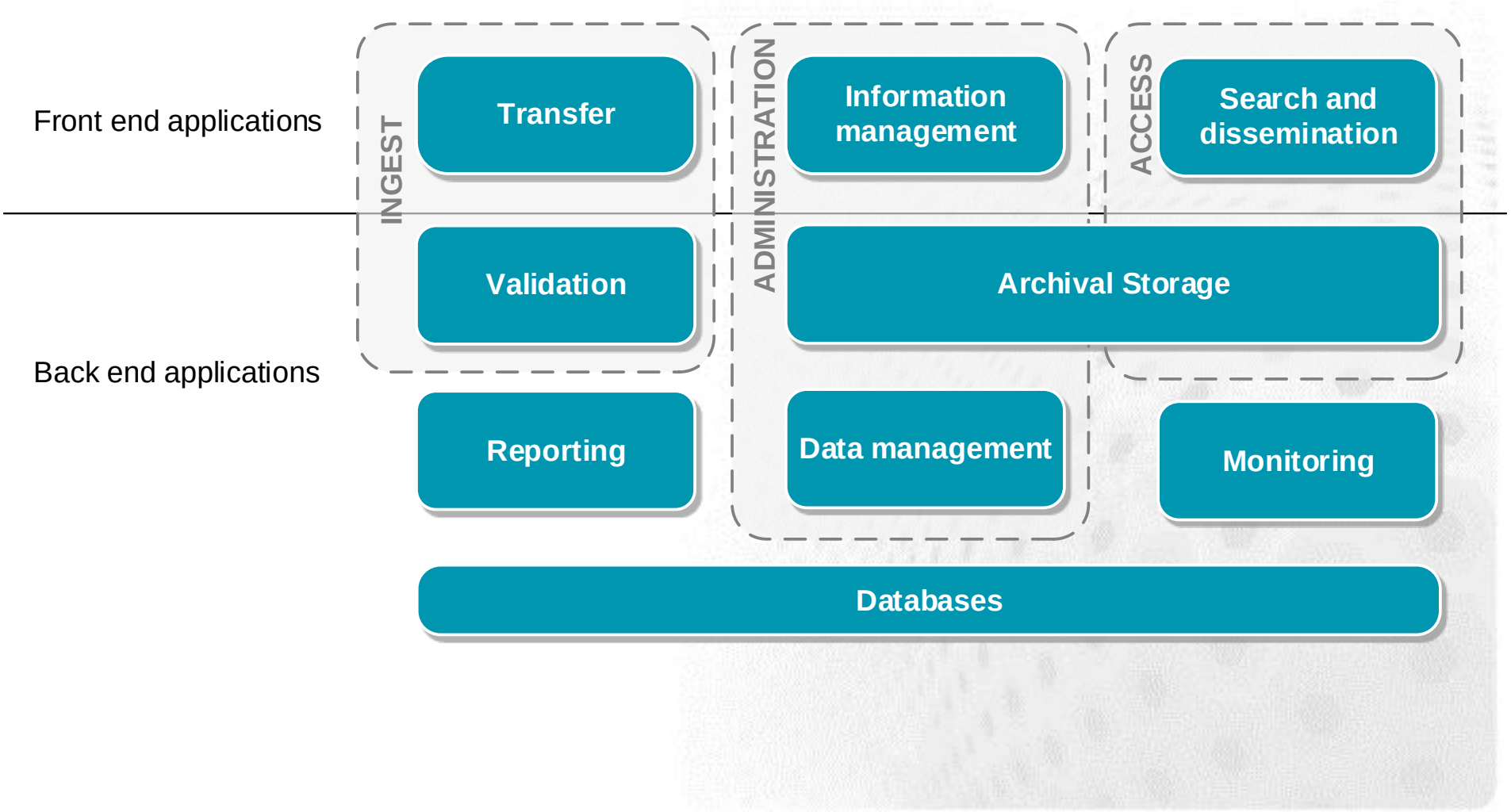
- Specifications available at (in Finnish)
  - http://www.kdk.fi/fi/pitkaaikaissailytys/maeaerittely-ja-dokumentit
  - Standard portfolio: 24.2.2014
  - Administrative and structural metadata and packaging material v1.4 (14.4.2014)
  - File formats: v1.3 (14.4.2014)
- Specifications are updated regularly especially in the early stages
  - …but those will be never ready…
- METS and SIP specifications were piloted with 10 memory organizations (archives, libraries, museums) in 2012.



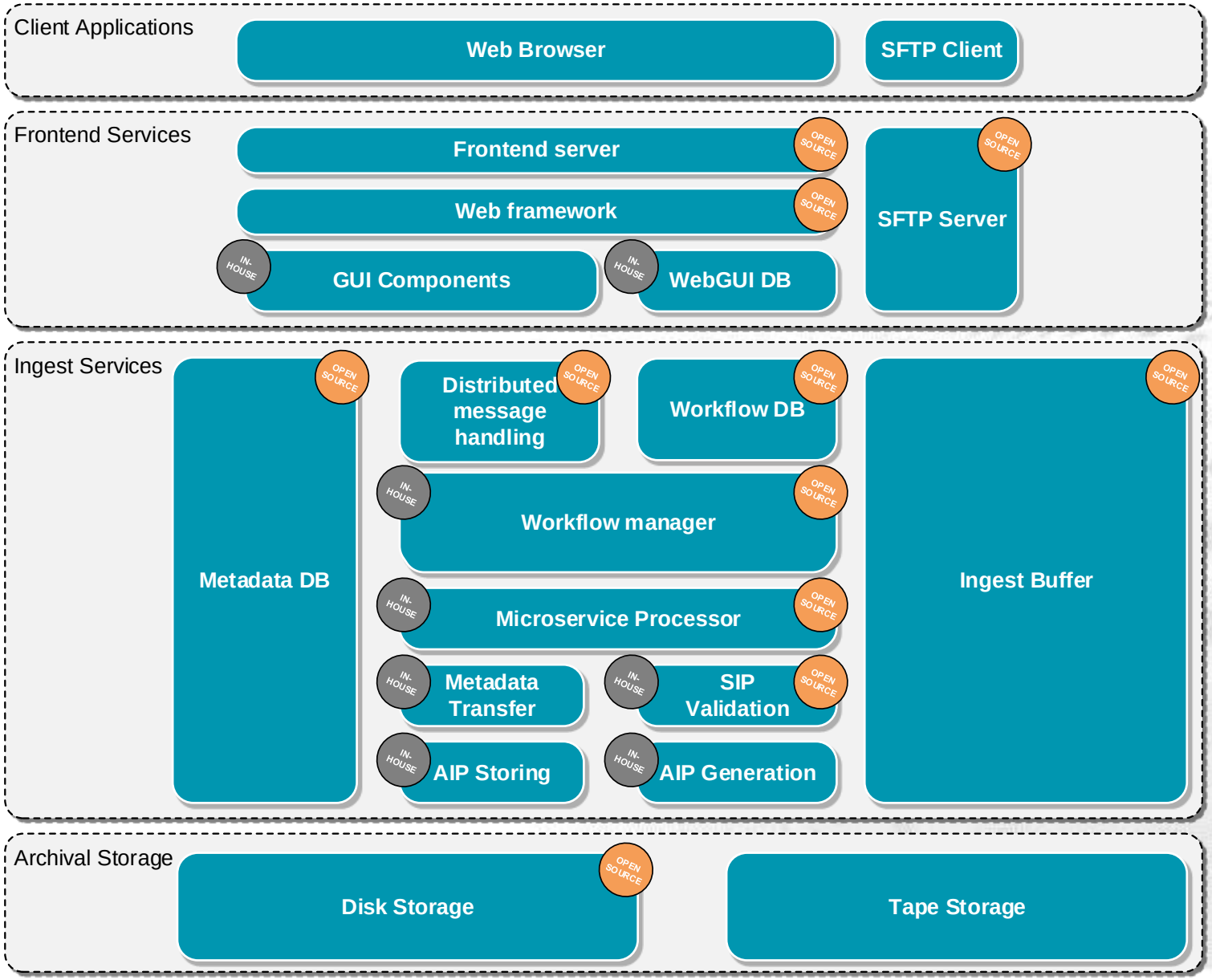
Amount of Data to be Preserved (2011)

	2010		2011		2015		2020	
	Number of files (millions)	Size (TB)	Number of files (millions)	Size (TB)	Number of files (millions)	Size (TB)	Number of files (millions)	Size (TB)
Documents	11,6	328	15,4	394	25,6	646	48,7	1301
Still Images	1,7	18	2,1	30	3,9	68	6,1	120
Digital Video	0,1	495	0,2	1143	0,8	3055	1,2	8020
Sound	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
Web Archive	496	20	646	27	1396	59	2300	97
Radio and TV Archive	0,8	95	1,2	142	2,9	327	5,0	558
TOTAL	**530**	**1 563**	**687**	**2 509**	**1458**	**5 575**	**2400**	**12 275**

Software Architecture of DP System



Ingest


Client Applications
Web Browser
SFTP Client
Frontend Services
Frontend server
Web framework
GUI Components
WebGUI DB
SFTP Server
Ingest Services
Metadata DB
Distributed message handling
Workflow DB
Workflow manager
Microservice Processor
Metadata Transfer
SIP Validation
AIP Storing
AIP Generation
Ingest Buffer
Archival Storage
Disk Storage
Tape Storage
OPEN SOURCE
IN-HOUSE

Thank you

http://kdk.fi/en