
Long-term preservation in French public institutions
Two case studies
Digital Preservation - The Planets Way
London; 9-11 February 2010
BnF
INES

Agenda

- French legal context for digital preservation
- Long-term preservation at CINES
  - Context of digital preservation at CINES
  - Architecture and processes
  - Project update and perspectives
- Long-term preservation at BnF
  - Context of digital preservation at BnF
  - Overall architecture and main features
  - Mid-term new developments

BnF

INES

French legal context

- Different institutions are involved at the end of the digital document lifecycle depending on the nature and context of production of the electronic objects to preserve
  - Archives départementales (local administrations, universities)
  - Archives Nationales (national administrations, other institutions for Higher Eduction & Research)
  - BnF (legal deposit: publications, Web)
  - CINES (PhD Theses)
- Digitisation programs don't fall in such context

  The digitising institution
  - keeps the original documents within its collections
  - has a choice of preservation solutions for the digital documents

BnF

CINES


Long-term preservation at CINES
Plateforme d'Archivage du CINES
(PAC)
BnF
CINES

Overview of CINES

Centre Informatique National de l'Enseignement Supérieur

- Based in Montpellier (Hérault, France)
- Created in 1999, formerly known as CNUSC (Centre National Universitaire Sud de Calcul) – created in 1980
- Administrated and funded by ministry of Higher education & research (MESR)
- Main areas of expertise
  - High Performance Computing – ranks 14th worldwide
  - Long-term preservation of digital documents

Cross-discipline activities : environment & server hosting

BnF

CINES

The mandate in long term preservation

In 2004 the CINES was given the mandate to provide long-term preservation capabilities for digital objects related to scientific and technical information

This mission has been confirmed by few decisions from the CINES administrative control :

- August 7th, 2006 : Arrêté relatif aux modalités de dépôt, de signalement, de reproduction, de diffusion et de conservation des thèses ou des travaux présentés en soutenance en vue d'un doctorat
- ➔ The CINES became the official preservation centre for electronic PhD theses
- February 12th, 2008 : Lettre de cadrage du ministère
- ➔ Reinforced the two mains activities of the CINES : high performance computing and long-term preservation of digital documents

BnF

CINES

The CINES preservation service

The PAC (Plateforme d'Archivage du CINES) project was initiated in 2004 to implement a generic platform dedicated to long-term preservation of electronic documents

Objectives : the rollout of an effective, high-performance, scalable, secure and inexpensive solution for the education and research digital heritage

Constraints

- Adherence to the OAIS model as well as other standards : Standard d'échange de données pour l'archivage électronique, DCMI, etc
- Support of standard file formats (limited set of formats accepted)

Focus on data :

- Scientific data – results of observations, measurements, etc.
- Cultural heritage – publications, pedagogics, etc.
- Administrative data – semi-current records

In due respect of the French archivistic legal context

BnF

CINES

Challenges for long-term preservation of digital objects

Challenge	Solutions
Knowledge of content	• Use of metadata (DCMI, etc) • Unique ID for stored documents (ARK)
File formats	• Use of a limited set of standard formats • Logical migration (conversion)
Medias	• Supervision, management of ageing of medias • Physical migration
Software and hardware obsolescence	• Technological watching activities, anticipation

BnF

INES

File formats supported

The file formats supported are :

- Open / published format
- Widely used format
- Standard format

Type	Format
Text	HTML, PDF, TXT, XML, ODT
Picture	GIF, JPEG, TIFF, PNG, SVG
Audio	WAV, AIFF, AAC, OGG (VORBIS)
Video	MPEG4, OGG (THEORA), MKV

The PAC platform uses Jhove, ImageMagick, DROID, mEncoder and ODF Toolkit libraries to

- Identify,
- Validate
- Characterize,

The format of transferred files

BnF

INES

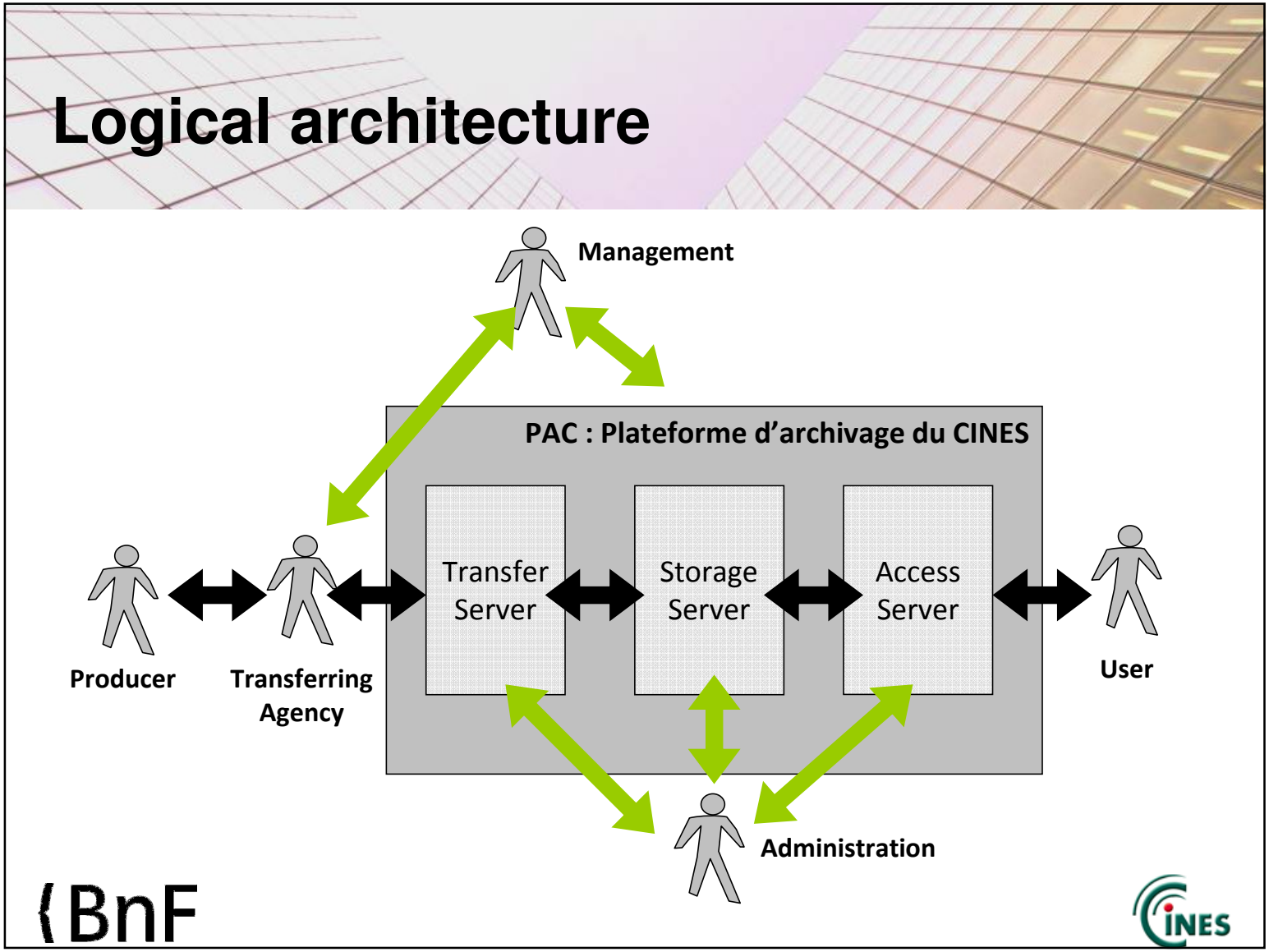
The PAC platform

PAC is built of three logical servers, as defined in the OAIS model

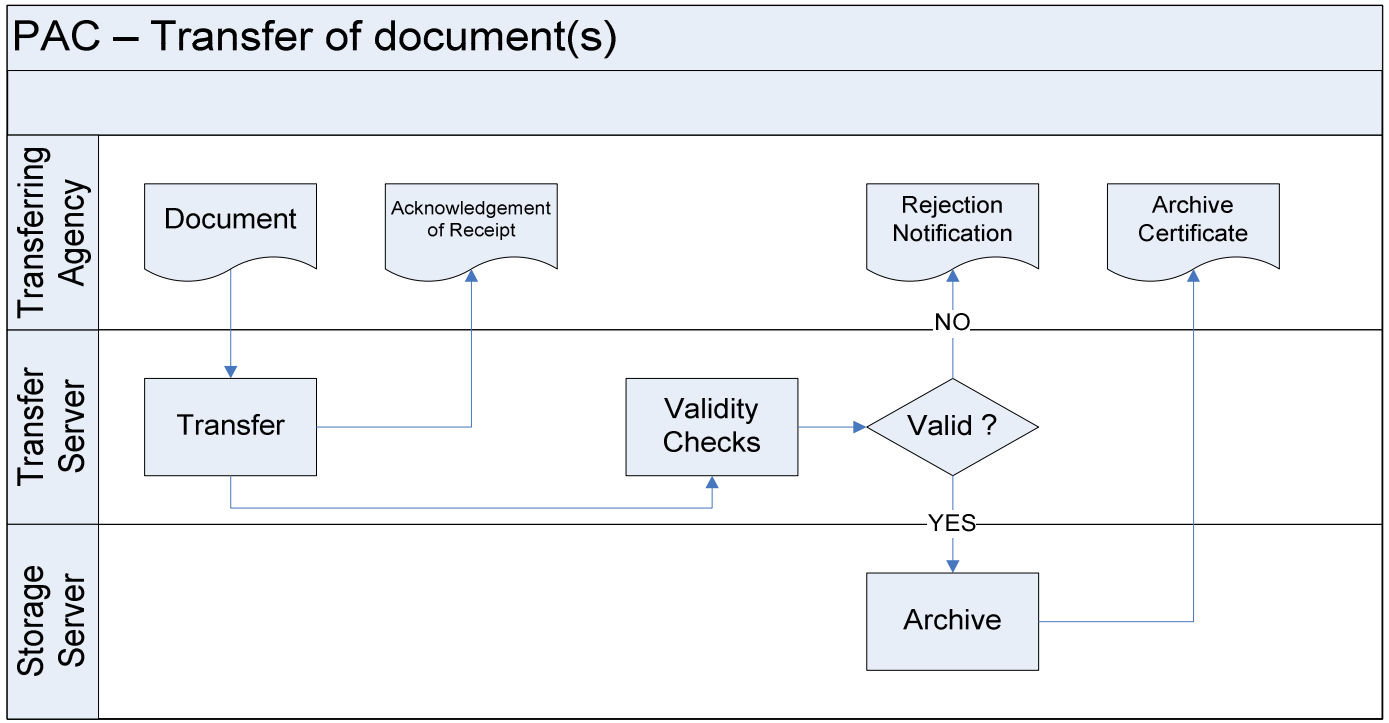
- A transfer server, where the archive producer can transfer his archives
  - Transfer of SIP (Submission Information Package)
  - Generation of acknowledgement receipt
  - Control of SIP – potential rejection
  - Creation of AIP (Archival Information Package)
- A storage server, where the archives are maintained
  - Multiple copy of AIP
  - Generation of archive certificate
  - Maintenance / migration operations
  - Reports
- An access server, where the producer and the authorized users can search, browse and retrieve the archives they need on line
  - Authentication of end-user
  - Communication of requested DIP (Dissemination Information Package)

BnF

INES


Logical architecture
Management
PAC : Plateforme d'archivage du CINES
Transfer Server
Storage Server
Access Server
Producer
Transferring Agency
User
Administration
BnF
INES

The « ingest » process



BnF

INES

The PAC project update

PAC v2.0 – extended storage capacity (40To)

- Based on domain standards
  - ISO 14721, SEDA, ISAD-G, ISAAR-CPF, DCMI metadata, ARK, SHA-256, etc.
- Limited set of file formats supported
  - Open / published, widely used or standard formats where possible
- Architecture based on SUN hardware, Arcsys software and open source libraries
  - Java, MySQL, Jhove, ImageMagick, DROID, ODF Validator, MPlayer
- Deployment in production May 2008
  - Following migration of documents archived on PAC v1.0

All the projects share the same infrastructure

- Pooling of projects on the preservation platform
- Generic “ingest” process
- Reduction of implementation and operation costs

BnF

INES

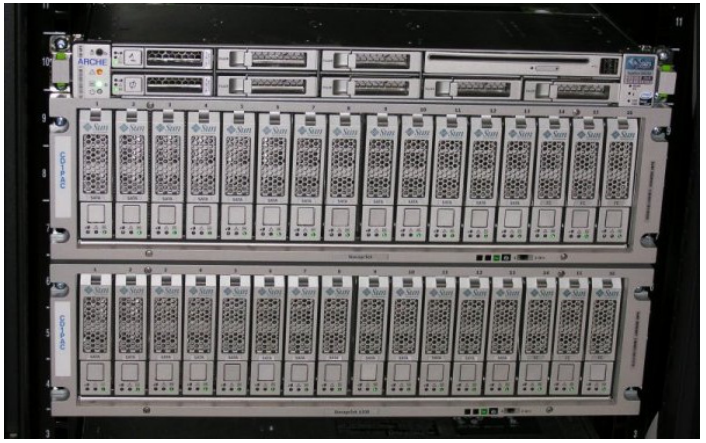
The PAC hardware

Application server

- SUN Fire X4150
  - Bi-processor Quad-core
  - 8Go RAM
  - Linux RedHat ES

Software

- Arcsys (Infotel)



Storage

- SUN Storagetek ST6140-4G
  - Application : 5 disks FC (146Go) – RAID 5 technology
  - Data : 50 disks SATA (1To) – RAID 5 technology

BnF

Sun microsystems

Infotel THE WEB-TO-DATABASE COMPANY

INES

The PAC hardware (continued)

Automated Cartridge System

- SUN-Storagetek 9310
- Capacity 6000 data cartridges (approx. 1,2Po)

Tape drives

- 11 tape drives SUN-Storagetek 9940
- Cartridge data capacity : 200Go (uncompressed), up to 800Go (compressed)
- Cartridges life cycle: 5-10 years in production

BnF

INES

Current projects

1. Two projects in production
   - Electronic PhD theses
   - Digitised Social Sciences & Humanities publications from the Persée program
2. Three projects in development
   - Audio documents produced as part of the exchange of linguistic data for speech research
   - Multimedia pedagogics / scholarly content from Canal-U production
   - Preservation of open archives HAL – Hyper Article en Ligne
3. Two projects in planning phase
   - Digitised Law & Economics Sciences documents from the CUJAS library
   - Digitised Medical Sciences documents from the BIUM library
4. One project in envisioning phase
   - Preservation of raw data produced by IMFT - Institute of Fluid Mechanics

BnF

INES

Perspectives – what's next ?

From a national perspective, the CINES is now one of the main actors of the digital preservation domain.

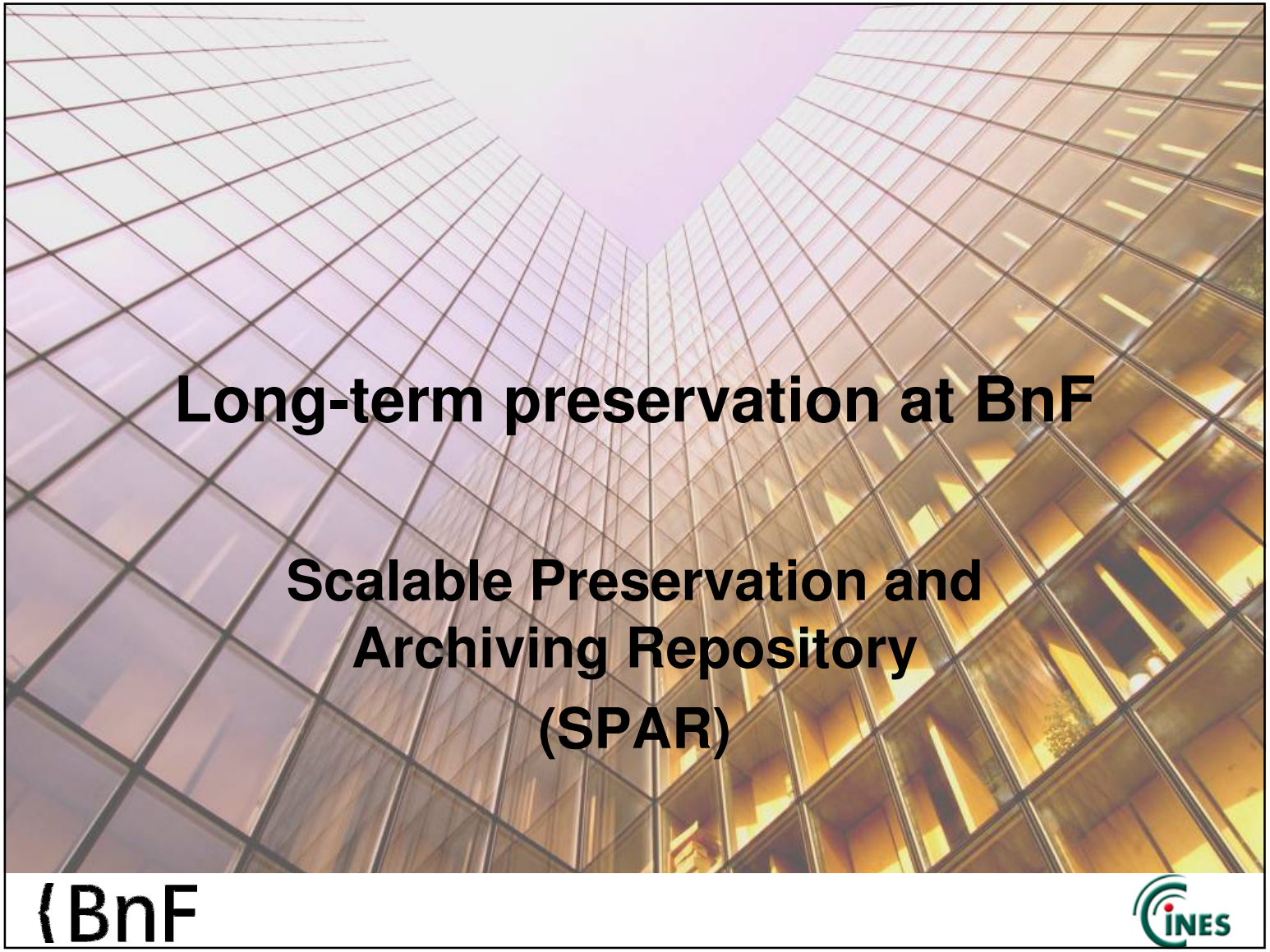
- National mandate for the preservation of electronic PhD these
- Expanded role in the national strategy for the preservation of the Education / Research digital heritage currently being put in place
- Involved a many national / international working groups or initiatives
  - France : PIN ; Europe : DPE, DSA, Alliance for Permanent Access, SHAMAN

Objectives 2009-2011 :

- Quality insurance and service improvement
  - Implement Representation Information library
  - Build mitigation plans as part of Risk Management Planning exercise
  - Document preservation processes
  - Data Seal of Approval (http://www.datasealofapproval.org/) accreditation in progress
  - Audit currently being run to identify strengths and weaknesses
- Certification of the department 2011

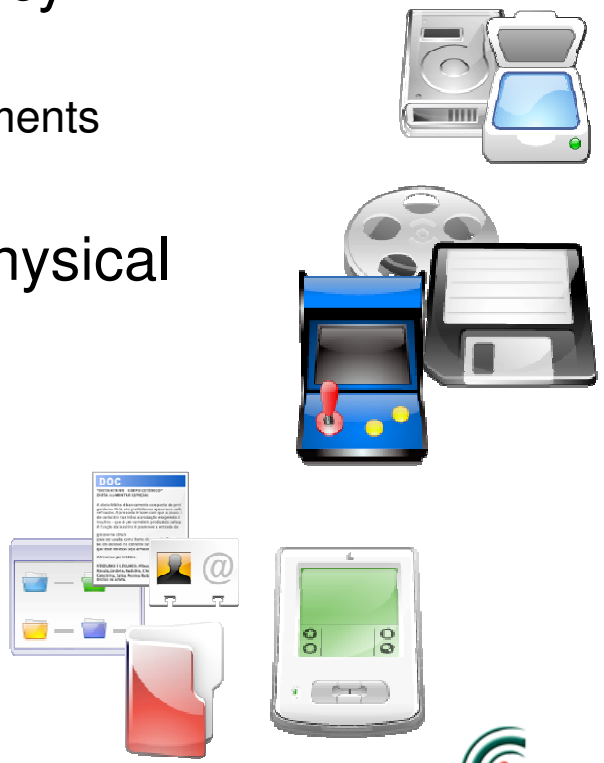
BnF

CINES


Long-term preservation at BnF
Scalable Preservation and Archiving Repository (SPAR)
BnF
INES

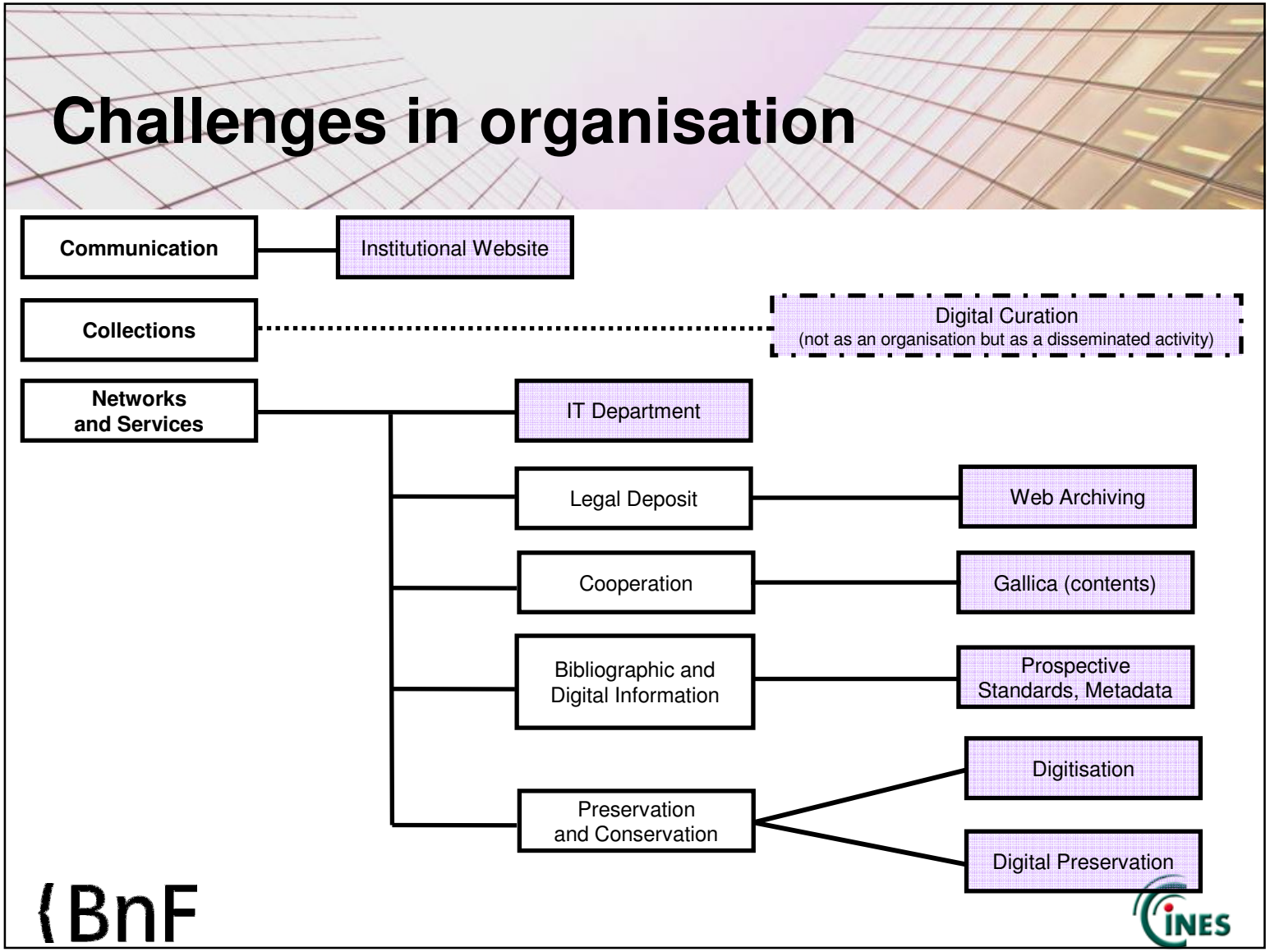
Challenges in scale

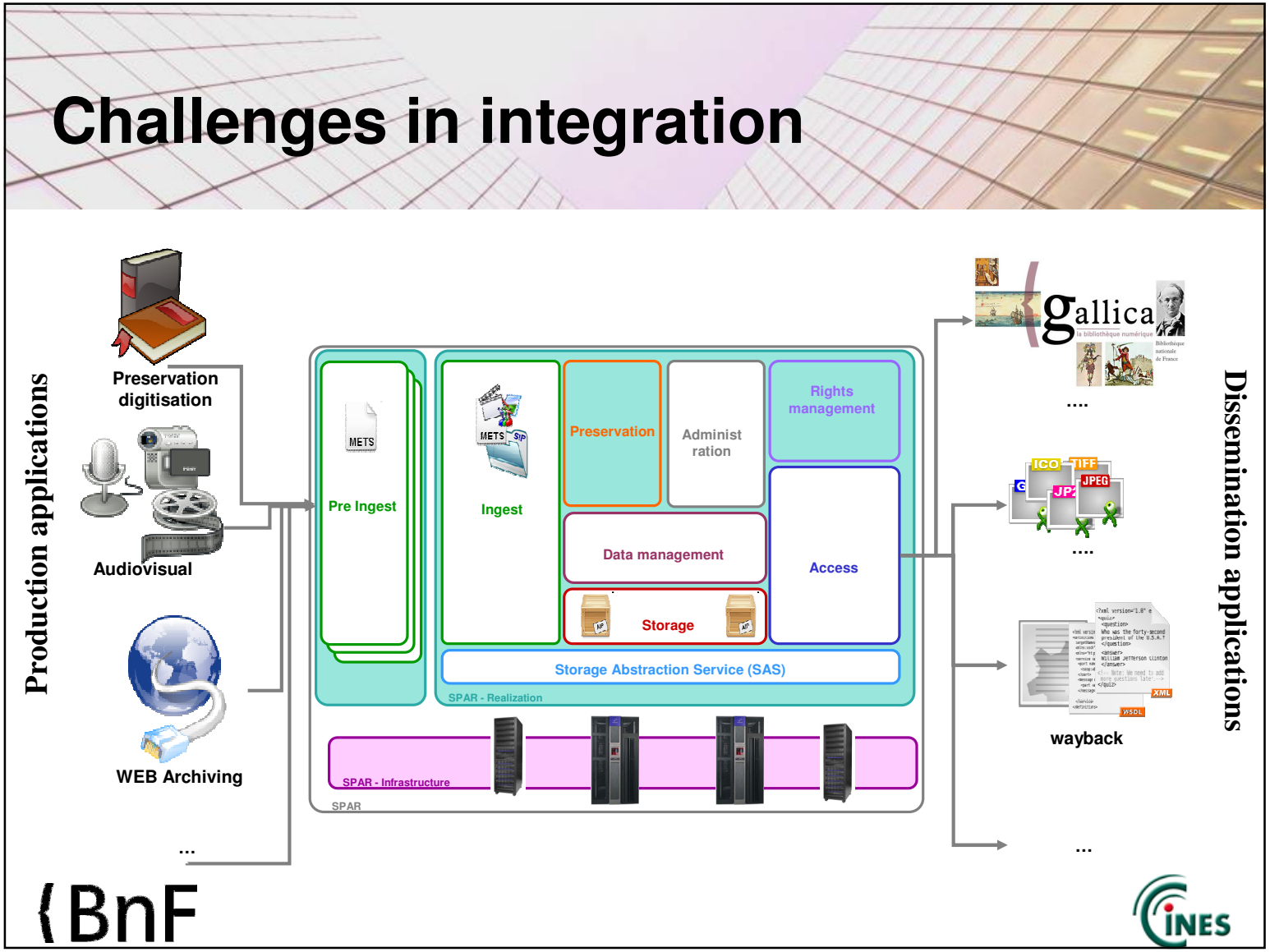
- Digitisation: from extra to policy
  - fast-growing digital library
  - rapid deterioration of original documents
  - decline of microfilm
- Digital as a substitution for physical
  - posters, regional press
- Born-digital documents
  - Web archiving
  - internal archives
  - e-journals? e-books? what next?



BnF

INES


Challenges in organisation
Communication
Institutional Website
Collections
Digital Curation
(not as an organisation but as a disseminated activity)
Networks and Services
IT Department
Legal Deposit
Web Archiving
Cooperation
Gallica (contents)
Bibliographic and Digital Information
Prospective Standards, Metadata
Digitisation
Preservation and Conservation
Digital Preservation
BnF
INES

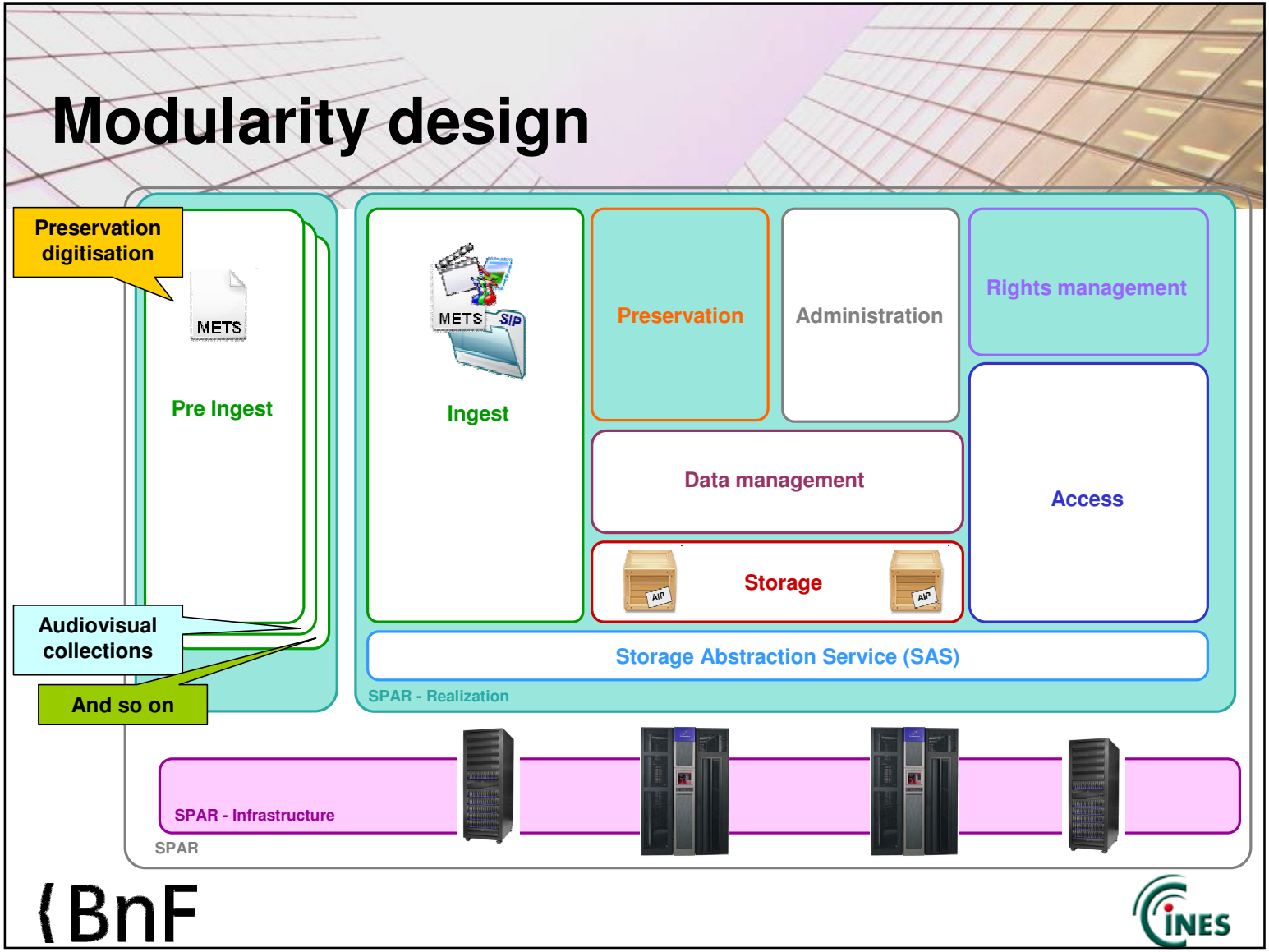

Challenges in integration
Production applications
Preservation digitisation
Audiovisual
WEB Archiving
...
Pre Ingest
METS
Ingest
Preservation
Administration
Rights management
Data management
Storage
Access
Storage Abstraction Service (SAS)
SPAR - Realization
SPAR - Infrastructure
SPAR
gallica
....
....
wayback
...
Dissemination applications
BnF
INES

Scalable Preservation and Archiving Repository (SPAR)

- Our goal: answering all of BnF's various digital preservation needs
- Key requirements:
  - OAIS compliance
  - modularity and distributivity
  - abstraction
  - use of well-known formats and standards
  - use of open-source technical building blocks

BnF

INES


Modularity design
Preservation digitisation
METS
Pre Ingest
Audiovisual collections
And so on
METS
SIP
Ingest
Preservation
Administration
Rights management
Data management
Access
Storage
AIP
AIP
Storage Abstraction Service (SAS)
SPAR - Realization
SPAR - Infrastructure
SPAR
BnF
INES

Decomposition in channels

- To deal with the variability and heterogeneity of the data, definition of *channels*
- Build on the relation between the digital objects and the archival system, independently of any given organisation:
  - Preservation digitisation
  - Audiovisual material
  - Negotiated legal deposit (deep Web, regional press)
  - Automatic legal deposit (surface Web)
  - Administrative production
  - Deposit / Third party archiving
  - Acquisition / Donation

BnF

INES

Model for describing a channel

- Each channel is formally defined by a reference package including:
  - a human readable description of the Service Level Agreement (SLA)
  - a machine actionable description of the SLA
    - links to accepted formats at various levels of commitment (stored, identified, known, managed)
    - maximum storage capacity
    - etc.

BnF

INES

Highlights from our information model: four categories of formats

00	Stored	No technical information Bit-stream preservation level
01	Identified	Format identified No preservation plan
10	Known	Format identified, documented, with tools and under monitoring by BnF experts
11	Managed	Format identified, documented, with tools BnF takes responsibility for this format

BnF

INES

Highlights from our information model: the seven deadly sections of METS

- Header
- DmdSec
- AmdSec
  - TechMD
  - DigiProvMD
  - SourceMD
  - RightsMD
- FileSec
- StructMap
- Structlink
- BehaviorSec

Descriptive and source metadata: qualified Dublin Core

Technical metadata: MIX for images, textMD for text

Provenance metadata: PREMIS

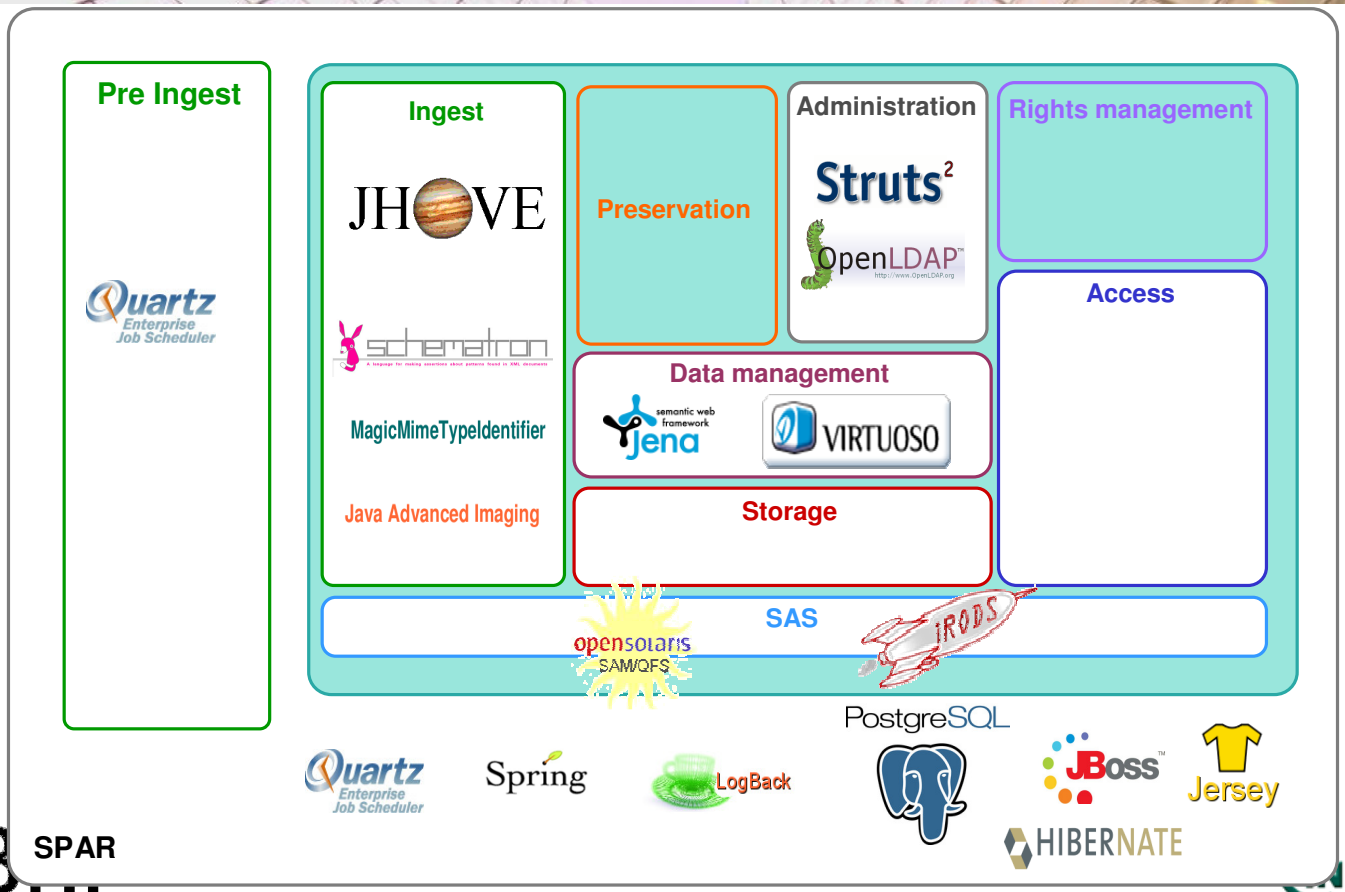
Structural metadata: METS

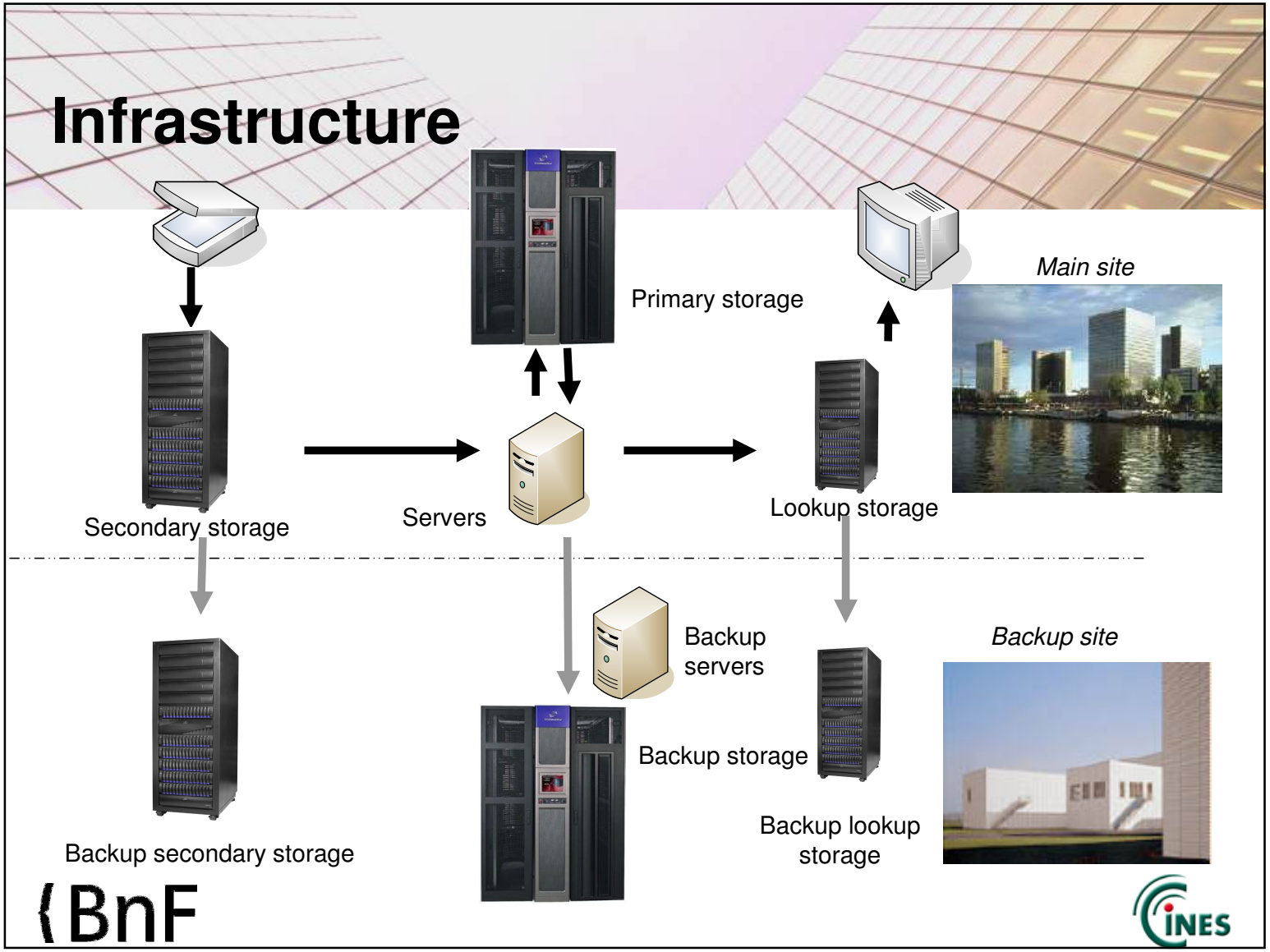
Not used in SPAR

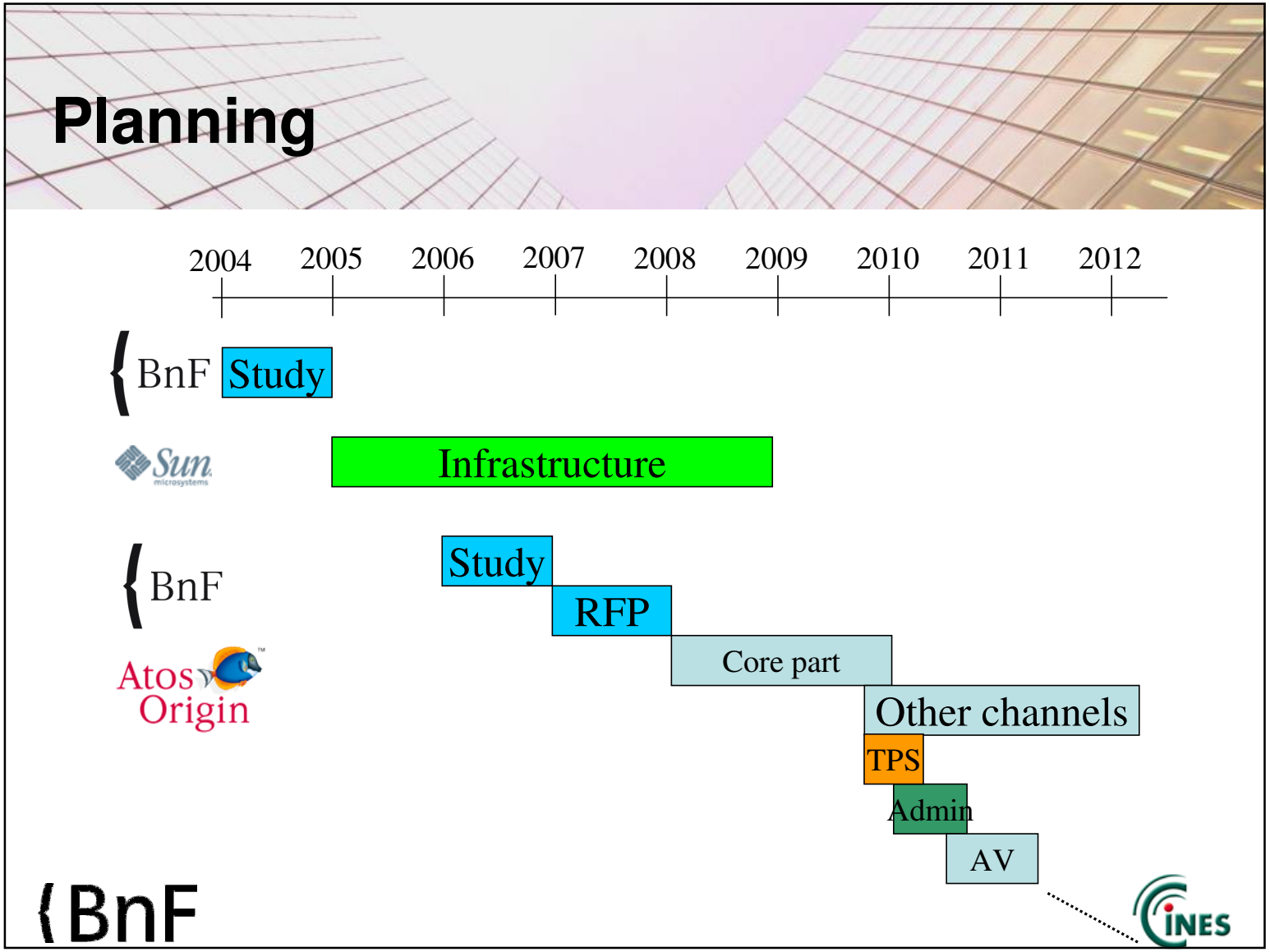
BnF

INES

Implementation: use of OpenSource frameworks




Infrastructure
Primary storage
Main site
Secondary storage
Servers
Lookup storage
Backup servers
Backup site
Backup storage
Backup lookup storage
Backup secondary storage
BnF
INES


Planning
2004
2005
2006
2007
2008
2009
2010
2011
2012
BnF
Study
Sun microsystems
Infrastructure
BnF
Study
RFP
Core part
Atos Origin
Other channels
TPS
Admin
AV
BnF
INES

Initial deployment

Ingest

Digital repository

Dissemination

gallica

Document digitisation

Ingest

In this first phase, the archive packages in SPAR won't be intensely disseminated → focus on ingesting and auditing

BnF

INES

Next step: new channels

- Introducing new channels means adjusting the model
- Each channel should integrate smoothly in the system and bring its own features
- Third party storage
  - accept any file, even non-recommended ones
  - bit-level preservation
  - deal with many producers and different ways of exchanging data

BnF

INES

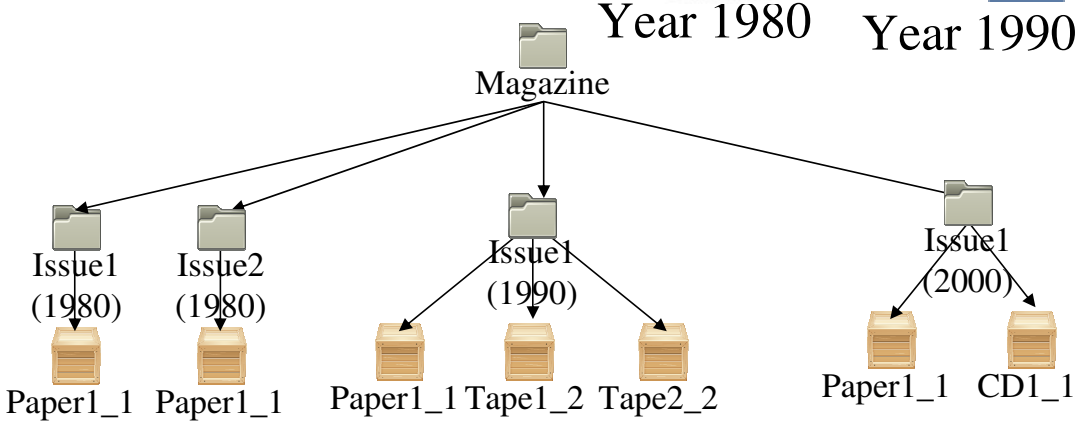
Challenges of new channels

- Audiovisual
  - complex hierarchical structure
    - Multimedia
    - multivolume

Year 1980

Year 1990

Year 2000



Each media produces its own digital artifact:
different workflow of digitisation

Challenges of new channels

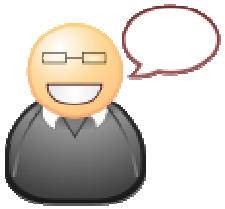
- Audiovisual
  - variety of formats

  - difficulty in the mandate
    - legal deposit: everything must be accepted
    - digitisation: production formats must be controlled
    - dissemination: need for known formats
  - introduction of 3 related channels
    - channel FIL_AUD_A: all legal deposit material
    - channel FIL_AUD_B: material controlled by the library (digitisation, extraction)
    - channel FIL_AUD_C: digital surrogate for dissemination (may need extensive work)
  - need for adequate validation tools
  - need for new technical metadata schemas

BnF

INES

What about preservation planning?

Digital curation	Administration of the repository	Preservation expertise
		
Curators + Producers of digital content	IT specialists + Digital stack attendants?	Preservation experts? + Metadata experts?

BnF

INES

Thank you for your attention

Any questions?

BnF olivier.rouchon@cines.fr louise.fauduet@bnf.fr CINES