Preserving French Scientific Data

Olivier Rouchon – Sun PASIG

Malta – June 25th 2009

Agenda

- Overview of the CINES
  - Location
  - Missions
  - Long-term preservation service
- Challenges
- The PAC platform
  - Logical architecture
  - Ingest process
  - Hardware
  - Project update
- Perspectives – what's next ?

Overview of the 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
- More information : http://www.cines.fr/

The mandate of the CINES 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

- May 2nd 2007 : Convention faisant suite à celle du 15 octobre 2003

➔ The CINES to archive SSH publications digitized as part of the Persée programme

- 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

The CINES long-term 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

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 standard formats • Logical migration (conversion)
Medias	• Supervision, management of ageing of medias • Physical migration
Software and hardware obsolescence	• Technological watching activities, anticipation

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 and ODF Toolkit libraries to

- Identify,
- Validate
- Characterize,

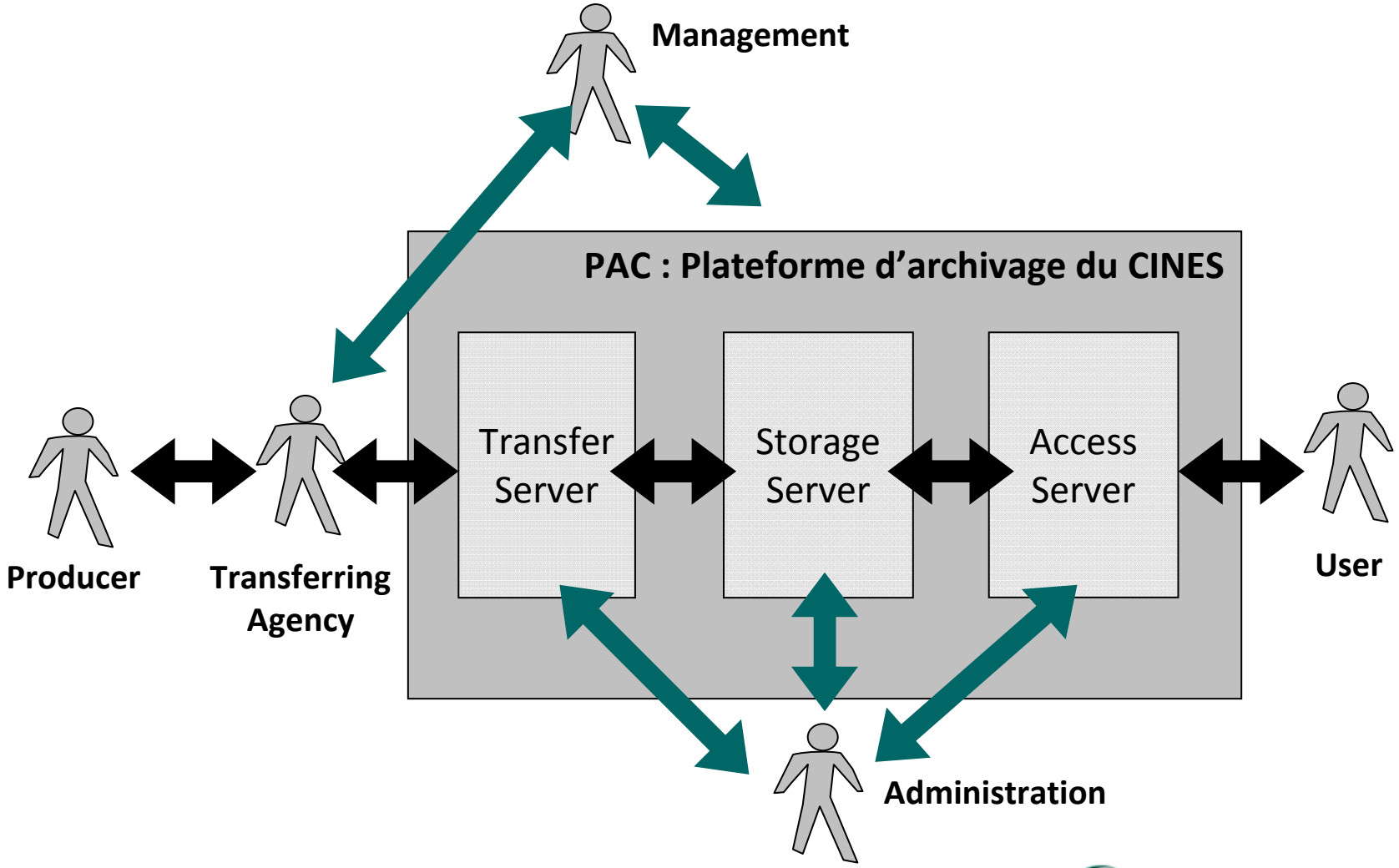
The format of transferred files

The PAC platform

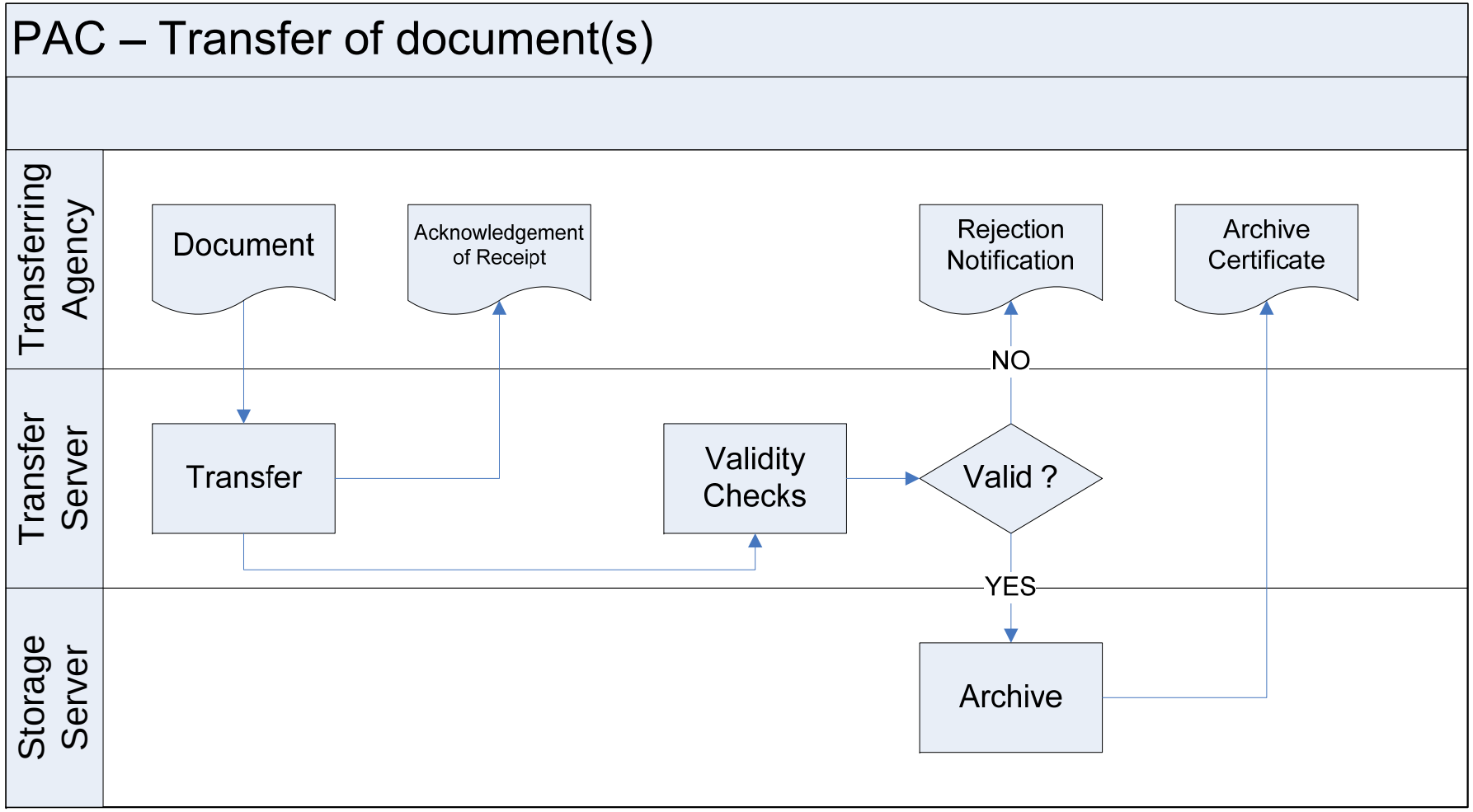
PAC is built of three logical servers, as defined in the OAIS model

- 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)

Logical architecture of the PAC platform



The document « ingest » process



The PAC project update (June 2009)

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

- Based on domain standards
  - ISO 14721, PSE, 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

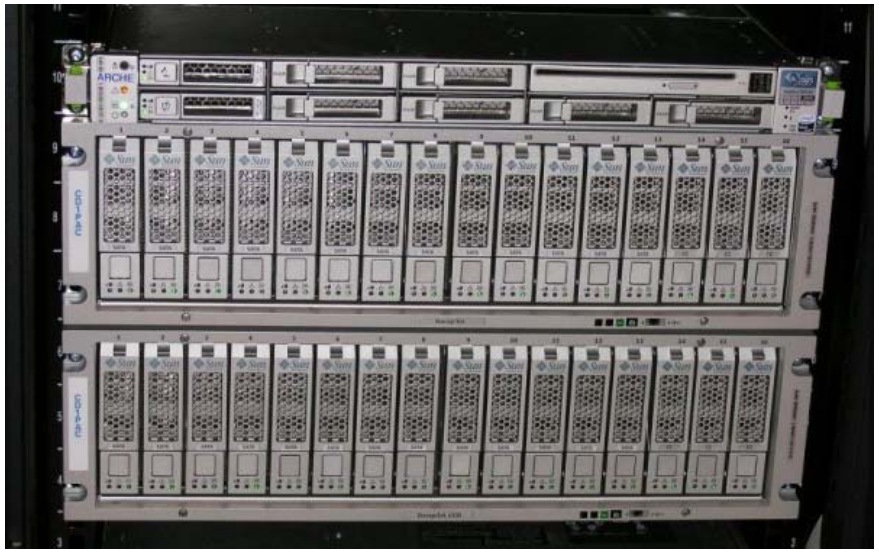
The PAC hardware

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

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

Current projects

1. Two projects in production
   - Electronic PhD theses
   - Digitized Social Sciences & Humanities publications from the Persée program
2. Two 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
3. Three projects in planning phase
   - Preservation of open archives HAL – Hyper Article en Ligne
   - Digitized Law & Economics Sciences documents from the CUJAS library
   - Digitized Medical Sciences documents from the BIUM library
4. One project in envisioning phase
   - Preservation of raw data produced by IMFT - Institute of Fluid Mechanics

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

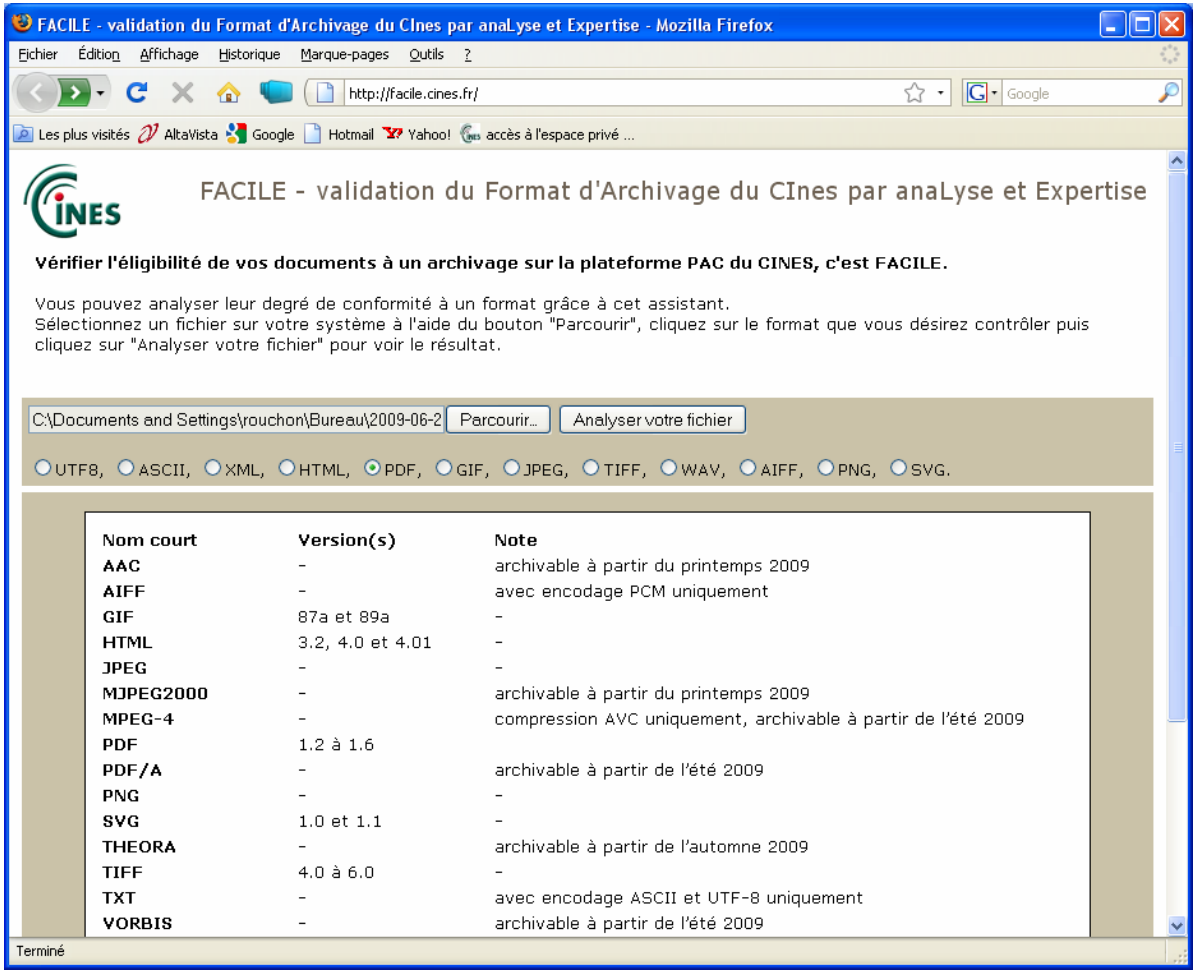
Objectives 2009-2010 :

- Quality insurance and service improvement
  - Publish source code of Ingest module
  - Implement Representation Information library
  - Build mitigation plans as part of Risk Management Planning exercise
  - Document preservation processes – based on Functional entities from OAIS model
  - Data Seal of Approval (http://www.datasealofapproval.org/) accreditation in progress
  - Audit currently being run to identify strengths and weaknesses
  - Certification of the department 2010

Questions & Answers ?

olivier.rouchon@cines.fr

FACILE web interface to file format validation tools



Nom court	Version(s)	Note
AAC	-	archivable à partir du printemps 2009
AIFF	-	avec encodage PCM uniquement
GIF	87a et 89a	-
HTML	3.2, 4.0 et 4.01	-
JPEG	-	-
MJPEG2000	-	archivable à partir du printemps 2009
MPEG-4	-	compression AVC uniquement, archivable à partir de l'été 2009
PDF	1.2 à 1.6	
PDF/A	-	archivable à partir de l'été 2009
PNG	-	-
SVG	1.0 et 1.1	-
THEORA	-	archivable à partir de l'automne 2009
TIFF	4.0 à 6.0	-
TXT	-	avec encodage ASCII et UTF-8 uniquement
VORBIS	-	archivable à partir de l'été 2009