Digital preservation in France

PILAE, a pilot digital preservation system

Françoise Banat-Berger and Olivier de Solan Direction des Archives de France

PIL@E

The context: e-Government

E-Government in France

• Ministry of finance / Directorate general for the modernisation of the State / Service for the development of e-Government

• A master plan to make processes digital in different areas of the Government

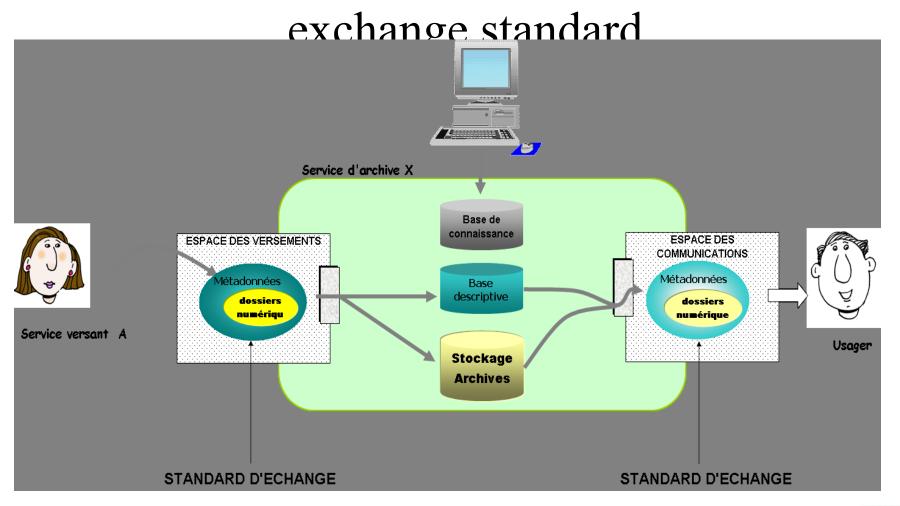
E-Government

- Law of 13 March 2000: new regime of legal evidence
 - Same legal value for paper and digital documents, if the author can be identified unambiguously and if the integrity of the document has been preserved
- Order of 8 December 2005
 - Application to public authorities
 - Legal definition of teleprocedures
 - Creation of the general interoperability framework and general security framework

Interoperability rules for archiving: security

- Survey on secure archiving
- Issued in July 2006 (http://www.ssi.gouv.fr/fr/confiance/archivage.html)

Interoperability rules for archiving:



Exchange standard

- Format and process to facilitate transfers between a creating agency and an archive
- It is the equivalent of a traditional transfer form: header, metadata, data embedded in an XML file
- When ingested by an archive, the metadata is incorporated into a descriptive database and the data is incorporated into a secure storage area

Exchange standard

- This standard is designed to be used by:
 - Creating agencies;
 - Business software vendors for exporting records;
 - Archives (public or private);
 - Storage services suppliers;
 - Archival software vendors for importing records.

Interoperability rules for archiving

Other rules about archiving are in the interoperability framework:

- About the lifecycle of records: to be integrated as soon as a new system is created
- File formats for preservation (images, structured or nonstructured documents, sound and video, etc.)
- Storage media

PIL@E Project

The context

PIL@E Project: the context

• Survey about the costs of a preservation system (Parker-Williborg), 2005.

• Rules defined in the general interoperability framework

PIL@E Project: enjeux

- New building for the National Archives in Pierrefitte-sur-Seine
 - The pilot system will be used in Fontainebleau (service Constance) 2008-2011

• Development of e-Government

PIL@E Project

- Relies on OAIS model
 - Functional description of the pilot system
 - Structure of data (SIP / AIP / DIP)
- Relies on description standards (ISAD-G)
- Relies on interoperability rules (security policy, exchange standard, file formats)

PIL@E

Functionalities

General view

• Objective:

- Receive, validate, store, disseminate born digital data and documents from the French Government
- Users are:
 - The archivists in the ministries, who send data
 - The National Archives, who receive, control, validate and store

• Not directly open:

- To the creators themselves
- To the public

Concerned data

- Be able to process several types of digital records
 - Data from databases
 - Office documents
 - Documents from EDMS
 - E-mails
 - Data exchnaged through teleprocedures...

PIL@E Project: organisational constraints

- Support by the director of French archives and the director of the National Archives, but...
- Little team (DAF/DGME) and IT department with limited resources
- Constance team with high experience in processing databases but discovering other types of digital documents
- Archivistes in ministries discovering digital preservation issues

PIL@E Project

We have chosen to:

- Outsource the development of the system
- Build a system with limited size but scalable in coming years
- Re-use existing tools
 - A secure storage software ("Coffre-fort électronique communicant" by Cecurity.com): hashcode, digital signature, timestamp, audit trail
 - Storage infrastructure by IBM

General View

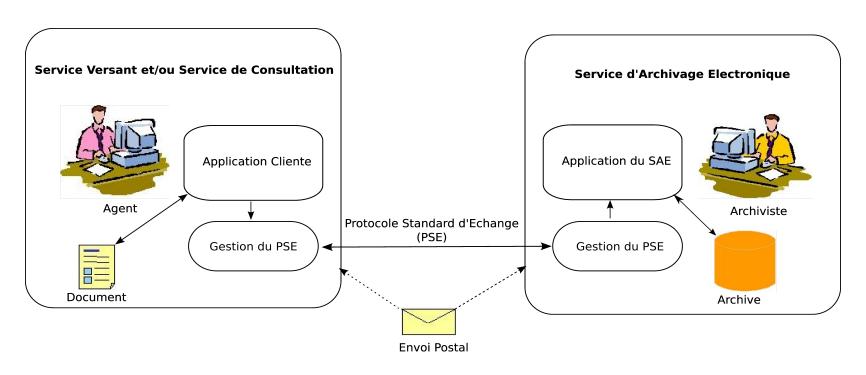
- Specific development to support:
 - The exchange standard
 - Archival functionalities (pour la prise en charge, la recherche, la consultation, les éliminations, la production de statistiques métier)
 - Limited retrieval features (information system of the National Archives to come)
 - Workflow between creating agencies and archives
 - Tools to identify and convert file formats

Some constraints

- Few business systems in creating agencies are able to deliver SIP compliant with the exchange standard.
- Creation of an interface to transfer records that have not been exported automatically

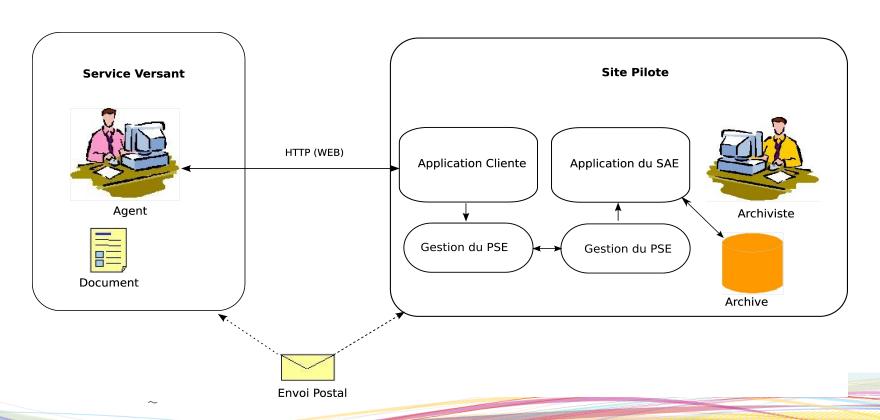
Software Architecture

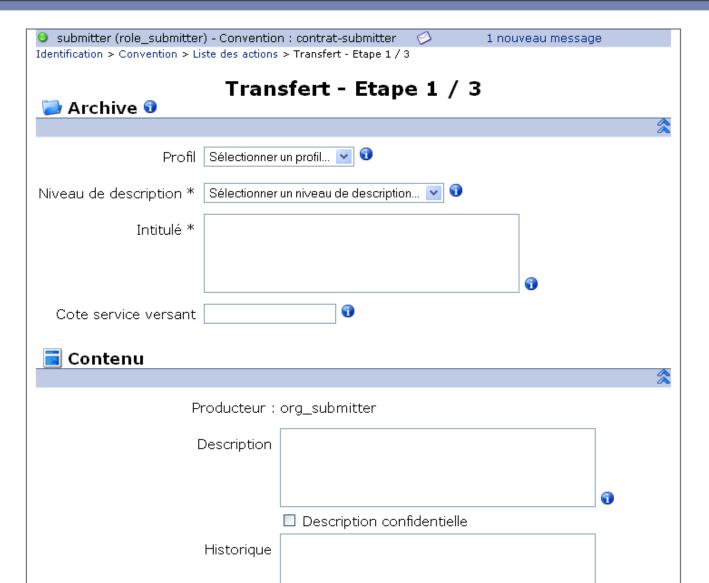
Target System

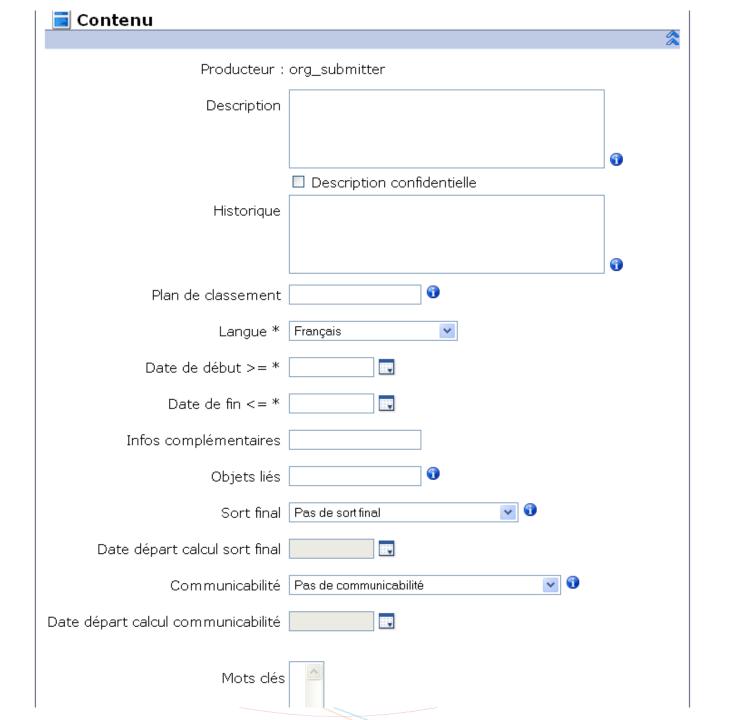


Software Architecture

Pilot System



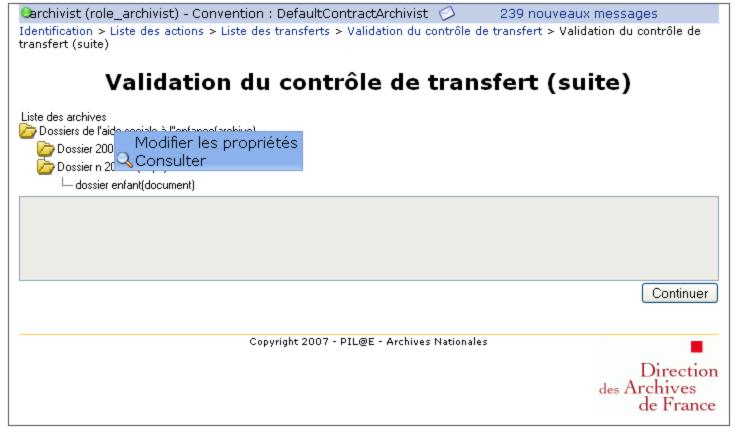


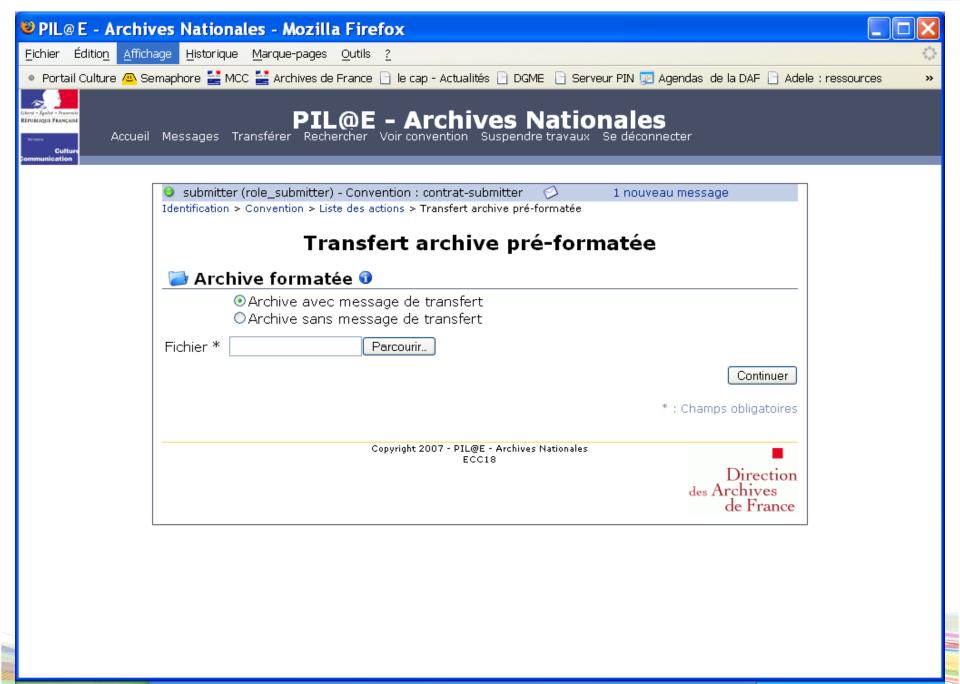


PIL@E - Archives Nationales Accueil Messages Gérer Rechercher Conventions Voir Suspendre travaux Se déconnecter

cgr (role_archivist_accept) - Convention : DefaultContractArchivist 🕗 9 nouveaux messages Identification > Liste des actions > Liste des transferts > Validation du contrôle de transfert > Validation du contrôle de transfert (suite) Validation du contrôle de transfert (suite) Liste des archives 🗁 Contrôle de légalité : délibération de la commune de Conseil général des Yvelines (archive) Pièces jointes(document) 눩 Accusé de réception de transmission d'un acte(objet) Accusé réception de transmission d'un acte (document) trace d_intégrite du fichier homonyme sans extension (document) 📂 Acte soumis au contrôle de légalité(objet). trace d_intégrite du fichier homonyme sans extension (document) trace d_intégrite du fichier homonyme sans extension (document). trace d'intégrite du fichier homonyme sans extension (document) Acte(document). - annexe 1(document) signature(document) Continuer Copyright 2007 - PIL@E - Archives Nationales Direction des Archives de France













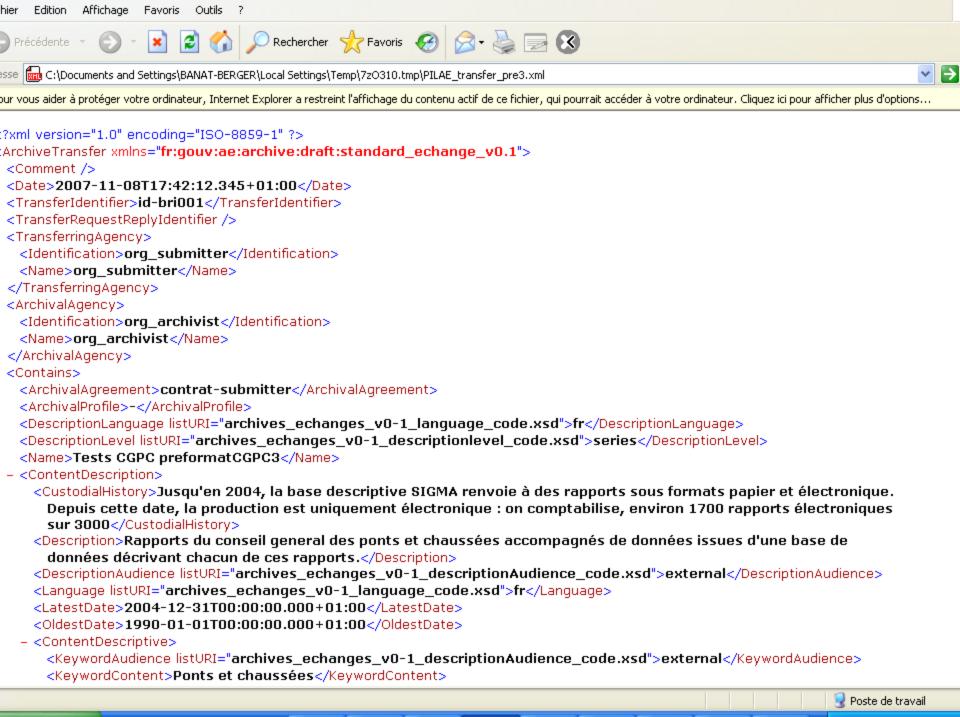


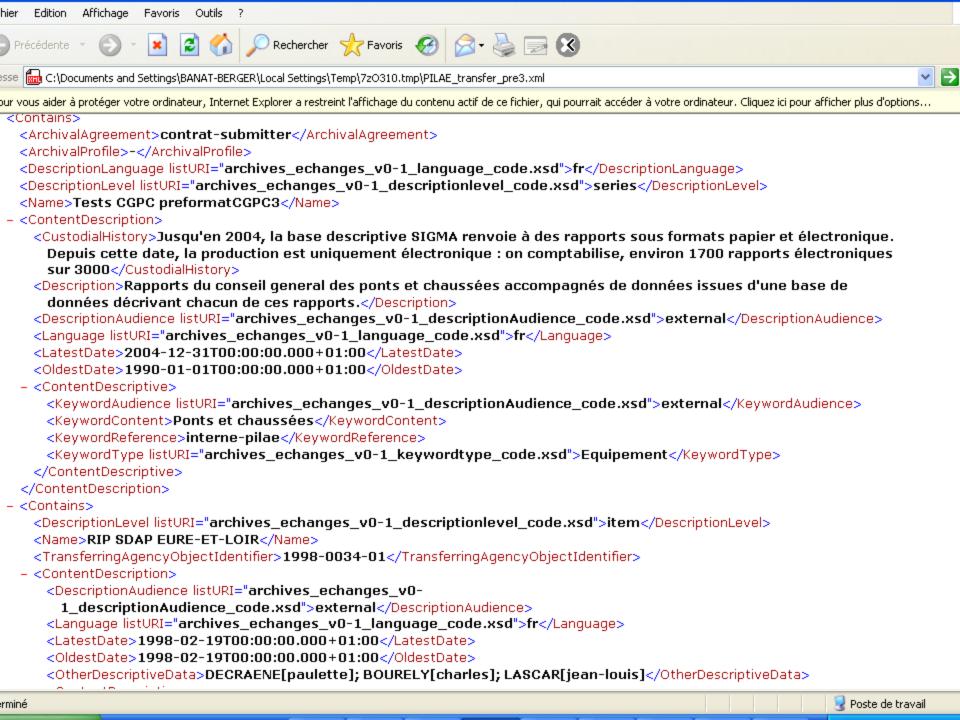












File formats: objectives

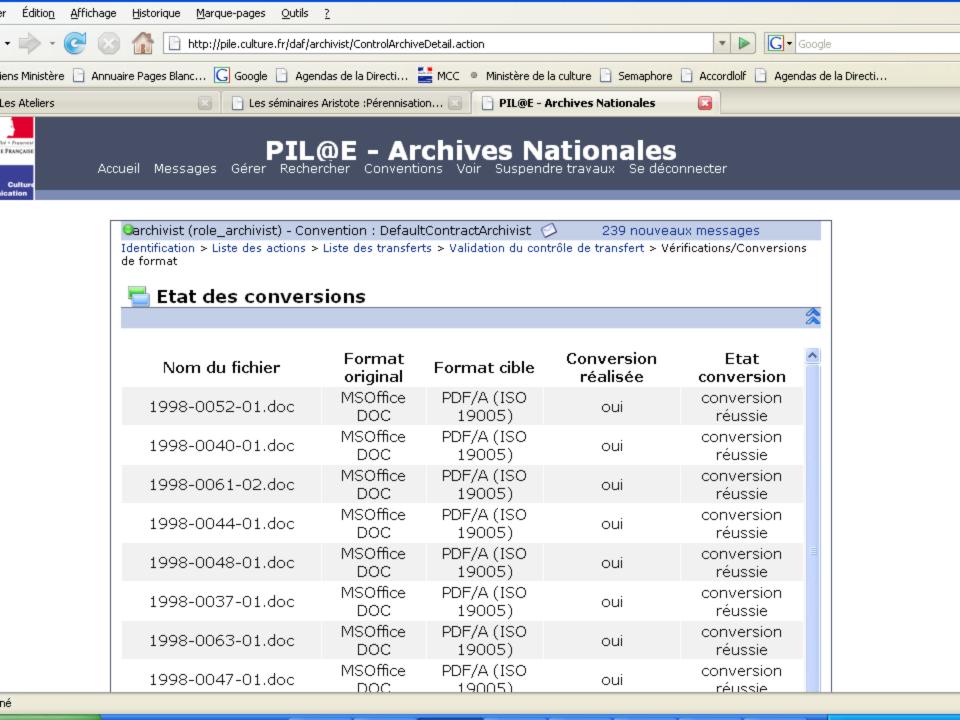
- Files are tested and converted as soon as they are ingested in the system
 - List of file formats for long-term preservation
 - List of file formats accepted for submission
 - Some conversion tools
 - Conversion rules
- Some choices
 - MS Office formats have not been selected as long-term preservation formats
 - Initial version, N version, N-1 version are kept together

List of long-term preservation file formats

- General Interoperability Framework
- Widely spread format
 - European or international standard
- Public and easily accessible specifications
- Format relatively stable
 - New version every 3 years or more
- Software:
 - Two vendor software products available on the French or European market
 - Or one open source software

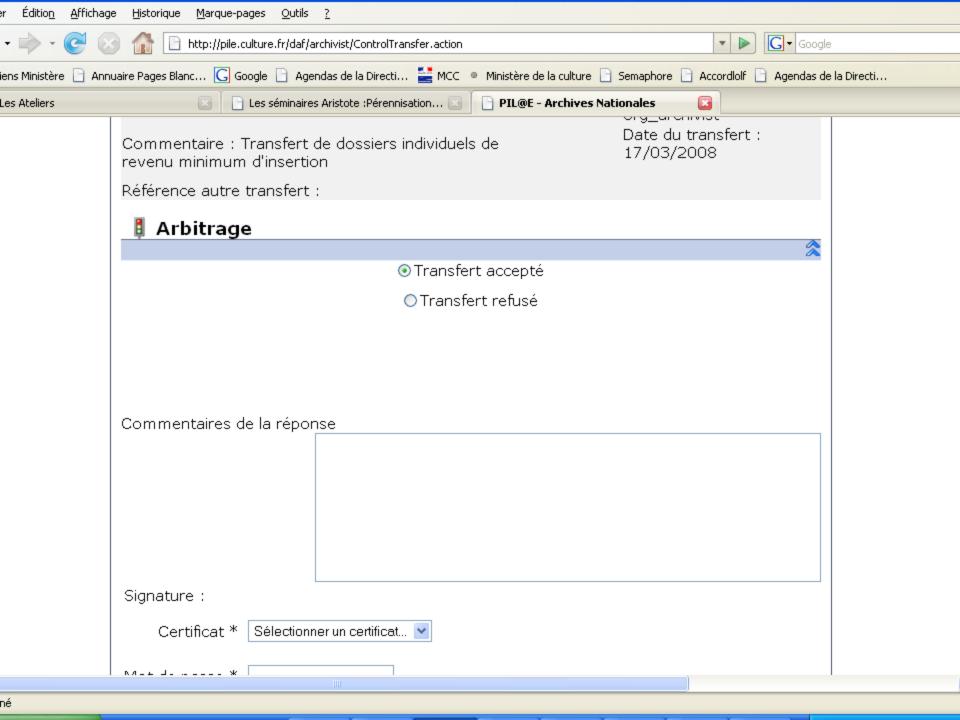
Type de données	Format accepté en entr	ée Version(s) supportée(s) Format cible d'archiva
Images fixes	PNG	1.0	PNG
	GIF	GIF89a	PNG
	JPEG 2000 (jp2)	ISO/IEC 15444:2000	JPEG 2000
	JPEG (jfif, jpg)	ISO 10918	JPEG
	TIFF	6.X	TIFF
	BMP	N/A	JPEG
	PCX	N/A	JPEG
Son et Vidéo	MPEG-2 Layer III (MP3)		MPEG-2 Layer III (MP3)
	WAV	i	WAV
	MPEG-2	ISO/CEI 13818	MPEG-2
	MPEG-4	ISO/CEI 14496	MPEG-4
	Ogg-Vorbis		Ogg-Vorbis
Documents structurées			
et non structurés	HTML	2/3.2/4.0/4.1	XHTML
	XHTML	1.0/1.1	XHTML
	OpenDocument	ISO 26300:2006	OpenDocument
	MSoffice DOC	97/2000/XP	OpenDocument
	Msoffice XLS	97/2000/XP	OpenDocument
	Msoffice PPT	97/2000/XP	OpenDocument
	Msoffice DOC	97/2000/XP	PDF/A (ISO 19005)
	Msoffice XLS	97/2000/XP	PDF/A (ISO 19005)
	Msoffice PPT	97/2000/XP	PDF/A (ISO 19005)
	PDF/A	PDF 1.4	PDF/A (ISO 19005)
	PDF	1.3	PDF
	RTF	1.5	PDF/A (ISO 19005)
	CSV	N/A	CSV
	TXT	N/A	TXT
Formats composites	ZIP	10.X	Format Original
	WinRAR	3.X	Format Original
	Sit		Format Original
	E-mail	RFC 2822 + RFC 2045	E-mail + pièces jointes
Documents structurés	XML	1.0	XML
	XSD	1.0	XSD
	XML validé par XSD	XML 1.0	XML
	SGBD	N/A	TXT
	Open DWG		Open DWG
Données graphiques	SVG	1.1	SVG
	DWG	R2007/R2004/R2000/R1	
	DXF	v.u.21.1.01/v.u.20.1.01/v	v.uSVG
	CGM	ISO 8632:1992	CGM
	STEP	ISO/10303	STEP

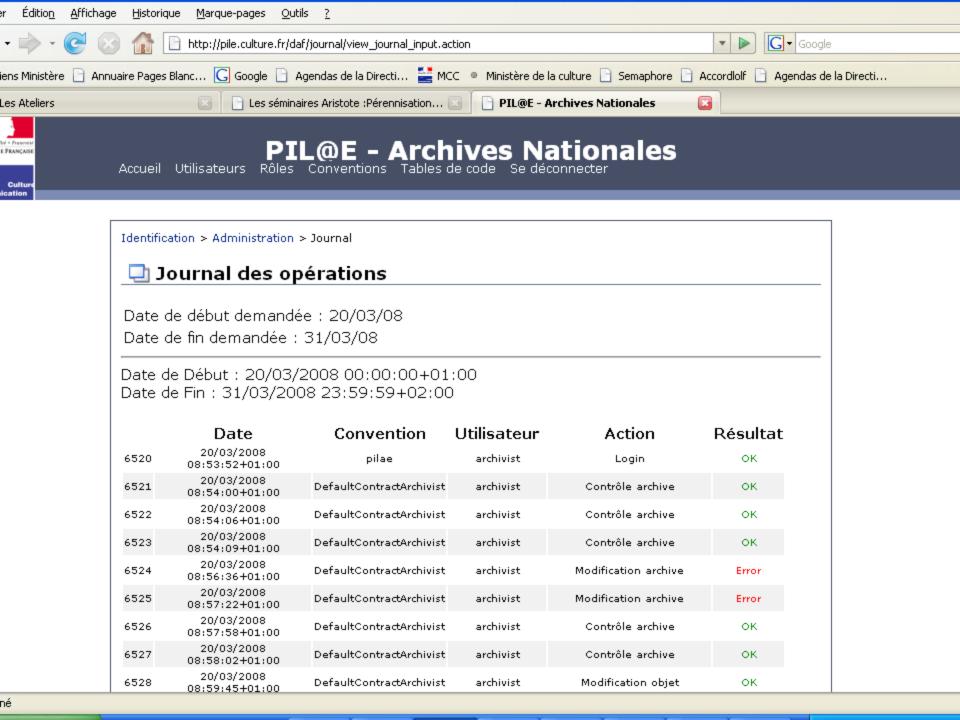




Data Security

- Integrity check at the ingestion time and then regularly triggered
- Timestamp features
- Copy of the data on a remote site (4 copies on 3 servers)
- Digital signature by the archivist who takes or rejects the custody of the material to be kept
- Traceability provided by the event log





- PIL@E will allow the National Archives to:
- Test some digital preservation features and to specify new features to come (unique identifier mechanism...)
- Define more precisely the access to the system by the public, the creating agencies...
- Decide whether data should be encrypted within the system
- Define more precisely how digitally signed documents will be archived
- Create a strong link with the general information system of the National Archives

38

- Three conditions
 - Actual involvement of the team in the National Archives
 - Actual involvement of the archivists in the ministries
 - Implementation of the exchange standard in the business systems

- Beyond the National Archives, PIL@E model may be used in other environments
 - For archival systems within creating agencies (semi-current records)
 - For archival systems in local authorities

- Other related projects
 - Archives départementales des Yvelines
 - Archives municipales de Limoges
 - Ministry of Finance
 - Archives départementales de l'Aube

- Main objective is to foster the creation of digital preservation systems
- Main activities in next months
 - Participate to the certification of services and software
 - ISO standard project ongoing
 - MoReq2 for records management software
 - Interesting private sector initiative (Coref)