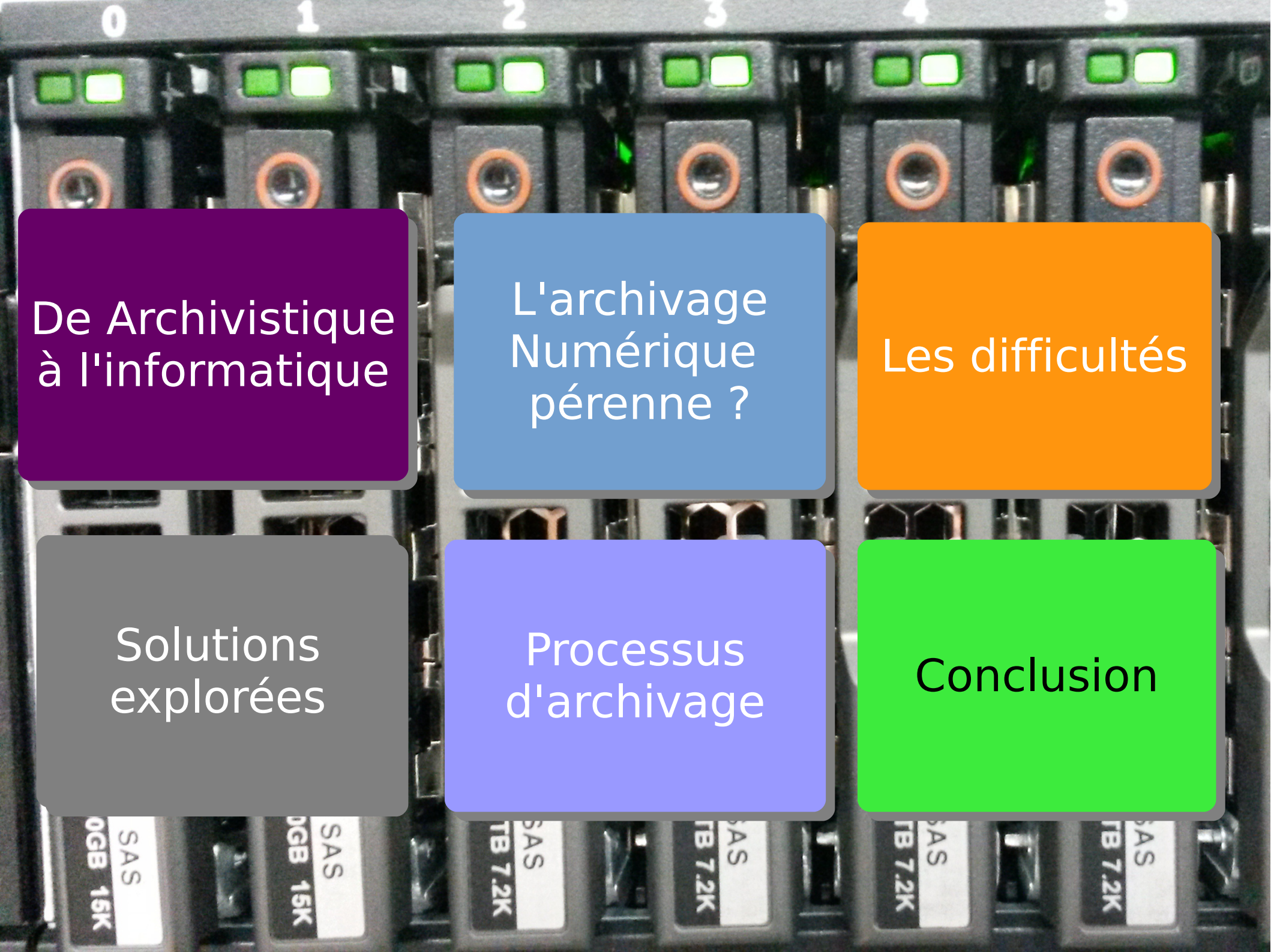


Pourquoi est-il si difficile d'archiver ?

Philippe Saby
Capitoul 26 février 2015



De Archivistique
à l'informatique

L'archivage
Numérique
pérenne ?

Les difficultés

Solutions
explorées

Processus
d'archivage

Conclusion

De l'archivistique à l'informatique

de Archivistique
à l'informatique



l'archivage
Numérique
pérenne

Les
difficultés

Solutions
explorées

Processus
d'archivage

Conclusion



UNIVERSITÉ
TOULOUSE III
PAUL SABATIER
Université
de Toulouse



Observatoire
Midi-Pyrénées
OMP

En Archivistique

Archives ?

- L'ensemble des documents, quels que soient leur date , leur lieu de conservation, leur forme et leur support, produits ou reçus par toute personne physique ou morale et par tout service ou organisme public ou privé dans l'exercice de leur activité (Code du Patrimoine, art. L. 211-1),
- les services et institutions qui se chargent de leur gestion,
- les espaces de stockage de ces documents.

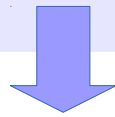
La notion d'archives est indépendante :

- de la date :
 - tout document est considéré comme archive dès son élaboration ou sa réception,
- de la forme :
 - document manuscrit, texte imprimé ou dactylographié, plan, affiche etc.
- du support :
 - papier, photographie, bande audiovisuelle, document numérique etc.

En Archivistique

Cycle de vie des documents ?

- **Dossiers actifs** (aussi appelés archives courantes ou archives administratives) qui regroupent les documents qui sont nécessaires à l'activité courantes des organismes qui les ont produits (créés ou reçus dans le cadre de leurs activités) ;
- **Dossiers clôturés** (archives intermédiaires) à conserver de façon intègre tant qu'ils ont une valeur probante (utilité administrative ou légale) pour la traçabilité des activités de l'organisme producteur ;
- **Dossiers inactifs** qui n'ont plus de valeur probante
 - Tri : éliminés ou archivés



archivage numérique pérenne

30 ans

De l'archivistique à ... l'informatique

OAIS

Open Archival Information System

- Modèle conceptuel destiné à la gestion, à l'archivage et à la préservation à long terme de documents numériques
- Entités fonctionnelles du modèle
 - les entrées : reçoivent les paquets d'information à verser et les transmettent au stockage,
 - le stockage : stocke et sauvegarde les paquets d'information archivés,
 - la gestion des données : met à disposition toutes les informations utiles au fonctionnement de l'archive,
 - l'administration : pilote le système,
 - la planification de la préservation : assure une veille technologique et propose les évolutions et les stratégies pour prévenir l'obsolescence,
 - l'accès : communique les paquets d'information diffusés aux utilisateurs.

De l'archivistique à ... l'informatique

OAIS

Open Archival Information System

- Modèle conceptuel destiné à la gestion, à l'archivage et à la préservation à long terme de documents numériques

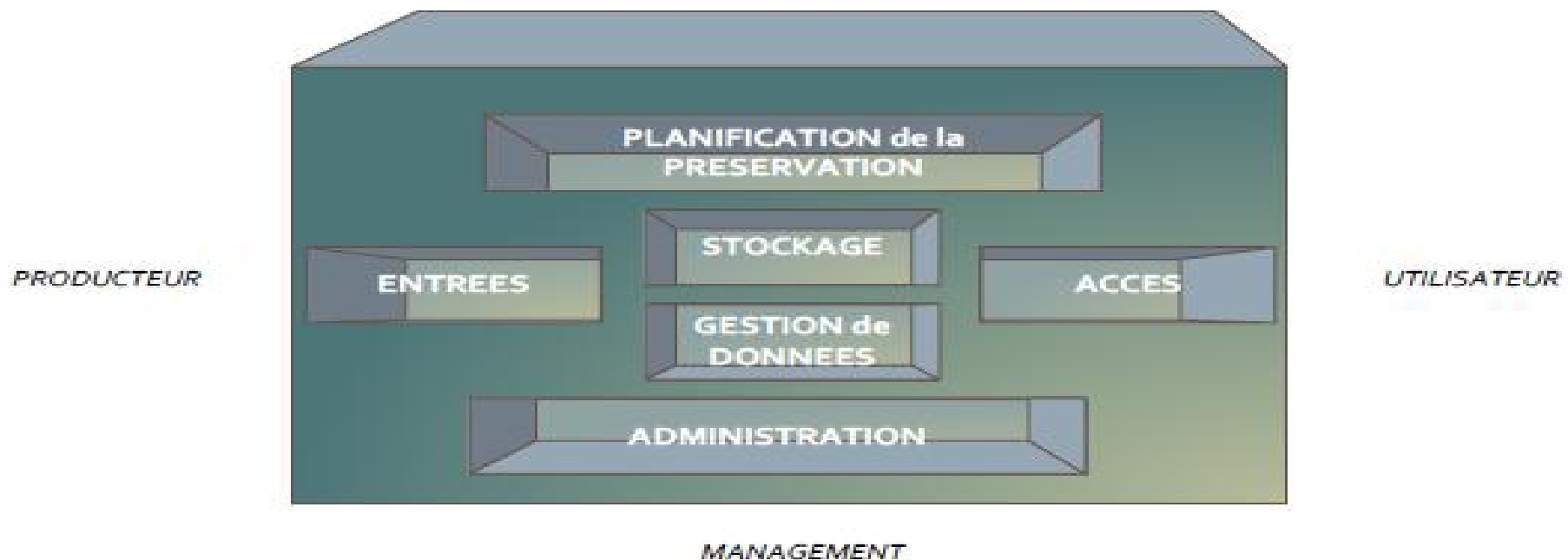


Figure 2 : Les 6 fonctions du modèle OAIS

L'archivage numérique pérenne ?

de Archivistique
à l'informatique

l'archivage
Numérique
pérenne



Les
difficultés

Solutions
explorées

Processus
d'archivage

Conclusion



UNIVERSITÉ
TOULOUSE III
PAUL SABATIER
Université
de Toulouse



Observatoire
Midi-Pyrénées
OMP

Stockage / Sauvegarde / Archivage

Stockage

- Données de travail
- Local (poste), lan (NFS, CIFS ...), WAN (cloud, objet ...)
- 1 copie



Sauvegarde Restauration

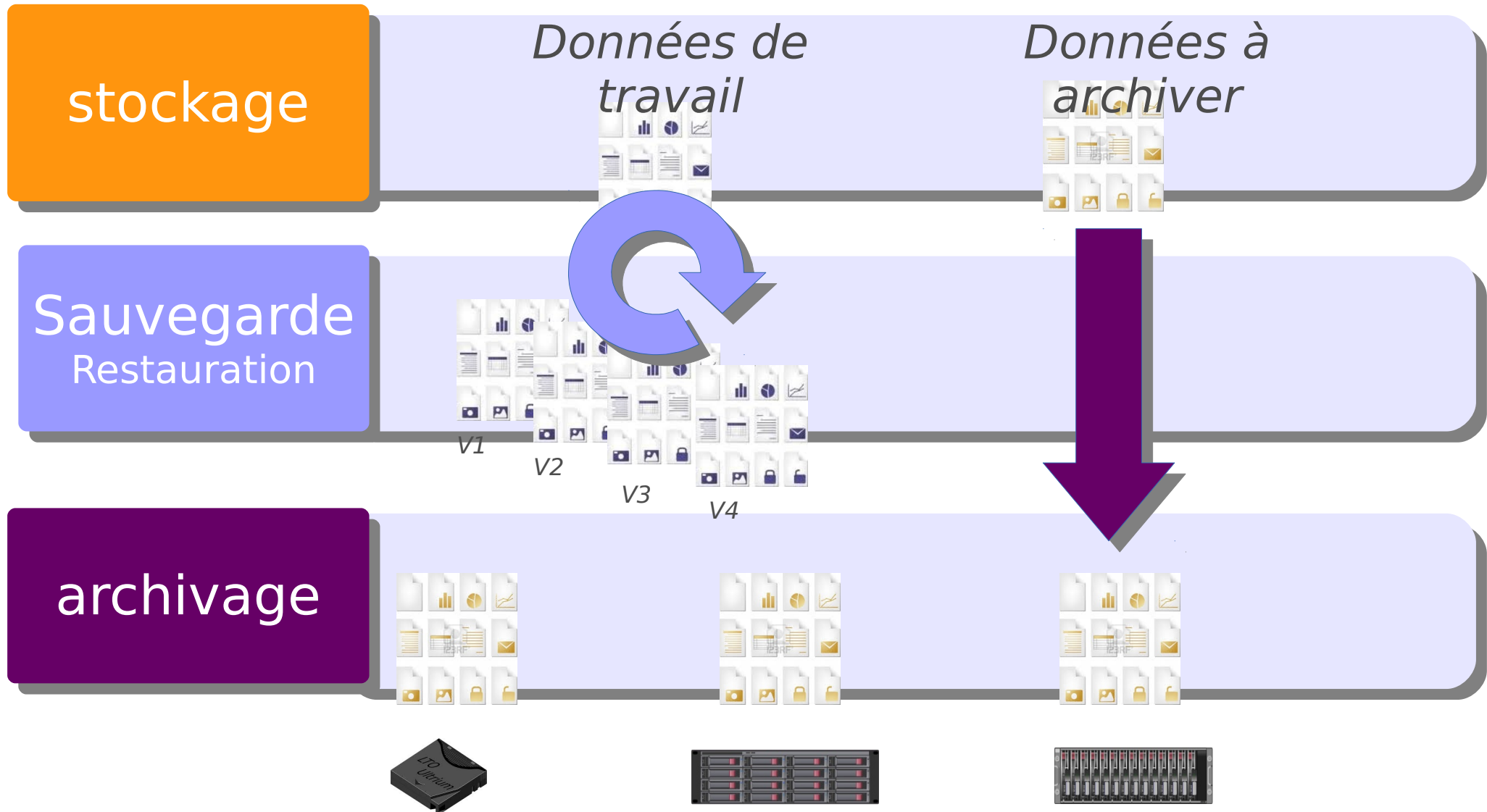
- Conserver plusieurs versions d'un fichier du stockage
- Processus automatique (système)
- Restauration sécurisée et autonome (utilisateur et/ou admin)



Archivage

- Mettre en sécurité (plusieurs copies) d'un fichier d'archive
- Processus manuel (utilisateur)
- Restitution sécurisée et autonome (utilisateur et/ou admin)

Stockage / Sauvegarde / Archivage



L'archivage numérique pérenne

a pour but

- Conserver le document
- Le rendre accessible
- En préserver l'intelligibilité

30 ans

ce n'est pas

- Une sauvegarde
- Un service de HSM
- La dernière étape avant l'oubli ou la perte définitive des données



Pourquoi est-il si difficile d'archiver ?

de Archivistique
à l'informatique

l'archivage
Numérique
pérenne

Les
difficultés

Solutions
explorées

Processus
d'archivage

Conclusion



UNIVERSITÉ
TOULOUSE III
PAUL SABATIER
Université
de Toulouse



Les difficultés

- l'obsolescence / panne matérielle,
 - Supports : Bandes , disques
 - Lecteurs : bandes (version LTO)
- l'obsolescence logicielle,
 - Changement version, arrêt maintenance
 - disparition société donc logiciel
- l'obsolescence du format de fichier,
 - Format propriétaire, local
 - Dépendance système disparu
- la perte de la signification du contenu,
 - Fichier seul, nom approximatif
- l'humain
 - J'ai pas l'temps de trier ...
 - Ah non j'peux pas effacer !

les solutions

- Plusieurs copies, plusieurs sites
- Rafrachir supports et migration
- Veille technologique, choix standards
- Virtualisation, émulation
- Formats « durables » et normes
- conversion
- Décrire -> Métadonnées
- Information, formation, motivation
- Pas de solution ...

Les difficultés

- l'obsolescence / panne matérielle,

- Supports : Bandes , disques
- Lecteurs : bandes (version LTO)

les solutions

- Plusieurs copies, plusieurs sites
- Rafrachir supports et migration

Sur quels critères puis-je agir ?

choix standards
on

» et normes

• Fichier seul, nom approximatif

• Décrire -> Métadonnées

- l'humain

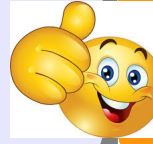
- J'ai pas l'temps de trier ...
- Ah non j'peux pas effacer !

- Information, formation, motivation
- Pas de solution ...

Les difficultés

- l'obsolescence / panne matérielle,

- Supports : Bandes , disques
- Lecteurs : bandes (version LTO)



- l'obsolescence logicielle,

- Changement version, arrêt maintenance
- disparition société donc logiciel



- l'obsolescence du format de fichier,

- Format propriétaire, local
- Dépendance système disparu



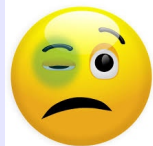
- la perte de la signification du contenu,

- Fichier seul, nom approximatif



- l'humain

- J'ai pas l'temps de trier ...
- Ah non j'peux pas effacer !



les solutions

- Plusieurs copies, plusieurs sites
- Rafraichir supports et migration

Progiciels ou outils standards

Veille technologique, choix standards
Virtualisation, émulation

Travail « métier » en amont

Formats « durables » et normes
conversion

Décrire -> Métadonnées

Moyens coercitifs pour la saisie

Information, formation, motivation

Best effort !

Les solutions explorées

de Archivistique
à l'informatique

l'archivage
Numérique
pérenne

Les
difficultés

Solutions
explorées



Processus
d'archivage

Conclusion



UNIVERSITÉ
TOULOUSE III
PAUL SABATIER
Université
de Toulouse



Cahier charges (fin 2013)

- Proposer un service d'archivage pour des données scientifiques à l'OMP
 - Types : divers et variés
 - Impact minimum sur les machines
 - Plusieurs dizaines utilisateurs
 - 100 et 200 To + 30 To /an

Travail réalisé en collaboration avec le LEGOS (B. Buisson)

Cahier charges (fin 2013)

- Proposer un service d'archivage pour des données scientifiques à l'OMP
 - Types : divers et variés
 - Impact minimum sur les machines
 - Plusieurs dizaines utilisateurs
 - 100 et 200 To + 30 To /an

Travail réalisé en collaboration avec le LEGOS (B. Buisson)

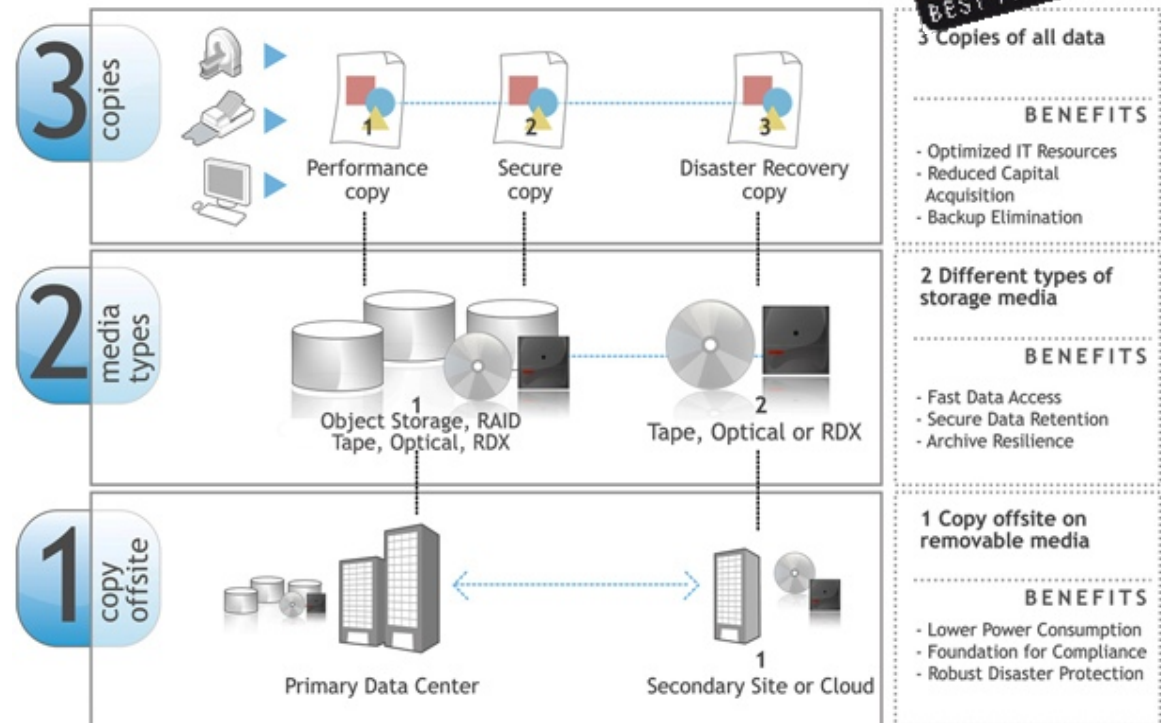
- Pistes explorées :
 - Externalisation
 - Solution interne

Cahier charges (fin 2013)

- Proposer un service d'archivage pour des données scientifiques à l'OMP
 - Types : divers et variés
 - Impact minimum sur les machines
 - Plusieurs dizaines utilisateurs
 - 100 et 200 To + 30 To /an

Travail réalisé en collaboration avec le LEGOS (B. Buisson)

- Pistes explorées :
 - Externalisation
 - Solution interne

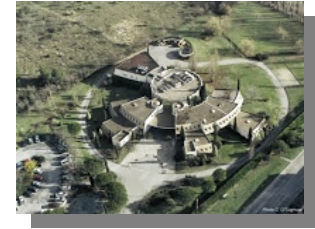


Externalisation

Le C.I.N.E.S. (Centre Informatique National de l'Enseignement Supérieur)

- trois missions stratégiques nationales :

- le calcul numérique intensif,
- **l'archivage pérenne de données électroniques,**
- l'hébergement de plates-formes informatiques d'envergure nationale



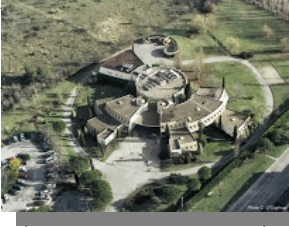
Une très bonne solution si nos formats sont supportés

Externalisation

Liste des formats archivables

Liste des formats archivables par la plateforme PAC :

Format	Version	Encodage	Identifiant Pronom	Type MIM
MPEG-4	ALL	AAC	fmt/199	-
AIFF	ALL	PCM	-	audio/x-aif
FLAC	ALL	FLAC	fmt/279	-
GIF	87a	-	fmt/3	image/gif
GIF	89a	-	fmt/4	image/gif
GEOTIFF	-	-	fmt/155	image/tiff
HTML	3.2	-	fmt/98	text/html
HTML	4.0	-	ftm/99	text/html
HTML	4.01	-	ftm/100	text/html
JPEG	Raw JPEG	-	fmt/41	image/jpeg
JPEG	1.0	-	fmt/42	image/jpeg
JPEG	1.01	-	fmt/43	image/jpeg
JPEG	1.02	-	fmt/44	image/jpeg
JPEG2000	-	-	fmt/151	-
MPEG-4	ALL	AVC/AAC	fmt/199	-
MPEG-4	ALL	AVC	fmt/199	-
MKV	ALL	AVC/FLAC	fmt/569	-

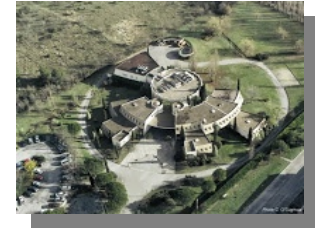
OGG	ALL	THEORA/ VORBIS	
ODT	1.0	-	
ODT	1.1	-	
ODT	1.2	-	fmt/291
PDF	1.4	-	fmt/18
PDF	1.5	-	fmt/19
PDF	1.6	-	fmt/20
PDF/A	1b	-	fmt/354
PDF/A	1a	-	fmt/95
PDF	1.7	-	fmt/276
PNG	1.0	-	fmt/11
PNG	1.1	-	fmt/12
PNG	1.2	-	fmt/13
SVG	1.0	-	fmt/91
SVG	1.1	-	fmt/92
TIFF	-	-	fmt/353
TXT	ALL	UTF-8	-
WAVE	ALL	Wave/PCM,data	fmt/6, fmt/141
XML	1.0	-	fmt/101
XML	1.1	-	fmt/101

Externalisation

Le C.I.N.E.S. (Centre Informatique National de l'Enseignement Supérieur)

- trois missions stratégiques nationales :

- le calcul numérique intensif,
- **l'archivage pérenne de données électroniques,**
- l'hébergement de plates-formes informatiques d'envergure nationale



Une très bonne solution si nos formats sont supportés ...

Un exemple : FITS (Flexible Image Transport System) est le format de fichiers le plus communément utilisé en astronomie.

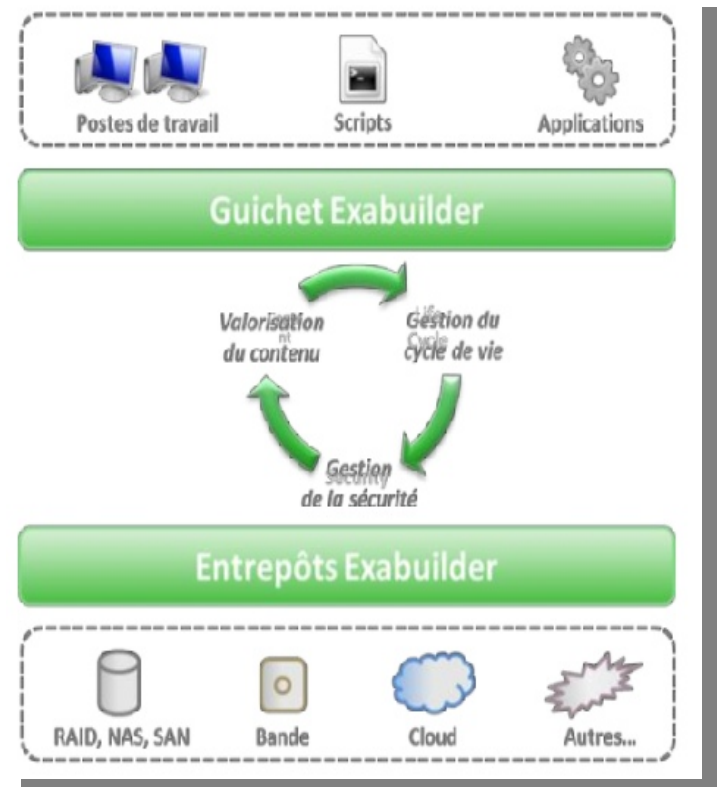
-> Une piste à explorer et un travail à faire sur les formats

Les solutions explorées

- Exabuilder
- ADA (ASG),
- TINA – module archivage (ASG),
- Serveur FTP sécurisé.

Les solutions explorées

- Exabuilder
 - Dédié archivage, entrepôts sur plusieurs supports avec migration automatique
 - Clients : légers (java), lourds (windows LINUX, Mac OS)
 - Erreur stratégie marketing -> liquidation judiciaire depuis juillet 2014 (reprise en cours ...)
- ADA (ASG),
- TINA – module archivage (ASG),
- Serveur FTP sécurisé.



Les solutions explorées

- Exabuilder

- Dédié archivage, entrepôts sur plusieurs supports avec migration automatique
- Clients : légers (java), lourds (windows LINUX, Mac OS)
- Erreur stratégie marketing -> liquidation judiciaire depuis juillet 2014

- ADA (ASG),

- TINA – module archivage (ASG),

- Serveur FTP sécurisé.

- Yacloud



Levée de fonds en cours

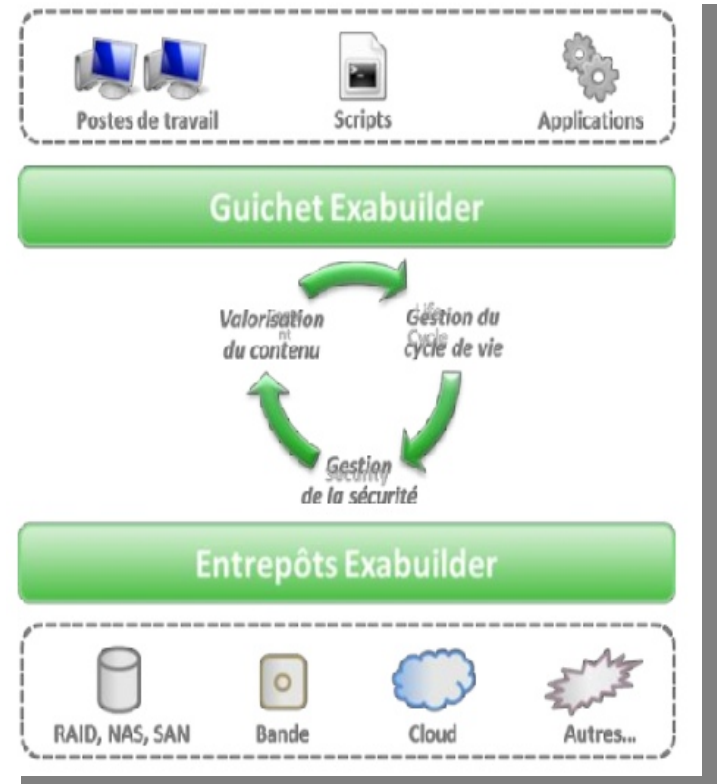
– 400 et 500 k€/an pour Exabuilder

- D. Vinay

– Investisseurs pour reprise

Renaissance 2016 ?

EXABUILDER



Les solutions explorées

- Exabuilder

- Dédié archivage, entrepôts sur plusieurs supports avec migration automatique
- Clients : légers (java), lourds (windows LINUX, Mac OS)
- Erreur stratégie marketing -> liquidation judiciaire depuis juillet 2014 (reprise

- ADA (ASG Digital Archive),

- Archivage et HSM + « stubbing »,
- Performant mais lié aux serveurs de fichiers du LAN
- Clients : légers (java) ou lourds (windows LINUX, Mac OS)
- Pas adapté besoin et prix

- TINA – module archivage (ASG),

- Serveur FTP sécurisé.

ADA SERVEUR SYSTEMES D'EXPLOITATION :

- Microsoft Windows, Linux, UNIX

INTERFACE UTILISATEUR SYSTEMES D'EXPLOITATION :

- Microsoft Windows, Mac OS, Java pour UNIX et plates-formes Linux

STOCKAGES ET MEDIAS SUPPORTES :

- Disque, Bande, Objet, Cloud, incluant :
Serveur de fichiers Windows, Mac OS, Linux & UNIX, HDS HCP, EMC Centera, NetApp, Amazon S3, EMC Atmos, DataDirect Networks WOS, Caringo Swarm, Scalify RING, ObjectMatrix, et autres

NAS SUPPORTANT LE STUBBING :

- NetApp FAS series, EMC Celerra, HDS FAP 3000 et 4000 series

Les solutions explorées

Pour 70 To
20 K€ + 10 %

- Exabuilder

- Dédié archivage, entrepôts sur plusieurs supports avec migration automatique
- Clients : légers (java), lourds (windows LINUX, Mac OS)
- Erreur stratégie marketing -> liquidation judiciaire depuis juillet 2014 (reprise en cours ...)

- ADA (ASG Digital Archive),

- Archivage et HSM + « stubbing »,
- Performant mais lié aux serveurs de fichiers du LAN
- Clients : légers (java) ou lourds (windows LINUX, Mac OS)
- Pas de connexion avec le LAN et prix

30 K€ + 10 %

- TINA – module archivage (ASG),

- Archivage, mots clés, dossiers d'archives
- Ancien, pas facile à utiliser
- Clients : lourds (windows LINUX, Mac OS)
- Inclus dans la licence TINA - POC en cours de déploiement

Licence TINA
20 K€ + 10 %

- Serveur FTP sécurisé.

- Permettre à chacun d'archiver ses données sans agent
- Client ftps + authentification LDAP + pureftpd
- Couplé avec archivage TINA

0 K€

Finalelement, un serveur FTP et un processus d'archivage...

de Archivistique
à l'informatique

l'archivage
Numérique
pérenne

Les
difficultés

Solutions
explorées

Processus
d'archivage

Conclusion



Les métadonnées

- carte d'identité d'un document
 - identifier, décrire, expliquer l'origine de sa création, son utilité et ses destinataires
- Trois types
 - **gestion**, pour accéder au document,
 - **description**, pour en comprendre le contenu,
 - **préservation**, pour garantir la pérennité de l'accès et de la compréhension du document.
- 15 éléments de description fondamentaux (Dublin Core) sont :

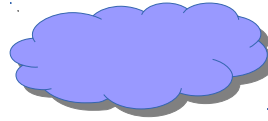
– title	– format
– creator	– source
– subject	– language
– description	– relation
– publisher	– coverage
– contributor	– rights
– date	– identifier
– type	
- + descripteurs spécifiques « métiers »

<http://dublincore.org/>

Le processus d'archivage

- Processus MANUEL réalisé par l'utilisateur

- Faire un tri
- Choisir répertoire à archiver



- Créer une archive + intégrité (checksum)
- Décrire le contenu



- Transférer l'ensemble vers le serveur



- Authentifier l'utilisateur
- Valider l'intégrité
- Acquitter réception

- Attendre acquittement transfert



- Acquitter réception

- Attendre acquittement archivage



- Mettre à jour base de contenu (métadonnées)
- Copier les N versions vers les entrepôts

- Effacement des données archivées



- Acquitter l'archivage



Conclusion

de Archivistique
à l'informatique

l'archivage
Numérique
pérenne

Les
difficultés

Solutions
explorées

Processus
d'archivage

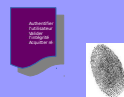
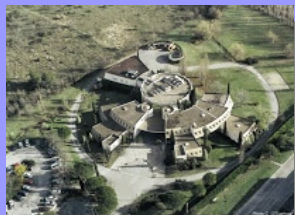
Conclusion



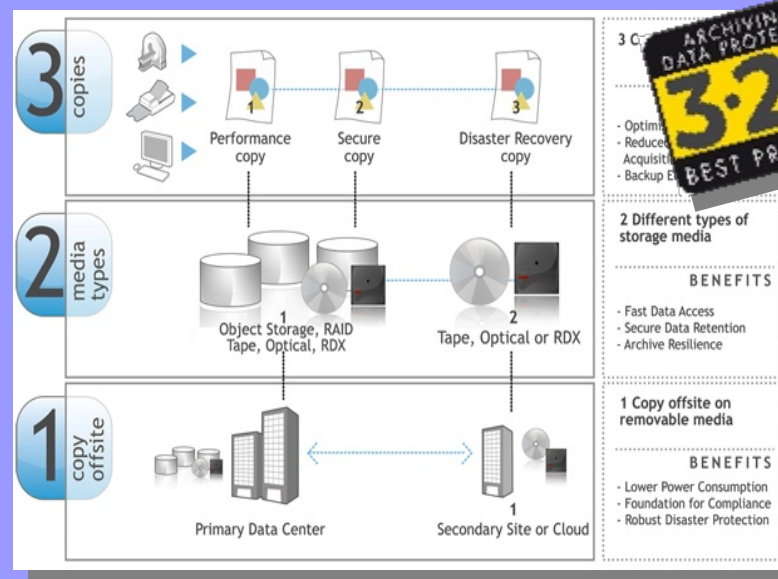
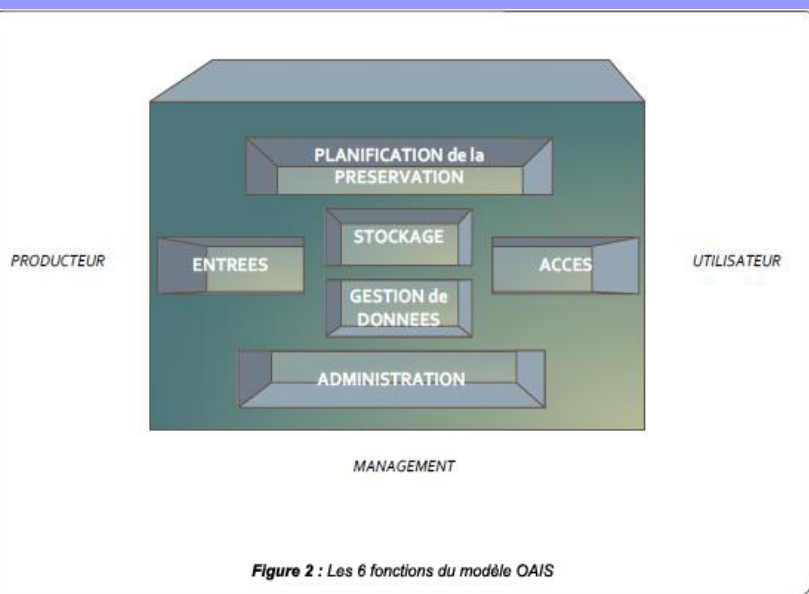
Pour archiver ...

- Nous devons
 - parler « l'archivistique », comprendre l'OAIS et les métadonnées,
 - conserver le document,
 - le rendre accessible,
 - en préserver l'intelligibilité.
 - Lutter contre
 - l'obsolescence (matériel, logiciel, format),
 - la perte de la signification du contenu
 - nos utilisateurs (parfois)
- comme l'archivage n'est pas « vendeur » nous n'avons pas ou peu de moyen





Pourquoi est-il si difficile d'archiver ?



Philippe Saby
Capitoul 26 février 2015

de Archivistique
à l'informatique

l'archivage
Numérique
pérenne

Les
difficultés

Solutions
explorées

Processus
d'archivage

Conclusion



UNIVERSITÉ
TOULOUSE III
PAUL SABATIER



Observatoire
Midi-Pyrénées
OMP

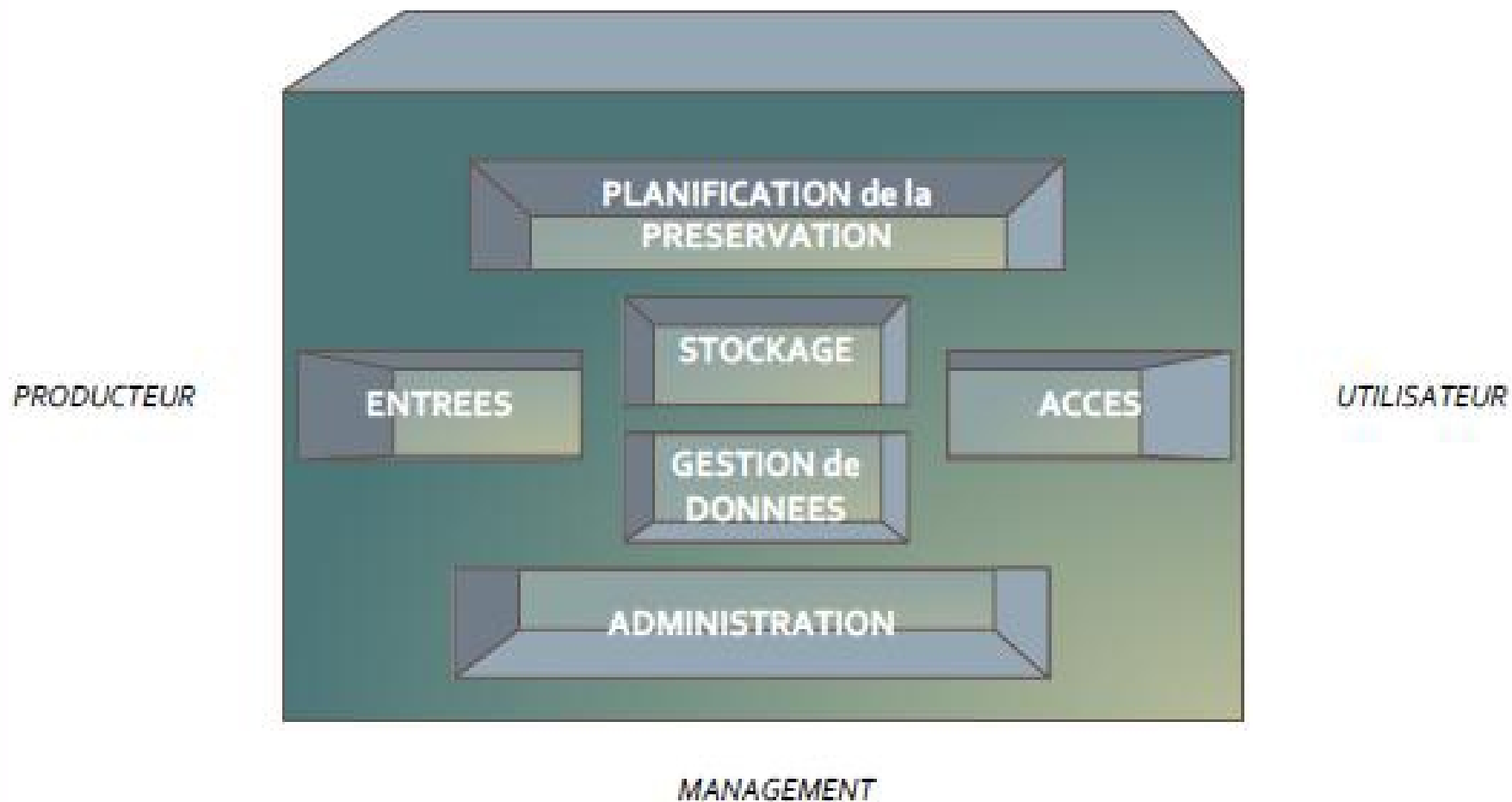
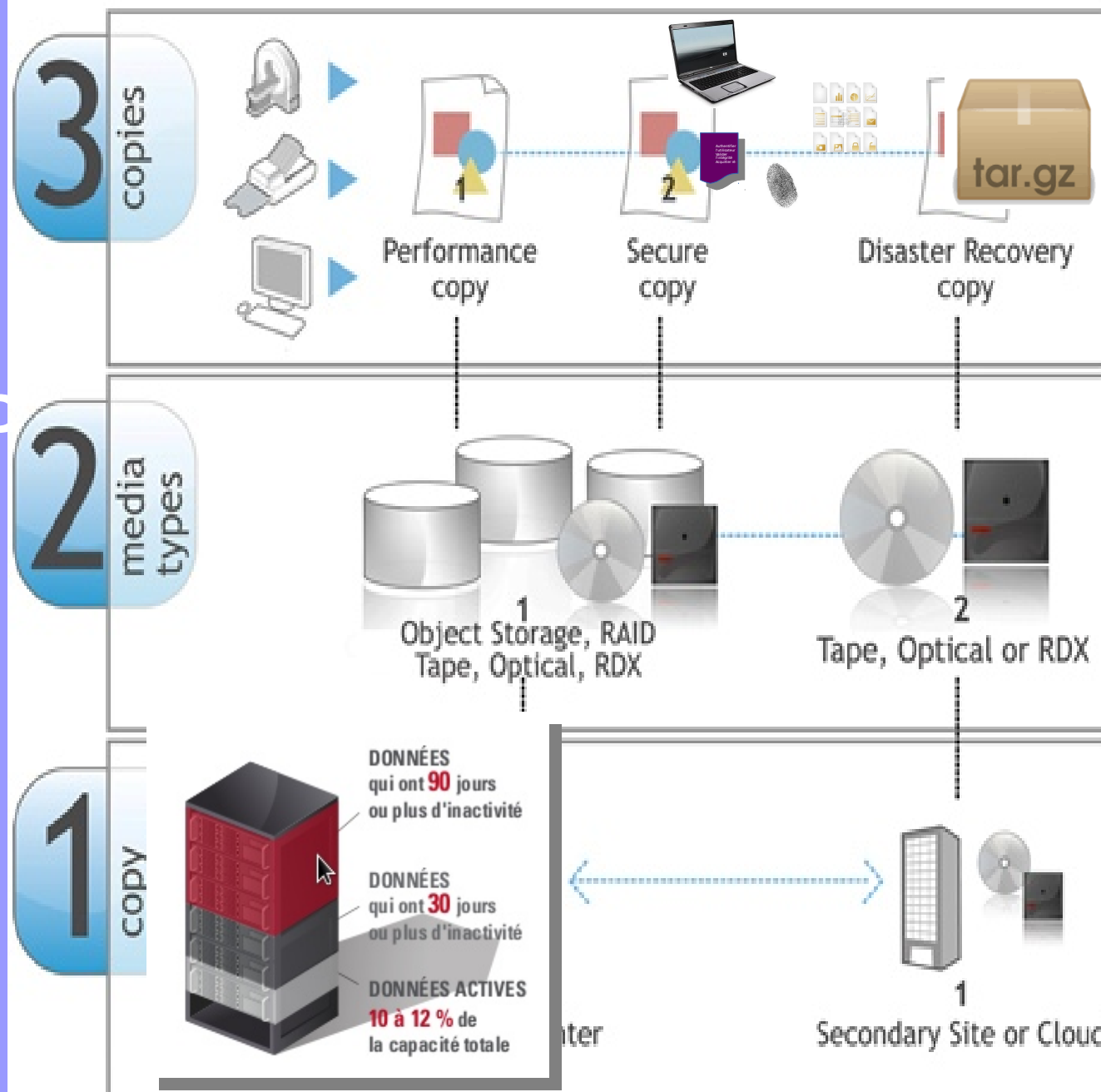


Figure 2 : Les 6 fonctions du modèle OAIS



3 Copies of all data



BENEFITS

- Optimized IT Resources
- Reduced Capital Acquisition
- Backup Elimination

2 Different types of storage media



- Fast Data Restoration
- Secure Data
- Archive Resilience

1 Copy offsite on removable media

BENEFITS

- Lower Power Consumption
- Foundation for Compliance
- Robust Disaster Protection