



« La question de l'archive n'est pas une question du passé. Ce n'est pas la question d'un concept dont nous disposerions ou ne disposerions pas déjà au sujet du passé, un concept archivable d'archive. C'est une question d'avenir, la question de l'avenir même, la question d'une réponse, d'une promesse et d'une responsabilité pour demain. »

Jacques Derrida

Philosophe (1930-2004)

Mal d'archive. Une impression freudienne



A person in a red shirt and blue jeans is skydiving over a city landscape. The background is a clear blue sky with some clouds. The image is overlaid with large, semi-transparent geometric shapes in shades of blue and purple. The title 'Archivage des bases de données' is written in a bold, purple font on a white rectangular background.

Archivage des bases de données

Arnaud Hulstaert
Grégory Ogonowski

Section Recherche

Table des matières

Introduction

- Enjeux de l'archivage
- Archivage : définition
- Normes et standards
- Difficultés et implications

Principe et méthodologie

- Archivage d'une base de données
- Méthodologie et organisation

Solutions technologiques

- Système d'archivage électronique
- Market overview

Conclusion





Enjeux de l'archivage

Enjeux

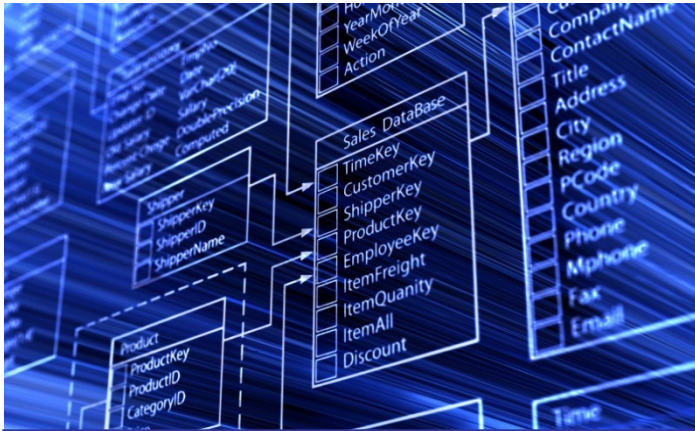


Légaux et réglementaires

- Obligations légales et réglementaires
 - **Santé** : conservation du dossier médical au moins 30 ans dans l'hôpital
 - **Données à caractère privée** : traçabilité des consultations pendant 10 ans
 - **Contrat de travail électronique** : 10 ans après le décès
 - **Nucléaire** : préservation de la documentation des centrales pendant toute leur durée de vie
 - **Aéronautique** : conservation de la documentation des avions jusqu'à ce que le dernier avion de ce type cesse de voler
- Protection juridique
 - gagner un procès
 - éviter un procès
 - éviter une amende



Enjeux



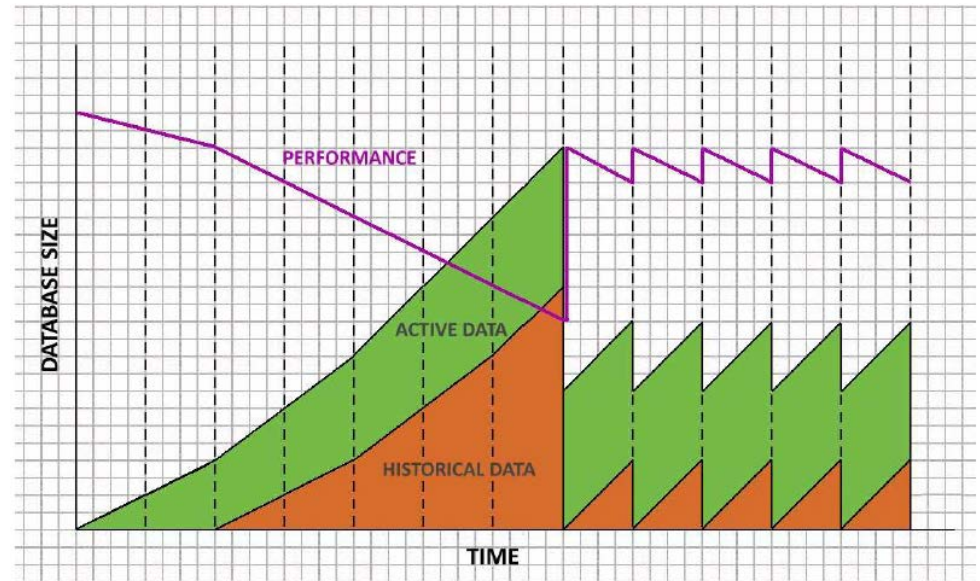
Volume et performance

- **Volume** en constante augmentation
- Difficultés croissantes :
 - **Disponibilité** des applications (99,9 %)
 - **Performances** des applications
 - Opérations de **maintenance** (backup, upgrade, migration, ...)

Gain estimé de l'archivage de bases de données, réduction

- de 89% pour l'exécution de requêtes
- de 48% des besoins de stockage
- de 37% du temps requis pour un backup complet

Source : HP



(source : IBM)

Enjeux



- Préservation des **droits des travailleurs et de l'état**
- Nécessité suite à la dématérialisation de certains documents
 - contrat électronique
 - fiches de paie
- Calcul des **pensions**
- Protection des **données à caractère personnelles ou médicales**



Enjeux



Scientifique

Coût estimé de la préservation :

- 5-7 \$ / Mb

Coût estimé pour reproduire les données :

- Secteur de l'ingénierie : 1250\$/Mb
- Secteur pétrolier : 10.000\$/m (de forage)

- Préservation du **patrimoine** pour les générations futures
- Enjeux sociétaux
 - archives des blessés de guerre
 - archives des déportés
- Disponibilité des données scientifiques exploitables vs coût de reproduction de ces données

International Radiology Archive

- Base de données sur les effets des radiations
- Années 80 : agrégation d'études européennes, américaines et japonaises
- > 300.000 cas
- Valeur estimée du contenu de la base : 2.000.000.000 €



Enjeux



Réduction des coûts

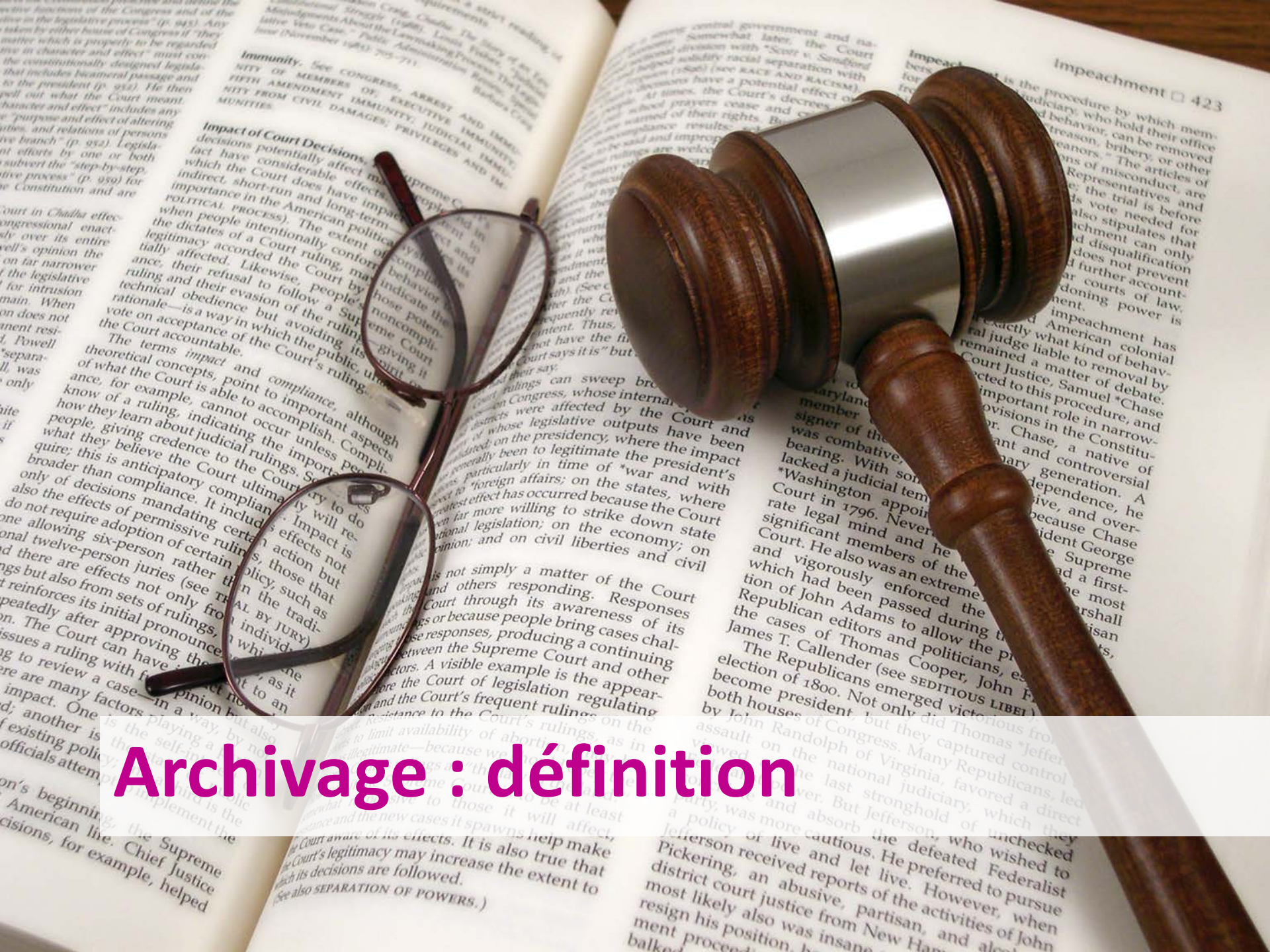
- **Maintien des applications obsolètes** pour pouvoir accéder aux données
- Coût important car
 - Technologies obsolètes
 - Maintien des compétences
 - Manque de documentation

« Conserver et consolider des applications obsolètes peut s'avérer très compliqué. Dans les faits, des coûts importants sont souvent nécessaires pour maintenir des applications non utilisées simplement pour s'assurer un accès aux données qui pourraient être ou ne jamais être nécessaires. »

Gartner



Difficulté lié au **calcul**
 du **ROI** de l'archivage
 = assurances



Immunity. See CONGRESS, ARREST AND IMMUNITY OF MEMBERS OF, EXECUTIVE IMMUNITY, FIFTH AMENDMENT IMMUNITY, JUDICIAL IMMUNITY FROM CIVIL DAMAGES, PRIVILEGES AND IMMUNITIES.

Impact of Court Decisions.

decisions potentially affect many people. The fact have considerable effects on people, which the Court does have impact on the indirect, short-run and long-term importance in the American political process. The extent of compliance when people intentionally conform to the dictates of a Court ruling, may indicate the legitimacy accorded the Court by those potentially affected. Likewise, people's noncompliance, their refusal to follow a Supreme Court ruling and their evasion of the ruling—giving it technical obedience but avoiding its spirit—rationale—is a way in which the public undermines the Court's authority.

The terms *impact* and *compliance*, although theoretical concepts, point to important aspects of what the Court is able to accomplish. Compliance, for example, cannot occur unless people know of a ruling, indicating the importance of how they learn about judicial rulings. So people, giving credence to the Court's authority, what they believe the Court ultimately will do; this is anticipatory compliance. It includes only of decisions mandating certain action but also the effects of permissive rulings, those that do not require adoption of certain policy, such as one allowing six-person juries rather than the traditional twelve-person juries (see *trial by jury*). and there are effects not only from the traditional reinforcements its initial pronouncements. The Court can have a significant impact on issues a ruling with far-reaching effects, as it is to review a case—in a way, but also impact. One factor is the self-interest of existing policy, officials attempting to implement the on's beginning, the Supreme American line. Chief Justice decisions, for example, helped

Impeachment 423

Impeachment is the procedure by which members of the judiciary, who hold their office for a fixed term, can be removed for misconduct. The articles of impeachment are the trial before the House of Representatives and the Senate. The Senate also stipulates that impeachment can only result in removal and disqualification. It does not prevent the impeached from further accounting for his or her actions in court. Impeachment has been used in American colonial history. Exactly what kind of behavior remained a matter of debate. In the case of Samuel Chase, a member of the Supreme Court, the House of Representatives brought charges against him. Chase, a native of Maryland, was combative, and controversial. With some support, he was appointed to the Supreme Court in 1796. Nevertheless, because Chase was an extreme Federalist, and overbearing, he was not popular. He also was an extreme Federalist, and overbearing, he was not popular. He also was an extreme Federalist, and overbearing, he was not popular.

Archivage : définition

Notion fondamentale : sécurité



« **Archiver** consiste à prendre un objet et à le transférer sous certaines conditions dans un système qui permettra d'en assurer la préservation pendant un certain laps de temps avec toute la sécurité requise » (M.-A. Chabin)

- Archivage = **SÉCURITÉ**
 - la modification des objets est interdite
 - la destruction est également interdite sauf sous contrôle strict
 - toute action effectuée sur l'objet doit être tracée
 - objectif : conserver la valeur de l'objet (légale et/ou informative)
- Archivage = « le stockage de vieux documents »
ne convient pas à l'environnement informatique





Ambiguïté lexicale (UK)

Archiving

supports de stockage

- transfert des objets sur des supports moins coûteux / plus lents
- refreshing des supports
- condition nécessaire mais pas suffisante

Records Management

gestion de l'information

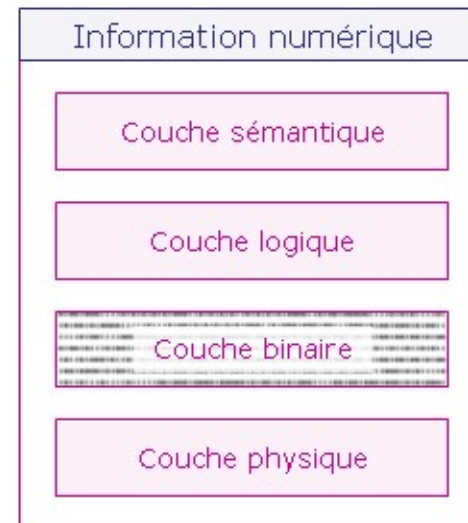
- actions, outils et méthodes
- pour réunir, identifier, sélectionner, classer et préserver
- de manière sécurisée
- pour exploiter et rendre accessibles dans le temps des objets
- que ce soit à titre de preuve ou à titre informatif



Archivage - définition

- En bref, l'archivage électronique, c'est
 1. **sélection** de ce qu'il faut archiver (en amont)
 2. **rattachement** des objets sélectionnés à des règles de gestion
 3. mise en **sécurité** dans un système dédié
 4. **préserver** les objets

*tous les couches de l'information,
de sa matérialisation physique
à sa compréhension sémantique*





Normes et standards

Paysage normatif

Normes managériales

ISO 3030x

ISO 15489

Normes conceptuelles et organisationnelles

ISO 14721
(OAIS)

ISO 20652
(PAIMAS)

Standards fonctionnels

Moreq2010

ICA-Req

NFZ42-013

Normes et standards metadata

ISO 23081

METS

PREMIS

Dublin
Core

ISAD(G)



ISO 15489



- Norme principale de l'archivage
- Management et démarche de records management
- Caractéristiques des objets archivés
- Outils et rôles pour l'archivage des objets
- Partie 1 : Principes directeurs
- Partie 2 : Guide pratique
- Complété depuis 2011 par les normes ISO 3030x



ISO 14721 (modèle OAIS)

- **Modèle conceptuel** qui fournit
 - Une terminologie commune
 - Une liste des questions à se poser
 - Une liste d'activités à assurer
- Explique
 - Ce qu'il faut faire pour préserver des documents numériques

Modèle fonctionnel

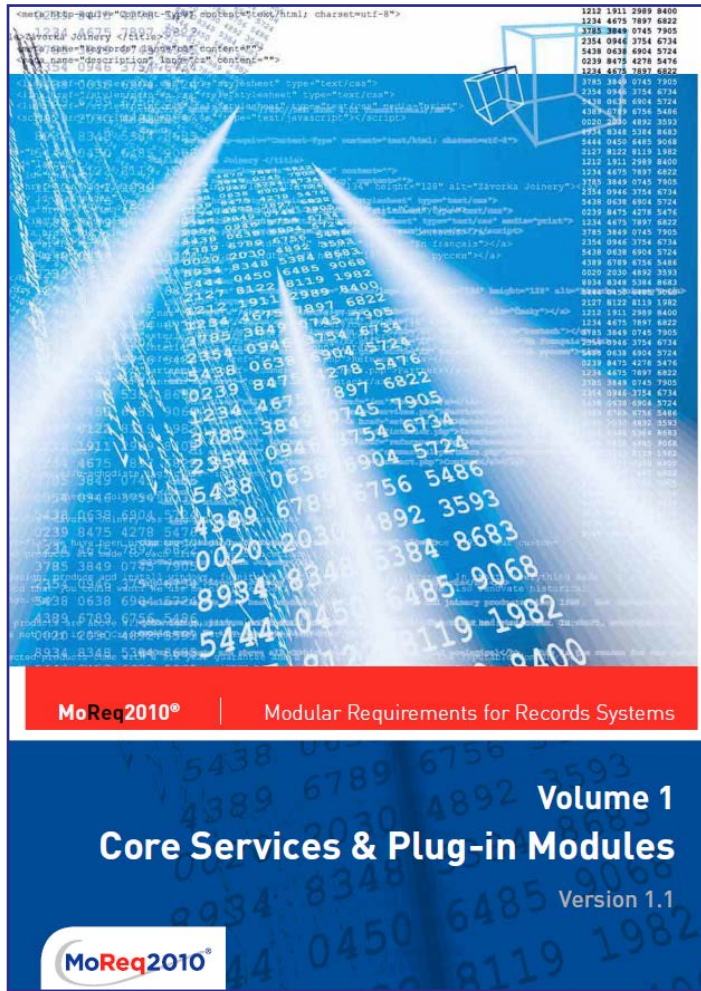
- Les types d'information à attacher aux objets en vue de leur préservation

Modèle d'information



MoReq2010

- Mise en oeuvre européenne de la norme ISO 15489
- Sponsor : Commission européenne
- MoReq2010 (publiée en 2011) : version actuelle v1.1
- Norme fonctionnelle et modèle de données
- Grande qualité mais ...
- Complexité des fonctions/entités
524 pages
253 requirements



Utilité des normes

- **Réflexion conceptuelle**

- enjeux de l'archivage
- procédures et organisation (management et fonctionnelle)
- difficultés

ISO 15489

ISO 14721
(OAIS)

MoReq2010

- **Guide fonctionnel**

- check-list quasi exhaustive
- audit d'une solution
- achat d'une solution via appel d'offres

ISO 14721
(OAIS)

MoReq2010

ICA-Req

- **Data model**

- guide conceptuel
- modèle pratique

ISO 14721
(OAIS)

MoReq2010

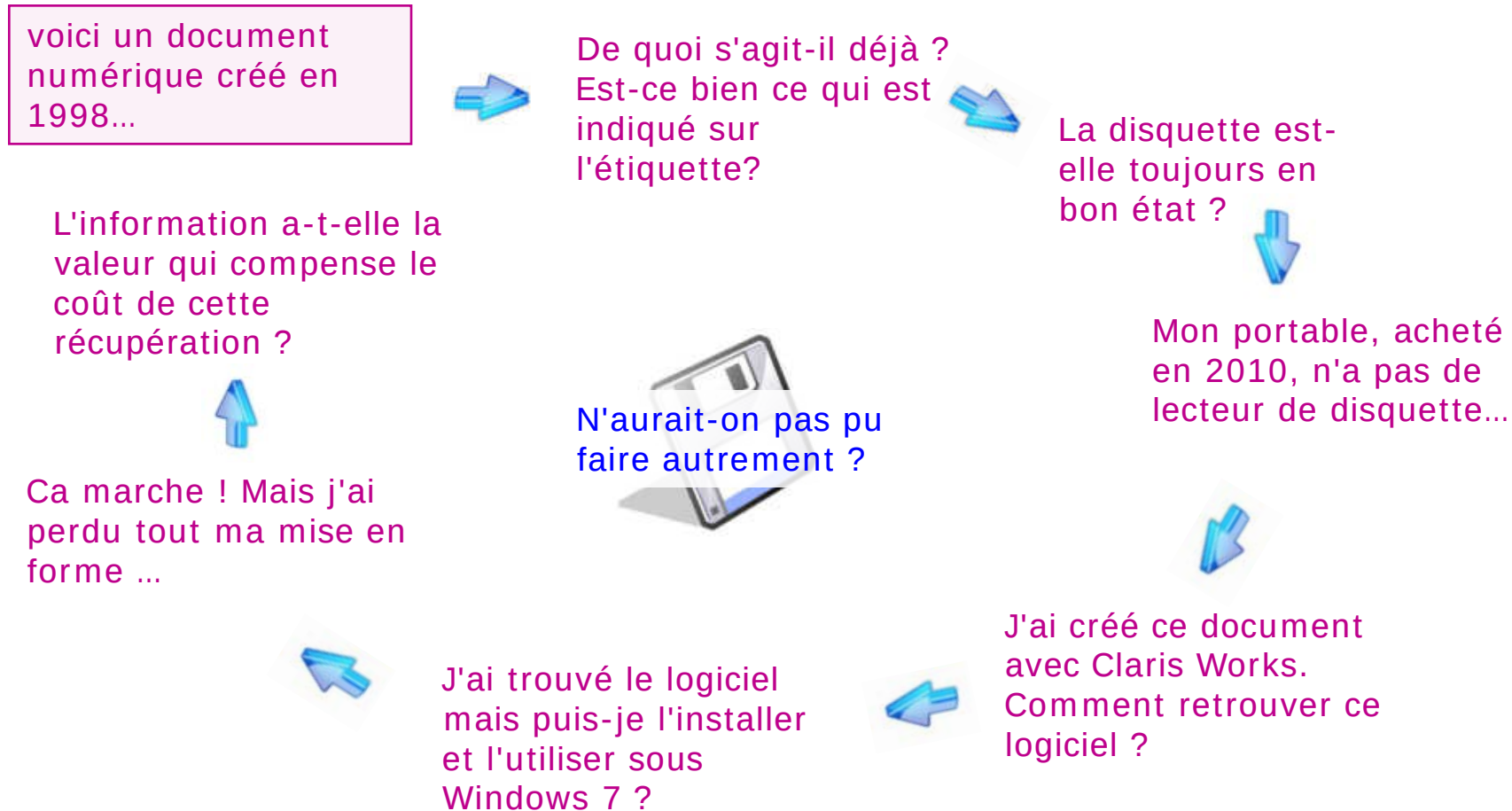
PREMIS



A photograph showing several vintage computer peripherals, including keyboards and mice, hanging from a wooden plank. The plank has handwritten numbers in black ink: '0242911839' at the top and '5' on the left. The keyboards are beige and appear to be from the late 1980s or early 1990s. The mice are also beige and have a corded design. The background is a dark, textured wooden surface.

Difficultés et implications

En bref

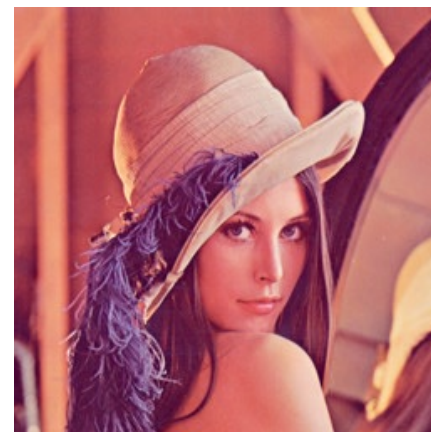


< E. Bernès (BnF)



Une image en format binaire

```
1010111000101001010101110110
1010101010101010000101010101
0101010010101001010101010101
0101010000001010111000101001
0101011101101010101010101010
0001010101010101010010101001
0101010101010101010000001010
1110001010010101011101101010
1010101010100001010101010101
0100101010010101010101010101
0100000000001010111000101001
0101011101101010101010101010
```

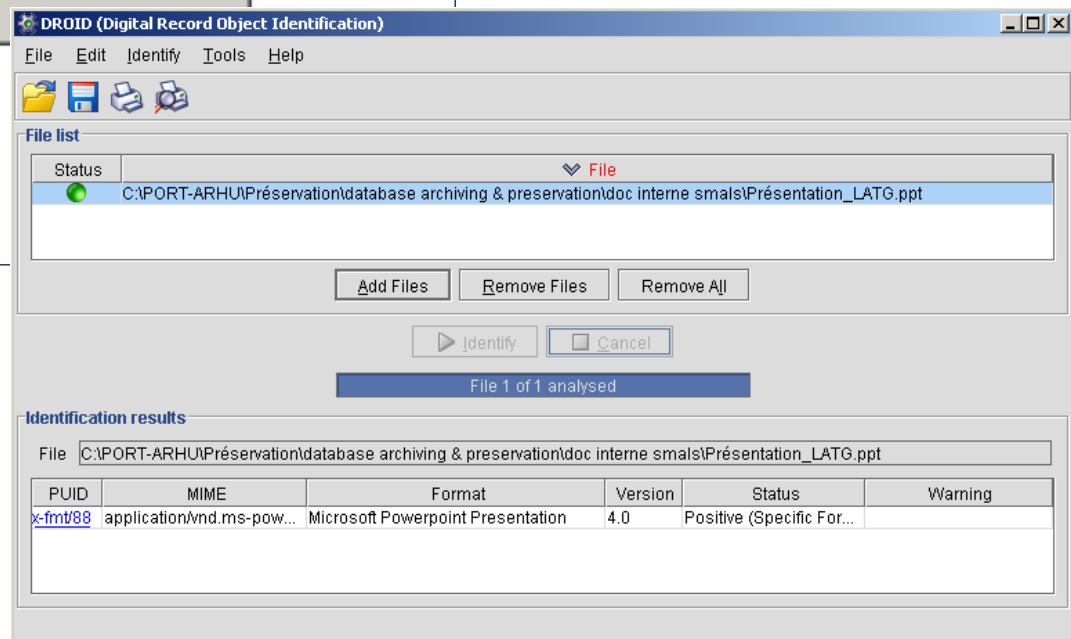
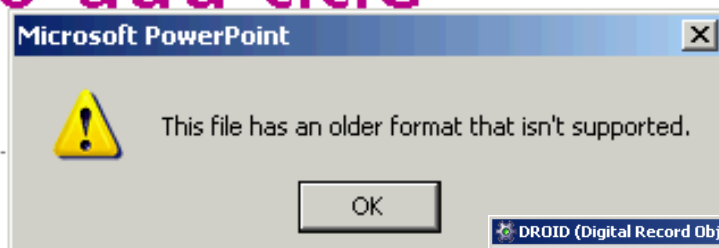


```
1010111000101001010101110110
1010101010101010000101010101
0101010010101001010101010101
0101010000001010111000101001
0101011101101010101010101010
0001010101010101010010101001
01010101010001010100000010101
1110001010010101011101101010
1010101010100001010101010101
0100101010010101010101010101
0100000000001010111000101001
0101011101101010101010101010
```



Un fichier créé avec powerpoint 97

Click to add title



Une base de données

120	58	5	1981-09-30T14:22:01.629Z	3898.70	2352.45	-1.66	-7.72	-18.61	-80.00	-80.00
120	58	5	1981-09-30T14:22:05.788Z	3894.64	2546.88	-1.49	-9.76	-18.96	-80.00	-80.00
120	58	5	1981-09-30T14:22:10.027Z	3925.05	2429.15	-1.49	-8.78	-18.27	-80.00	-80.00
120	58	5	1981-09-30T14:22:18.426Z	3946.38	2500.82	-1.71	-8.42	-19.33	-80.00	-80.00
120	58	5	1981-09-30T14:22:22.585Z	3964.01	2591.92	-1.76	-8.83	-19.86	-80.00	-80.00
120	58	5	1981-09-30T14:22:35.222Z	4021.69	2406.25	-1.62	-7.75	-18.24	-80.00	-80.00
120	58	5	1981-09-30T14:22:39.381Z	4007.80	2474.73	-1.69	-8.05	-18.88	-80.00	-80.00
120	58	5	1981-09-30T14:22:43.620Z	4037.11	2329.44	-1.74	-6.70	-18.23	-80.00	-80.00
120	58	5	1981-09-30T14:22:52.018Z	4099.12	2399.84	-1.33	-8.46	-16.84	-80.00	-80.00
120	58	5	1981-09-30T14:22:56.177Z	4120.44	2447.08	-1.46	-8.25	-17.44	-80.00	-80.00
120	58	5								)
120	58	5	1 58 5 1981-09-30T14:22:01.629Z	3898.70	2352.45	-1.66	-7.72.. ...			)
120	58	5								)
120	58	5	3 58 5 1981-09-30T14:22:05.788Z	3894.64	2546.88	-1.49	-9.76			)
120	58	5								)
120	58	5	5 58 5 1981-09-30T14:22:10.027Z	3925.05	2429.15	-1.49	-8.78			)
120	58	5								)
120	58	5	7 58 5 1981-09-30T14:22:18.426Z	3946.38	2500.82	-1.71	-8.42			)
120	58	5								)
120	58	5	120 58 5 1981-09-30T14:22:22.585Z	3964.01	2591.92	-1.76	-8.83			)

Date de l'observation
Année/mois/jours
Heure/minute/seconde

Fréquence de plasma
Unité = kHz
Type réel



Une preuve



- contrat
- actes authentiques
- document de travail
- brevets
- autres



Implications

voici un document
numérique créé en
1998...

L'information a-t-elle la
valeur qui compense le
coût de cette
récupération ?

Financement



Ca marche ! Mais j'ai
perdu tout ma mise en
forme ...

Intégrité - authenticité



De quoi s'agit-il déjà ?
Est-ce bien ce qui est
indiqué sur l'étiquette?

Métadonnées descriptives

N'aurait-on pas pu
faire autrement ?

Organisation



J'ai trouvé le logiciel
mais puis-je l'installer
et l'utiliser sous
Windows 7 ?

Système d'exploitation



La disquette est-
elle toujours en
bon état ?

Support



Mon portable, acheté en
2010, n'a pas de lecteur
de disquette...

Environnement matériel

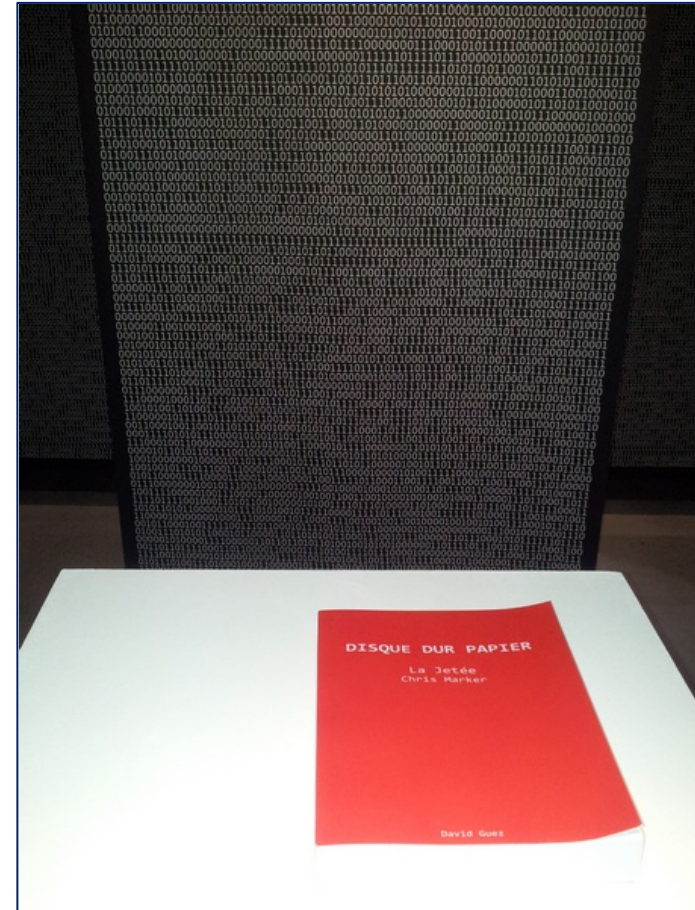


J'ai créé ce document
avec Claris Works.
Comment retrouver ce
logiciel ?

Environnement logiciel



Ceci est-il une solution ?

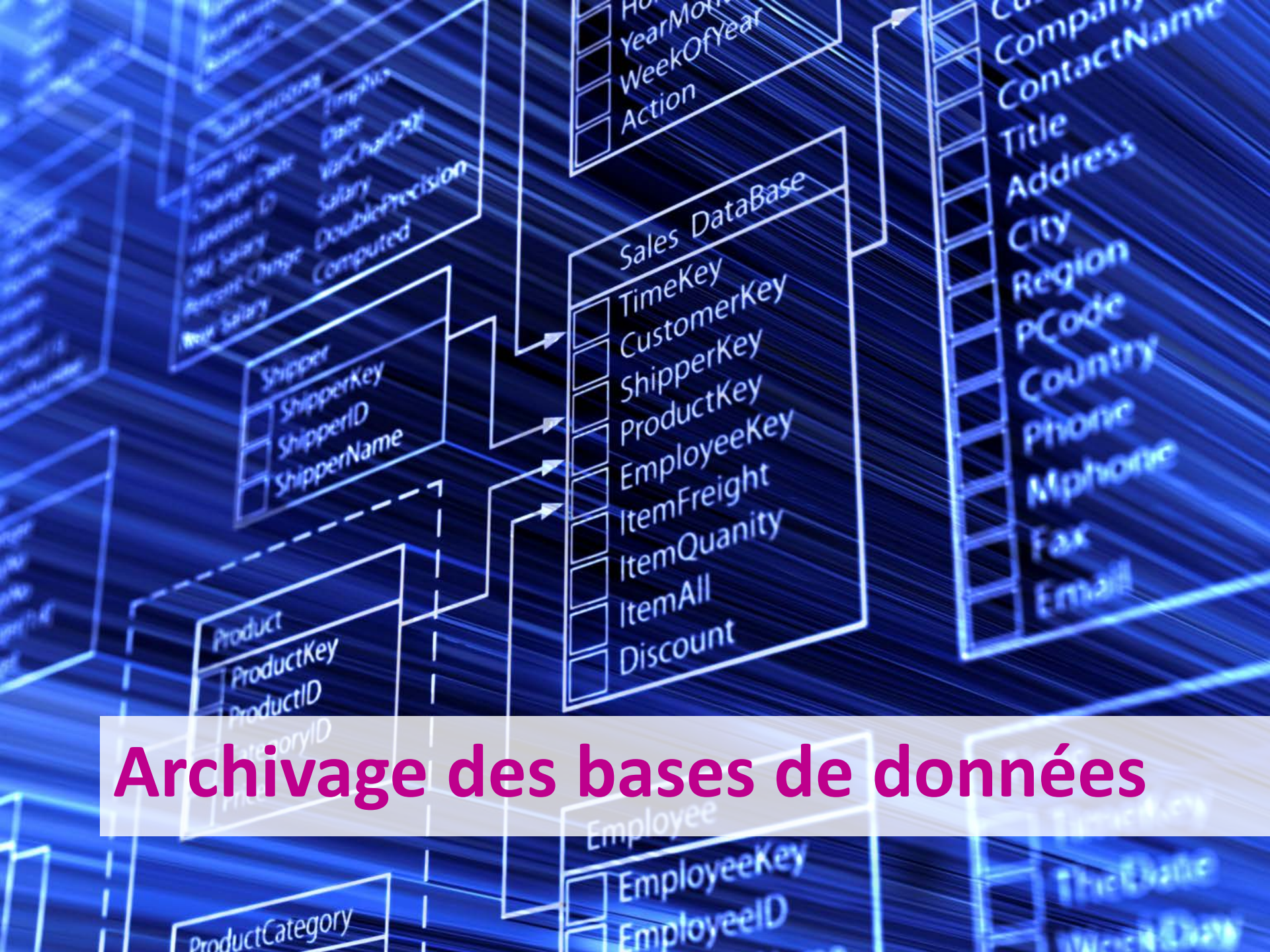


David Guez - Un disque dur papier

<http://www.disquedurpapier.net/>

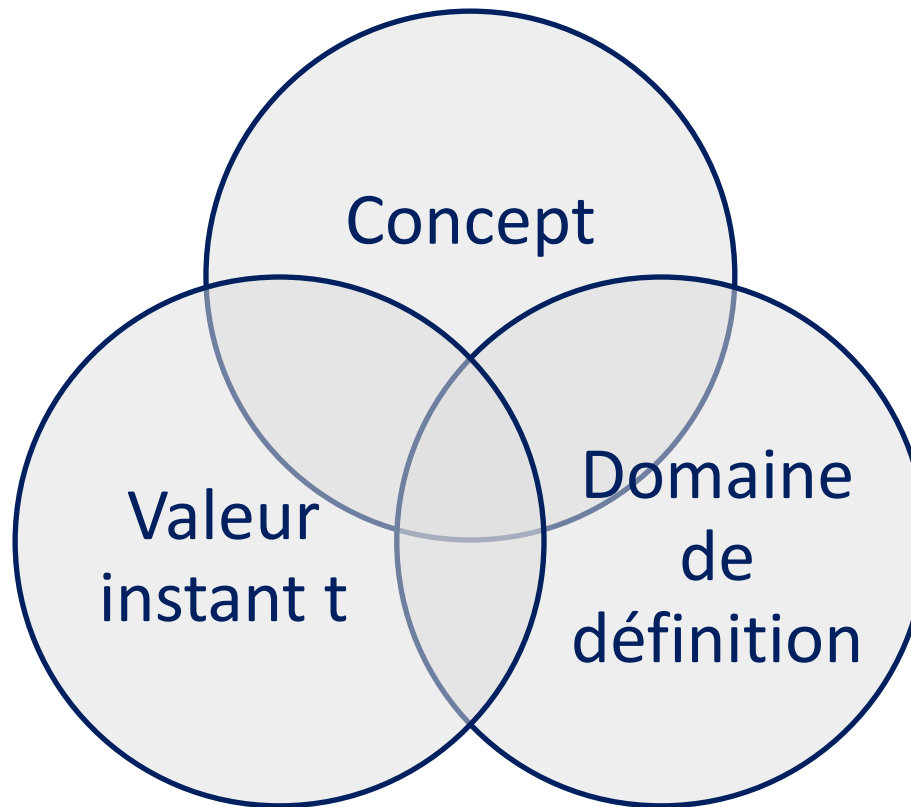


Archivage des bases de données



Qu'est-ce qu'une donnée ?

Ex : salaire mensuel



Ex : 3000

Ex : « valeur
numérique incluse
entre 1000 et
100.000 »



Qu'est-ce qu'une base de données ?

Database : a collection of related data with

- a logically coherent structure
- some inherent meaning
- a specific purpose
- a large varying size
- a scope or content of varying breadth
- a physical organization of varying complexity
- a persistence over a long period of time

< Elmasri & Navathe

A database management system (DBMS) is a collection of programs that enables users to create and maintain a database.



Base de données administrative

- Modifications législatives fréquentes et complexes
→ gestion des versions
- Force probante des données → historique des records
- Volume de données important
 - Ex : DmfA, par trimestre, environ : 4 millions d'enregistrements
- Incidences sociales et financières considérables



Points spécifiques aux bases de données

- Structure des données potentiellement complexe (parfois > 100 tables)
- Extraction d'un objet potentiellement distribué et corrélé
- Veiller à la consistance des données restantes
- Archivage des données brutes pas suffisant
 - Capture d'informations complémentaires (metadata)
 - contexte permettant d'exploiter les données
 - rules, procédures PL/SQL qui définissent la manière dont les données sont utilisées
- Capture des modifications successives (si d'application)
- Volume important
- Exploitation des données



Types de bases de données

- Plusieurs **types** de bases de données existent
- Le modèle relationnel est le plus fréquent
- Un standard existe pour l'interrogation des bases de données relationnelle : SQL
- Autres types de bases de données :
 - Hiérarchiques
 - Réseau
 - Objet
 - Documents
 - XML
 - NoSQL



G. Ogonowski, *NoSQL – Hype ou innovation ?*, Section Recherches, Smals, octobre 2011.

<http://documentation.smals.be/>



Types de bases de données

- Approche et stratégies proposées :
indépendantes du type de DB
- Impact au niveau des logiciels
 - Prise en compte des types de bases de données – cf. infra
 - Capacité d'extraction
 - Consultation des données
- Travail en amont différent :
 - Préparation des objets
 - Conversion vers SQL ?



Archivage en amont

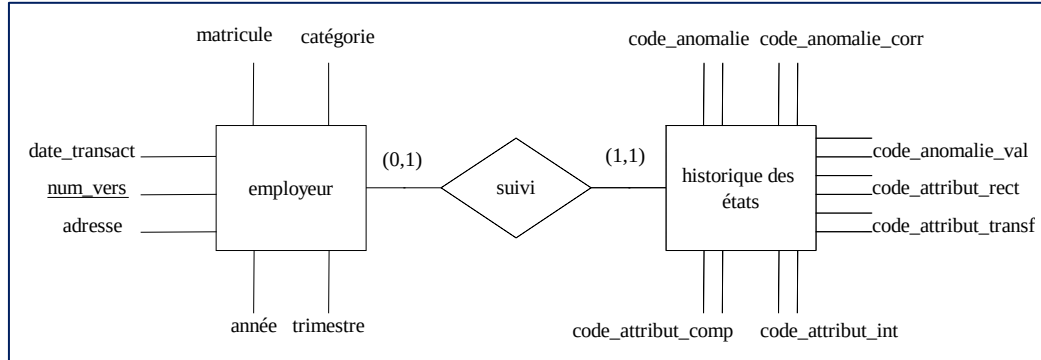
- Record (data)
 - trace une action qui s'est produite ou décrit un objet
 - est un élément à un instant t
- Si le record à un instant t a une **valeur**, il ne peut être modifié ou effacé **sans traçabilité**.
- En cas de mise à jour, nécessité d'enregistrer **l'historique des modifications**



L'archivage se prévoit dès la conception de l'application : l'enregistrement d'un historique des modifications doit être pensé dès le design de la base de données.



Archivage en amont



source : Isabelle Boydens

Chaque mise à jour pour l'entité employeur est enregistrée et tracée de manière à pouvoir retracer tout l'historique des modifications.

Références

Voir les travaux d'Isabelle Boydens
 e.a. : I. Boydens et al., *Gestion intégrée des anomalies*, Section Recherches, mars 2011.

I. Boydens, *Data Quality : Best Practices*, Section Recherches, mai 2006.

I. Boydens, *Evaluer et améliorer la qualité des bases de données*, Section Recherches, janvier 1998.

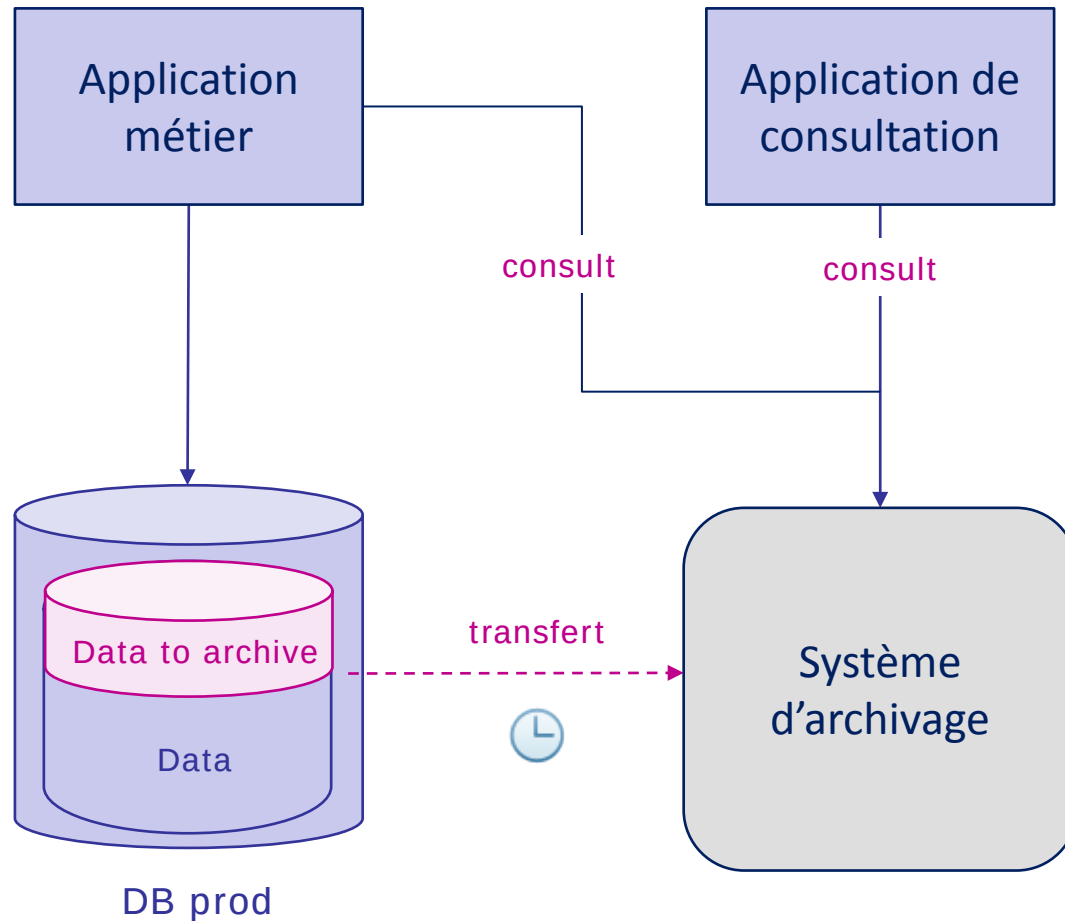
I. Boydens, *Informatique, normes et temps*, Bruxelles : Éditions E. Bruylant, 1999.

<http://documentation.smals-mvm.be/>

<http://www.ulb.ac.be/cours/iboydens/>



Transfert des données

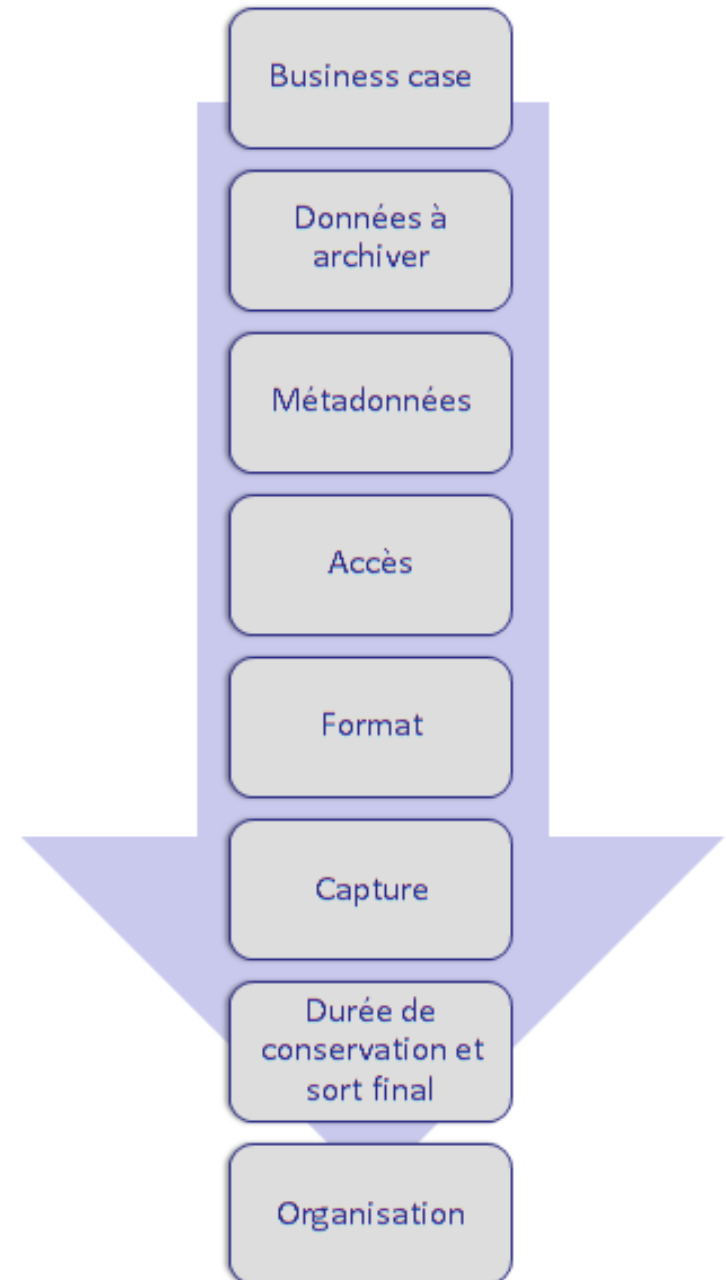




Méthodologie et organisation

Méthodologie

1. Business case
2. Données à archiver
3. Métadonnées
4. Accès
5. Format
6. Capture
7. Durée de conservation et sort final
8. Organisation



1. Business case

Timestamping

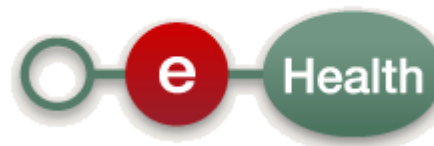
Horodatage des transactions effectuées dans les applications eHealth.

Volume [futur] : 300.000/jour

CoBRHA

Source authentique IAM

Volume : 20.000 / jour

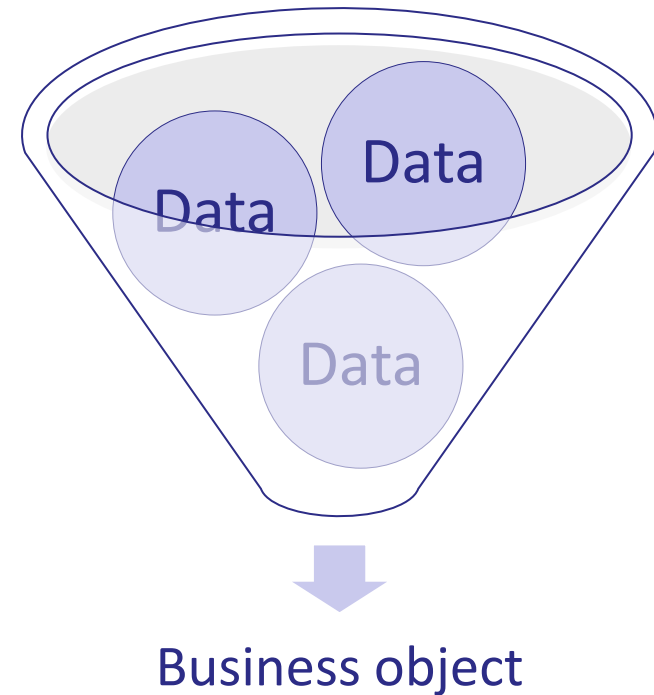


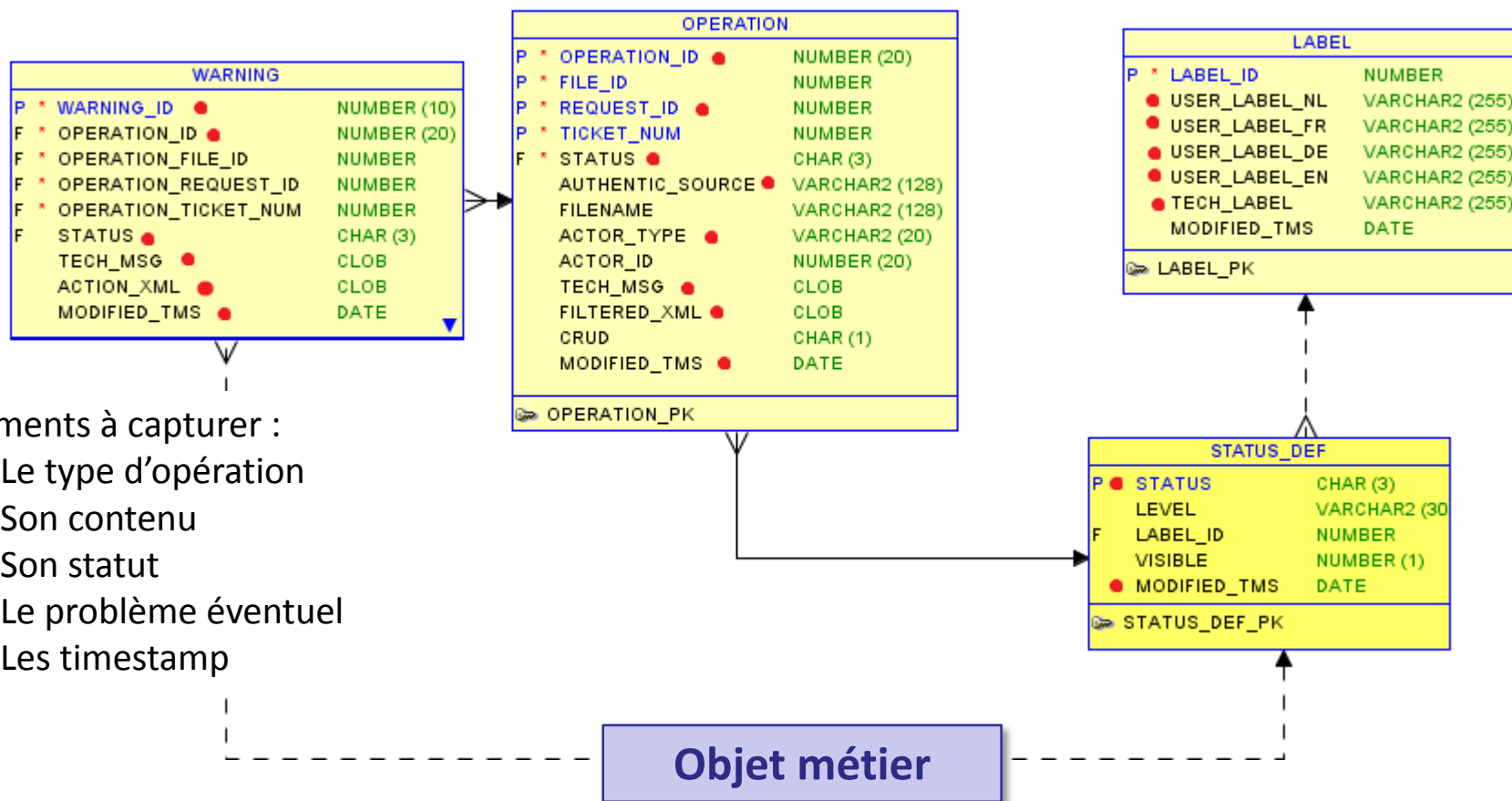
1. Conserver les données sur du long terme (> 10 ans) pour des raisons légales et réglementaires
2. Eviter des problèmes de performance des applications (environnement 99,9 %)
3. Réduire temps du passage à une « next release » trimestriel



2. Données à archiver

- Toutes les données ne doivent pas être archivées
- **Sélection** des données à archiver
- Constitution d'un **objet métier**
 - *Données qui décrivent un évènement (transaction) ou une entité business persistante (J. Olson – voir bibliographie)*
 - Définition d'une table ROOT et définition des relations et des données à archiver
 - Travail délicat, collaboration entre
 - Data architect / analyst
 - Personne métier
 - Juriste



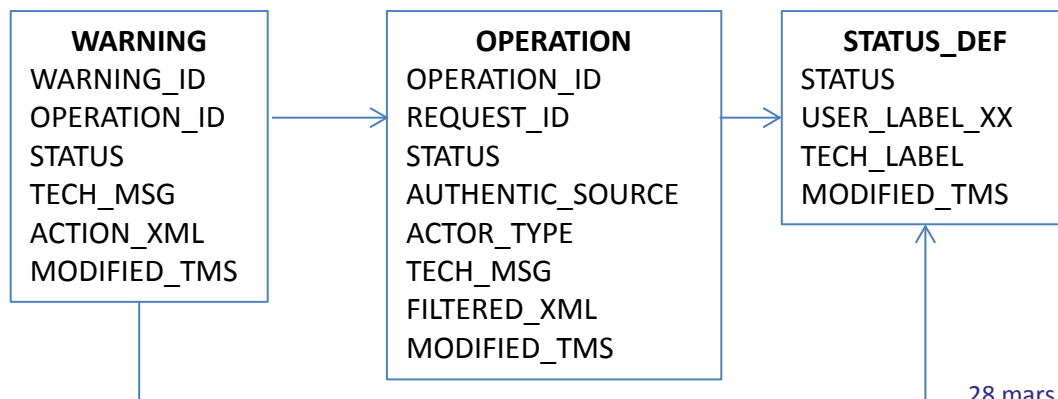


Éléments à capturer :

- Le type d'opération
- Son contenu
- Son statut
- Le problème éventuel
- Les timestamp

eHealth - CoBRHA

Archivage des opérations effectuées sur la table audit_service



3. Métadonnées



Statuette d'homme avec gobelet à libation

Portant une longue barbe et coiffé d'un bonnet, l'homme est sans nul doute un haut personnage, notable ou plus vraisemblablement prêtre, dont la physionomie bienveillante souligne la fonction protectrice.

« Futur antérieur », une exposition ludique sur le 21^e siècle vue de l'an 4005 -Exposition conçue et réalisée par le musée romain de Lausanne-Vidy, 23/11/2010 au 13/03/2011



3. Métadonnées

Une métadonnée est une information qui décrit les différents aspects d'un élément d'information pour améliorer son utilisation tout au long de son cycle de vie.

Source : Gartner

- Dépend des besoins des futurs utilisateurs (« communauté des utilisateurs » (terme ISO 14721))
- Evolution de la communauté des utilisateurs
- Veiller à l'indépendance par rapport aux applications sources.

A. Hulstaert, *Préservation à long terme de l'information numérique. Rendre l'information accessible durablement*, Section Recherches, Smals, février 2010.

<http://documentation.smals-mvm.be/>

C. Rectem, *La pérennisation digitale dans le secteur public : étude critique de plusieurs modèles ouverts de méta-information*, Mémoire en Sciences de l'information et de la documentation, ULB, Bruxelles, 2009.



3. Métadonnées

Exemple : DmfA

3 composants au minimum :

- Données DmfA
- Glossaires + Instructions aux SSA expliquant la syntaxe et sémantiques des données
- Business Rules qui contrôlent les DmfA

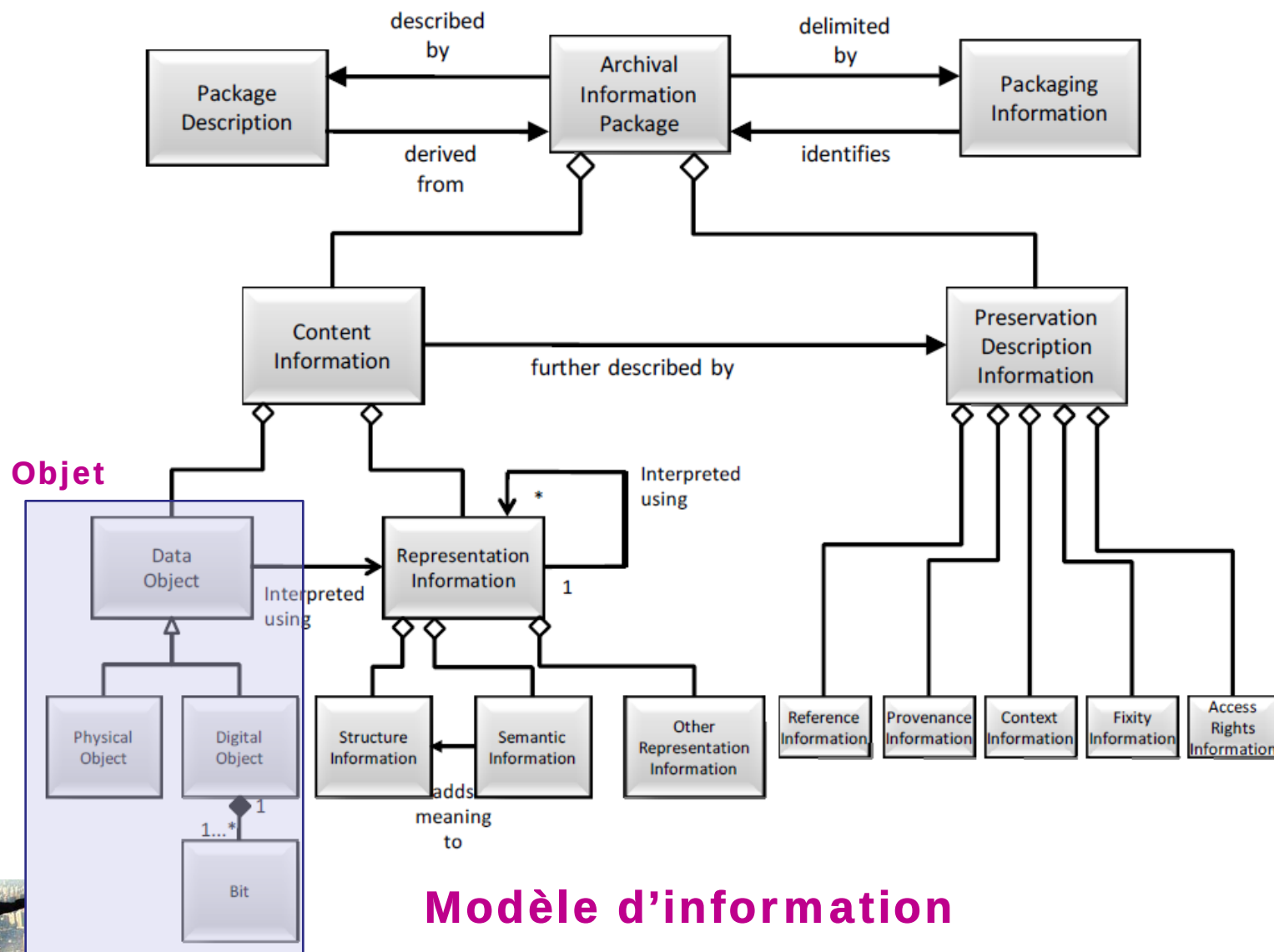
Exemple : Requête PL/SQL

```
DELIMITER //
CREATE PROCEDURE `proc_TOT_T1` (IN max_occur INT, OUT
total INT)
BEGIN
    DECLARE a, b, c INT;
    DECLARE cur1 CURSOR FOR SELECT col1 FROM table1 ORDER
BY col1;
    DECLARE CONTINUE HANDLER FOR NOT FOUND SET b =
max_occur;
    OPEN cur1;
    SET b = 0;
    SET c = 0;
    WHILE b < max_occur DO
        FETCH cur1 INTO a;
        SET c = c + a;
        IF a > 1000 THEN
            SET b = b + 2;
        ELSE
            SET b = b + 1;
        END IF;
    END WHILE;
    CLOSE cur1;
    SET total = c;
END //
```



3. Métadonnées

ISO 14721



4. Accès

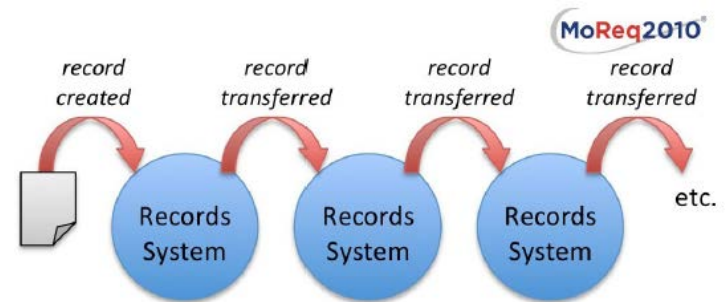
- Qui peut accéder aux données ?
- Fréquence d'accès ?
- Volume lors d'un accès ?
- Mêmes données accédées ?
- Traitement souhaité sur les données ?
- Arbitrage « intégrité versus consultation »
 - modification de format / de structure au moment du versement



5. Format

• Critères importants

- Indépendance vis-à-vis des applications de production
- Indépendance technologique (OS, DB specific)
- Possibilités d'archiver des paquets d'information : (original) data + metadata
- Possibilités de recherche
- Accessibilité à partir des applications
- Restauration des données



La durée de vie des données archivées est supérieure à la durée de vie des solutions d'archivage et des technologies.



5. Format

- Format
 - ouvert
 - documenté
 - no vendor-specific
- Point d'attention : blob file
- Si nécessaire, **conversion**
 - arbitrage « pérennité versus intégrité »

Formats recommandés

- SQL standard (SQL 1999)
- XML
- PDF/A
- ASCII / Unicode
- SQLite

A. Hulstaert, *Préservation à long terme de l'information numérique. Rendre l'information accessible durablement*, Section Recherches, Smals, février 2010.

A. Hulstaert, *Préserver l'information numérique : Codage et conversion de l'information*, Section Recherches, Smals, juin 2008.

<http://documentation.smals-mvm.be/>



6. Capture

- Déterminer une **périodicité d'extraction**
- Dépend
 - Des **besoins en consultation** à partir de l'application d'origine
 - De la **fréquence** de consultation
 - Des **volumes** et de leur impact sur les performances
- **Processus de capture**
 - Responsabilité
 - Validation de la demande (?)
 - Informations transmises au producteur



CoBRHA

Tout log vieux de plus de 31 jours peut être archivé.

En pratique, extract calqué sur processus de Next-Release (3 mois).



6. Capture

- Critères de sélection
 - Sélection des objets susceptibles d'être archivés
 - Automatisable
 - Exprimé sous forme de texte ET SQL (si relationnel)
- Exemples de critères
 - Time since ...

90 jours après réception paiement

```
IF DAYS(CURRENT_DATE - LOAN_PAYMENT.TRANSACTION_DATE) > 90
```
 - Completion status

90 jours après clôture du dossier

```
IF CASE_STATUS=CLOSED AND DAYS(CURRENT_DATE - CASE_CLOSED_DATE) > 90
```



7. Durée de conservation et sort final

- Durée de conservation
 - Dans le système d'archivage
 - Durée temporelle
 - Start time /event ?
- Dépend de la réglementation et des risques
- Sort final
 - Ensemble des opérations à exécuter au terme de la durée de conservation
 - Destruction / Transfert / Review
- Evènements susceptibles d'interrompre le processus ?



eHealth

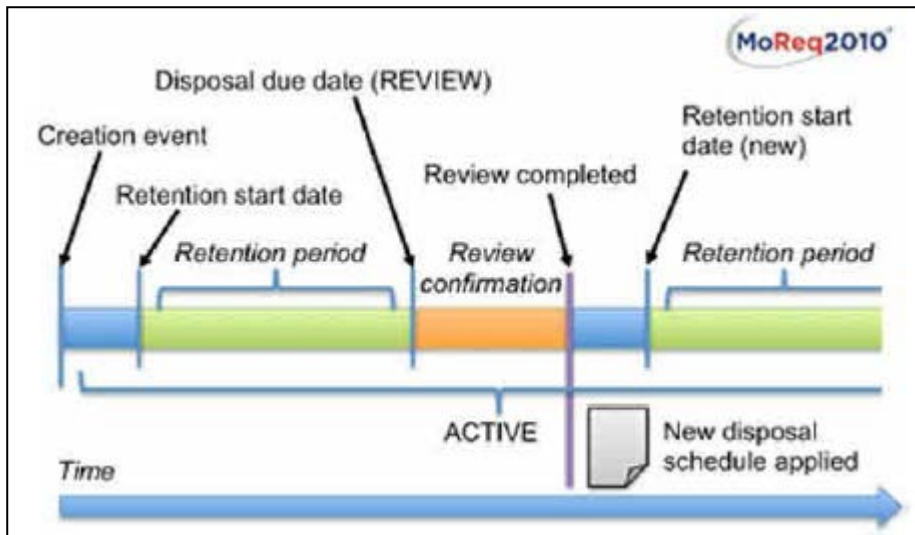
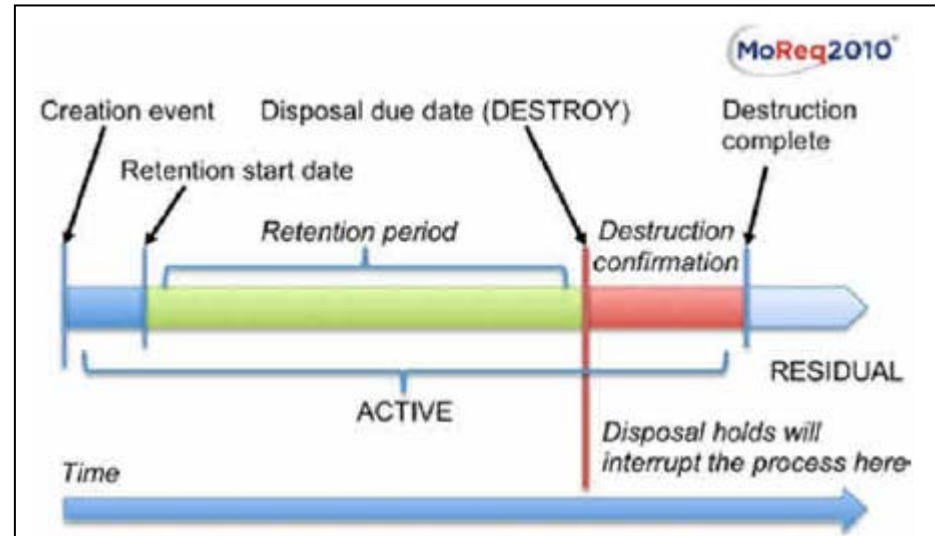
Données médicales : 10 ans
 Log : 10 ans à partir de la
 génération du log.



Sort final : exemples

- Destruction de l'objet
- La durée de conservation commence suite à un évènement (event)
- Après destruction de l'objet, des traces résiduelles sont éventuellement gardées

DESTRUCTION



REVIEW

- Au terme de la durée de conservation, la durée et le sort final doivent être révisés.
- Après validation du review, une nouvelle durée de conservation et de sort finale est définie.
- Relativement complexe à gérer.



Rôles et responsabilité

Spécifiques

- Data Producer
- Business Data Analyst
- Juriste

Transversaux

- Manager
- Expert archivage
- Administrateur système
- Gestionnaire de stockage
- Team Development
- Expert sécurité



Rôles et responsabilité

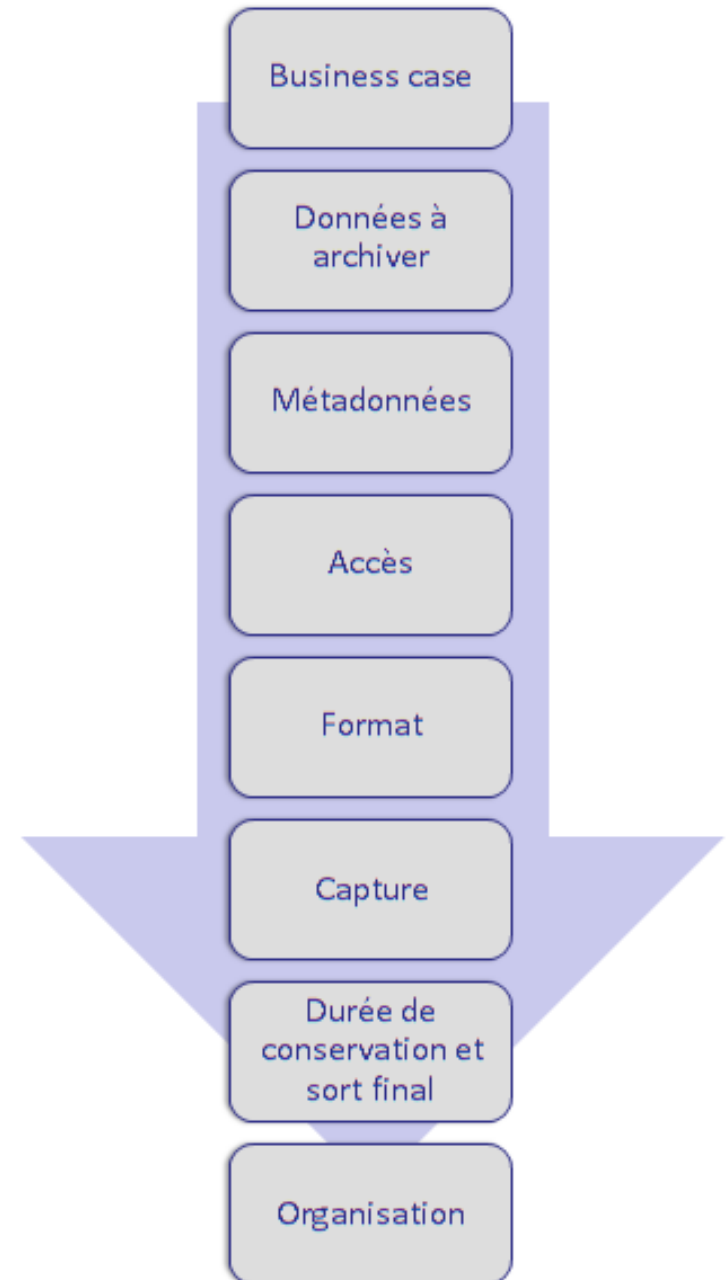
Archiviste de l'Etat

- Législation fédérale
 - Loi du 24 juin 1955 et du 6 mai 2009
 - A.R. 18 août 2010
- Toute destruction d'archive publique (sens large) est soumise à l'approbation de l'Archiviste général du royaume et de ses délégués
- Formule des avis, des conseils et des recommandations
- Détermine
 - la durée de conservation historique
 - le sort final au terme de la conservation « administrative »



Méthodologie

1. Business case
2. Données à archiver
3. Métadonnées
4. Accès
5. Format
6. Capture
7. Durée de conservation et sort final
8. Organisation



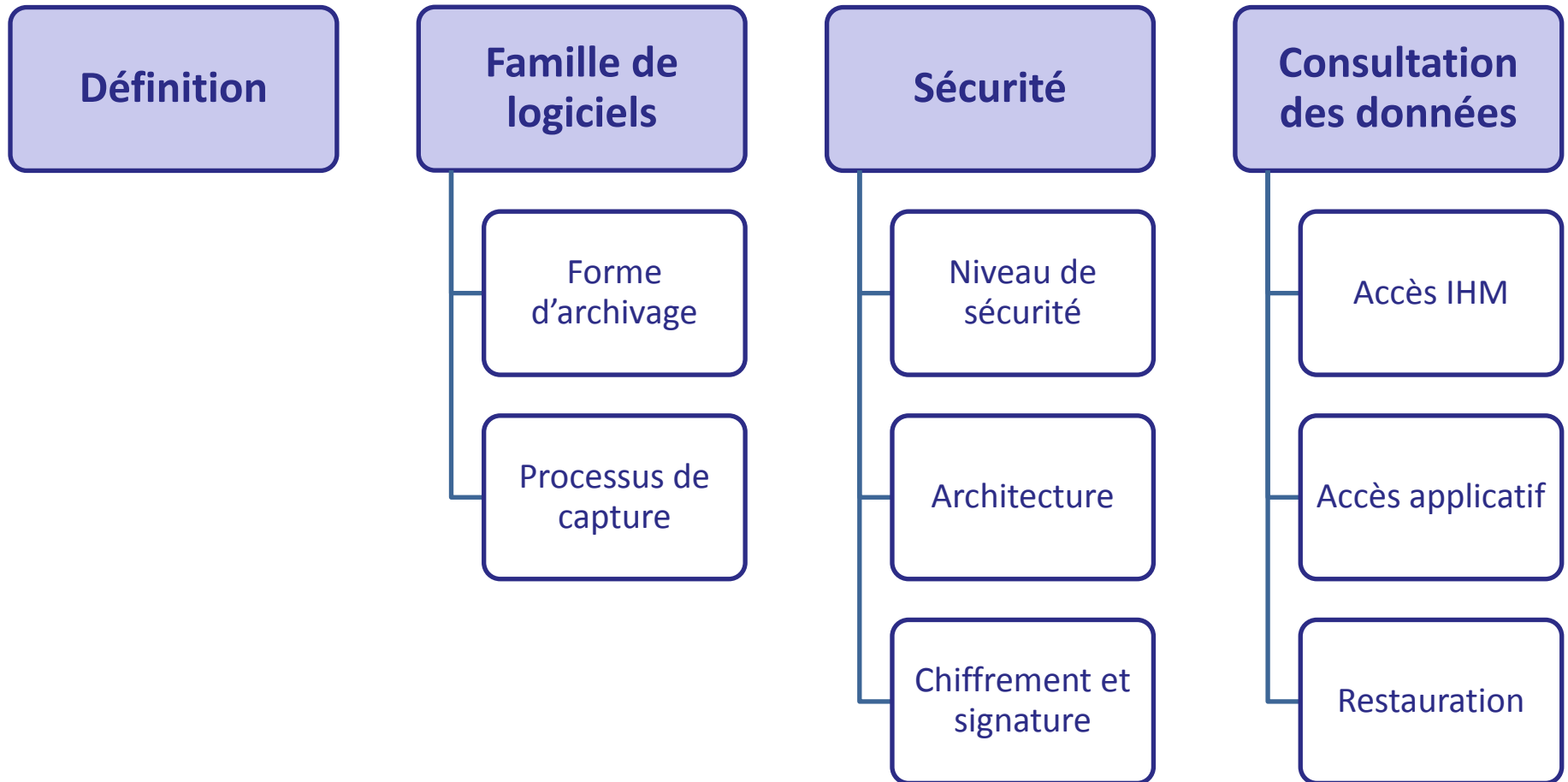
Pause



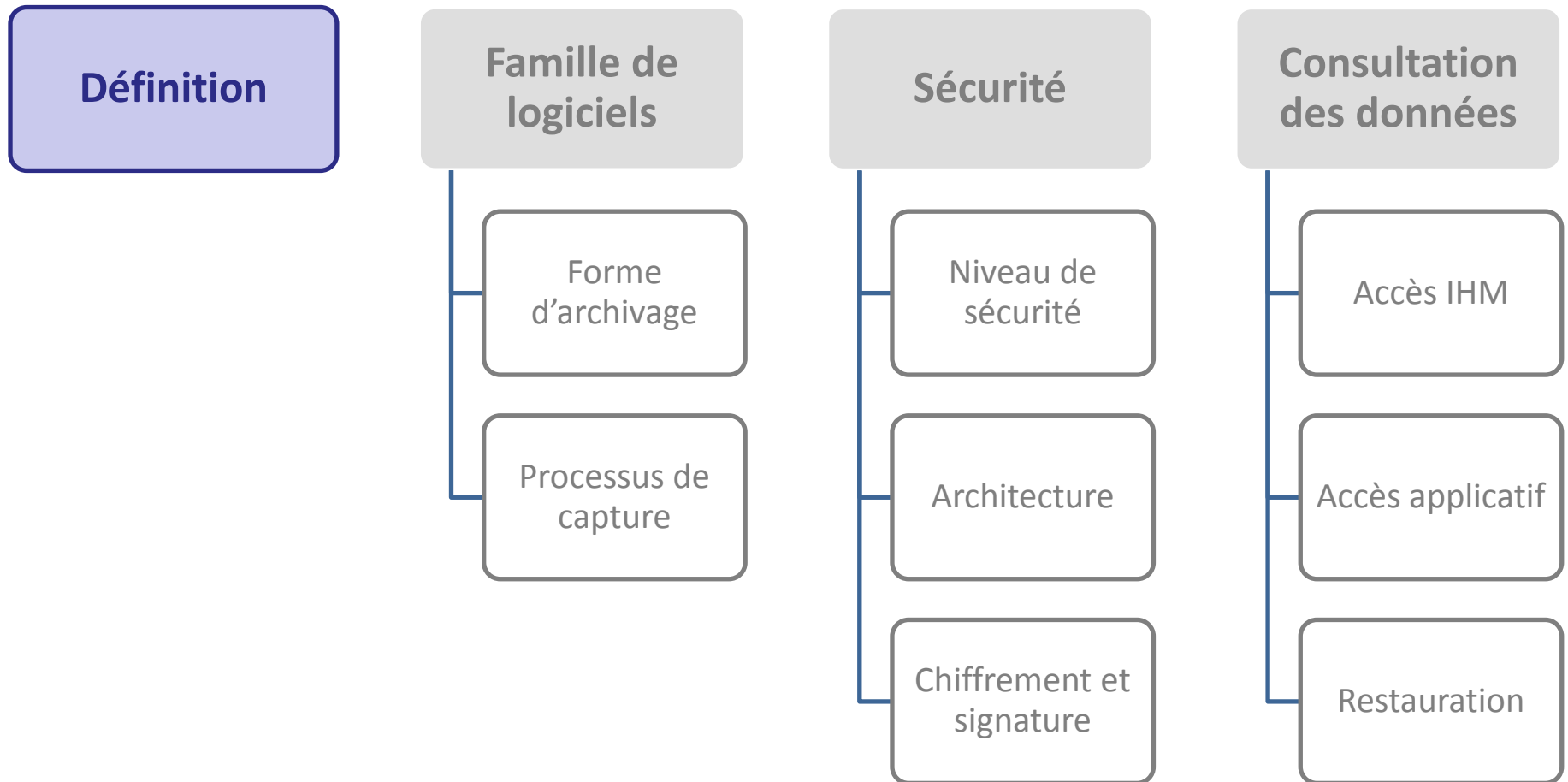


Système d'archivage électronique

Structure



Structure



Définition

Ensemble des **matériels**, **logiciels** et **procédures** qui organise et contrôle la capture, la conservation et la destruction (...) des objets (M-A chabin)

- Un SAE doit comporter certaines caractéristiques
 - **sécurité** du système (software, hardware et physique)
 - **conformité** (à la réglementation)
 - **exhaustivité** du périmètre (caractère systématique)
 - **disponibilité**
 - **traçabilité**

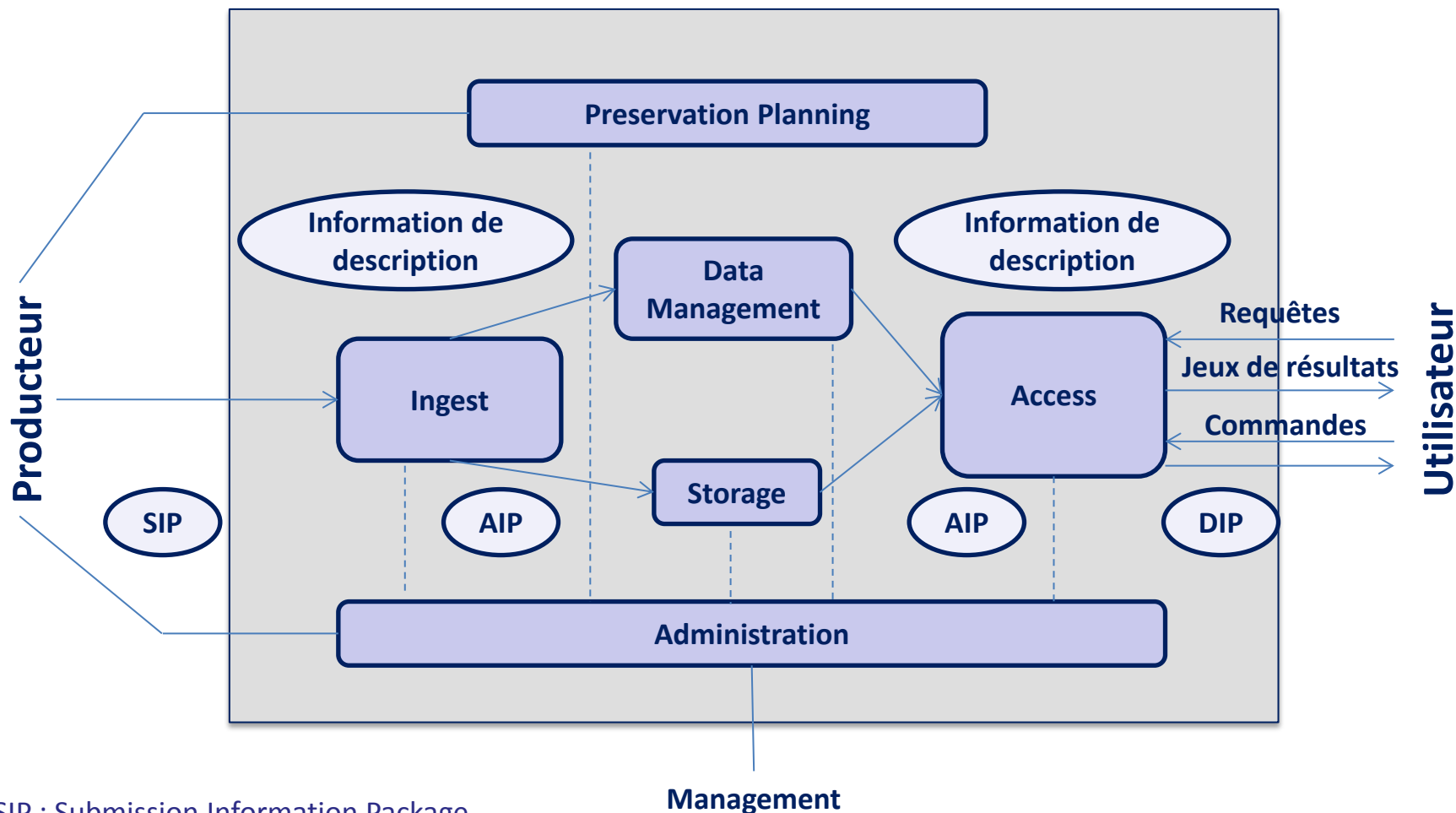


Définition

- Techniquement, stockage de données électroniques qui
 - est pérenne (long terme)
 - garantit l'intégrité des données
 - ne stocke pas uniquement le contenu mais aussi la structure, les relations et les données contextuelles
 - offre des méthodes d'accès aux données
 - permet un contrôle étroit du cycle de vie des données, y compris le versement et l'accès
 - permet le remplacement complet du matériel et des logiciels
 - s'appuie sur du matériel existant (disques et bandes)

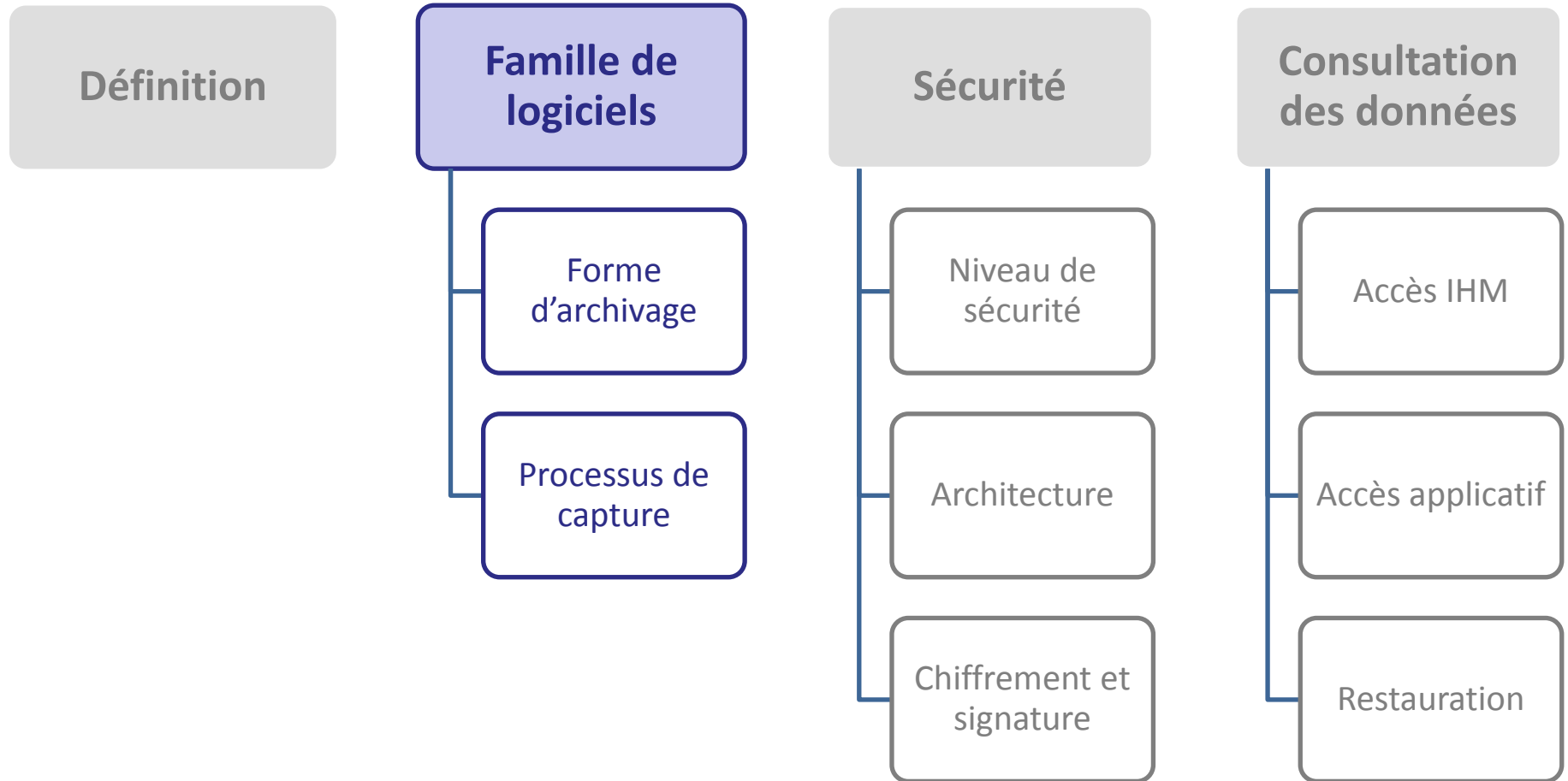


Modèle fonctionnel (OAIS)



SIP : Submission Information Package
AIP : Archival Information Package
DIP : Dissemination Information Package

Structure



Famille de logiciels

Distinction essentiellement

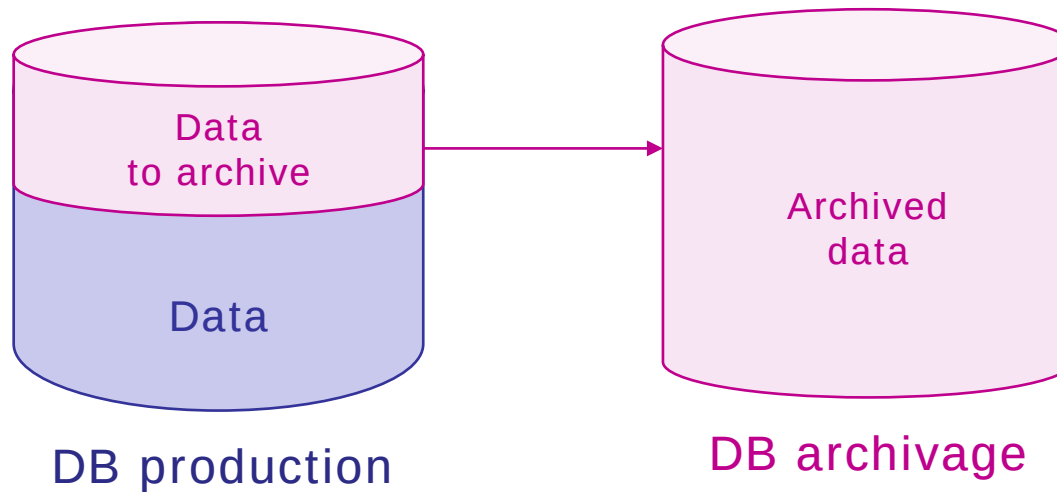
- au niveau de la **forme d'archivage**
- au niveau du processus de **capture**
- au niveau de la gestion de **l'intégrité**



Forme d'archivage

Système
d'archivage

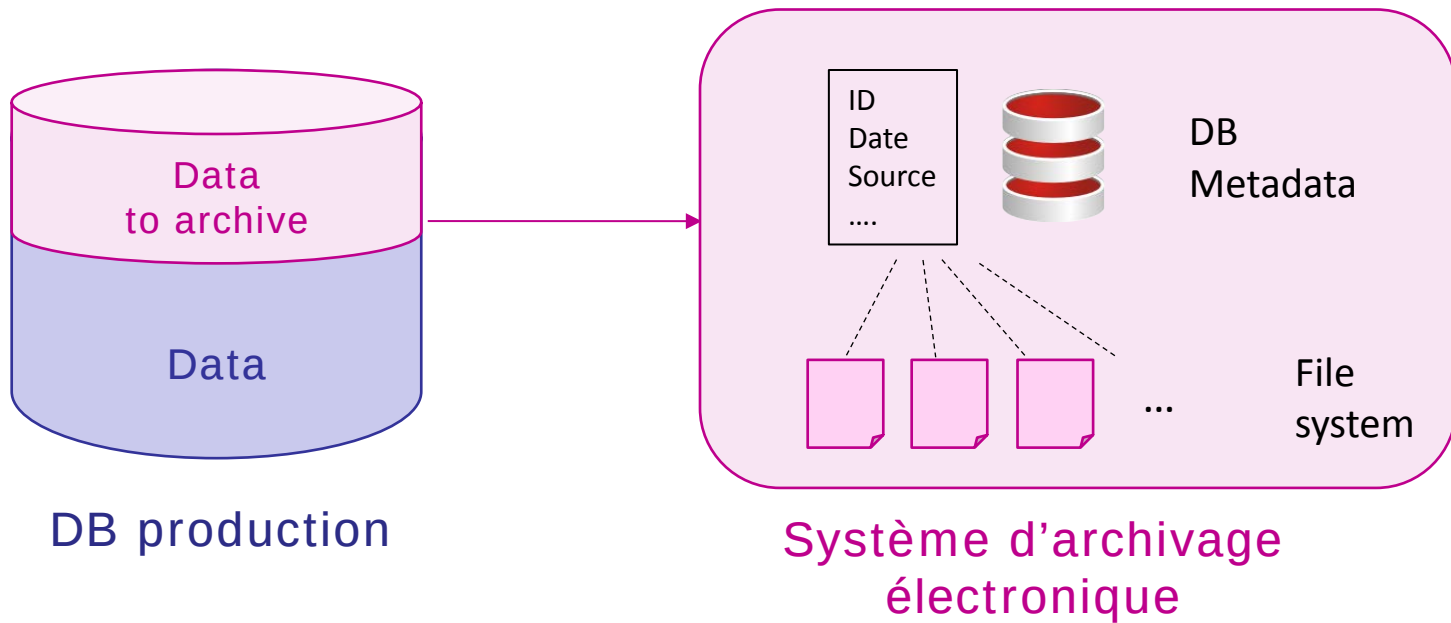
1. Archivage sous forme de base de données
conçue pour l'archivage



Forme d'archivage

Système
d'archivage

2. Archivage sous forme de fichiers



Forme d'archivage

Base de données d'archivage

- Ne permet pas de rendre les données indépendantes.
- Pas de gestion de l'évolution de la structure des données :
 - chaque modification de structure dans la db de production est répercutée dans la DB d'archivage
 - non respect de l'intégrité
- Facilite la consultation (query sur un DBMS).
- Pas une solution pour le long terme

Fichiers

- Permet de rendre les données indépendantes
- Respect de l'intégrité
- Système transversal
 - gère tout type d'objet (suite binaire)
- Consultation moins aisée et plus lente
 - soit via connecteur (couche applicative d'accès)
 - soit consultation via IHM solution d'archivage (fichiers textes)
 - soit via upload dans une base de données temporaire via composant d'extraction (rare)



Processus de capture

PUSH

- Les données sont extraites de la source pour être versées dans le système d'archivage
- Indépendant de la source de données
- La création des objets business doit être effectuée en amont

PULL

- Le système d'archivage doit procéder lui-même à l'extraction des données
- Le système d'archivage doit connaître la source et pouvoir communiquer avec elle
- La création des objets business est réalisée par le système d'archivage



Processus de capture

PUSH

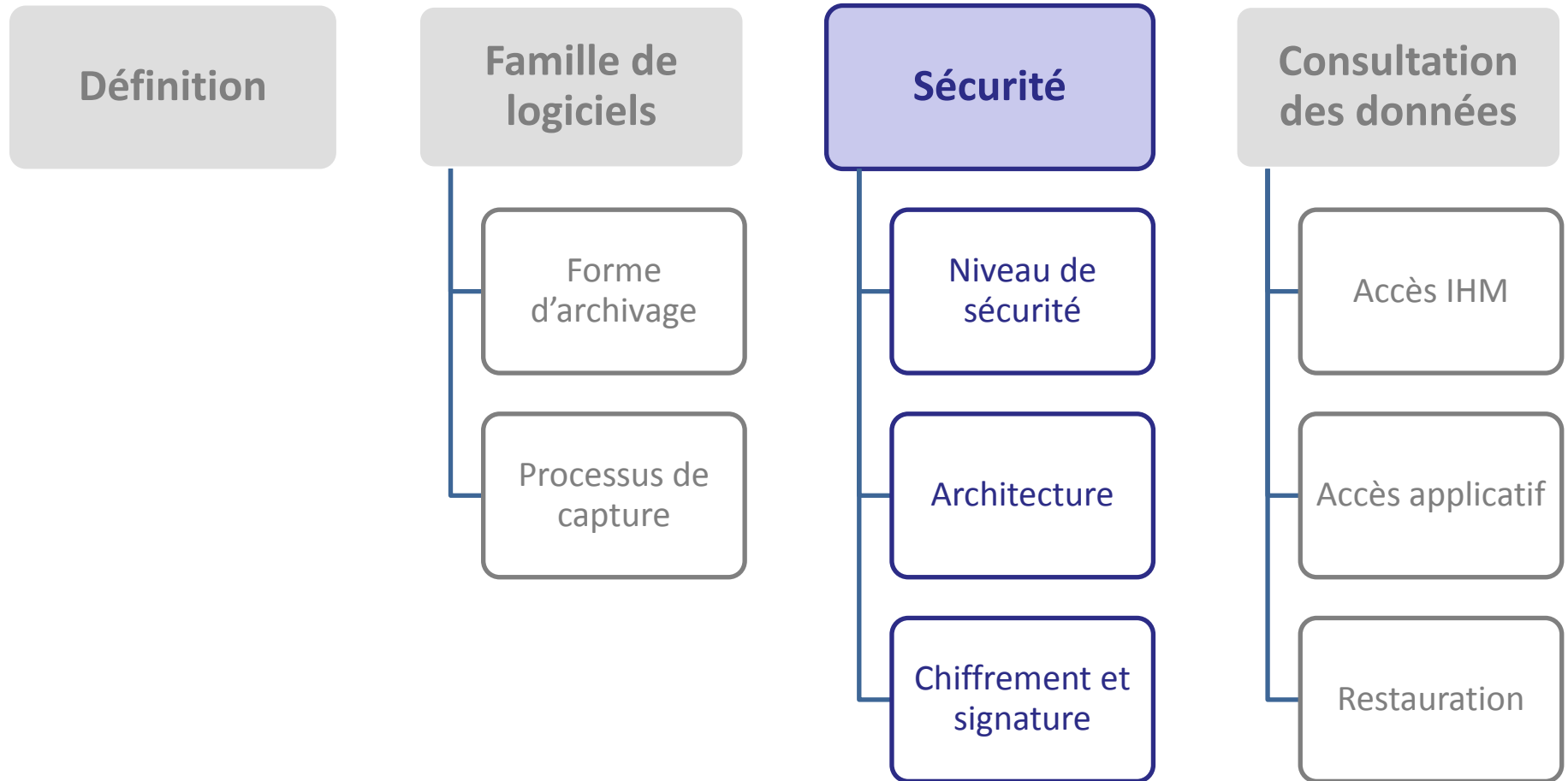
- Si le modèle de données change, pas d'adaptations requises au niveau du système d'archivage
-
- Le système d'archivage n'a pas besoin d'avoir un accès aux systèmes en production
 - Peu, voire pas, d'impact sur la sécurité d'un système

PULL

- Si le modèle de données change, le processus d'extraction doit être revu
-
- Le système d'archivage doit avoir un accès aux données en production (lecture, suppression)
 - Risque de suppression (volontaire ou non) de données de production

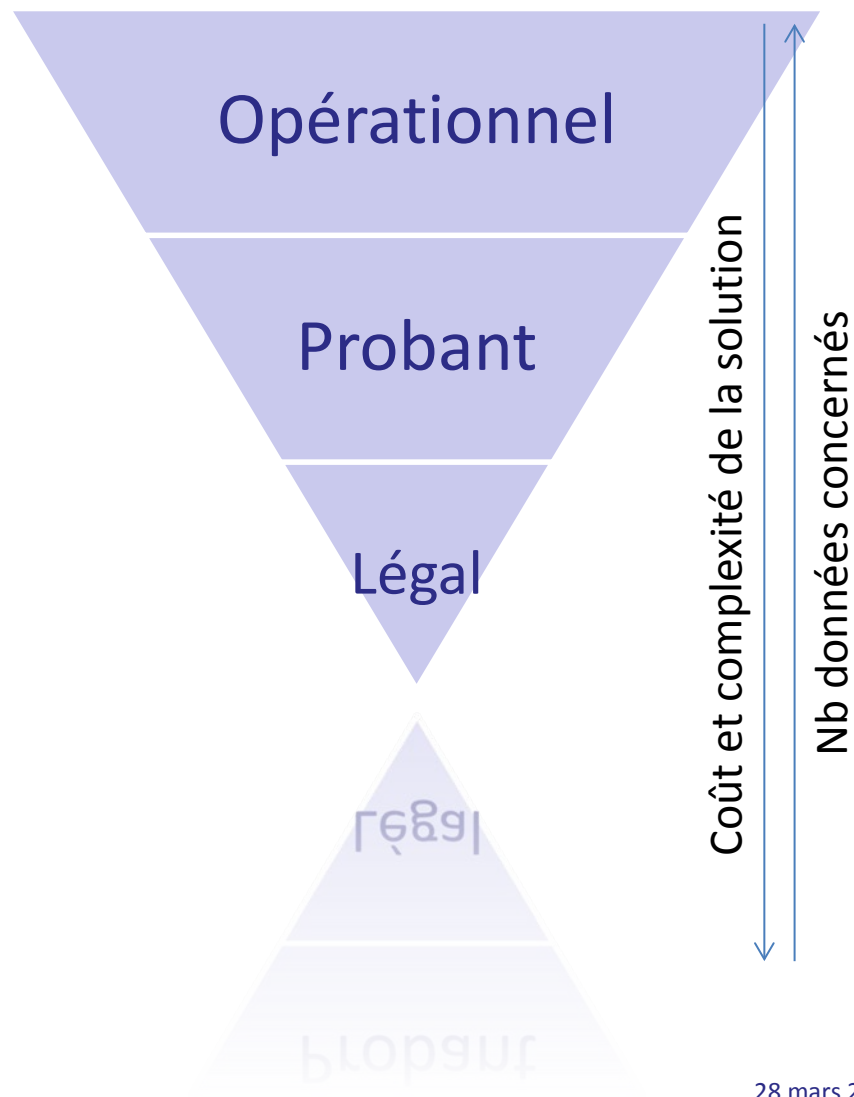


Structure



Niveau de sécurité

- Les besoins de **sécurité** en archivage varie
 - selon la **valeur des données**
 - du **risque** acceptable
- Définition d'un « **niveau de sécurité** »
- Elaboration d'une **architecture** correspondante
- En pratique, 9 projets sur 10 requièrent au minimum un niveau probant.



Architecture - Niveau probant ou légal

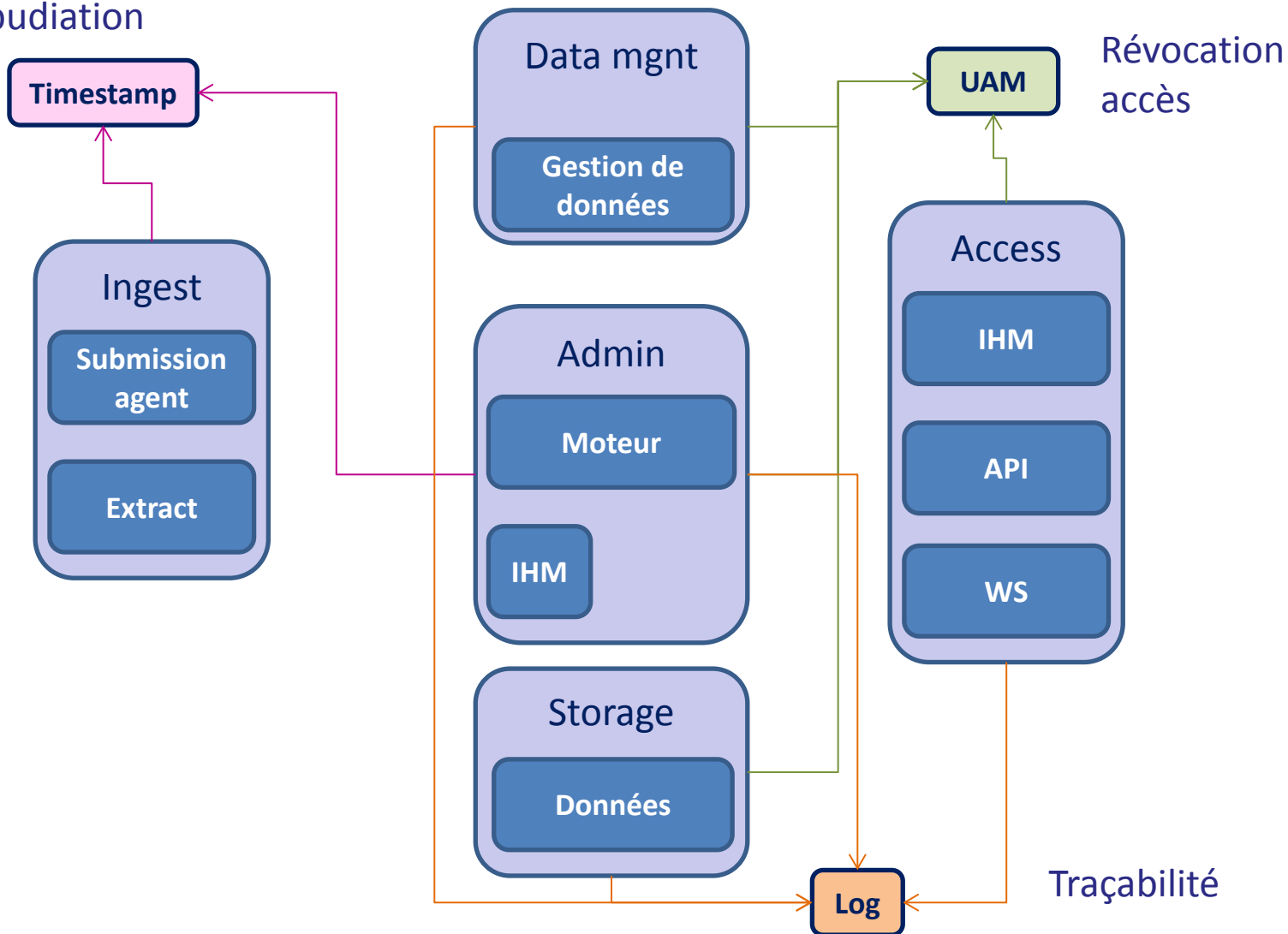


- Objectifs :
 - Retrouver des données en cas de besoin (litige)
 - Garantir qu'elles n'ont pas été altérées suite à un problème logique ou physique ou suite à une manœuvre intentionnelle
 - Prouver qu'une donnée bien précise dont le propriétaire est connu a existé à un moment donné



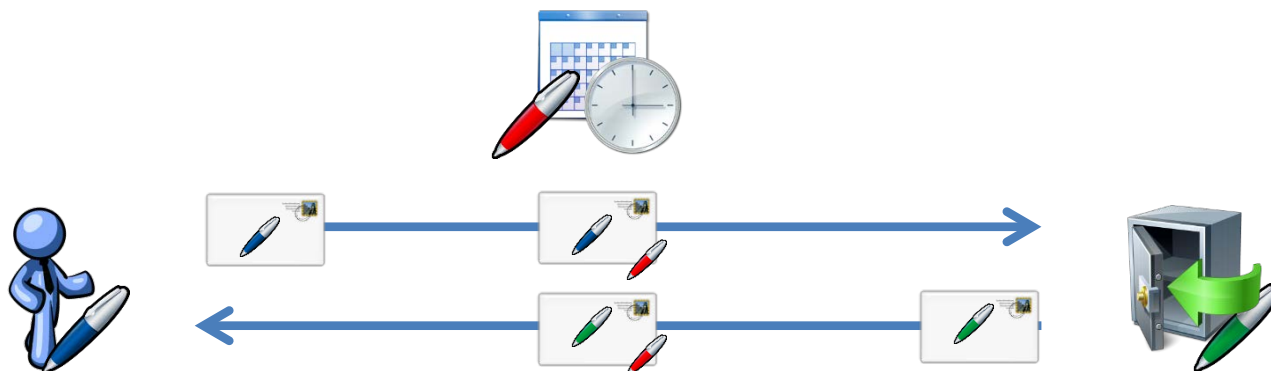
Architecture - Niveau probant ou légal

Non répudiation

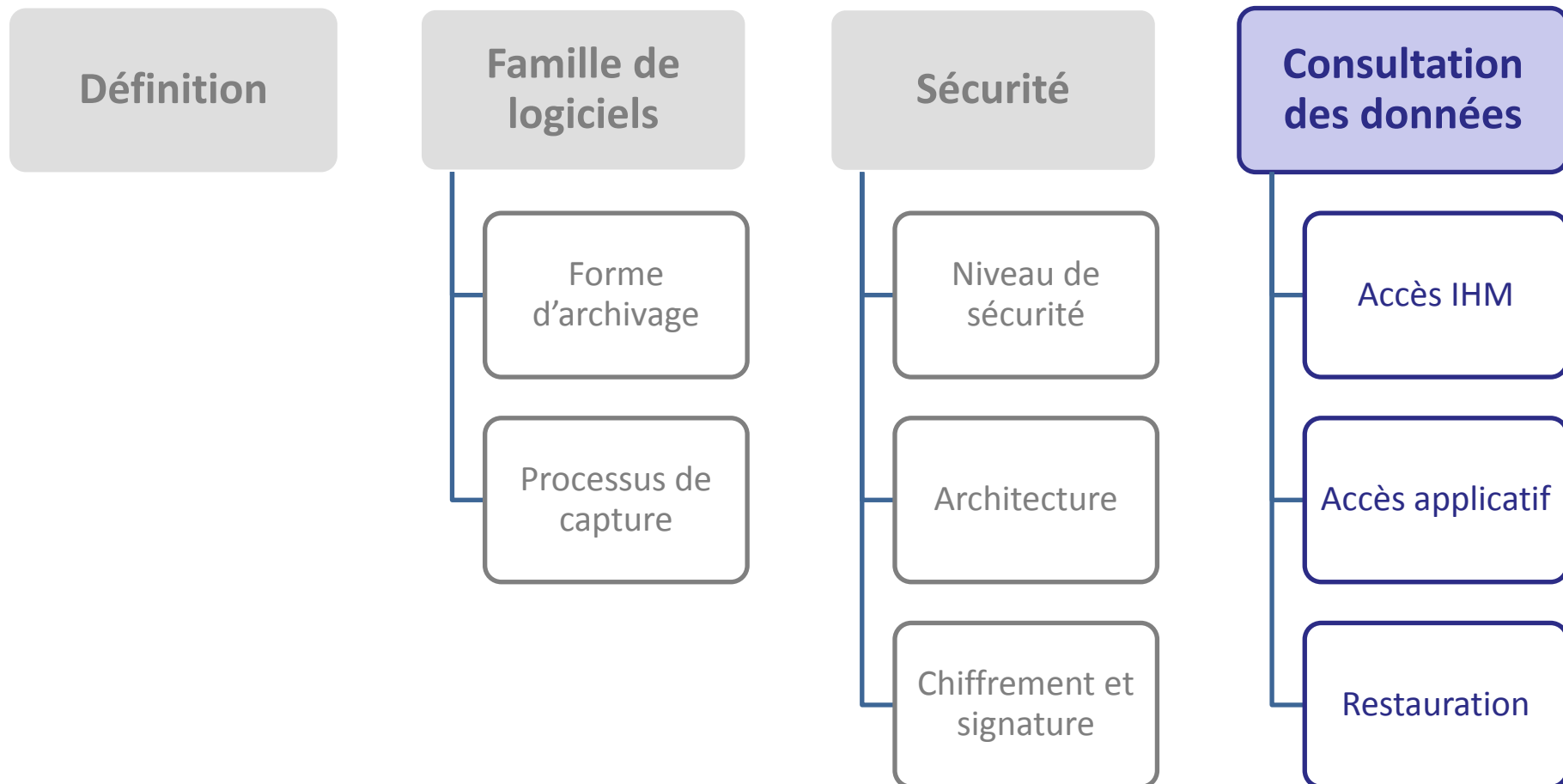


Sécurité – signature digitale

- La non-répudiation n'est pas facile à gérer en pratique
- Une signature est requise
- Un système de timestamping externe est souhaitable
- Idéalement, la validité de la signature apposée par le timestamp devra être supérieur à la durée de conservation des archives...
- ... si ce n'est le cas, il faudra resigner les données



Structure

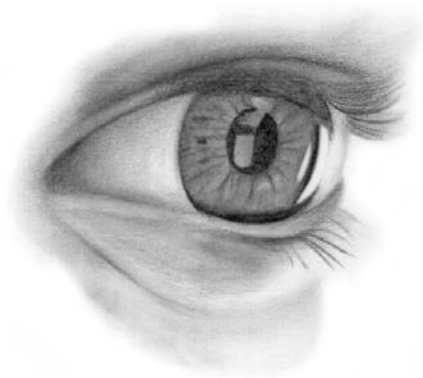


Accès IHM

- Via interfaces de la solution d'archivage
 - Utilité si accès
 - Peu fréquent
 - Par personne n'ayant pas (ou pas besoin) d'un accès via une application
 - Indépendamment d'une quelconque application (indépendance)
 - Indexation via moteur intégré ou via moteur externe



Accès applicatif



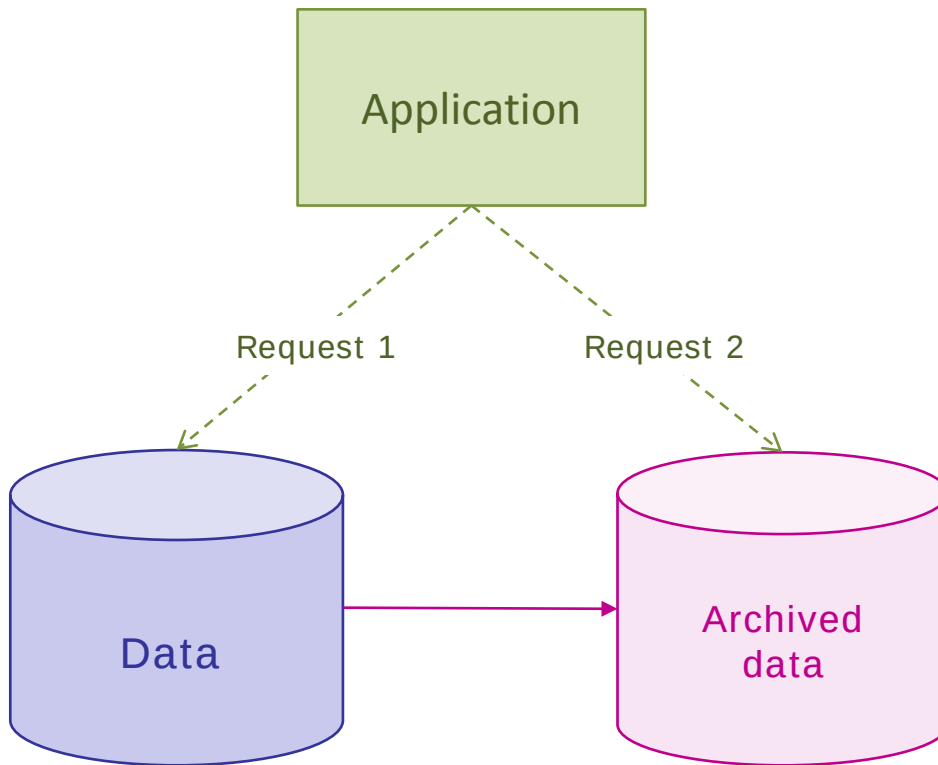
Nécessaire si

- une application est requise ou souhaitée pour consulter les données (e.a. pour déterminer les critères de recherches)
- une recherche fédérée dans les données de production et les données archivées est souhaitée



Accès applicatif

1. Archivage dans une base de données



Processus

1. L'application interroge la DB de production.
2. Si elle ne trouve pas, elle interroge la DB d'archivage

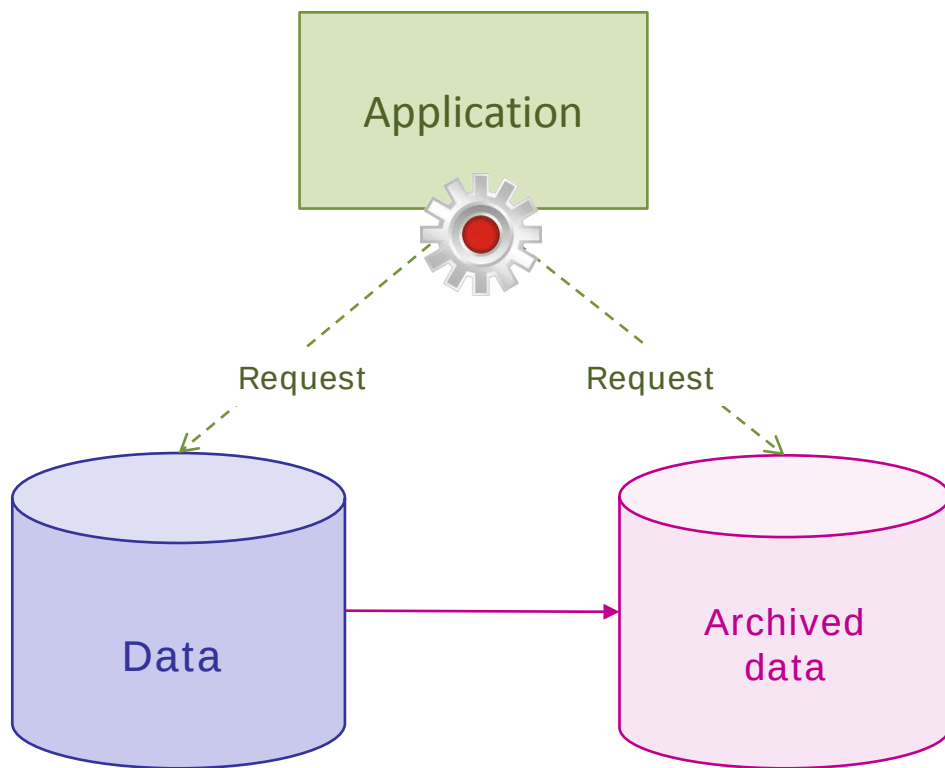
Remarques

- 2 requêtes (si la requête 1 échoue)
- Si accès « archived data » fréquent, impact important sur les performances



Accès applicatif

1. Archivage dans une base de données



Processus

1. L'application connaît la logique de l'archivage.
2. Elle sait donc interroger directement la bonne source.

Remarques

- 1 seule requête
- Logique à coder (ou recours à BRE) dans l'application.
- Logique parfois difficile à formaliser.

BRE = Business Rules Engine

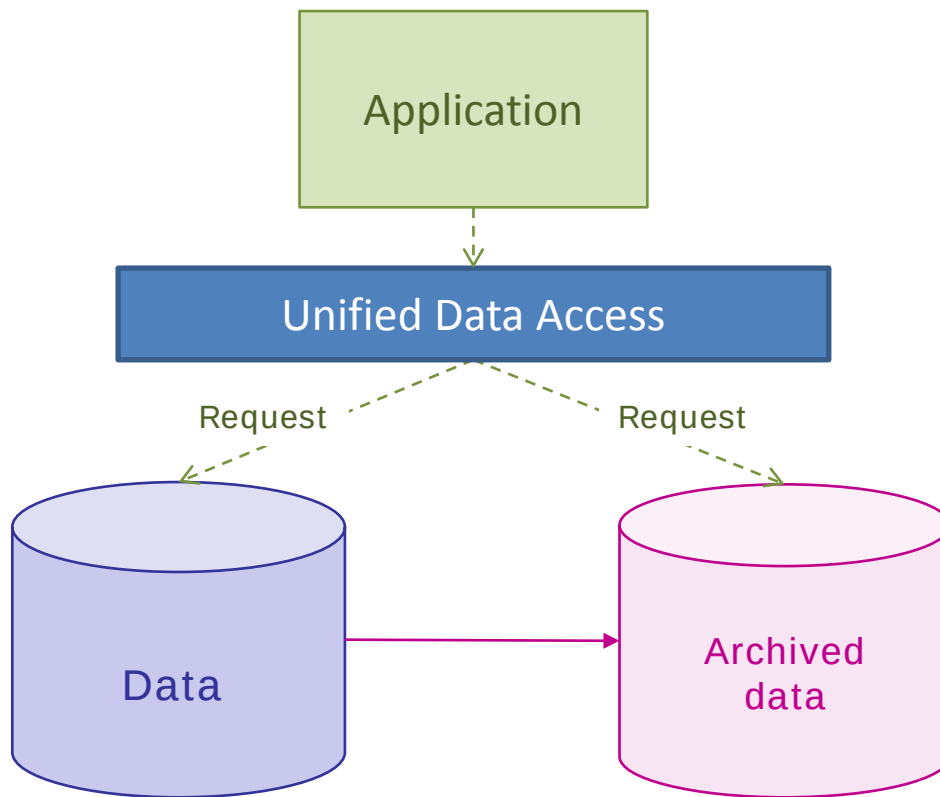
G. Ogonowski, *Business Rules Technologies & web sémantique : La gestion des règles business*, Section Recherches, Smals, février 2008.

<http://documentation.smals.be/>



Accès applicatif

1. Archivage dans une base de données



Processus

1. Une couche « Unified Data Access » gère l'interrogation des sources
2. Elle renvoie une vue unique à l'application

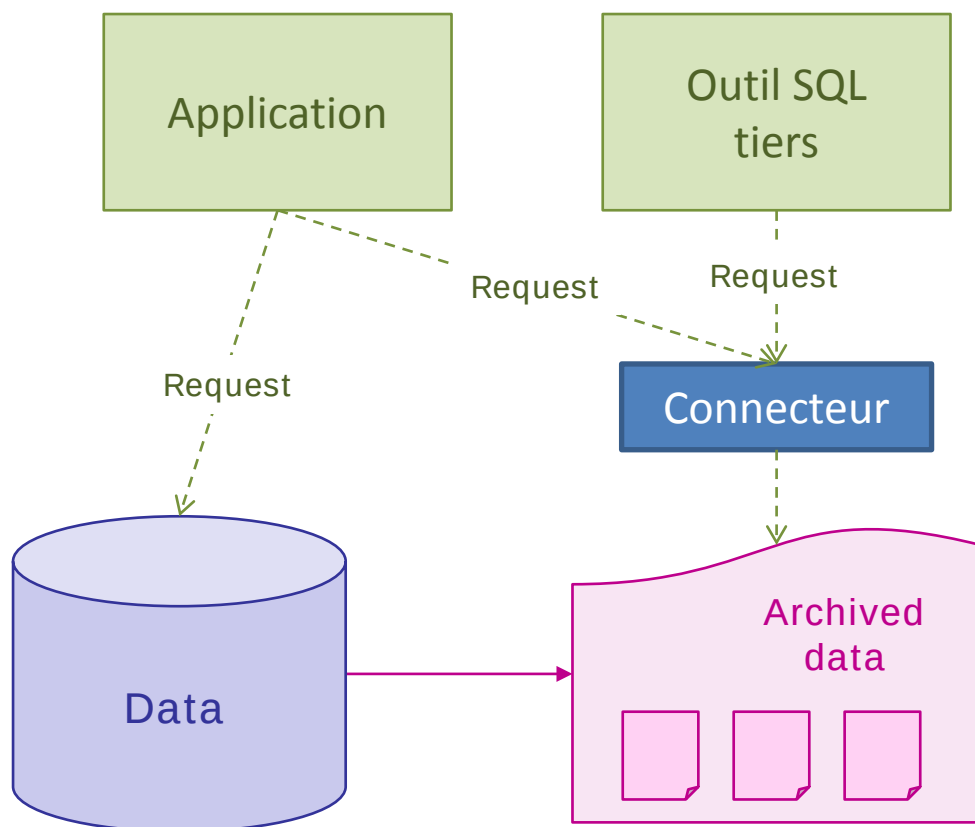
Remarques

- Très paramétrable
- Souplesse en cas de multiplication des sources
- Indépendance par rapport à l'application source (ODBC/JDBC)
- Couche à acheter/développer



Accès applicatif

2. Archivage sous forme de fichiers



Processus

- L'application contacte le connecteur qui gère l'interrogation des fichiers.

Remarques

- Permet exploitation complexe des données archivées.
- Pas de vue unifiée sur les données.

Format Archived Data

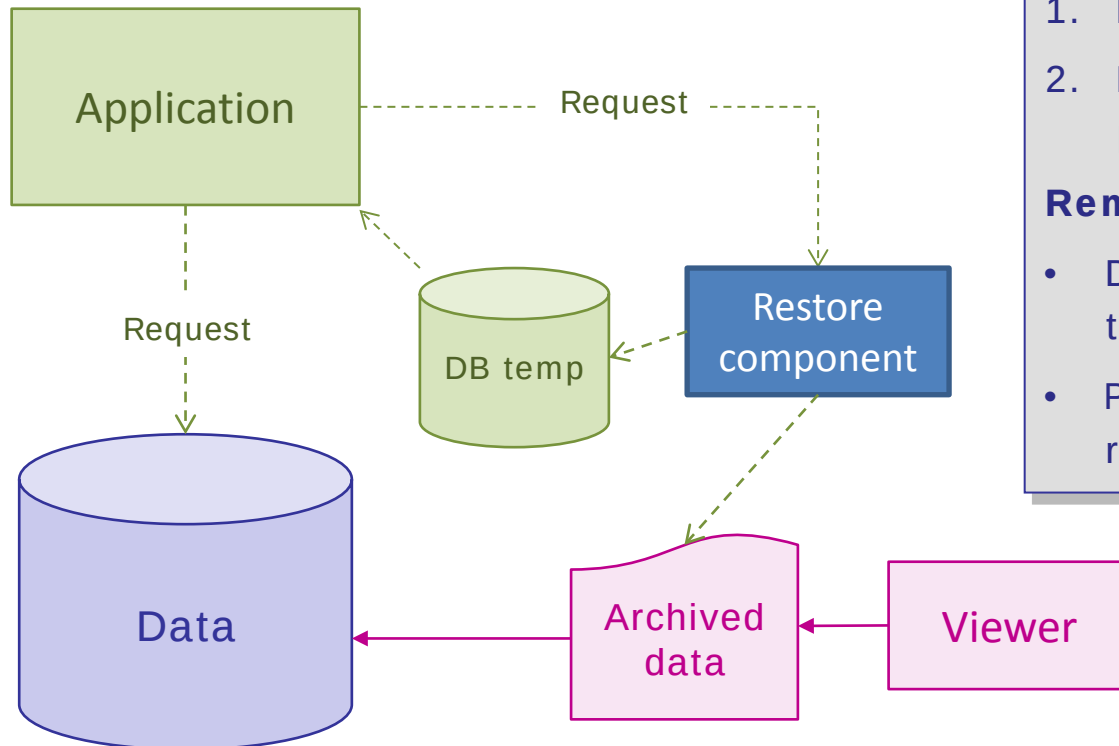
1. XML/CSV
2. SQLite

Connecteur

1. Transformation query SQL en Xquery
2. Librairie SQLite



Restauration



Processus

1. Retrouver les objets business
2. Les restaurer dans une DB pour exploitation

Remarques

- DB de production ou DB temporaire
- Possible mais peu recommandé





Market overview

Tableau récapitulatif

	IBM Optim	Arcsys	HP AIO-TRIM	Informatica
Scope				
Type de capture				
Extraction				
Forme d'archivage				
Intégrité				
Gestion cycle de vie				
Consultation				
Format des archives				



Market overview

4 solutions analysées :

- IBM Optim (Pull)
- Arcsys (Push)
- HP AIO - Trim (Pull)
- Informatica (Pull)

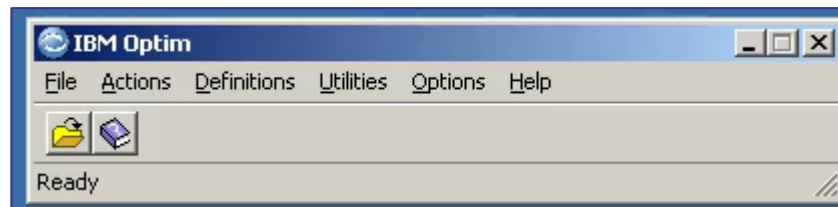


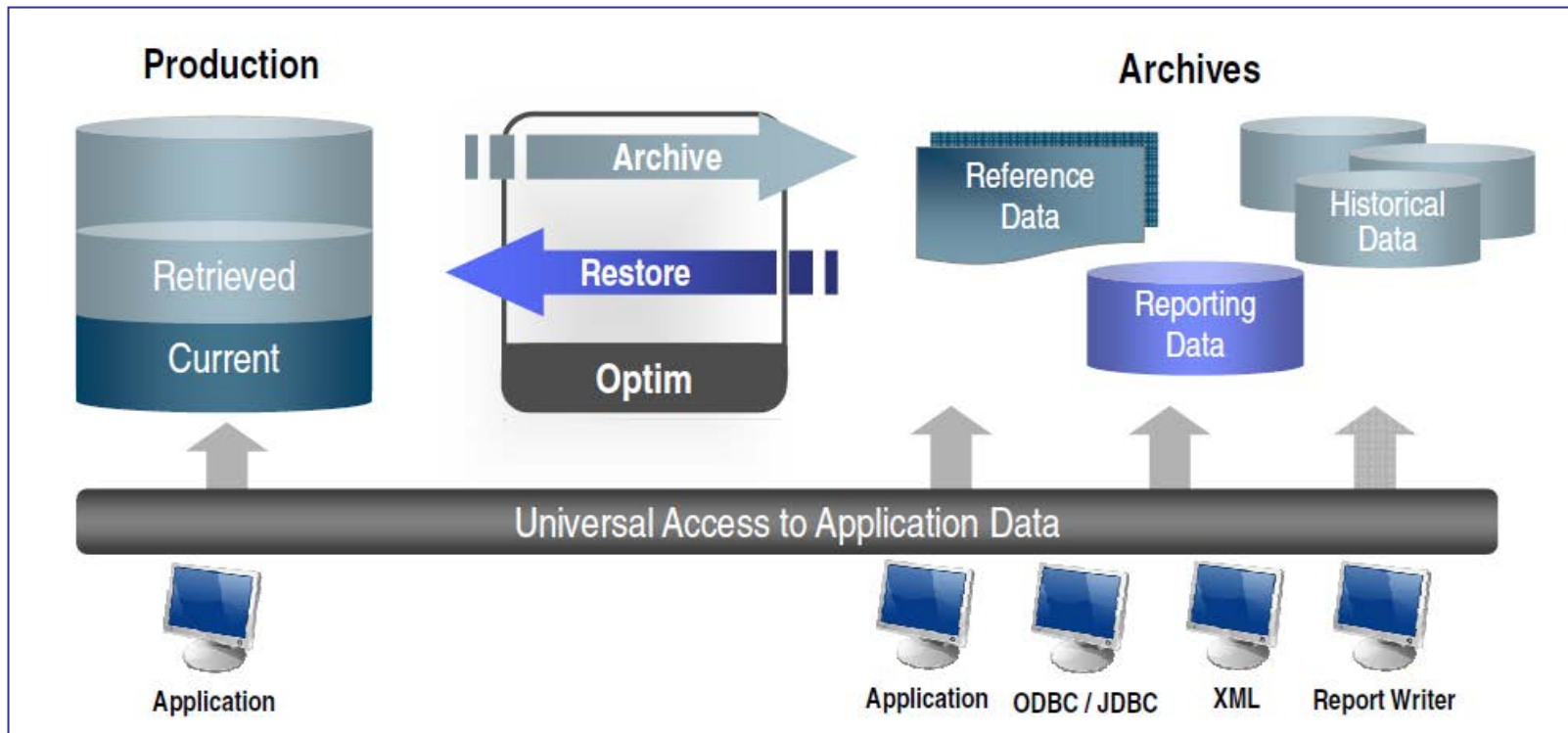
... d'autres solutions existent sur le marché



IBM Optim (Data Growth Solution)

- Objectif :
 - alléger les bases de données de production
- Caractéristiques :
 - Extraction de données depuis une ou plusieurs bases de données
 - Archivage de ces données dans un format compressé et propriétaire
 - Consultation des données à partir du logiciel d'archivage et support ODBC / JDBC

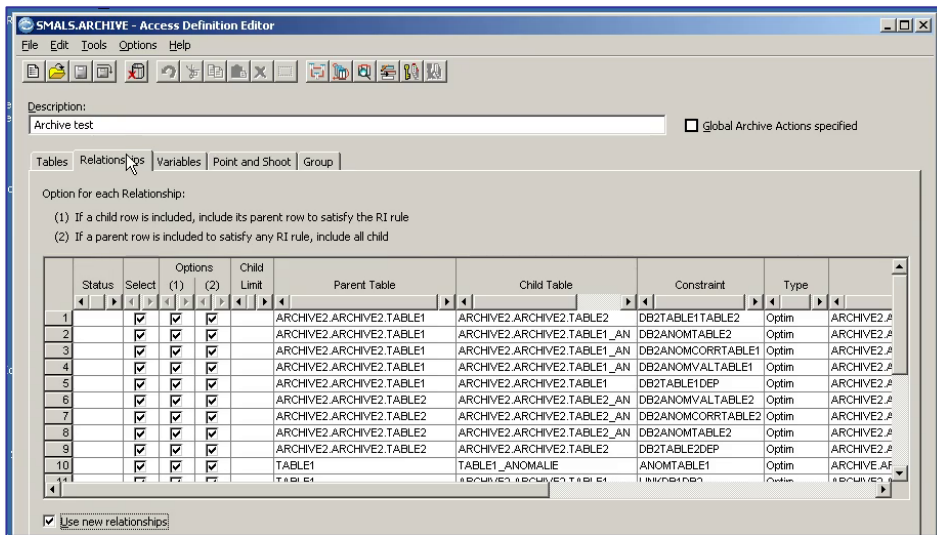




IBM Optim

- Architecture relativement simple
- Protections logicielles contre la modification

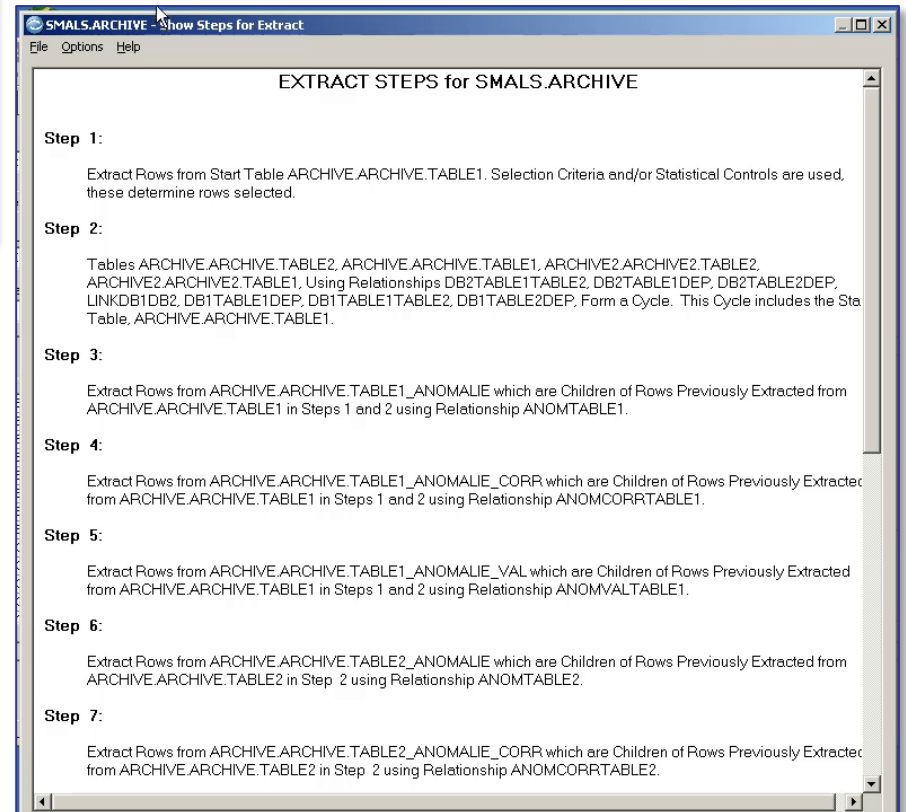
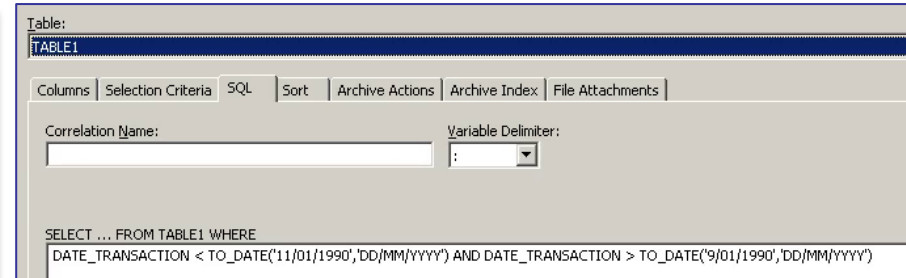




IBM Optim

Sélection des objets à archiver :

- Définition de la procédure d'extraction complexe
- Mécanismes de parcours de tables assez puissants



IBM Optim (Data Growth Solution)

- Constatations :
 - Possibilités d'extraction évoluées (modèle PULL)
 - Analyse du processus d'extraction affichée
 - Interfaces peu conviviales
 - Intégrité à gérer au niveau hardware
 - Pas de gestion du cycle de vie (durée de conservation)



Arcsys



- Objectifs :

- Préservation des données et vérification de l'intégrité
- Traçabilité de toutes les opérations effectuées

Actions	Status	Date
	Initialized	27/02/2012 13:36:52
	Error	27/02/2012 13:36:53
		27/02/2012 13:36:57

Code=[4017/188a7d5] - Le serveur de transfert Arcsys a rencontré un echec sur le test d'intégrité [SHA1] pour le lot [8] et séquence [2].

- Caractéristiques :

- Utilisation d'un format ouvert et documenté pour le conteneur des archives : tar.gz
- La description des données archivées est stockée au format XML
- Archives exploitables même sans le logiciel

```
<OBJECT ID="268">
  <TYPE>FILE_OF_DIRECTORY</TYPE>
  <OBJECT_PARENT_ID>267</OBJECT_PARENT_ID>
  <FILE_EXTENSION>PDF</FILE_EXTENSION>
  <IS_NATIF>true</IS_NATIF>
  <SIZE>65</SIZE>
  <TAR_NUMBER>2</TAR_NUMBER>
  <HASH_FUNCTION>SHA1</HASH_FUNCTION>
  <HASH_VALUE>88640c4d10b34121185cb715fce95bed0b96b97f</HASH_VALUE>
  <LOGICAL>/home/recherche/For_Archives/manual_model_oss-fr.pdf</LOGICAL>
  <PHYSICAL>file1</PHYSICAL>
  <OBJECT_DESCRIPTOR>file1.arcsys</OBJECT_DESCRIPTOR>
  <LAST_MODIFIED>2012/02/27 13:05:32</LAST_MODIFIED>
  <UID>1000</UID>
  <GID>100</GID>
  <RIGHTS_ACCESS>420</RIGHTS_ACCESS>
  <LAST_ACCESS>2012/02/27 13:06:18</LAST_ACCESS>
  <LAST_CHANGING_STATUS>2012/02/27 13:06:00</LAST_CHANGING_STATUS>
  <FILE_TYPE>R</FILE_TYPE>
  <META_DATA_LIST/>
</OBJECT>
```



localhost:8080/webarcsys/TitRecupListeObject.jspa?acces=init

Arcsys V4.5.3.0 Archivage et Restitution

Object List the objects

arcsyssuper

« Back to home

Actions

Select an agent
Archive
Restore
View requests

Administration

Profiles
Keywords
Masks
DB Operations

Entities

Classification scheme
Collections
Lots
Objects

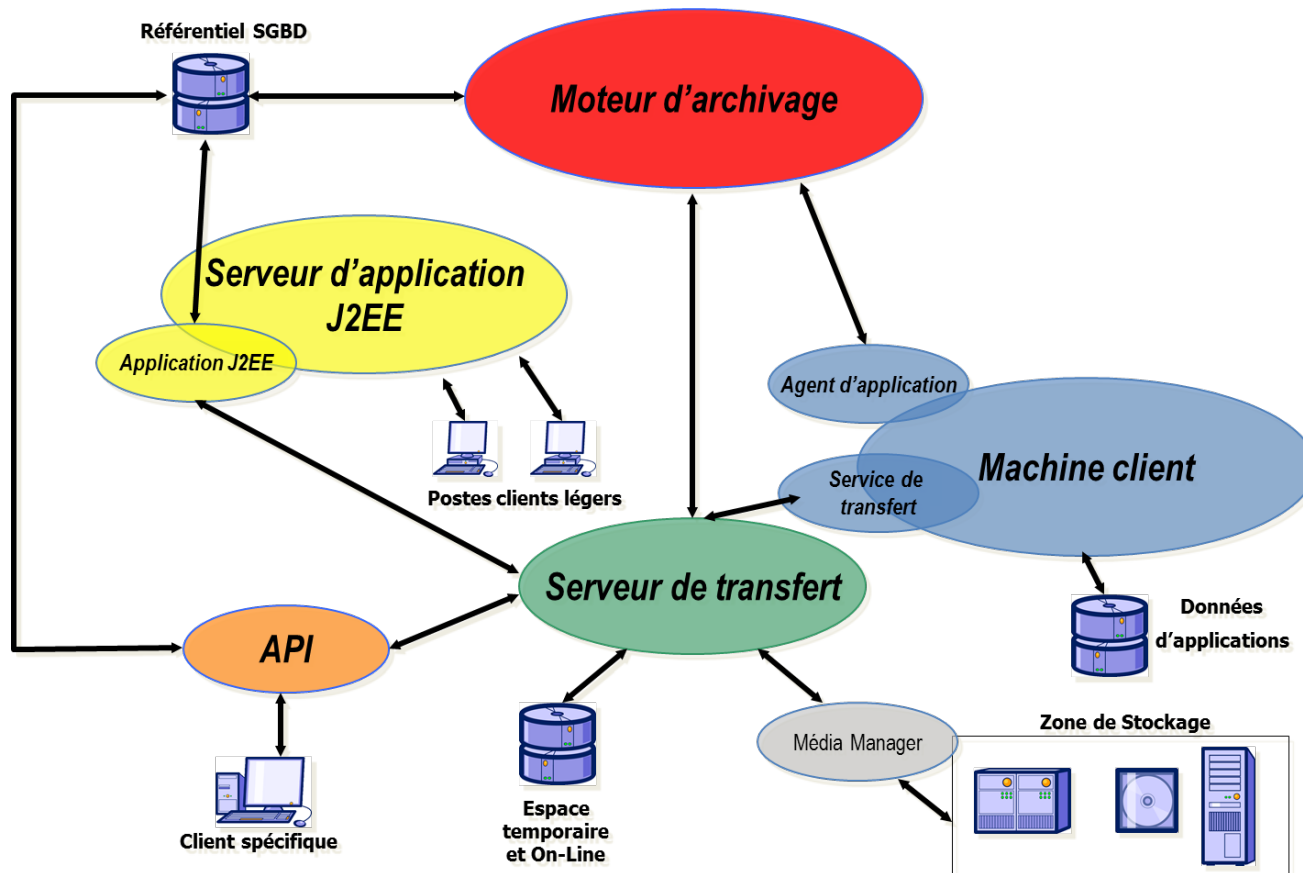
Folder /home/recherche/For_Archives

Id	Path	Type	Size	Metadata	Delete
267	/home/recherche/For_Archives	Folder	36769 kB		
Files archived of this folder					
Id	Path	Type	Size	Native	
269	/home/recherche/For_Archives/arcsys_manuels/Arcsys-administration-4.5.3-fr.pdf	File	1920 kB	yes	
277	/home/recherche/For_Archives/arcsys_manuels/Arcsys-api-4.5.3-fr.pdf	File	910 kB	yes	
270	/home/recherche/For_Archives/arcsys_manuels/Arcsys-compatibility-matrix-4.5.3-fr.pdf	File	259 kB	yes	
272	/home/recherche/For_Archives/arcsys_manuels/Arcsys-error-4.5.3-fr.pdf	File	330 kB	yes	
276	/home/recherche/For_Archives/arcsys_manuels/Arcsys-functional-description-4.5.3-fr.pdf	File	499 kB	yes	
271	/home/recherche/For_Archives/arcsys_manuels/Arcsys-installation-4.5.3-fr.pdf	File	9254 kB	yes	
273	/home/recherche/For_Archives/arcsys_manuels/Arcsys-presentation-4.5.3-fr.pdf	File	5849 kB	yes	

Arcsys – Interface des versements

- Versement adhoc via interface Web
- Industrialisation via zone de dépôt et WS

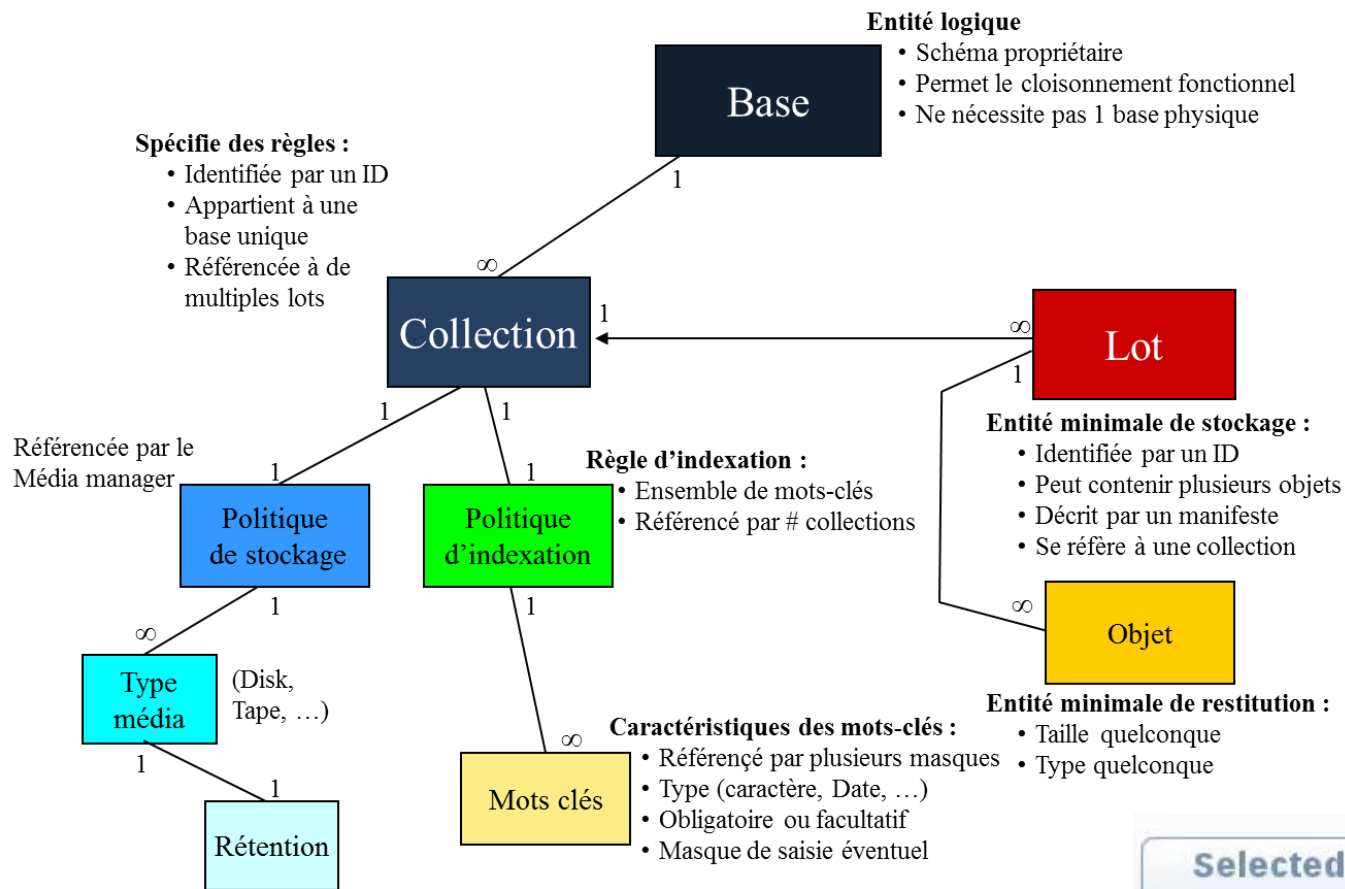




Arcsys – Architecture et composants

Nombreux composants à répartir sur diverses machines en fonction du niveau de sécurité souhaité





Arcsys – Data Model

Modèle de données bien décomposé

Compréhension de ce modèle requise pour utiliser le logiciel

Selected elements

Base	AdminBase
Coll	LogDefault
Lot	Lot Greg 007
Age	linux-jv7x.site:50030



Arcsys

- Constatations :
 - Logiciel complet (modèle Push)
 - Modèle Pull avec des DB disponible, mais mécanismes d'extraction assez limités
 - Archivage sous forme de fichiers uniquement
 - Versement via interface Web fastidieux, processus industrialisable via l'injecteur paramétrable ou les services Web
 - Recherche sur base de meta-data ou full-text
 - Consultation via outils tiers (pas de viewer intégré)





HP AIO - TRIM

- AIO : Application Information Optimizer
- TRIM : Logiciel de Record Management
- Objectif :
 - Extraire des données de bases de données
 - Permettre la recherche dans les données archivées
 - Garantir l'intégrité des données
- Caractéristiques :
 - Stockage des données dans une DB ou dans des fichiers compressés
 - Recherche possible dans les fichiers au moyen d'un connecteur ODBC/JDBC
 - Possibilité d'ajouter des scripts à tous les niveaux du processus d'extraction



The screenshot displays the Smals application interface. At the top, there is a menu bar with options: Orders, Orders_D2F_SRMS, *Orders_D2F_BF_SRMS, Undo_Orders_DW_D2F_BF, Undo_Orders_D2F_BF, and Orders_DW_D2F_BF. On the left, a sidebar contains a 'Select' button and several icons with labels: Archive, Reload, Upload, IDOL Index, Groovy Script, Condition, and Interrupt. The main workspace shows a workflow diagram with three steps: a 'Start' node (green circle with a downward arrow), an 'Orders_D2F_SRMS Copy' node (database icon with a green arrow), and a 'Call_SRMS_L' node (document icon with a pencil). A 'Groovy Script' dialog box is open on the right, titled 'Insert a Groovy script.' It contains a 'Name' field with the value 'Call_SRMS_Loader' and a 'Code' field with the following Groovy script:

```
//This call loads the output generated by all cartridges
//within the current business flow into HP TRIM.

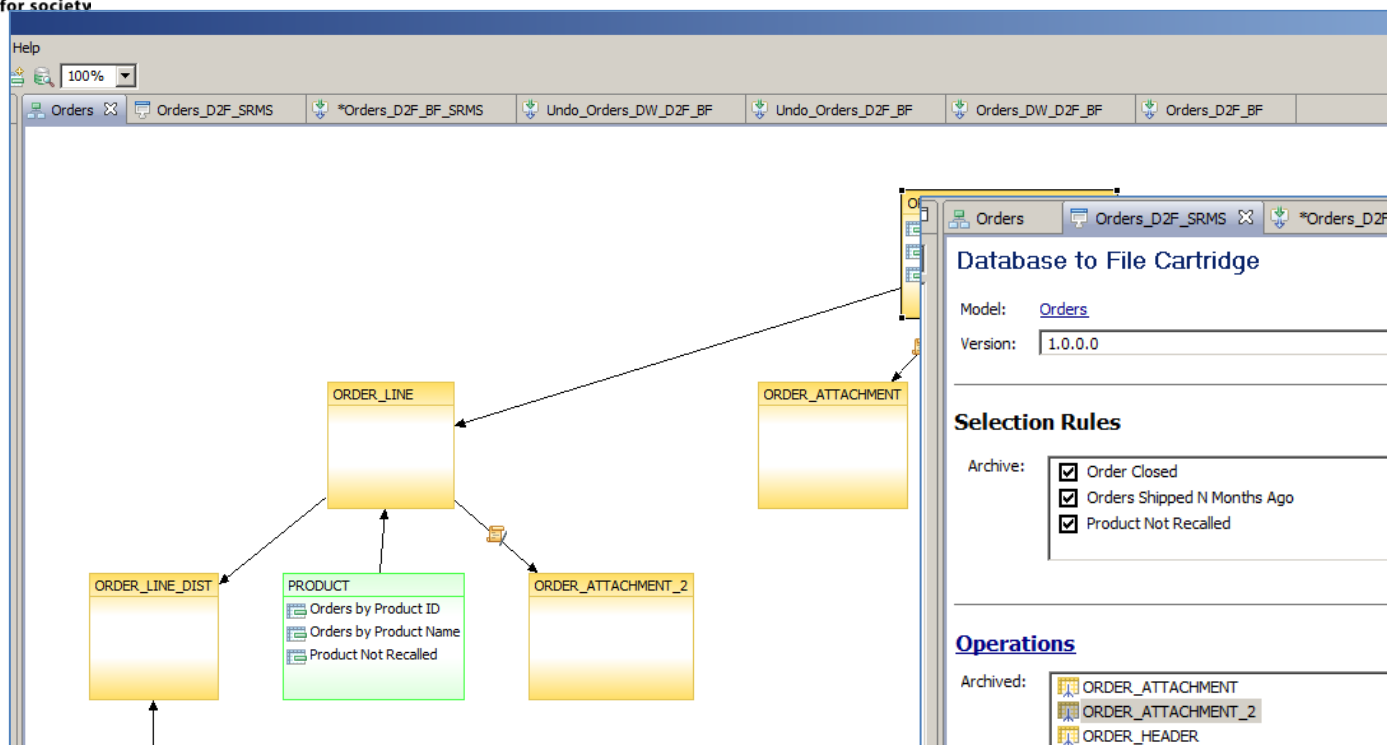
import groovy.sql.Sql
import com.hp.ilm.db.extensions.srms.infrastructure.*

SrmsMetadata.sendToSRMS("TRIM", CURRENT_GROUP_RUN_ID,
    ENVIRONMENT_NAME, REPOS_DB,
    SrmsMetadata.getDefaultConfigFile(),
    "SYNCHRONOUS", "DEFAULT_PATH", "LOCAL")
```

HP AIO : Processus d'extraction

Ajout d'un script Groovy





HP AIO - Sélection des objets à archiver

- Sélection des tables aisée
- Définition de règles avancées (ex : anonymisation de certains champs)

Database to File Cartridge

Model: [Orders](#)

Version: 1.0.0.0 [Annotation...](#)

Selection Rules

Archive:

- ☒ Order Closed
- ☒ Orders Shipped N Months Ago
- ☒ Product Not Recalled

Operations

Archived:

- ☒ ORDER_ATTACHMENT
- ☒ ORDER_ATTACHMENT_2
- ☒ ORDER_HEADER
- ☒ ORDER_LINE
- ☒ ORDER_LINE_DIST
- ☒ ORDER_TAX

Purged: No Table Uses Selected.

Data Masking

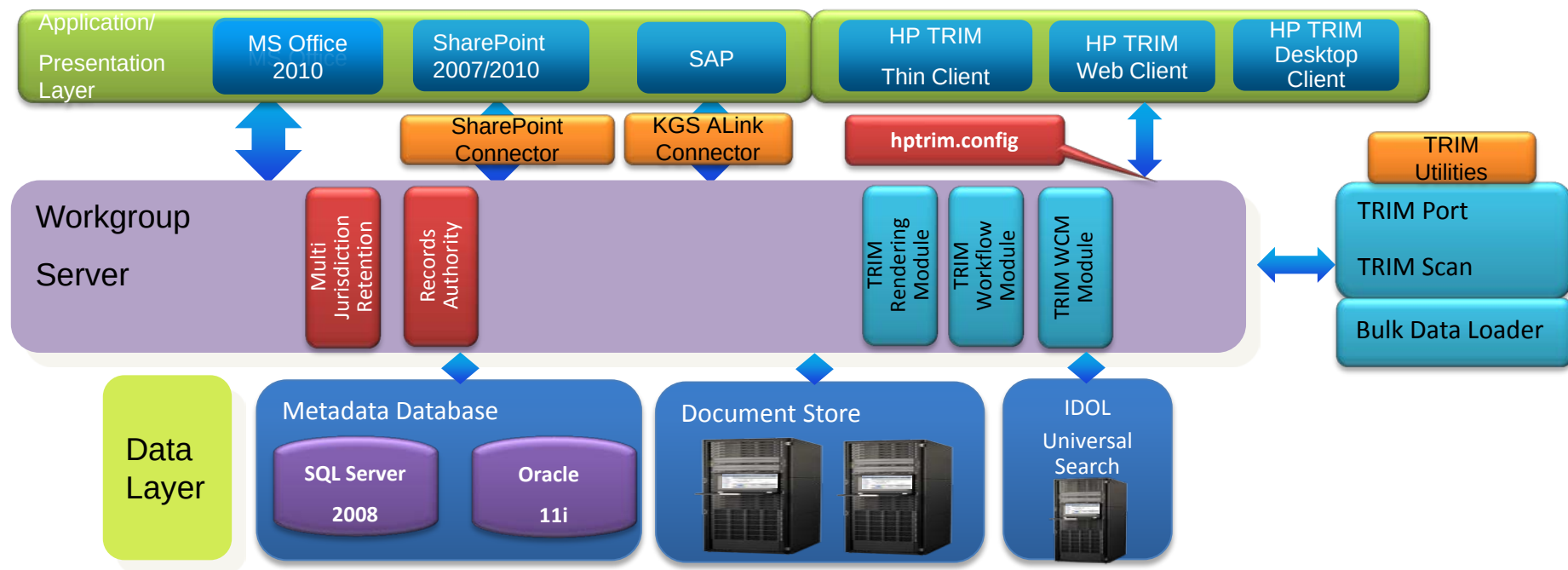
Columns with Masks	Data Masks
.... No data masking. (default)	

Validation

.... No columns validated. (default)

Indexes





HP TRIM - Architecture et composants

Architecture assez complète et paramétrisable
Peut s'intégrer avec des outils tiers



HP AIO - TRIM

- Constatations :
 - Solution complète (modèle Pull) mais nécessité de combiner deux logiciels HP
 - Possibilités de paramétrisation de l'archivage très évoluées (scripts Groovy)
 - Possibilité d'interroger les fichiers archives au moyen de SQL
 - Consultation des données possible directement depuis l'outil
 - Logiciel modulaire (ex : Rendering Module)
 - Format ouvert pour les fichiers archives (XML)



Informatica Data Archive

INFORMATICA® 9

INFORMATICA® Information Lifecycle Management Informatica Administrator, Log Off | Help

Home Accelerators Workbench Administrator Jobs Help

Define Retention and Access

Current Location: **Workbench > Retirement Workbench**

LOYALTY RETIREMENT ENTITY

Retention

Policy: New Policy ★ Name Corp Data Retention Description Standard policy to retain c ★ Retention Schedule 5 years ★ Entity Table TICKET ★ Retention Column ISSUE_DATE

Access Roles

Select Role Add Role

★Role Name	★Description	★Valid From	Valid Until	Delete
Auditor	Internal audit role	24-Oct-2009		✕
Customer Inquiry	Access to customer data	24-Oct-2009		✕

du processus
d'extraction

Access Roles

Select Role Add Role

★Role Name	★Description	★Valid From	Valid Until	Delete
Auditor	Internal audit role	24-Oct-2009		✕
Customer Inquiry	Access to customer data	24-Oct-2009		✕



The screenshot displays the Informatica Data Architect interface. On the left, a tree view shows the project structure, including 'Informatica Products', 'Custom Apps Only', and 'LOYALTY TICKETS'. The 'LOYALTY_TICKETS' entity is selected, and its details are shown on the right. The entity details include a list of attributes: TICKET, TICKET_COUPON, and TICKET_COUPON_SAVE. The 'TICKET' attribute is expanded, showing its properties: TICKET NUMBER, ISSUE DATE, CUST ID, PASSENGER NAME, LOCATION, I ADDED A COLUMN, and CUP OF COFFEE. The 'TICKET_COUPON' attribute is also expanded, showing its properties: TICKET NUMBER, COUPON, FROM CITY, TO CITY, and IS USED. The 'TICKET_COUPON_SAVE' attribute is expanded, showing its properties: TICKET NUMBER, COUPON, and FROM CITY. The 'XC_7_LOYALTY1' attribute is expanded, showing its properties: PURGEABLE FLAG, STATS DATE, ORG ID, TICKET NUMBER, and IS_SEG_COMPLETE.

Below the entity details, a navigation bar shows three tabs: 'General Information' (selected), 'Identify Entities', and 'Manage Execution'. The 'General Information' tab is active, showing the 'LOYALTY_TICKETS' entity configuration. The configuration includes a dropdown for 'Entities' set to 'LOYALTY_TICKETS', a dropdown for 'Candidate Generation Report Type' set to 'Summary', and a checkbox for 'Enabled' which is checked. The 'Policy' is set to 'Retention Policy for 20 Years'. The 'Role' is set to 'None'. The 'Operator' is set to 'older than' and the 'Value' is set to '140 Months'. The 'Earliest Archive Date' is set to 'None'. The 'CUSTOMER_ID' is set to 'Equals' and the 'Value' is set to 'None'.

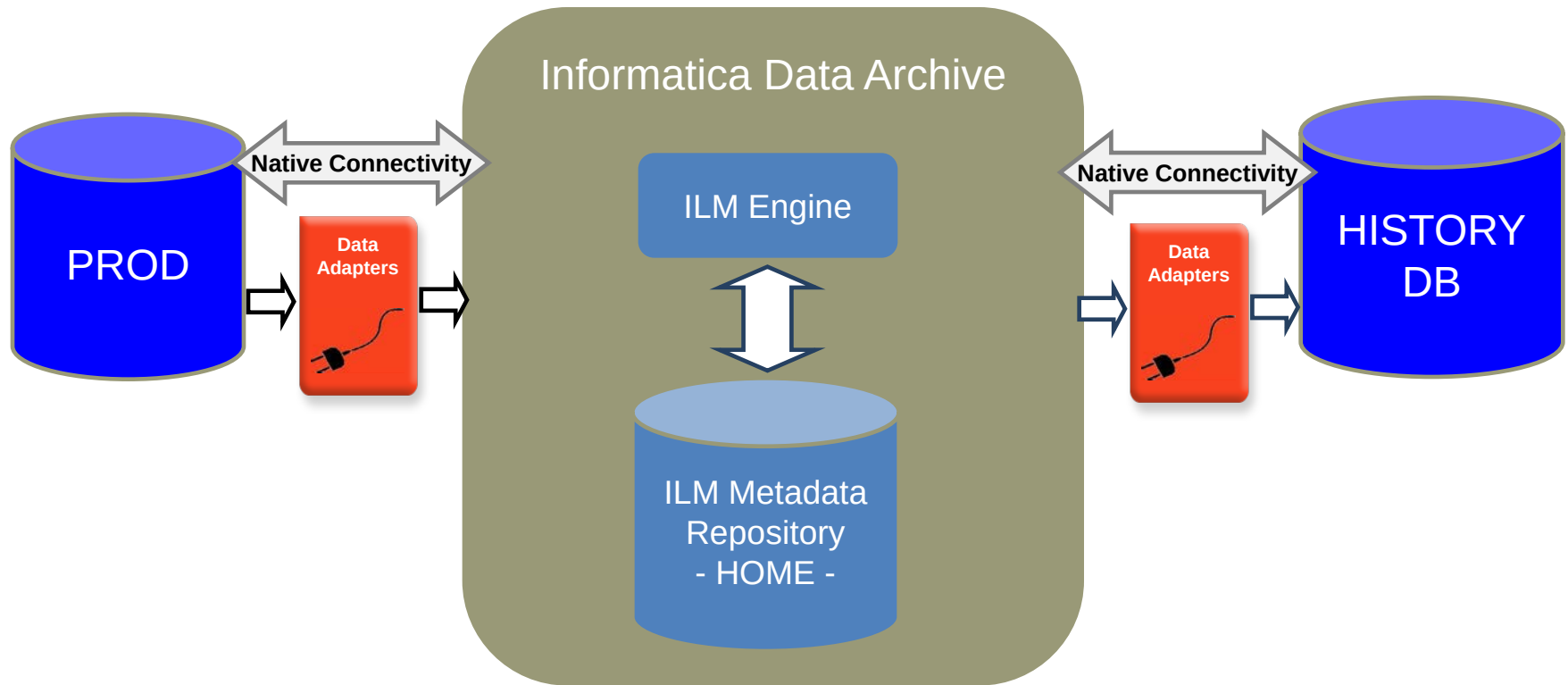
At the bottom of the interface, a table shows the deletion of application tables:

ID	Action	Table Name	Table Name
10	Delete Apps tables	XC_7_LOYALTY1	TICKET_COUPON
11	Delete Apps tables	XC_7_LOYALTY1	TICKET

Informatica - Sélection des objets à archiver

- Sélection des tables aisée
- Possibilité de simuler un versement

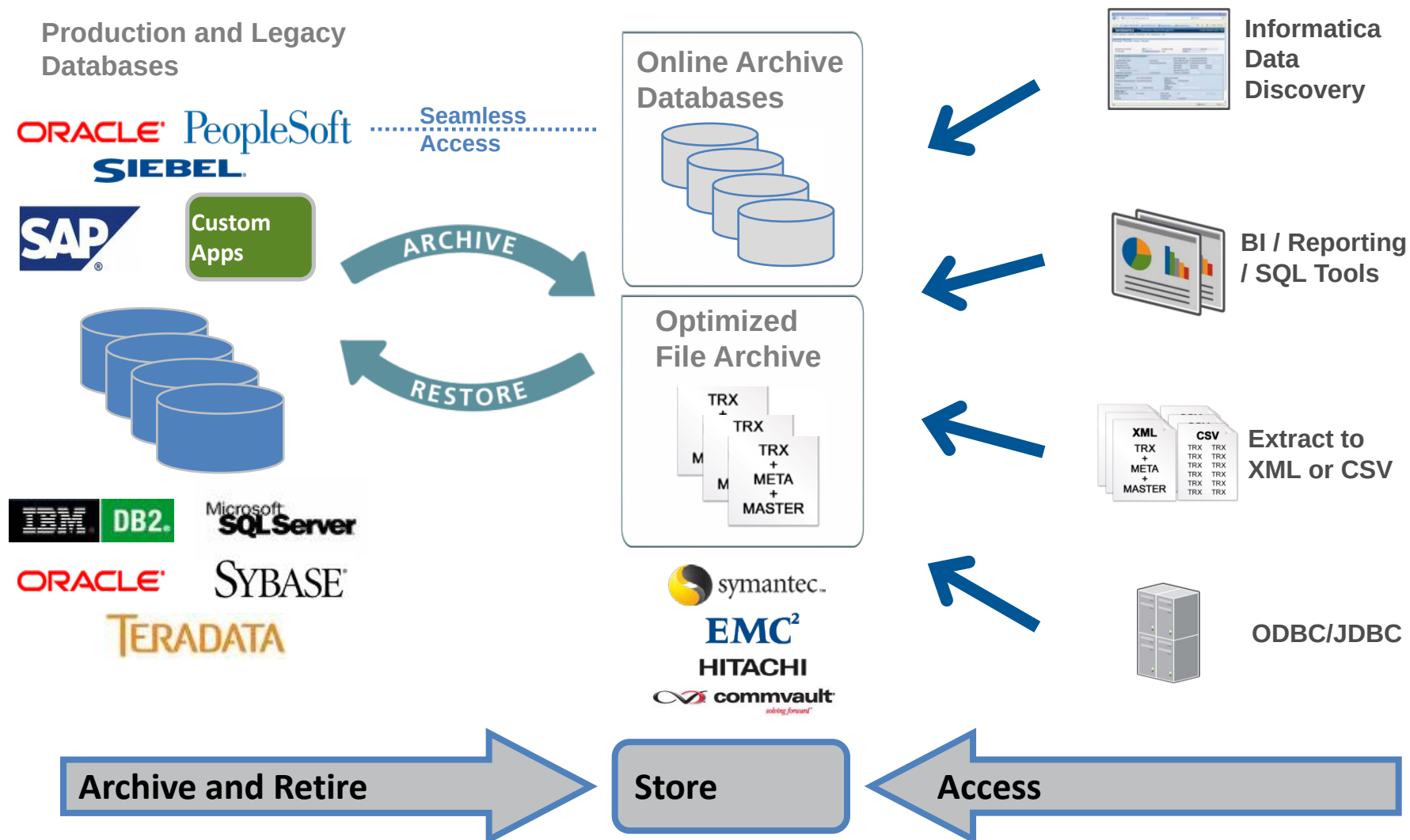




Informatica - Architecture et composants

- Architecture assez complète et paramétrisable





Informatica Data Archive



Informatica Data Archive

- Constatations :
 - Fonctionne suivant le modèle Pull
 - Intégrité hors scope de la solution
 - Possibilité d'interroger les fichiers archives au moyen de SQL
 - Possibilité de greffer de nombreux outils tiers
 - Format de données propriétaire et fermé



Tableau récapitulatif

	IBM Optim	Arcsys	HP AIO-TRIM	Informatica
Scope	Performance	Intégrité	AIO = performance TRIM = intégrité	Performance
Type de capture	PULL	PUSH-PULL	AIO = PULL et PUSH vers TRIM	PULL
Extraction	+++	+	+++ (AIO)	+++
Forme d'archivage	File	File	DB / File	DB / File
Intégrité	Hors scope	+++	Scope TRIM	Hors scope
Gestion cycle de vie	Hors scope	+++	Scope TRIM	Hors scope
Consultation	Via Optim et ODBC/JDBC	Via Arcsys ou outil tiers	Via HP et ODBC/JDBC	Via Informatica et ODBC/JDBC
Format des archives	Propriétaire et fermé	Format ouvert (TAR.GZ + XML)	Format ouvert (XML)	Propriétaire et fermé



Politique de licence

IBM Optim

- Modèle unique :
 - Achat d'un volume à archiver (minimum 1 To)

HP

- AIO :
 - Licence par source : 1, 5, 25, +25
 - Archive Query Server
- TRIM : licence par user nominatif
 - 1/user pour base
 - 1/user/module additionnel

Arcsys Software

- Modèle unique :
 - Licence unique pour un volume entrant annuel (To)
 - Licence unique pour module complémentaire
 - Support annuel
 - Mutualisation des instances (To)

Informatica Data Archive

- Standard Edition Base Licence
 - Archive to database
 - Achat d'un volume archivé (2 To)
- Advanced Edition Base Licence
 - Archive to file
 - Achat d'un volume archivé (2 To)
- Volume supplémentaire/base



$$x = 2A_1 + A_2 = 2 \frac{1}{2} (2ms) + 1m = (4ms) F_{max}$$

$$m_R V_0 = (m_R - m_B) V^2 + m_B (V_0 + V_{rel})$$

$$m_R V_0 - m_B (V_0 + V_{rel}) = N m_B (V_0 + V_{rel})$$

$$N m_B (V_0 + V_{rel})$$

Conclusion

Conclusion

- Enjeux divers
 - Légaux et réglementaires
 - Volume et performance
 - Sociétaux et scientifiques
 - Financiers
- Archivage = sécurité
- S'appuyer sur les standards et recommandations existantes
- Difficultés inhérentes
 - Obsolescence du matériel et des couches logicielles
 - Obsolescence des formats
 - Préservation de la signification des données
 - Authenticité et intégrité des données



Conclusion

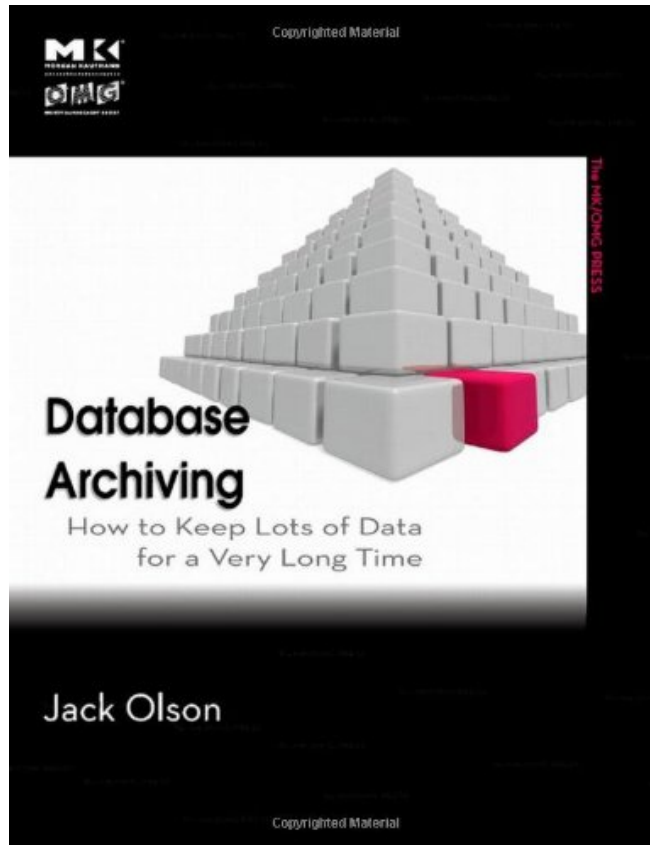
- Archivage d'une base de données : transfert des données vers un autre système
 - DB : solution à court terme et non respect de l'intégrité
 - Fichiers : solution à long terme, consultation moins aisée
- Méthodologie rigoureuse
 - Bien définir les objectifs souhaités
 - Travail d'analyse important (et conséquent)
 - Projet multidisciplinaire
- Organisation et compétences
- Solutions techniques
 - Permettre la préservation des couches d'information
 - Stratégie de sortie (format ouvert nativement ou via export)





Bibliographie et annexe

Bibliographie



J. OLSON, *Database Archiving: How to Keep Lots of Data for a Very Long Time*, Morgan Kaufmann, 2009.

ISBN-13: 978-0123747204



Bibliographie

- I. Boydens et al., *Gestion intégrée des anomalies*, Section Recherches, mars 2011.
- I. Boydens, *Data Quality : Best Practices*, Section Recherches, mai 2006.
- I. Boydens, *Evaluer et améliorer la qualité des bases de données*, Section Recherches, janvier 1998.
- I. Boydens, *Informatique, normes et temps*, Bruxelles : Éditions E. Bruylant, 1999.
- A. Hulstaert, *Préservation à long terme de l'information numérique. Rendre l'information accessible durablement*, Section Recherches, Smals, février 2010.
- A. Hulstaert, *Préserver l'information numérique : Codage et conversion de l'information*, Section Recherches, Smals, juin 2008.
- G. Ogonowski, *NoSQL – Hype ou innovation ?*, Section Recherches, Smals, octobre 2011.
- G. Ogonowski, *Business Rules Technologies & web sémantique : La gestion des règles business*, Section Recherches, Smals, février 2008.
- C. Rectem, *La pérennisation digitale dans le secteur public : étude critique de plusieurs modèles ouverts de méta-information*, Mémoire en Sciences de l'information et de la documentation, ULB, Bruxelles, 2009.





Arnaud Hulstaert

02.787.51.91

arnaud.hulstaert@smals.be

Grégory Ogonowski

02.787.45.11

gregory.ogonowski@smals.be

More on Smals Research

Website Smals : www.smals.be

Blog : <http://blogresearch.smalsrech.be/>

Twitter : <https://twitter.com/smalsresearch>

