La migration des fichiers du dépôt légal de l'Internet à la BnF

Clément Oury

Service du dépôt légal numérique

Bibliothèque nationale de France

clement.oury@bnf.fr

Plan de l’intervention

- Contexte et enjeux de la migration
- Les outils de migration
- Approche et méthodologie

Contexte et enjeux

Quelques mots de contexte…

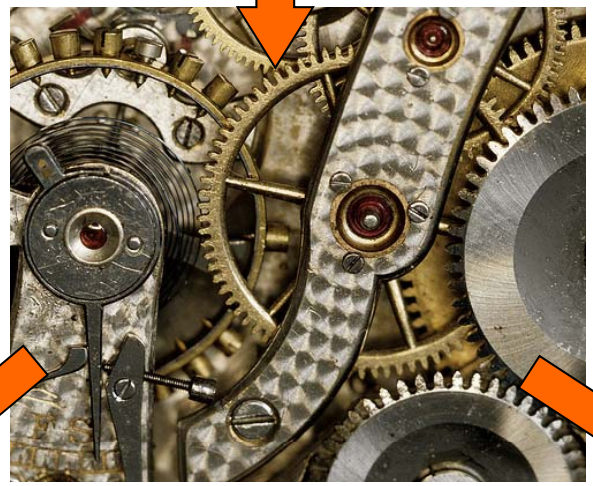
- Un objectif : l'archivage de contenus Internet à des fins patrimoniales
- Un cadre juridique et scientifique : le dépôt légal
  - Une mission inscrite dans le code du Patrimoine, partagée avec l'INA
  - Deux modifications fondamentales…
  - … mais un objectif scientifique similaire : ne pas juger de la qualité
- Une voie privilégiée: les outils de collecte automatique, les « crawlers »
- Un projet qui repose largement sur la coopération internationale et sur le consortium IIPC

Au cœur de tout le dispositif : les robots de collecte

- Logiciel appelé robot de collecte, « aspirateur », « araignée » ou « moissonneur » de sites
- Part d'une liste d'adresses URL « graines »
- Extrait les liens dans le code des pages, les suit comme un internaute automatique
- Copie les éléments qu'il trouve et qui font partie du périmètre de la collecte

Le circuit du document

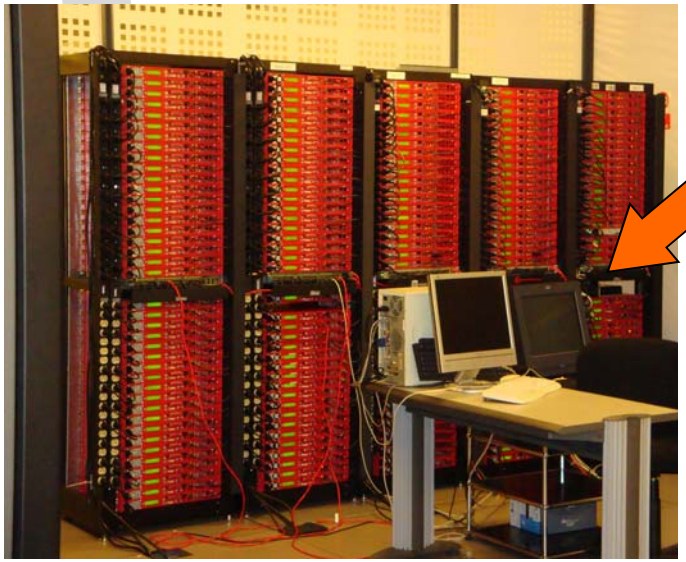
Collecte



Indexation

Préservation

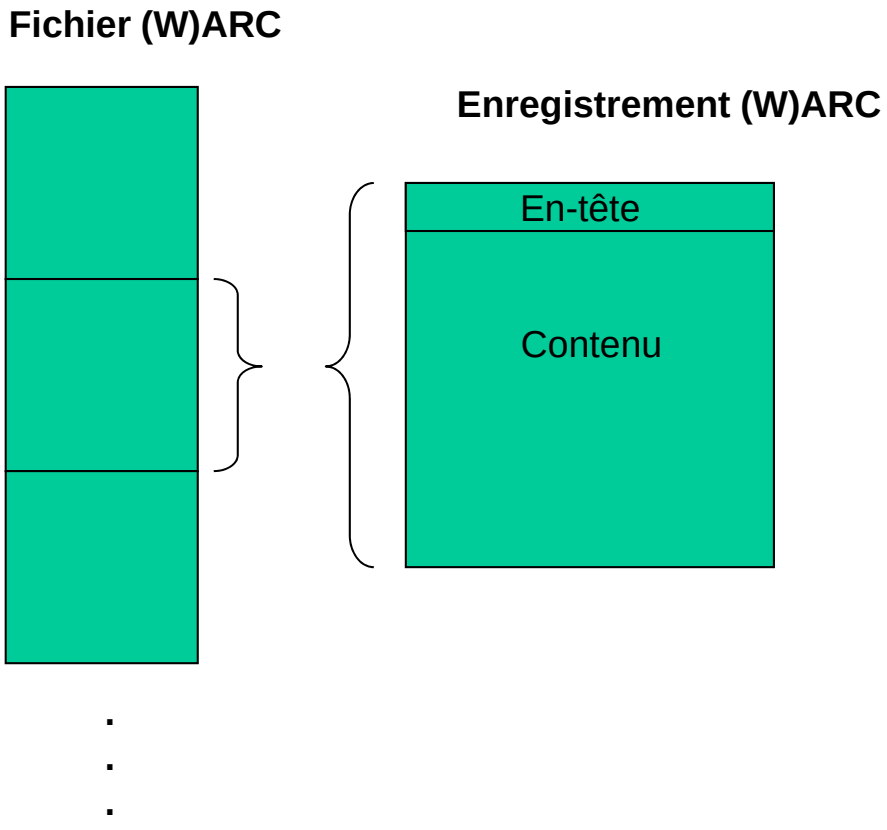
Accès



Les données produites

- Les données collectées sont regroupées au sein de fichier *container* ARC
  - Un fichier ARC regroupe de multiples enregistrements ARC
  - Un enregistrement ARC est le fichier collecté sur le Web, accompagné de métadonnées
- Caractéristiques du format ARC (1996)
  - format auto descriptible
  - format extensible
  - format compressé
  - taille limitée arbitrairement à 100 Mo
  - organisation des données délibérément aléatoire
- Il y a d'autres types de fichiers
  - Documentation de la collecte : fichiers de configuration, *logs* et rapports d'activité
  - Fichiers d'index

Anatomie d’un fichier (W)ARC



Un enregistrement ARC

Métadonnées:

URL IP-address Archive-date Content-type Archive-length

http://desirsdavenir-mosaique.over-blog.com/ 195.20.15.131 20070416172243 text/html 132721

Objet numérique collecté :

```
HTTP/1.1 200 OK
Date: Tue, 16 Apr 2007 17:22:56 GMT
Server: Apache/2.0.58 (Unix) mod_ssl/2.0.58 OpenSSL/0.9.7e PHP/4.4.2
X-Powered-By: PHP/4.4.2
Last-Modified: Tue, 16 Apr 2007 17:22:56 GMT
Content-Type: text/html

<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" lang="fr" xml:lang="fr">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1" />
<title>Comité Mosaïque</title>
<meta name="description" content="desirsdavenir-mosaique.over-blog.com
hébergé par over-blog.com Le comité de la diversité riche de ses différences
qui soutient Ségolène Royal à la présidence de la République." />
...
```

Le format WARC

- Héritier du format ARC
- Normalisé au sein d'IIPC et de l'ISO : ISO 28500:2009
  - Ce qui correspond à une attente forte des institutions
- Un format ARC étendu :
  - Ajout d'un identifiant unique pour chaque enregistrement
  - Gestion des doublons
  - Gestion d'un nombre supplémentaire d'informations (informations sur la collecte, archivage des requêtes du robot, possibilité d'ajouter des métadonnées)
  - Gestion des migrations de format des fichiers contenus
  - Taille cible : 1 Go
  - Aucune régression par rapport au format ARC.
- Des enrichissements qui justifient une migration

Les enjeux d'une migration

- Comment appréhender la masse de données à traiter ?
  - Les collections de la BnF : 150 To, 13 milliards de fichiers contenus
  - Relative hétérogénéité : différents outils de collecte au fil du temps
- Quelles mesures de validation adopter ?
- Doit-on conserver les données originales ?
- Peut-on enrichir les données ?
- Quelles ressources sont nécessaires ?
  - Temps, machines, hommes

Des outils pour un format

Une suite complète d'outils

- Tous les outils développés au sein d'IIPC sont capables de traiter des fichiers WARC
  - Pour la production : le robot de collecte Heritrix
  - Pour l'indexation plein-texte : le logiciel NutchWAX
  - Pour la visualisation : la Wayback Machine
- Outils de traitement et de manipulation : les WARC tools

Les « WARC tools »

- Outils libres développés par la société Hanzo Archives, membre d'IIPC…
- …et avec des fonds d'IIPC
- Objectif : promouvoir l'adoption du format WARC
- Ne pas dépendre d'une seule implémentation logicielle

Deux séries d’outils

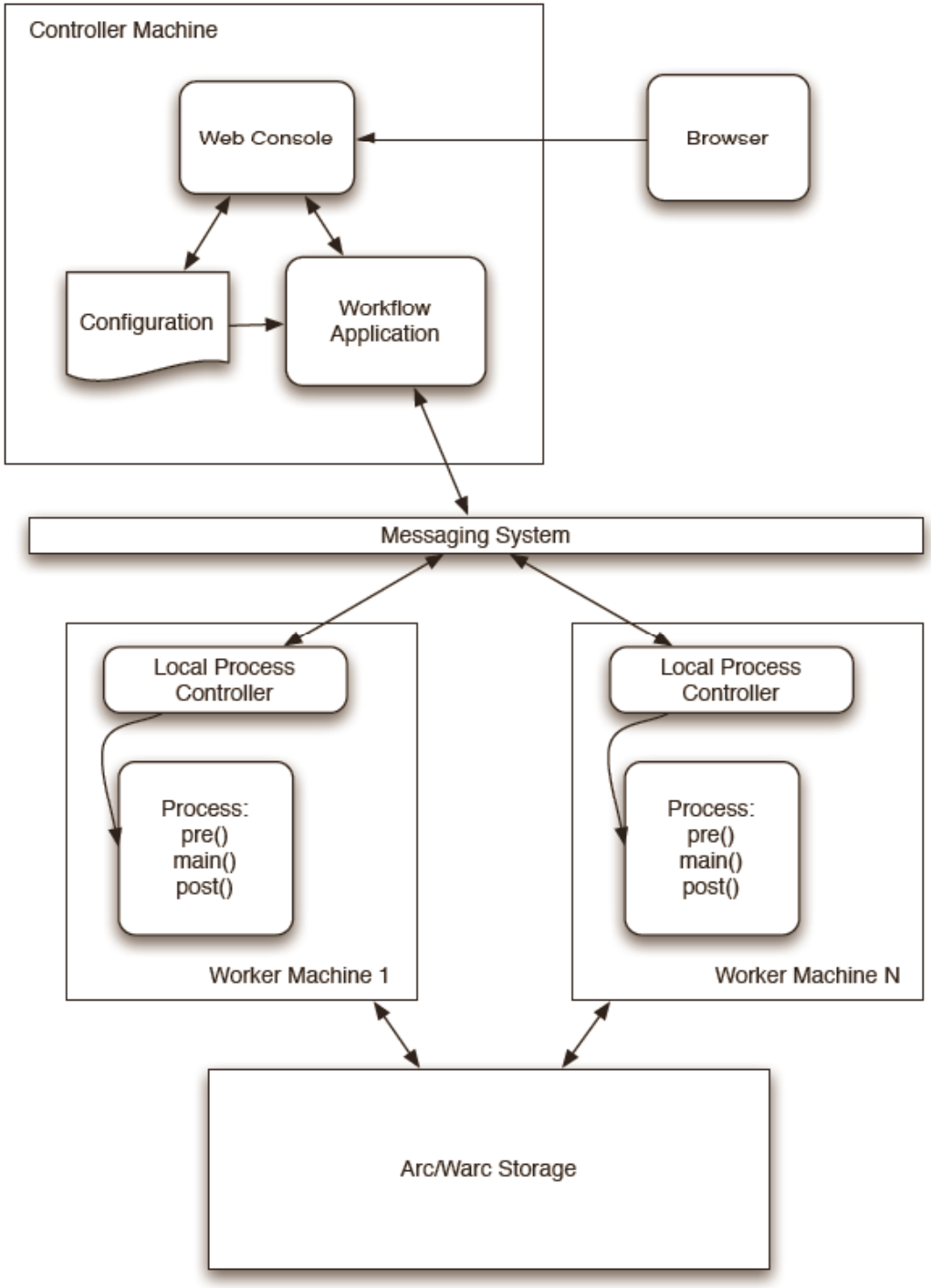
- Première série (phases I & II) : une bibliothèque d’outils destinés à accomplir des actions unitaires
  - outils d’écriture / de harvesting
  - outils de lecture
  - outils d’identification / de validation
  - outils de migration

- Deuxième série (phase III) : intégrer les outils des phases précédentes dans des applications professionnelles
  - Analyse et extraction des données: « reporting », « repackaging »
  - Migration de grande échelle
  - Livraison octobre 2010, validation février 2011

L’application de migration

- Une application de workflow intégrant :
- Une interface graphique pour définir la « stratégie de migration »
- Un outil de migration
  - Capable d’appeler des outils extérieurs
- Un outil de validation
- Une console pour piloter le processus (choix des fichiers à traiter, pause, restaurations…)
- Spécifications non fonctionnelles
  - Scalabilité (!)
  - Distribution de la charge sur plusieurs machines
  - Compatibilité avec un environnement Java

Vue générale du système



Enrichir les données lors de la migration ?

- Le recours à des outils extérieurs
  - Choix du système d'identifiants uniques
  - Outils d'identification ou de caractérisation
  - Outils antivirus

- Dédupliquer les données ?

Approche et méthodologie

WARC : un format plus complexe

- Spécifications du format ARC
  - 4 pages
  - 2 types d'enregistrement
  - 9 champs d'en-tête
- Norme WARC
  - 8 types d'enregistrement
  - 17 champs d'en-tête
  - … le tout sur 28 pages

- La norme WARC explique comment créer un fichier WARC valide…
- … mais elle laisse de nombreux choix ouverts
- Des enjeux critiques d'interopérabilité

Les « WARC implementation guidelines »

- Juin-septembre 2009 : élaboration d'un guide d'application de la norme
- Groupe de travail international (IIPC) regroupant des utilisateurs (bibliothécaires et ingénieurs) et les développeurs des outils

- Une grande partie concerne la migration :
- Règles de « mapping » ARC / WARC
- Choix des identifiants uniques, nommage des fichiers migrés
- Génération et enregistrement de métadonnées

- http://www.netpreserve.org/publications/WARC_Guidelines_v1.pdf

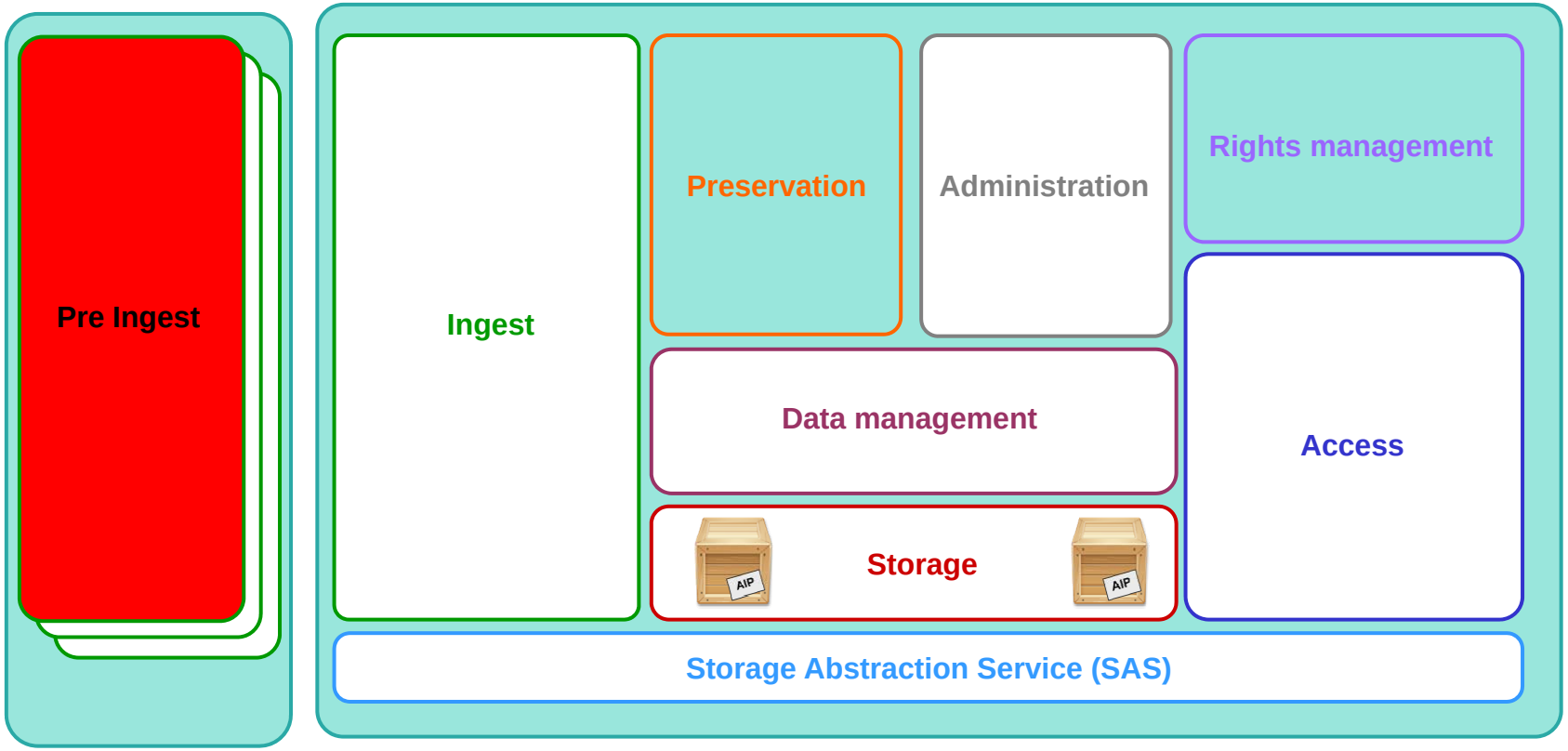
Gros ou petit bout ?

- Commencer par le courant ou le rétrospectif ?
- La production de WARC « natifs » semble aujourd'hui prématurée
- Les outils ne correspondent pas nécessairement à nos besoins
- Il serait difficile de gérer en même temps deux générations de format
- Il faut s'appuyer sur nos partenaires internationaux
  - Robots de collecte, outils d'accès : Internet Archive
  - Outil de gestion côté bibliothécaire : netarchive.dk

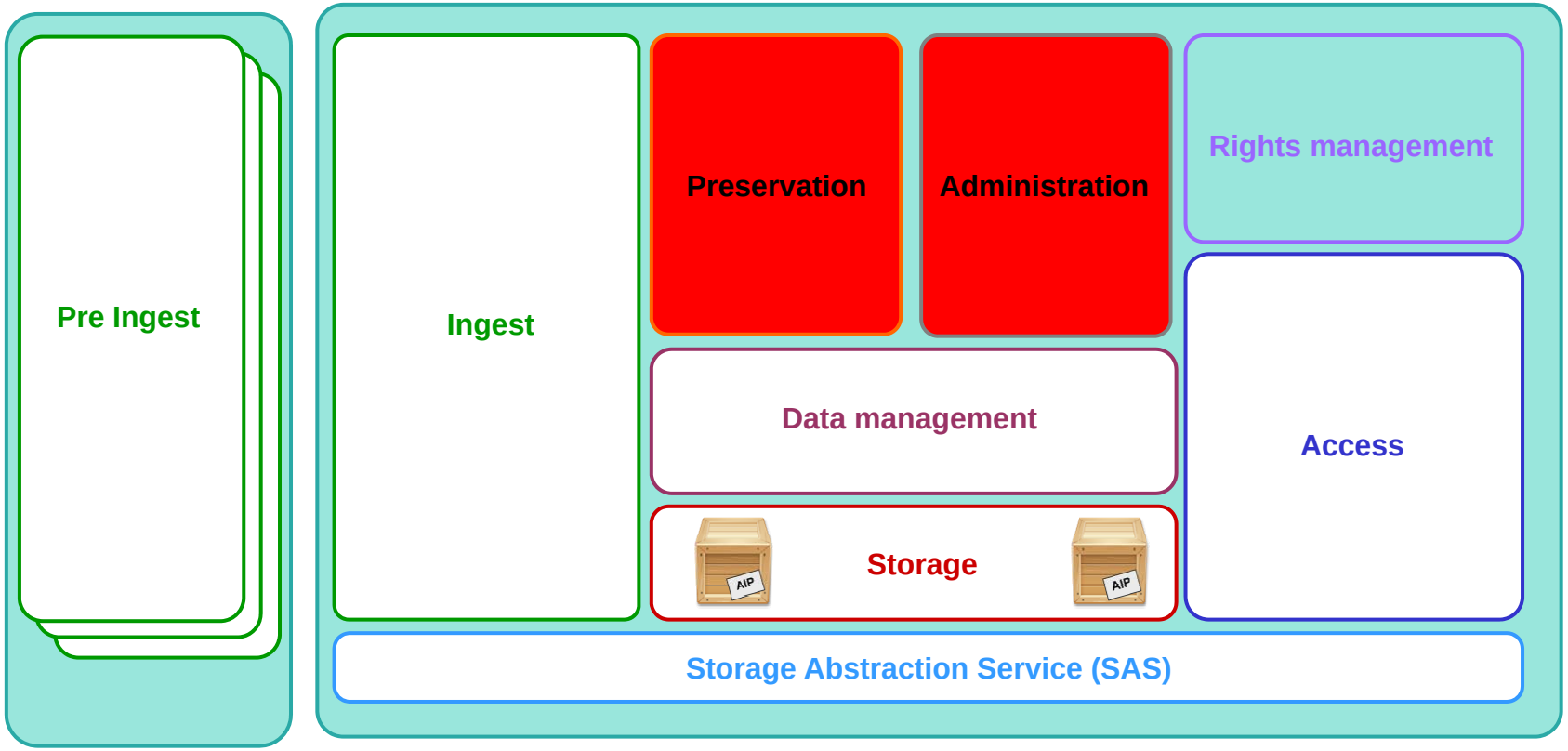
La migration rétrospective

- Migrer les données historiques…
- … et faire de la migration une étape – provisoire – du circuit de production
- Impact moindre sur le circuit actuel…
- … mais coût de stockage plus important
- Un banc de test pour préparer la conversion de l'ensemble du workflow
- Pourrait être considéré comme une étape de l'entrée dans SPAR

La migration ARC/WARC dans SPAR



La migration ARC/WARC dans SPAR



En conclusion…

- La migration ARC / WARC est avant tout un enjeu de préservation
- Les objectifs et les enjeux sont définis
  - La similitude des deux formats facilite la migration
  - Le guide d'application précise les contours de la migration
- Mais des difficultés demeurent
  - La masse, encore et toujours
  - Une forte dépendance vis-à-vis des outils actuellement disponibles
- La concertation internationale représente à la fois une chance et une nécessité

Merci de votre attention !

Crédits photographiques

http://www.r-geek.com/2007/08/28/i-robot/

http://www.avl-importexport.com/Location.html

http://www.flickr.com/photos/villeneuve53/1808995620/

http://www.preparationmariage.com/IMG/jpg/Fotolia_120123_S.jpg

http://www.niffylux.com/components/com_virtuemart/shop_image/product/Oeuf__1_4b21145d94340.jpg

http://www.flickr.com/photos/papalars/2197212826/