Qualité et conservation des données

Olivier Rouchon – CINES

8ème École Inter-Organismes

« Qualité en Recherche et en Enseignement Supérieur »

8 Septembre 2010

QuaRES
Association pour la Qualité en Recherche et Enseignement Supérieur

Sommaire

- Rappel : qu'est-ce que l'archivage pérenne ?
- Pourquoi la qualité ?
- Qualité technique
  - Les métadonnées
  - Les formats de fichier
  - Le stockage
- Qualité organisationnelle
  - La formalisation des processus
  - La gestion des risques
  - La certification

QuaRES
Association pour la Qualité en Recherche et Enseignement Supérieur

Qu’est-ce que l’archivage pérenne ?

L’archivage pérenne des documents électroniques consiste à conserver le document et l’information qu’il contient :

- Dans son aspect physique comme dans son aspect intellectuel,
- Sur le très long terme soit 30 ans et au-delà,
- De manière à pouvoir le rendre accessible et compréhensible.

Or, la plupart des fichiers informatiques de plus de 10 ans sont aujourd’hui illisibles :

- Connaissance perdue du contenu des fichiers,
- Format de fichier inconnu,
- Support physique détérioré,
- Logiciel ou matériel de lecture disparu

QuaRES
Association pour la Qualité en Recherche et Enseignement Supérieur

Les défis de l'archivage pérenne

Contrainte	Solutions
Connaissance du contenu	• Utilisation de métadonnées • Identification unique et pérenne des documents archivés
Format de fichier inconnu	• Privilégier les formats durables • Identification, validation des formats • Migration logique (conversion de formats)
Support physique détérioré	• Gestion du vieillissement des médias • Migration physique (changement de support)
Logiciel ou matériel de lecture disparu	• Veille technologique et anticipation

QuaRES
Association pour la Qualité en Recherche et Enseignement Supérieur

Pourquoi la qualité ?

La qualité recouvre deux domaines :

1. La qualité technique

- Qualité des métadonnées = capacité à garder la connaissance des contenus
- Qualité des formats de fichiers = capacité à convertir à de nouveaux formats
- Qualité du stockage = capacité à conserver le train de bits constituant les fichiers

Pourquoi la qualité ?

2. La qualité organisationnelle

- Documentation des processus métiers = répétabilité et amélioration des mécanismes de conservation
- Gestion des risques = maintient d'un niveau de qualité acceptable en identifiant de façon proactive les évènements pouvant impacter la conservation et les plans d'actions à mettre en place
- Démarche de certification = validation des actions entreprises et constitue un levier pour l'obtention de budgets auprès des décideurs

L'adoption de normes / standards facilite la démarche qualité pour la conservation

QuaRES
Association pour la Qualité en Recherche et Enseignement Supérieur

La qualité des métadonnées

Les métadonnées permettent de préserver les informations décrivant les objets numériques :

- Métadonnées / informations de pérennisation
  - Descriptives (du contenu de l'information),
  - Source (provenance de l'information),
  - Historique (versions successives)
- Métadonnées / informations de représentation
  - Techniques (aspect de l'information),
  - Structure (forme de l'information),
  - Administratives (droit d'accès à l'information)

QuaRES
Association pour la Qualité en Recherche et Enseignement Supérieur

La qualité des métadonnées

Plusieurs contrôles de qualité peuvent être effectués :

- Contrôle du format de la métadonnée par l'adoption d'un standard
  - Métadonnées génériques pour la description des ressources numériques : ex. Dublin Core (ISO 15836)
  - Métadonnées spécifiques à un domaine : ex. commerce électronique ebXML, données géographiques (ISO 19115)
  - Métadonnées techniques : préservation (PREMIS, METS),
  - Métadonnées administratives : propriété intellectuelle et droits d'auteur (indecs, MPEG-21)
- Contrôle de la valeur des métadonnées par une logique applicative métier
  - Liste de valeurs autorisées, etc.

QuaRES
Association pour la Qualité en Recherche et Enseignement Supérieur

La qualité des formats de fichier

Pour permettre le contrôle de la qualité d'un fichier, celui-ci doit être dans un format identifié et vérifiable :

- Format publié ; ex. WAVE, SVG
- Format largement utilisé ; ex. XML, MPEG4
- Format normalisé si possible ; ex. PDF (ISO 32000-1:2008), PNG (ISO 15948:2004)

Pour pouvoir être lisibles dans le temps, et convertibles, les fichiers doivent respecter les spécifications de leur format - des outils libres permettent une identification, validation et caractérisation des formats

Type	Format
Texte	HTML, PDF, TXT, XML, ODT
Image	GIF, JPEG, TIFF, PNG, SVG
Audio	WAV, AIFF, AAC, VORBIS
Vidéo	MJPEG2000, MPEG4, THEORA

JHOVE

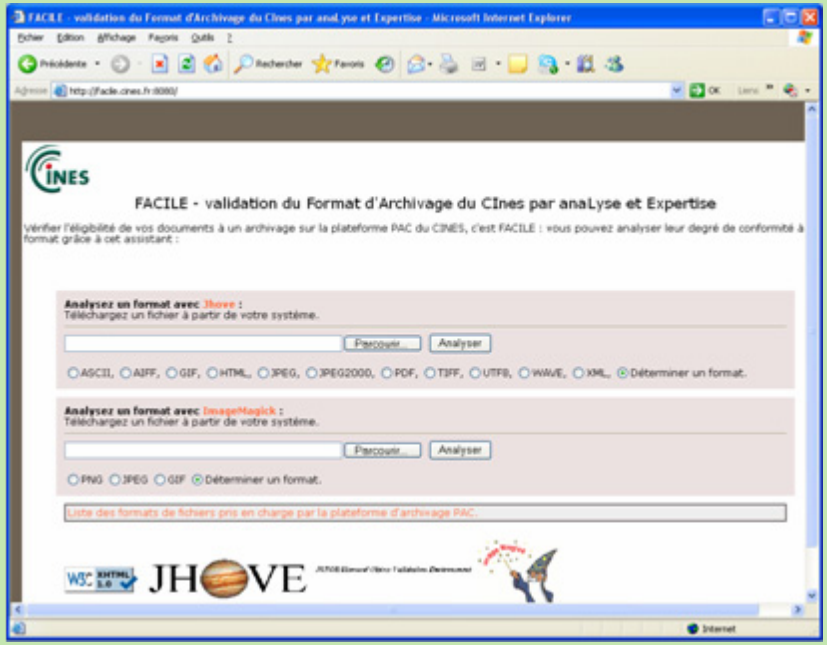
ImageMagick

DROID

QuaRES Association pour la Qualité en Recherche et Enseignement Supérieur

Un outil de contrôle qualité

FACILE – validation du Format d'Archivage du CInes par anaLyse et Expertise



- Outil en ligne permettant de valider les fichiers par rapport aux spécifications de leur format
- Projet initié dans le cadre de l'archivage des thèses électroniques
- Développement en langage Java par l'équipe du Département Archivage et Diffusion du CINES
- Code open source disponible pour intégration dans des applications

QuaRES
Association pour la Qualité en Recherche et Enseignement Supérieur

Un outil de contrôle qualité

FACILE

- Intègre la même liste de formats supportés et les mêmes outils (Jhove, Imagemagick, DROID, ODF Validator,) que la plateforme d'archivage du CINES
- Les contrôles effectués sont les mêmes que ceux effectués lors d'un dépôt de document
- Permet une validation des fichiers avant dépôt de la part du producteur avec alerte en cas de non-conformité
- Une assistance de 2ème niveau par l'équipe du CINES est possible pour résoudre les problèmes de non-conformité plus complexes

http://facile.cines.fr/

QuaRES
Association pour la Qualité en Recherche et Enseignement Supérieur

La qualité du stockage

La qualité du stockage garantit la conservation du train de bits composant les fichiers de données

- Copies multiples (>2)
- Indépendance des supports de copies (mix disques/bandes, localisation géographique)
- Audit fréquent de l'intégrité des copies

Plusieurs études sur la fiabilité des support de stockage :

- Etudes des Archives de France sur les CD-R / DVD-R et graveurs du marché (2008).
- Etudes de Carnegie Mellon University, Google et Univ. Wisconsin-Madison sur les disques durs

QuaRES
Association pour la Qualité en Recherche et Enseignement Supérieur

La qualité du stockage

Le contrôle de l'intégrité des fichiers permet d'anticiper la corruption de l'information.

Il peut se faire :

- Au niveau matériel
  - Vérification CRC par les contrôleurs de disques, contrôleurs réseau etc.
- Au niveau logiciel
  - Vérification des sommes de contrôle (en anglais *checksum*)
  - Calcul des empreintes numériques par échantillonnage et comparaison avec l'empreinte initiale
  - Utilisation d'algorithme de hachage (MD5, SHA-256), etc.

QuaRES Association pour la Qualité en Recherche et Enseignement Supérieur

La formalisation des processus métiers

Rappel des intérêts de la démarche BPM pour l'archivage :

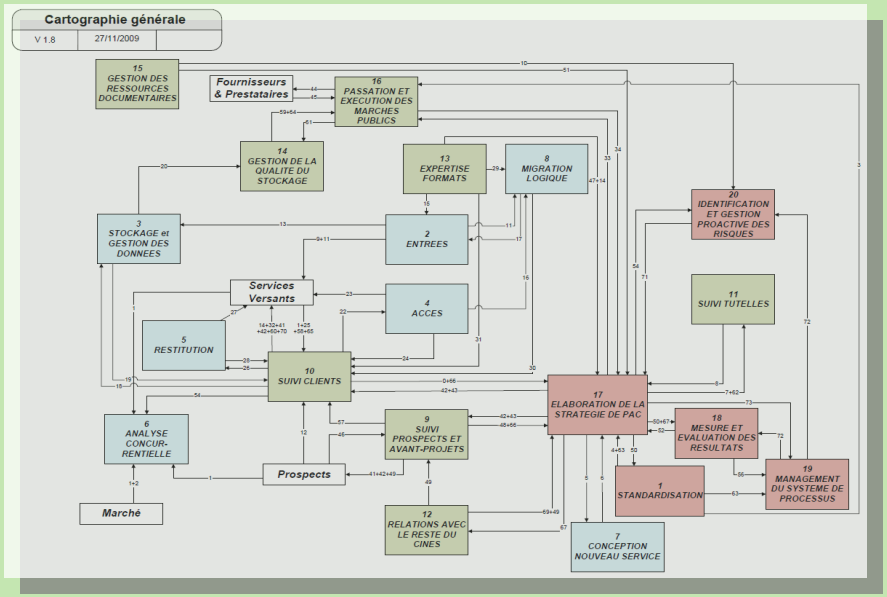
- Communication interne : diffusion + actualisation de la connaissance métier ;
- Communication externe (prospects, clients, ministère, auditeurs externes, comparaison avec d'autres organismes…) ;
- Présentation et documentation de l'activité sous un nouvel angle → nouvelle vision et nouveaux constats envisageables (amélioration continue) ;
- Elément constituant la colonne vertébrale de la documentation, pré-requis pour une certification ;
- Déploiement structuré de la stratégie ;
- Amélioration permanente facilitée (auto-évaluation).

QuaRES
Association pour la Qualité en Recherche et Enseignement Supérieur

La formalisation des processus métiers

La démarche :

1. Décomposition des fonctions métier de l'archivage en processus, sous-processus, activités
2. Etablissement d'une cartographie générale des processus
3. Détail de chaque processus identifié – à rapprocher des groupes fonctionnels OAIS
    - 33 fonctions théoriques
    - 22 processus identifiés et applicables

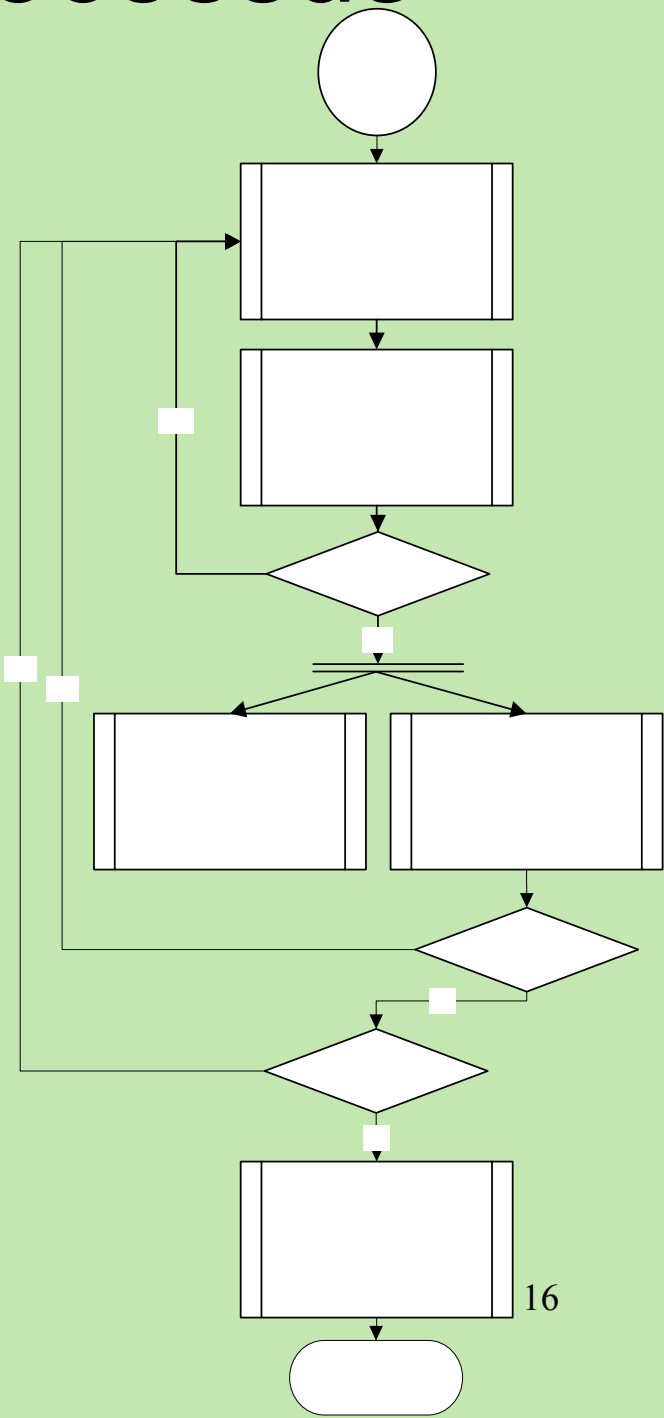


La formalisation des processus métiers

Les étapes de la description d'un processus:

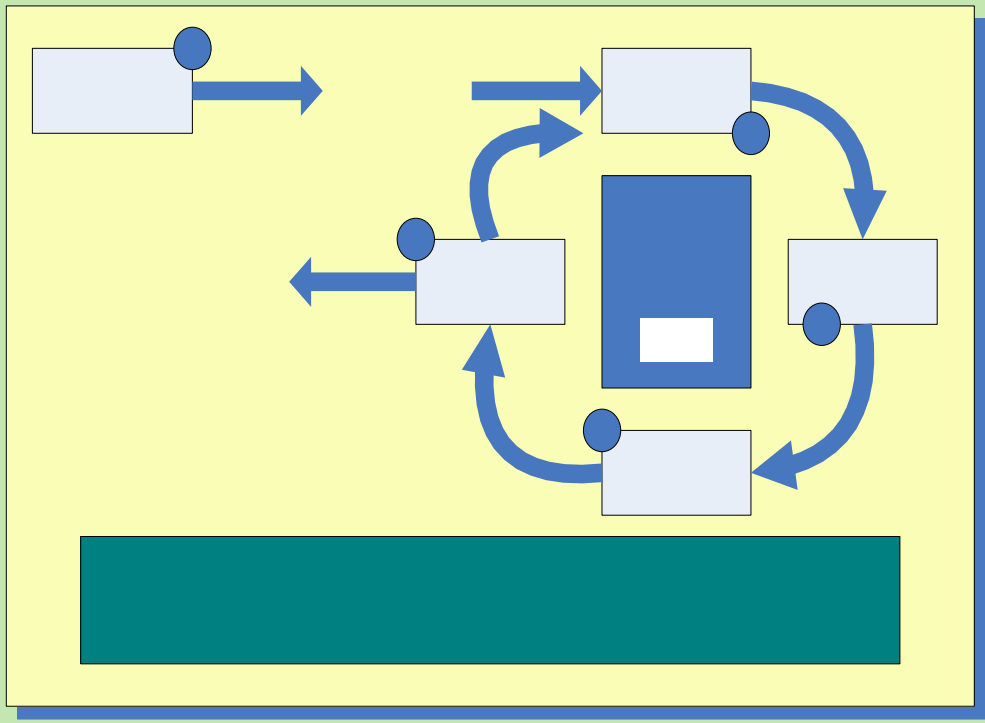
1. Formalisation par l'expert (pilote du processus) : entretien, réalisation de la cartographie ;
2. Validation de la cartographie et caractéristiques associées par la hiérarchie ;
3. Validation par l'équipe après corrections éventuelles ;
4. Veille sur cette cartographie.

Cette étape requiert l'adhésion et l'implication de tous les agents impliqués.



La gestion des risques

1. **Définition du contexte**
   - Fixer les objectifs de la gestion des risques
2. **Identification et catégorisation des risques**
3. **Evaluation des risques**
   - Analyser la probabilité et l'impact de chaque risque dans le temps
4. **Prise de décision**
   - Identifier les risques prioritaires, les moyens de leur traitement et le plan d'action
5. **Maîtrise des risques**
   - Mettre en place les actions nécessaires pour diminuer le niveau de risques
6. **Itération**



La certification

La certification est :

- L'aboutissement de la consolidation d'un organisme/service
- La reconnaissance de la qualité et du professionnalisme, donc un moyen d'instaurer des relations de confiance avec les communautés d'utilisateurs
- Un levier pour obtenir des budgets auprès des organismes de tutelle

Plusieurs types de certification sont envisageables :

- Certifications généralistes : ISO 27000 (sécurité informatique), ISO 9000 (qualité), CMMI (ingénierie), ITIL (services), etc.
- Certifications spécifiques à l'archivage pérenne :
  - Accréditations : DSA (bonnes pratiques) DRAMBORA (gestion des risques), TRAC (liste de critères)
  - Certifications en cours d'élaboration : AFNOR (Z42-013) avec le SIAF, ISO 16363 (European Audit Framework) avec le CCSDS

QuaRES
Association pour la Qualité en Recherche et Enseignement Supérieur

La certification

La démarche de certification est un projet conséquent

- Investissement humain
  - Conduite du projet,
  - Pilotage des changements requis
- Investissement financier
  - Audits externes réguliers à prévoir

D'où la nécessité de bien identifier le type de certification qui aura le plus d'impact

- Sur la communauté d'utilisateurs
- Sur les organismes de tutelle

QuaRES
Association pour la Qualité en Recherche et Enseignement Supérieur

Conclusion

- **De la qualité des objets numériques dépend grandement la facilité de leur préservation dans le temps ;**
  - Il ne faut pas réduire cette qualité au seul aspect technique des objets numériques ;
  - La qualité des processus de conservation qui leur sont appliqués est toute aussi cruciale ;
- **Une démarche qualité pour la conservation à long terme représente un investissement conséquent et immédiat dont les effets ne seront perceptibles qu'à long terme ;**
  - Les indicateurs / métriques sur la qualité ne sont pas encore clairement définis ;
  - Une telle initiative – et son aboutissement qu'est la certification – requiert l'adhésion de toutes les ressources humaines impliquées dans le processus de conservation.

Rendez-vous dans 30 ans ?

QuaRES
Association pour la Qualité en Recherche et Enseignement Supérieur

Questions ?

olivier.rouchon@cines.fr

QuaRES
Association pour la Qualité en Recherche et Enseignement Supérieur

Quelques liens utiles

Groupe PIN (Pérennisation de l'Information Numérique) :

http://www-pin.aristote.asso.fr/doku.php

Etudes des Archives de France sur les CD/DVD:

http://www.archivesdefrance.culture.gouv.fr/gerer/archives-electroniques/stockage/

Etude Google sur les disques durs :

http://static.googleusercontent.com/external_content/untrusted_dlcp/labs.google.com/fr//papers/disk_failures.pdf

Etude Carnegie Mellon sur les disques durs :

http://www.cs.cmu.edu/~bianca/fast07.pdf

Etude Wisconsin Madison sur les disques durs :

http://www.cs.wisc.edu/wind/Publications/corruption-fast08.pdf

QuaRES
Association pour la Qualité en Recherche et Enseignement Supérieur

Quelques liens utiles

Accréditation DSA d'un système d'archives :

http://www.datasealofapproval.org/

Certification ISO 16363 d'un système d'archivage :

http://wiki.digitalrepositoryauditandcertification.org/bin/view

Le modèle OAIS :

http://public.ccsds.org/publications/archive/650x0b1.pdf

Le site web du CINES :

http://www.cines.fr/

FACILE :

http://facile.cines.fr/

QuaRES
Association pour la Qualité en Recherche et Enseignement Supérieur