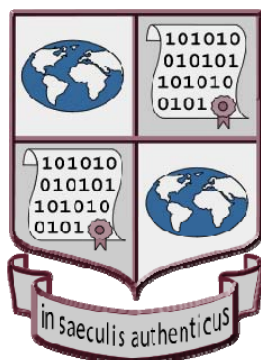


Projet InterPARES 2

International Research on Permanent Authentic Records in Electronic Systems



Sélection de formats de fichiers numériques pour préservation à long terme

Tracey P. Lauriault

XXXVIe congrès annuel

Association des archivistes du Québec (AAQ)

D'hier à demain : Quarante ans de théories et de pratiques archivistiques au Québec – quelles pistes pour l'avenir?

1 juin 2007



InterPARES Project

Tracey P. Lauriault présente le rapport de Evelyn Peters McLellan
Rapport d'étude générale 11

Table des matières

- Les chercheuses
- Méthodologie de recherche
- Qu'est-ce qu'un format de fichier?
- Critères de sélection
- Implication pour les politiques
- Recommandations – mise en œuvre des politiques



Les chercheuses

- Evelyn Peters McLellan, co-enquêtrice pour InterPARES 2 est chef de recherche pour ce rapport
- Les assistantes de recherche, étudiantes supérieures à la School of Library, Archival and Information Studies de l'Université de la Colombie-Britannique
 - Tracey Krause
 - Yvonne Loiseau
- Étude générale 11



Méthodologie de recherche

- Examen qualitatif de la documentation disponible sur les sites internet de:
 - 20 dépôts
 - quatres groupes de collaboration multi-institutionnels
- Recherche conçu pour examiner les critères appliqués afin de choisir les formats de fichiers numériques
 - manque de terminologie uniforme



Institutions Examinés

1. Art Institute of Chicago, Department of Architecture
2. Arts and Humanities Data Service, United Kingdom
3. California Digital Library
4. Cornell University Library
5. Digital Image Archive of Medieval Music (DIAMM), Universities of Oxford and London
6. Florida Center for Library Automation
7. Bibliothèque et Archives Canada
8. Library of Congress
9. Massachusetts Institute of Technology
10. National Archives of Australia
11. National Archives of the United Kingdom
12. U.S. National Archives and Records Administration
13. Netherlands Institute for Scientific Information Services

1. Ohio Electronic Records Committee
2. On-line Computer Library Center, United States
3. Public Records Office of Victoria
4. State and University Library, Aarhus, and the Royal Library, Copenhagen
5. Technical Advisory Service for Images, United Kingdom
6. UC Berkeley Art Museum/Pacific Film Archive
7. UK Data Archive
8. DAVID Project (Digitale Archivering in Vlaamse Instellingen en Diensten, or Digital Archiving in Flemish Institutions and Administrations)
9. Digital Preservation Coalition, United Kingdom
10. EU-US Working Group on Spoken-Word Audio Collections
11. Research Libraries Group and Digital Library Federation



InterPARES Project

Tracey P. Lauriault présente le rapport de Evelyn Peters McLellan
Rapport d'étude générale 11



Qu'est-ce qu'un format de fichier?



Définitions

La plupart ne définissent pas les termes de fichier

- structure d'organization des données dans un dossier (*déf. Typique*)
- «l'organisation des données dans les fichiers, habituellement conçu pour faciliter le stockage, l'extraction, le traitement, la présentation et/ou la transmission des données avec des logiciels» (*InterPARES 2*)
- «structures spécifiques préétablie pour l'organisation d'un fichier numérique ou d'un train de bits» (*Data Dictionary for Preservation Metadata PREMIS*)



Les particularités

- Certains types de codages sont spécifiquement associés à des formats de fichiers
 - MP3
 - plein texte: ASCII, EBCDIC, Unicode
- Stockage et transmission exige souvent une compression
 - TIFF
 - WAVE
- TIFF, WAVE, AVI sont des Formats envelopeurs / conteneurs et non des formats de fichiers
- XML et GML sont des métalangages



Recommandations

- Les institutions doivent établir des lignes directrices sur les formats de fichiers
 - préciser les formats acceptables
 - stipuler si les «formats» sont:
 - des formats de fichiers
 - des formats enveloppeurs précisant les codages des trains de bits constituants
 - ex. XML – format du fichier, le codage et le schema XML ou la définition du type de document
 - ou des fichiers balisés
 - inclure les versions



Formats de fichiers « ouverts »

- Généralement un format de fichier ayant plusieurs caractéristiques:
 - ces spécifications sont publiées
 - spécifications sont largement diffusées
 - format créé par un logiciel non exclusif
 - protégé par un brevet?



Les particularités

- Spécification connue n'est pas synonyme avec format de fichier ouvert
- Format ouvert et non exclusif - sont elles distinctes?
- Fichier standard ouvert
- Spectre de l'ouverture
- Format exclusif + spécification diffusée
- Formats ouverts
 - codes sont diffusés et peuvent être modifiés



Formats de fichiers « standard »

- Formats de fichiers acceptés ou recommandés?
- Des formats largement adoptés, avec spécifications publiés, interopérable, n'exigeant aucune compression, et qui soutient des metadonnées préservable.
- “*de facto* standards”
- Norme de l'industrie / Utilisation générale
- Recommandé par ANSI ou ISO
- Format neutre



Formats de fichiers « stables »

- Plusieurs parlent de stabilité de format
- Semble être rétrocompatible et bien soutenu par l'industrie des logiciels
- Spécifications du format soient stable et non susceptibles de changer constamment avec le temps
- Rétrocompatibilité



Les particularités

- AVI est devenue *de facto* sauf que Microsoft a retiré leur appui – donc la conversion est recommandée



Normalisation des termes

- Manque de définitions uniformes ou simplement manque de définition
- Définir:
 - ouvert
 - standard
 - stable
- Nombreuses variantes dans la terminologie





Critères de sélection



Caractéristiques nécessaires

Formats de fichiers généralement considérés nécessaire pour la préservation

- largement adoptés
- non exclusifs
- soutenus par une solide documentation
- indépendants des plates-formes / interopérable
- non comprimés ou avec une technique ne causant aucune perte de données



Utilisation largement répandue

- 18/24 institutions
- Format largement acceptés et non un petit créneau
- Microsoft Word et PPT, Lotus 123 car si largement utilisés
- Tend à perpétuer le soutien par l'industrie des logiciels
- Moins susceptible à la désuétude
- La création d'outil de migration et d'émulation par l'industrie est plus vraisemblable



Les particularités

- La détermination de ce qui constitue une utilisation largement répandue est une démarche subjective
 - largement adopté par de vastes communautés et des gros groupes
 - utilisé par une vaste communauté pour une longue période ex. MP2 et non MP4
 - accepté par d'autres institutions d'archives ex. TIFF et PDF car elle sont largement utilisés pour fin d'archives



Origine non exclusive

- 17/24 institutions
- Soit des formats *ouverts*
- *Sans devoir verser des redevances,*
- *Des frais de licence ou*
- *Des droits exigibles en vertu d'un brevet*



Les particularités

- Exceptions pour certains formats tel que :
 - utilisation largement répandue
 - diffusion de leur spécifications
 - possibilités de convertir les fichiers dans des formats non exclusifs
 - non exclusif mais utilise pdf comme format de préservation



Large diffusion des spécifications

- 17/24 des institutions
- Documentations ou spécifications
- Développement approfondi avec la documentation correspondante
- Publié ou bien diffusé
- Peuvent être exclusif ou non mais il faut de la documentation ou des spécifications publiées
- Possibilité qu'une archive ait accès et non le public



Les particularités

- Documentations et spécifications ne sont pas des synonymes
- Sources ou les spécifications puissent être inspectées ou un format étayé par des documents
- L'inexistence de spécifications publiées n'est pas nécessairement un obstacle à la préservation



Indépendance des plates-formes (interopérabilité)

- Mentionné par 13/24 institutions
- Soutenue par une vaste gamme de systèmes
- Indépendant du matériel, des systèmes d'exploitation et d'autres logiciels
- La capacité d'échanger des documents avec d'autres utilisateurs et systèmes de TI
- Dépendances externes comme critère
- XML conforme aux définitions de types de document (DTD)



Les particularités

- TIFF car c'est un format compatible et transférable pour les fichiers d'images
- XML et GML sont préconisés comme ouverts mais ils ne sont pas une panacée pour résoudre des problèmes d'échange de données ex. format basé sur XML ou fondé sur des DTD



Compression

- Mandats et politiques d'acquisition varient
- Préférence marquée pour des fichiers non comprimés
- Stipulation que seulement certain types de fichiers peuvent être comprimés



Les particularités

- TIFF non comprimé, ou ceux comprimés sans pertes
- JPEG comprime avec perte
- Compression sans perte visuelle
- Recommander des techniques ou les normes
- Comprimé avec algorithmes publiques



Analyse des critères

L'utilisation largement répandue des formats et le fait qu'ils ne sont **pas exclusifs**, que leurs **spécifications sont publiées**, qu'ils sont **interopérables** et **sans compression** (ou avec compression sans perte de données) semblent être les plus importants facteurs de sélection de formats de fichiers numériques pour préservation à long terme, si variables que soient les termes employés pour les décrire.



Les particularités

- Manque de formats réunissant toutes ces caractéristiques
- Besoin d'équilibrer les exigences de qualité, stabilité, longévité et d'acceptation de l'industrie
- Si possible préciser des formats non exclusifs ou standard ou *de facto*
- *Des règles générales*



Les Sciences

- *Formats idéaux dans les domaines scientifique et artistique sont difficiles à trouver*
- Interopérabilité pour les données scientifiques est très important
- Common Data Format Markup Language (CDFML) pour les données astronomiques
- Des langages de balisages – XML, GML, XSIL, CML, MatML et DDI
- Formats standard de transfert dans les dépôts



Les arts numériques

- TIFF est le plus largement accepté pour la conservation des images
- Fichiers audionumériques – le meilleur objet à préserver fait appel à un codage conservant un maximum de profondeur de bits avec la fréquence d'échantillonnage la plus élevée et un minimum de compression
- Cinématographie – conformer aux désirs de l'artiste et une stratégie médiatique variable



Secteurs gouvernementaux

- Plus prometteur en ce qui concerne la production de document d'archives
 - taille de l'organisation
 - les éléments communs des programmes
 - systèmes de traitement de textes et de base de données
 - plate-forme de service normalisé basé sur Internet
 - interopérabilité interne
 - normalisation des formats exclusifs en comparaison avec les sciences et les arts numériques





Implication pour les politiques



Implications pour les politiques

- Limiter le nombre de formats de fichiers acceptable pour la préservation?
- Facteurs commerciaux pour les dépôts?
- Refus de certains formats?
- Qu'arrivera t'il pour les sciences et les arts numériques?
- Accepter des formats sans promesse de préservation?



...les politiques

- Fichiers éphémères dès leur création?
- Tenté de répondre au besoins de diverses populations?
- Activement encourager l'applications des normes?
- Et les sciences et Arts?
- Archives gouvernementaux devraient elles maîtriser certains formats de fichiers préservables?
- Les sciences peuvent dicter les caractéristiques désirables pour la préservation
- Relation entre le créateur et les archives



Les particularités


- Dif. degrés de préservation en fonction du format.
- 3 degrés de préservations
 - soutenu
 - connu
 - non soutenu
- Préservation des trains de bits pour formats nouveau ou inconnus
- 3 degrés en fonction du format
 - préservation et accès de base
 - préservation accrue
 - préservation et accès maximaux



...particularités

- Sentiment de sécurité injustifié
- Problèmes stratégiques
- Problèmes éthiques
- Obligations des archives gouvernementales
- Charge de travail indéterminé sans aucune garantie
- Mandats, relations avec institutions mère, organisations donateurs et relations juridiques avec les créateurs





Recommandations pour l'élaboration et la mise en œuvre des politiques



1. Clarifier la terminologie :

déterminer ce qu'on entend par des termes comme ouvert, standard, stable et étayé par une solide documentation, et définir ces termes dans les politiques



2. Faire une distinction

entre les formats de fichiers eux-mêmes, les formats enveloppeurs (ou conteneurs) et les formats balisés, comme les fichiers balisés en XML, et veiller à ce que les versions et les autres caractéristiques telles que le codage soient comprises et entièrement spécifiées.



3. Dans le cas des fichiers en XML

exiger qu'ils soient bien formés, valides et accompagnés des DTD ou des schémas pertinents



4. Sélectionner des formats largement utilisés

Dans toute la mesure du possible, sélectionner des formats largement utilisés, non exclusifs, indépendants des plates-formes et ayant des spécifications largement diffusées.



5. Préciser si les fichiers comprimés

sont acceptables et, si oui, préciser le type de compression permise. Dans toute la mesure du possible, opter pour des techniques de compression sans perte de données et conformes aux normes internationales acceptées.



6. Si pas #4 opter pour les formats préservés

S'il est impossible de sélectionner des formats ayant les caractéristiques énumérées dans la recommandation 4, opter pour des formats préservés dans d'autres dépôts de documents numériques et collaborer avec ces dépôts pour concevoir des plans de préservation à leur égard.



7. Travailler de concert avec les créateurs

Dans toute la mesure du possible, travailler de concert avec les créateurs de documents d'archives pour veiller à ce qu'ils utilisent des logiciels capables de créer ces documents dans des formats répondant aux critères énumérés dans la recommandation 4.



- <http://www.interpares.org/>
- Rappports:
 - http://www.interpares.org/ip2/ip2_case_studies.cfm?study=35
- Études générales
 - http://www.interpares.org/ip2/ip2_general_studies.cfm

