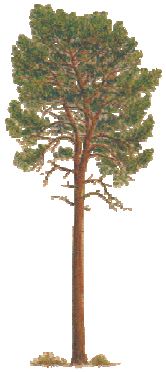
Les répertoires de formats



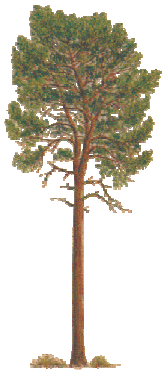
Etat de l'art

Réunion du groupe PIN
25/01/2006

Emmanuelle Bermès - emmanuelle.bermes@bnf.fr
Bibliothèque nationale de France
Département de la bibliothèque numérique

PLAN

- Introduction
  - → questions sur les formats
  - → caractéristiques des répertoires de formats
- Le répertoire des MIME media types
- Divers répertoires existants
- OCLC Inform methodology
- L'initiative de la Library of Congress
- GDFR
- Pronom
- Conclusion
  - → comparatif
  - → actions ?



Questions sur les formats

- Sélection — **PRODUCTION**
  - → j'ai un contenu, dans quel format le représenter ?
- Identification — **INGEST**
  - → j'ai un objet numérique, dans quel format est-il ?
- Validation
  - → J'ai un objet numérique censé être en format X, est-ce exact ?
- Caractérisation
  - → J'ai un objet au format X, quelles sont ses propriétés ?
- Evaluation — **PRESERVATION**
  - → J'ai un objet au format X avec des propriétés Y, quel est le risque d'obsolescence ?
- Traitement
  - → j'ai un objet au format X avec des propriétés Y, comment réaliser l'opération Z sur ce format ?

Caractéristiques d'un répertoire de formats

- Etre **ouvert, accessible, interopérable** (gestion de droits pour les spécifications devant rester secrètes)
- Savoir gérer les **niveaux de granularité** de description de format (format + version + caractéristiques) et les relations (format dépendant, format associé)
- *Inclusive* ou *self-descriptive* c'est à dire **contenir toutes les données nécessaires** pour fonctionner indépendamment d'autres ressources

Caractéristiques d'un répertoire de formats

- Actionnable **par machine** : disposer d'outils qui l'interrogent de manière automatisée
- Digne de **confiance** :
  - → être sous la responsabilité d'un organisme neutre et reconnu
  - → être respectueux des informations propriétaires
- **Objectivité** : ne pas mêler des évaluations de pérennité aux informations factuelles et aux spécifications

Caractéristiques d'un répertoire de formats

- Informations à rassembler concernant un format :
  - → Noms canoniques et variantes
    - PDF, Adobe PDF, Portable Document Format
  - → Signatures internes et externes
    - .pdf, magic number
  - → spécifications
    - http://partners.adobe.com/public/developer/en/pdf/PDFReference16.pdf
  - → Auteurs, titulaires de droits, maintenance
    - Société Adobe
  - → classifications et relations
    - PDF <has subtype> PDF 1.4, PDF/A, PDF/X…
  - → systèmes, services et outils
    - Adobe Acrobat Reader...

Le répertoire des MIMETYPE

- Maintenu par l'IANA (http://www.iana.org/assignments/media-types/)
- Insuffisant pour la préservation…
  - → se contente d'énoncer des types de format
    - ❖ image/tiff et non TIFF V.6
  - → ne donne pas de moyen d 'énoncer les caractéristiques
    - ❖ image/tiff et non TIFF V.6 non compressé profondeur 24 bits
  - → ne donne pas toujours les spécifications des formats
  - → pas d'outils de validation ni d'évaluation

Divers répertoires existants

- NSRL/NIST Reference Data Set (http://www.nsrl.nist.gov/)
  - → par le département de la justice américain (objectif = lutte contre le piratage…)
  - → base de données de référence pour l'identification des formats (signature) mais pas de spécifications
  - → diffusion sous forme de CD à 90$ par an
- Wotsit (http://www.wotsit.org/)
  - → à destination des programmeurs
  - → classement par types (texte, image etc.)
  - → spécifications disponibles en lien ou en téléchargement
  - → 918 formats décrits

OCLC Inform methodology

- Le but est de donner des indicateurs pour évaluer la « préservabilité » d'un format.
  - → gestion des risques
  - → propriétés du format
  - → fonctionnalités du format
- utilisation de notes de 1 à 5 combinant probabilité et impact
- méthodologie de notation et politique d'action
- Cf : *Assessing the Durability of Formats in a Digital Preservation Environment, The INFORM Methodology* par Andreas Stanescu, Dlib Magazine, novembre 2004 : http://www.dlib.org/dlib/november04/stanescu/11stanescu.html

L'initiative Library of Congress

- Un site web à but informatif :
  - →http://www.digitalpreservation.gov/formats
- Informations, publications, ressources
- Facteurs de longévité pour évaluer les formats
- Description de formats classées par type (texte, image, audio, video)

L'initiative Library of Congress

- Les facteurs d'évaluation :
  - → **ouverture** (disponibilité des specs)
  - → **adoption** (étendue de l'utilisation par des acteurs diversifiés)
  - → **transparence** (simplicité du codage, sens de lecture, absence de cryptage ou de compression, éventuellement lecture humaine)
  - → **formats auto-documentés** (inclusion de métadonnées)
  - → **indépendance technique** (hardware, software, inclut par ex. des périphériques spécifiques)
  - → **impact des brevets logiciels**
  - → présence ou possibilité de mettre en œuvre des **mesures de protection techniques** (profil d'application du format)

L'initiative Library of Congress

- Le contenu des descriptions de formats
  - → identification du format et ses émanations, relations
  - → « local use » usage à la LoC
  - → fiche descriptive reprenant les critères d'évaluation précités
  - → éléments d'identification (signature)
  - → liens vers les spécifications du format

GDFR

- Global Digital Format Registry (http://hul.harvard.edu/gdfr/)
- Une initiative internationale pour créer une fédération de répertoires de formats
- Objectifs :
  - → fournir des **identifiants uniques** pour identifier les formats
  - → fournir la **documentation** utile sur ces formats
- Mise en place d'un modèle de données et d'un modèle de service

GDFR

- Les points forts :
  - → un ou plusieurs identifiants dont l'un est déterminé comme « canonique »
  - → relations entre les formats
  - → liens externes permettant de fédérer plusieurs répertoires
  - → copie locale des spécifications
  - → organisme de gouvernance
- Vient de recevoir un financement de 2 ans de la Mellon Foundation

GDFR, JHOVE et FRED

- Un outil développé par Harvard et Jstor :

  JHOVE

  http://hul.harvard.edu/jhove/

  - → est une application des principes de GDFR
  - → notamment
    - identification
    - validation
    - caractérisation
  - → ajouts de formats sous forme de modules plug-in (pas relié à un répertoire)
  - → logiciel libre GNU GPL
- Et un prototype de démonstration :

  FRED

Pronom

- Archives nationales UK (http://www.nationalarchives.gov.uk/pronom)
- Une base de données
  - → environ 550 formats
  - → modèle de données aligné avec GDFR
  - → informations sur les formats :
    - description et relations
    - documentation (référence)
    - information d'identification (signatures)
    - compression, encodage de caractères
    - droits
    - **identifiant unique PUID**
- invitation à soumettre de nouveaux formats

Pronom

- Un outil :

  DROID

  http://www.nationalarchives.gov.uk/aboutapps/pronom/droid.htm

  → identification automatique des formats

  → utilise les signatures internes et externes référencées dans la base Pronom

- Mais…

  → beaucoup de formats sont justes signalés et pas décrits

  → les spécifications sont externes (= problème pour la préservation)

Comparatif des initiatives

	Sélection	Identification	Validation	Caractérisation	Evaluation	Traitement
IANA	Non	Oui	Non	Non	Non	non
RDS	Non	Oui	Oui	Oui	Non	Non
Wotsit	Non	Non	Non	Oui	Non	Oui
OCLC	Non	Non	Non	Non	Oui	Non
LOC	Oui	Oui	Non	Oui	Oui	Non
GDFR	Non	Oui	Oui	Oui	Non	Oui
Pronom	Non	Oui	Oui	Oui	Non	Non

Phase de production Phase d'ingest Phase de préservation

Que faire ?

- Entrer dans Pronom ? Dans GDFR ?
- Utiliser les modèles fournis par GDFR pour créer un répertoire ?
  - → De sélection ?
  - → D'identification ?
  - → D'évaluation ?
- Plusieurs répertoires ? Qui ?
- Utiliser et adapter des outils « partiels » comme JHOVE ?
- ...

Références complémentaires

- Stephen Abrams, conférence IPRES (sept. 2005), *digital formats and preservation*
  - → http://rdd.sub.uni-goettingen.de/conferences/ipres/programme
- Page concernant les formats sur le site de la BnF :
  - → http://bibnum.bnf.fr/conservation/formats.html
- GDFR, détail de la proposition d'action
  - → http://hul.harvard.edu/gdfr/documents/Proposal-2005-09-29.doc