Quelques applications Open Source pour l'archivage numérique

Fabrice Lecocq (lecocq@inist.fr)

Institut de l'Information Scientifique et Technique

CNRS

17 Septembre 2009

Sommaire

- D-Space, Fedora Commons, DuraSpace
- LOCKSS
- Autres outils
- Evaluation

D-Space

www.dspace.org

17 Septembre 2009

D-Space

- 2000 : Collaboration HP Labs / bibliothèque du MIT pour les besoins du MIT
  - Auto-archivage (dépôt, indexation, recherche, affichage)
  - Workflow peu contraignant
  - Interface facile à utiliser
  - Publication dans un contexte Open Access (OAI)
  - Préservation des contenus numériques
- 2002 : v 1.0 mise en Open Source
- 2009 : v 1.5.2 dernière version stable

Couverture fonctionnelle

- Système d'archivage d'objets numériques
  - gestion, distribution et préservation à long terme d'objets numériques.
  - organisation en collection
- Base de données des métadonnées associées
  - création, indexation, recherche…
- Plate-forme de dépôt et de diffusion de publications électroniques
  - Workflow de publication
  - OAI v2 (metadata), OpenURL (nommage)
- Droits, délégation, alertes, espace personnalisé

Quelques exemples d’usage

- Archives Institutionnelles (Instituts, Universités)
- Thèses en ligne
- Plateforme de revues en ligne
- Plateforme de E-learning
- Banques d’images, vidéo

Modèle de données

D-Space

- Communauté
- Sous-communauté
- Collection
- Item (métadonnées)
- Bundle
- Bistream

Exemple I-Revues (INIST)

- Revue
- Volume
- N°
- Article
- Fichiers

Attributs

- Chaque communauté, sous-communauté et collection possède sa propre page d'accueil
- Pour les collections
  - Chaque collection peut définir son propre workflow
  - Chaque collection peut posséder son ou ses administrateurs
- Items : niveau auquel sont associées des métadonnées descriptives ; peut appartenir à plusieurs collections
- Bistream : niveau le plus fin avec des formats divers
  - Supported (soutenu) / Known (connu) / Unsupported (non soutenu)

Acteurs et workflow

- E-peoples et Groupes

- Gestion du workflow
  - Processus de dépôt est défini pour chaque collection
  - Les tâches sont assignées par des alertes mail
  - Les tâches se font via une interface web

- Les étapes & les actions
  - 1 - Acceptation ou rejet du dépôt
  - 2 - Édition des métadonnées, acceptation/rejet du dépôt
  - 3 - Finalisation des métadonnées
  - 4 - Transmission pour stockage et mise en ligne

Fonctions d’administration

- Gestion des déposants
  - Édition, suppression, ajout de E-peoples/Groupes
- Gestion des items
  - Éditer les métadonnées d’un item
  - Retirer un item de la collection
- Lien avec d’autres collections
  - Inclure des items d’une autre collection

L’interopérabilité

- Descriptif des métadonnées en Dublin Core Qualifié
- DSpace METS Document Profile for Submission Information Packages (SIP)
  - http://wiki.dspace.org/index.php/DSpaceMETSSIPProfile
- Supporte Unicode

- Dissémination des contenus par OAI-PMH v2.0
- Identifiant pérenne
  - un préfixe handle par instance D-Space (handle.net)
  - Peut-être renforcé par un DOI

- - Module d’ingestion/diffusion propriétaire

D-Space 2 (2010)

- Architecture modulaire, design avec Web Services
- Scalabilité :
  - Fichiers de plus de 2 Go
  - Nombres d'items portés à 10 millions ou plus
  - Taille globale : plusieurs To
  - Augmentation du nombre d'utilisateurs simultanés
- Replication / Mirroring / Virtualisation du stockage
- Implémentation de nouveaux standards
  - SRU/W Search pour l'interrogation ?
  - Schéma de Métadonnées plus riche (RDF ?)
- Internationalisation

La communauté D-Space

- First User Group Meeting – MIT (mars 2004)
- 2005-2006 - Mise en place d'un conseil de gouvernance : Dspace Foundation
- Sept 2009 : 600 sites D-Space sur 70 pays

Fedora Repository

Flexible Extensible Digital Object Repository Architecture

www.fedora-commons.org

17 Septembre 2009

Historique

- Développement initié à la Cornell University (1997)

- 2003 : Fedora v1.0
- 2009 : Fedora v3.2

- Sept 2009 : 165 projets répertoriés

Particularités

- Architecture de type SOA avec Web Services
  - API disponibles permettant de s'interfacer avec une autre application (par exemple CMS)
- Architecture distribuée
  - Les fichiers peuvent se trouver sur un autre serveur
- Chaque objet peut avoir un ou plusieurs *disséminateurs*, c'est-à-dire un service externe qui fournit des vues extensibles des objets
- Normalisation interne : Fedora Object XML, RDF

DuraSpace

www.duraspace.org

17 Septembre 2009

Historique et objectif

- Rencontre PASIG 2008 à San Francisco (mai)
- Annonce en Mai 2009

- Coordination des 2 communautés (700 sites)
  - Partage des innovations et des développements
  - Standards communs pour interopérabilité
  - Promotion

- DuraCloud : projet pilote testant les technologies de Cloud Computing en matière de préservation

LOCKSS

Lots of Copies Keep Stuff Safe

www.lockss.org

17 Septembre 2009

Historique

- 1999 : Développement d'un prototype (Stanford, Harvard, Columbia, Berkeley, Tennessee, Los Alamos)
  - bâtir une architecture de préservation robuste, à un coût très bas
  - doit être transparent pour les utilisateurs
  - ne doit pas générer des surcoûts chez les producteurs
- Avril 2004 : v1 (logiciel Open Source)
- Août 2009 : v1.39.2
- Actuellement : équipe de 10 personnes + 5 chercheurs associés (HP Palo Alto, Intel Berkeley, Harvard, Sun)

Fonctionnement - collecte

- LOCKSS va chercher la forme publiée sur le site web du producteur (crawl)
  - Le producteur n'a pas à préparer ses contenus ou à donner accès à ses systèmes de production
  - Le site LOCKSS n'a pas à régénérer les pages web à partir des données source
  - Le producteur a la garantie que ses contenus primaires ne seront pas ré-exploités, seule la forme web peut éventuellement être réutilisée
- Le producteur doit autoriser l'accès pour la collecte par le crawler LOCKSS. La permission est donnée globalement, et non institution par institution. Ceci pour garantir la redondance et faciliter le mécanisme de réparation

Fonctionnement - stockage

- Pré-requis : 1 PC entrée/milieu de gamme
- Comparaison continuelle entre le contenu collecté et le même contenu collecté par d'autres plateformes LOCKSS ; réparation et resynchronisation en cas de différence
- Mutualisation entre sites serveurs
  - Il ne suffit que d'une seule copie par site serveur, ce sont les autres sites serveurs qui assurent la redondance (6 sites de préférence. Des mécanismes de comparaisons analysent continuellement l'état de chaque copie.
  - Pas besoin de mettre en œuvre des procédures locales d'administration et d'audit

Fonctionnement - accès

- Système transparent de type proxy : quand un utilisateur veut accéder à un contenu, LOCKSS intercepte la requête et la propage sur le site du producteur. En cas de non réponse, c'est la copie préservée qui est présentée.

Tâches d’administration

- Interface d’administration locale
  - liste des contenus,
  - vérifier l’état des collections,
  - définir les droits d’accès
- L’administrateur local LOCKSS doit définir l’URL où est le contenu numérique ainsi que les limites du crawl (ce paramétrage peut être hérité si une autre institution a déjà fait la déclaration)
- C’est aux sites serveurs de gérer les accès aux archives locales avec leur politique habituelle (IP, LDAP, Ezproxy)

LOCKSS et OAIS

- Le système supporte 3 types d'Information Packages :
  - Submission Information Package = publisher manifest, page à la charge du producteur
  - Archival Information Package = décrit le contenu, les métadonnées extraites du Manifest, les entêtes HTTP ainsi que des métadonnées externes décrites en XML
  - Dissemination Information Package, qui en fait reprend les informations du SIP

Lockss.stanford.edu/technicalspecificationsOAIS.htm

Préservation

- LOCKSS préserve les contenus dans le format web avec lequel ils ont été publiés. Migration à la volée des contenus si le browser ne supporte plus le format (dialogue http)

Description détaillée : Transparent Format Migration of Preserved Web Content ; David S. H. Rosenthal and al ; Stanford University Librairies ; D-Lib Magazine ; janvier 2005
www.dlib.org/dlib/january05/rosenthal/01rosenthal.html

Bilan : les manques

- Outils d'administration :
  - Gestion des accès assez rudimentaire (par IP)
  - Pas adapté pour gérer diverses communautés en parallèle
  - Pas de notion de collections virtuelles

- Pas de fonction avancée
  - Moteur de recherche
  - Alertes sur les contenus archivés

Bilan : les plus

- Logiciel libre, JAVA, XML, OpenBSD
- Logique OAIS
- Fonctionnalités simples, transparentes et efficaces
- Machine autonome (crawl, audit, réparation)
- Applicable à tout contenu web
- Assistance utilisateur de Stanford
- Communauté (négociation éditeurs, plug-in)

- Economique
- Oblige à travailler en réseau

Quelques projets basés sur LOCKSS

- Meta Archive Project : Library of Congress + 6 universités US (réseau privé thématique) ; culture et histoire de l'Amérique du Sud
- Alabama Digital Preservation NETwork : 8 institutions (réseau privé régional)
- Association of Southeastern Research Libraries : 8 Universités US (réseau ouvert) : thèses
- Government Printing Office Pilot : journaux électroniques fédéraux : 18 universités + GPO + bibliothèques
- Alaska State Publications Program

CLOCKSS - Controlled LOCKSS

- Le projet est lancé en janvier 2006 avec une liste « fermée » de partenaires : 7 bibliothèques et 10 éditeurs + 1 (Elsevier qui participe aux discussions et au financement).

- Durée : deux ans (prototype)
- Version de « production » : 2009

- Basé sur la technologie LOCKSS dont le principe fondateur est la mise en place d'une architecture d'archivage distribuée, répliquée et économique

UK LOCKSS Pilot Programme

- Partenariat entre le JISC (Joint Information Systems Committee) et CURL (Consortium of research Libraries in the British Isles)
- 24 bibliothèques UK sélectionnées
- Pilote du 1er mars 2006 au 31 juillet 2008
  www.jisc.ac.uk/whatwedo/programmes/preservation/lockss
- Bilan : « Evaluation of the JISC UK LOCKSS Pilot, Birmingham City University, Mai 2008 » (www.ebase.bcu.ac.uk) ; principale recommandation : établir des liens avec les résolveurs de liens

LOCKSS UK Alliance

- Depuis Août 2008, participation payante et ouverture à de nouvelles bibliothèques :
  - http://www.jisc-collections.ac.uk/catalogue/lockss
- PLN - Private LOCKSS Networks
  - Lancé en juin 2008 lors de la Joint Conference on Digital Librairies (JCDL) à Pittsburg

Autres outils

Autres communautés Open Source

- E-Prints, CDSware Invenio, Greenstone
- Les outils : JHOVE, DROID, Parser XML
- Les initiatives européennes
  - Planets (Preservation and Long-term Access through Networked Services) (2006-2010) (exemple : Plato)
  - Alliance pour Permanent Access (données scientifiques)
  - Caspar (Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval)
  - DPE (Digital Preservation Europe)
  - WePreserve = coordination Caspar/DPE/Planets

Evaluation

OAIS et son implémentation dans les logiciels disponibles

- Tous les logiciels se revendiquent de l'OAIS…
  - Qu'en est-il réellement ?

- Etude BNF dans le cadre du projet SPAR (2005)

  Mise en place de 2 évaluations
  - Évaluation à « grosse maille » : large nombre de solutions explorées / nombre limité de critères
  - Evaluation à « petite maille » : évaluation vis-à-vis des contraintes BnF des solutions préalablement sélectionnées

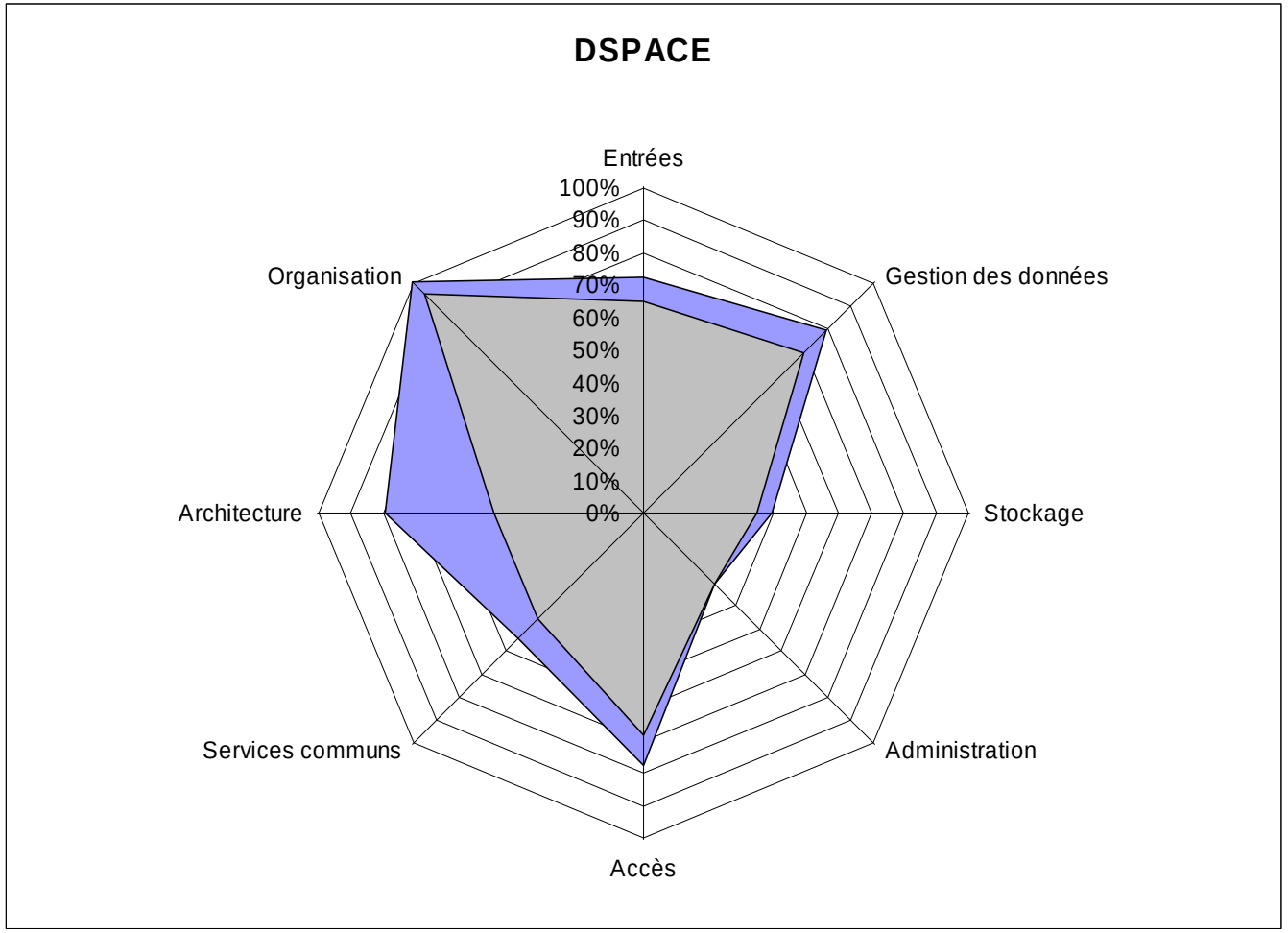
vds.cnes.fr/pin/presentations/2007/Presentation_SPAR.pdf

Evaluation approfondie

Grille en 132 critères répartis en 9 catégories

1. L'entité « Entrée »
2. L'entité « Gestion de données »
3. L'entité « Stockage »
4. L'entité « Administration »
5. L'entité « Planification de la pérennisation »
6. L'entité « Accès »
7. Les services communs
8. L'architecture
9. L'organisation

Exemple de synthèse



Bâtir un référentiel ?

- Faire de la veille et tester des plateformes est un travail lourd, qui monopolise les équipes techniques et les équipes métier
  - Le nombre de logiciel à évaluer est assez conséquent
  - Il y a souvent de nouvelles versions
- Constat : des organismes ont déjà investi un temps important dans ces évaluations (BNF, CINES, DAF, INIST...)
- Il faudrait pérenniser ces évaluations et partager les résultats : monter un référentiel ?

Survol du CLIR : Council on Library and Information Resources (2006)

- E-Journal Archiving Metes and Bounds: A Survey of the Landscape - Anne R. Kenney, Richard Entlich, Peter B. Hirtle, Nancy Y. McGovern & Ellie L. Buckley – Sept. 2006

  http://www.clir.org/pubs/reports/pub138/pub138.pdf

- Etude sponsorisée par le CLIR (Washington) sur la période février-juin 2006

- Synthèse présentée lors de la 8ème réunion de l'International Coalition of Library Consortia (Rome) : notes de Pierre Carbone (Couperin) – 12 Initiatives analysées

Autres survols

- 2007 : « Preserving scientific electronic journals:a study of archiving initiatives » ; The Electronic Library - Vol.26 - No.1 – 2008 www.emeraldinsight.com/0264-0473.htm
  - JSTOR, Portico, E-Prints, Lockss, OCLC Digital Archive, JICS, Pub Med Central, KB e-Depot

- 2008 : Long Term Preservation – Survey Research Results ; Sarah Durrant - ALPSP (Association of Learned and Professional Society) ; Juillet 2008 www.alpsp.org
  - Une vision plutôt éditeur
  - PubMedCentral, Portico, LOCKSS, CLOCKSS, KB e-Depot
  - C'est du ressort des bibliothèques nationales et des centres de dépôt légal d'assurer l'archivage pérenne

Critères d'évaluation

- Positionnement des solutions : 3 critères
  - Conformité OAIS
  - Neutralité vis-à-vis de la technologie : la solution doit pouvoir être implémentée sur n'importe quelle nouvelle technologie
  - Indépendance du domaine : la solution doit pouvoir être utilisable dans des contextes variés, dans des organisations du secteur public ou du secteur privé

Audit et certification

- OCLC / RLG - Digital Archive Attributes Working Group
- RLG / NARA Task Force on Digital Repository Certification (www.rlg.org)
- DPC (Digital Preservation Coalition) Technology Watch Report, Brian F. Lavoie (OCLC), janvier 2004 (http://www.dpconline.org/docs/lavoie_OAIS.pdf)
- Drembora (2007, DCC)

Pour finir

- Imaginer le pire !
  - Pérennité de la solution, rôle de la communauté, aptitude à participer ou à reprendre le développement
  - Comment sortir d'un logiciel ?

- Travailler dans un réseau d'acteurs
  - Archives distribuées, partagées…
  - Abandon / reprise / fusion d'archive