La préservation de l'accès aux revues numériques - 2

Fabrice Lecocq (lecocq@inist.fr)

Institut de l'Information Scientifique et Technique
CNRS - UPS 076

Présentation Groupe PIN
10 mai 2006

LOCKSS
Lots of Copies Keep Stuff Safe

- Objectifs et historique
- Fonctionnement
- Situation actuelle
- Les initiatives utilisant LOCKSS
- Bilan

Stuff = chose, truc

Constat (1)

- Le principal frein à la préservation de contenus numériques est une limitation économique :
  - personne n'a les moyens, seul, de tout préserver
  - plus la solution sera économique, plus il y aura de contenus préservés, et ce pour une période plus longue

Constat (2)

- Plus un système de préservation est utilisé, et ce de manière aisée (web), plus les contenus sont vérifiés continuellement.
- Ceci est à opposer aux systèmes « archives noires » où les documents sont confiés uniquement pour préservation et ne sont pas accessibles. Peut-on faire une confiance aveugle à de tels systèmes ?

Pourquoi choisir LOCKSS ?

C'est un plus pour les éditeurs

C'est un plus pour les utilisateurs

Avantages pour les éditeurs

- Pas de contraintes techniques :
  - LOCKSS va chercher la forme publiée sur le site web de l'éditeur (crawl). L'éditeur 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 éditeurs
  - L'éditeur a la garantie que ses contenus primaires ne seront pas réexploités, seule la forme web peut éventuellement être réutilisée

Avantages pour les usagers

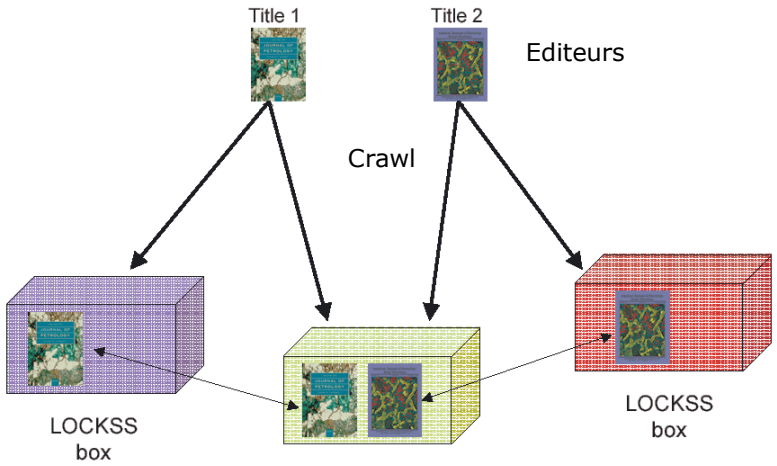
- Pas de contraintes aux utilisateurs
  - Système transparent de type proxy : quand un utilisateur veut accéder à une revue, LOCKSS intercepte la requête et la propage sur le site de l'éditeur. En cas de non réponse, c'est la copie préservée qui est présentée
- Mutualisation entre sites serveurs
  - Il ne suffit que d'une seule copie par site serveur, ce sont les autres sites serveurs qui assurent la redondance. 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

Le projet LOCKSS - Historique

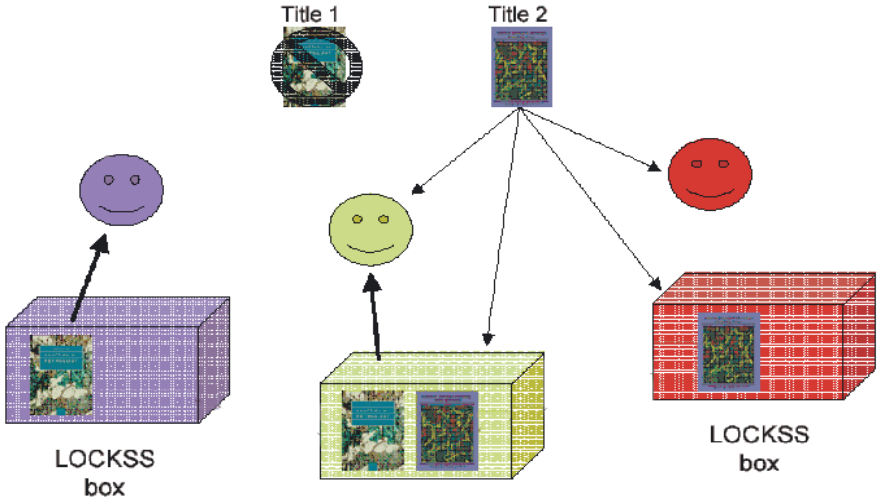
- 1999 : Développement d'un prototype (Stanford, Harvard, Columbia, Berkeley, Tennessee, Los Alamos)
- 2000-2002 : Test entre 50 bibliothèques du monde entier
- 2002 : Financement par la Andrew W. Mellon Foundation du développement par l'Univ. de Stanford d'une version de production
- 2003-2004 : Test de la version de production entre plusieurs universités américaines
- Avril 2004 : V1 (logiciel Open Source)
- 2005 : protocole avancé pour la réplication et le vote
- Actuellement : équipe de 5 ingénieurs + 5 chercheurs associés (HP Palo Alto, Intel Berkeley, Harvard, Sun)

Fonctionnement

Acquisition des contenus



Accès aux contenus



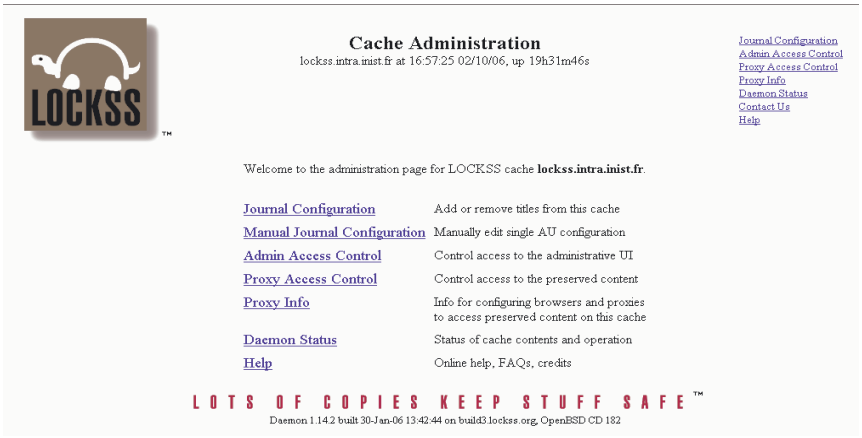
Fonctionnement

- Pré-requis : 1 PC entrée de gamme
- Scan, par utilisation d'un web crawler, des nouveaux titres des journaux auquel on est affilié
- Comparaison continuelle entre le contenu collecté et le même contenu collecté par d'autres plateformes LOCKSS ; votation ; réparation et resynchronisation en cas de différence
- Interface d'administration locale pour définir la liste des titres, vérifier l'état des collections et définir les droits d'accès
- Conseil : il vaut mieux qu'il y ait au moins 6 sites qui acceptent de préserver 1 même titre pour garantir la pérennité

Tâches d'administration

- L'éditeur doit autoriser l'accès pour la collecte par le crawler LOCKSS
- L'administrateur local LOCKSS doit définir l'URL où est la revue ainsi que les limites du crawl (ce paramétrage peut être hérité si une autre bibliothèque a déjà fait la déclaration = LOCKSS « plug-in »)

Interface d'administration

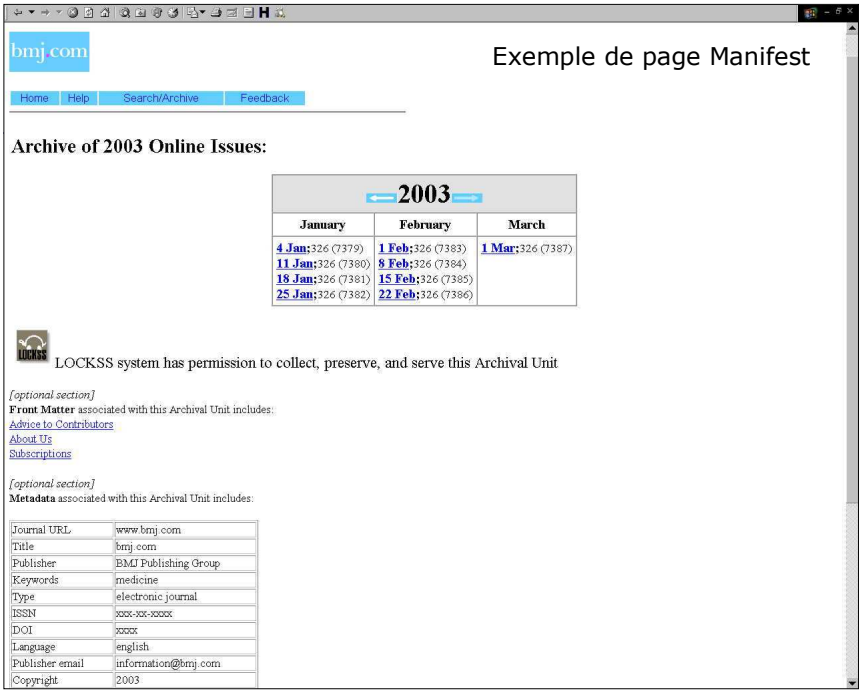


Définition des accès

- Management des revues :
  - Métadonnées
  - Cluster (Archival Unit)
- Gestion des alertes (surveillance des collections archivées)
- Gestion des accès :
  - Déclaration par adresse IP
  - LDAP serait supporté (à confirmer)
  - Ezproxy supporté (testé)
  - Shibboleth à l'étude (spécifications SAML ; fédération d'Identité et Web Services)

Les éditeurs - Fichier Manisfest

- Les éditeurs donnent la permission du crawl par LOCKSS globalement, et non institution par institution. Ceci pour garantir la redondance et faciliter le mécanisme de réparation. C'est généralement la LOCKSS Alliance (association) qui fait cette demande
- C'est aux sites serveurs de gérer les accès aux archives locales avec leur politique habituelle (IP, LDAP, Ezproxy)



bmj.com

Exemple de page Manifest

Home | Help | Search/Archive | Feedback

Archive of 2003 Online Issues:

← 2003 →		
January	**February**	**March**
4 Jan;326 (7379)	**1 Feb**;326 (7383)	**1 Mar**;326 (7387)
11 Jan;326 (7380)	**8 Feb**;326 (7384)	
18 Jan;326 (7381)	**15 Feb**;326 (7385)	
25 Jan;326 (7382)	**22 Feb**;326 (7386)	

LOCKSS system has permission to collect, preserve, and serve this Archival Unit

[optional section]
Front Matter associated with this Archival Unit includes:
Advice to Contributors
About Us
Subscriptions

[optional section]
Metadata associated with this Archival Unit includes:

Journal URL	www.bmj.com
Title	bmj.com
Publisher	BMJ Publishing Group
Keywords	medicine
Type	electronic journal
ISSN	xxx-xx-xxxx
DOI	xxxx
Language	english
Publisher email	information@bmj.com
Copyright	2003

Les contenus : Plug-in éditeur

- Pour que LOCKSS puisse crawler les revues, nécessité d'un plug-in par éditeur (en fait, 1 plug-in par Archival Unit = volume)
- Le plug-in peut être écrit
  - Par l'éditeur lui-même
  - Par un des sites d'archivage
  - Par la LOCKSS Alliance

Les contenus - Plug In éditeur

- Le plug-in décrit :
  - Le site web éditeur : méthode d'authentification, gestion des erreurs, limites de crawling
  - Les caractéristiques et la structure des revues : fréquence de publication, URL de départ,
  - Le contenu dans les revues

  Ecrit en code XML…

  … mais il existe un outil java de génération de plug-in

Configuration

- Machine dédiée
  - CPU 2,5 Ghz (mini 600 Mhz)
  - RAM 512 Mo (mini 128 Mo)
  - Lecteur de cédérom
  - Lecteur de disquette
  - Disque : 250 Go (mini 60 Go)
  - Réseau Ethernet (avec IP fixe)
- Soit des configurations économiques à 1000-1500€

Logiciel

- Le serveur LOCKSS tourne avec un Système d'Exploitation préconfiguré (Open BSD)
- Lors du processus d'installation, les paramètres du site serveur sont générés et un cédérom bootable en lecture seule est créé.
- Le site serveur tourne avec le Système d'Exploitation directement sur le cédérom

- LOCKSS est écrit en java (portabilité)

Situation actuelle

Institutions participantes

- Etats-Unis : près de 60 universités + Bibliothèque du Congrès, centre de Los Alamos
- Canada : 2 universités + CISTI
- Europe : UK (7 universités dont Cambridge + British Library), RFA (6), Pays-bas (3), Italie (2) puis Belgique, Finlande, Grèce, Norvège, Portugal, Espagne, Suède, Suisse (tous 1)
- Australie (1), Nouvelle-Zélande (1), Chine (5), Singapour (1), RSA (4), Israël (2), Brésil (1)

Les éditeurs acceptant LOCKSS

- BioOne (20 titres)
- Institute of Physics (2)
- Johns Hopkins University Press (30 titres)
- Indiana University Press (15 titres)
- Oxford University Press (15 titres)
- SAGE Publications (3 titres)

- Revues en Open Access (35)

- Publishers partners : BioMed, Blackwell, Emerald Groups, Lippincott, Nature PG, Springer, HighWire Press
- Annonce Elsevier du 10 janvier 2006

Pb potentiels (1)

- DOI (Digital Object Identifier)
  - A priori stable. Il est supposé que si l'éditeur change l'organisation de ses collections, il devra mettre à jour ses contenus et leur DOI et le signaler aux LOCKSS Box
- Migration de contenu
  - LOCKSS préserve les contenus dans le format web avec lequel ils ont été publiés. LOCKSS converti de manière transparente les contenus pour suivre l'évolution des readers ; 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

Pb potentiels (2)

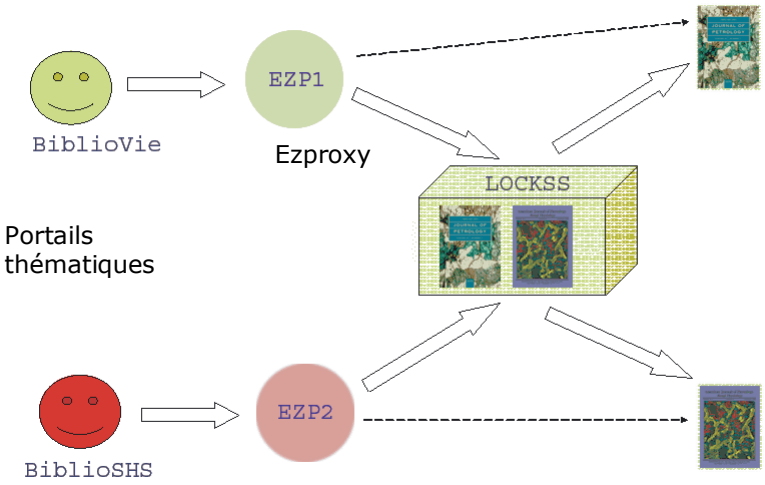
- On n'a pas la main sur la machine et les changements doivent être coordonnées avec l'équipe de Stanford

  Les tests montrent que finalement cela se passe très bien
- Machine up tout le temps, pas de RAID, pas de backup local. Confiance en la communauté…
- Certains éditeurs ne donnent accès à LOCKSS que pendant une courte période ; il faut collecter dès que les contenus sont mis en ligne

Pb potentiels (3)

- Outils d'administration :
  - Gestion des accès assez pauvre (par IP)
  - Pas adapté pour gérer diverses communautés en parallèle (le faire à l'extérieur)
  - Pas de notion de collections virtuelles
- Pas de fonction avancée (search, alertes sur les contenus…) ; comment se fera l'interopérabilité avec la couche au-dessus (portail) ? La recherche fédérée peut-elle fonctionner avec de telles archives…

LOCKSS à l'INIST



LOCKSS et OAIS (1)

LOCKSS et la description des contenus :

- Content Information : extraite du Header http, en particulier les types MIME
- Preservation Description Information : décrit la provenance (l'URL initiale), le contexte (liens à d'autres objets), des références (URL et tout identifiant ou métadonnées utiles pour la recherche) et des infos sur l'état du contenu (Fixity) déduite du protocole de réplication et d'auto-surveillance des LOCKSS Box
- Packaging information

LOCKSS et OAIS (2)

- Le système supporte 3 types d'Information Packages :
  - Submission Information Package = publisher manifest, page à la charge de l'éditeur
  - Archival Information Package = instance Java décrivant le contenu, les métadonnées extraites de la page 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

Les initiatives utilisant LOCKSS

Modèle économique

- LOCKSS, logiciel libre
  - On peut l'utiliser librement pour ses propres contenus ou pour les revues en Open Access proposées sur le site lockss (200 titres)
- Que faire dans les autres cas ?
  - Aller voir ses éditeurs
  - Faire le paramétrage (écriture des plug-in)
  - Se grouper à plusieurs pour la réplication

LOCKSS Alliance

- Avantages :
  - support technique avancé de Stanford
  - Informations en primeur
  - Influence sur le développement / la gouvernance
- Dans les faits, on ne peut installer des collections payantes si on n'appartient pas à l'Alliance
- Coût : de 1100 $ (collège) à 11000 $ (Université)
  Ira à la baisse en fonction du nb de participants

Controlled LOCKSS (janvier 2006)

- Exploiter l'originalité de LOCKSS, mais créer des « archives noires » sous contrôle des éditeurs
  - Contenus : American Medical Association, American Physiological Society, Blackwell, Nature Publishing Group, OUP, SAGE Publications, Springer, Taylor and Francis, John Wiley & Sons (Elsevier en support financier)
  - Sites : Edinburgh University, Indiana University, New York Public Library, Rice University, Stanford University, University of Virginia
  - www.lockss.org/clockss

Autres utilisations de LOCKSS

- JISC (Joint Information System Commitee) / CURL (Consortium of Research Librairies) / 20 Universités anglaises (décembre 2005)
  - Monter un service+, au dessus de LOCKSS et à l'échelle nationale (www.jisc.ac.uk)
- MetaArchive (Bibliothèque du Congrès + 5 bibliothèques américaines)

Initiatives concurrentes

- JSTOR Electronic Archiving Initiative
- Portico – Mellon Foundation
  - 2 premiers éditeurs à l'origine : Elsevier et Oxford University Press
  - Actuellement 8 éditeurs
  - Service payant sur un modèle proche de celui de la Bibliothèque Royale des Pays-Bas (auquel participe aussi Elsevier)

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)

- Excellent logiciel
- Economique

Bilan : les moins

- Liste de revues disponibles limitée pour l'instant
- Timidité des grands éditeurs (ils sont là en « observateur »)
- Interface utilisateur rustique, bien loin du niveau fonctionnel de EJOS/SDOS (commercial) ou de logiciel type D-Space (Open Source)

Les tendances

- LOCKSS, future couche basse d'autres logiciels ? (exemple couplage D-Space / LOCKSS à l'étude)
- Reprise en main des éditeurs (CLOCKSS)
- Organisation de fédérations soit nationales (UK) soit thématiques

INIST - Plateformes utilisées

Selon les contrats passés avec les éditeurs :

- SDOS (Science Direct On Site) / EJOS (Electronic Journal On Site)
  - cible : EPST
  - Contenus : la majorité des gros éditeurs (5000 titres)
- D-Space : plutôt dans un contexte Open Access et pour des accès au-delà des EPST
- LOCKSS : demandé par certains éditeurs

Bibliographie LOCKSS

- Site LOCKSS : www.lockss.org
- www.diglib.org/preserve/stanfordfinal.pdf
  LOCKSS: A Distributed Digital Archiving System - Progress Report For The Digital Library Federation Preservation Web Site ; Mellon Electronic Journal Archiving Program ; Stanford University Libraries ; 8 October 2002
- www.istl.org/02-fall/article1.html
  LOCKSS As A Cooperative Archiving Solution for E-Journals ; Victoria A. Reich ; Stanford University ; 2002
- www.diglib.org/preserve/introduction.pdf
  Archiving Electronic Journals ; The Digital Library Federation ; Council on Library and Information Resources ; Washington DC ; 2003

Bibliographie LOCKSS

- A Fresh Look at the Reliability of Long-Term Digital Storage ; Mary Baker, Mehul Shah, David S.H. Rosenthal, Mema Roussopoulos, Petros Maniatis, TJ Giuli, Prashanth Bungale ; HP Labs, Palo Alto – Stanford University – Harvard University – Intel Research, Berkeley ; dépôt sur arXiv – août 2005
- Preserving Peer Replicas by Rate-Limited Sampled Voting ; Mary Baker, David S.H. Rosenthal, Mema Roussopoulos, Petros Maniatis, TJ Giuli, Yanto Muliadi ; Stanford University ; 2005