

Préservation d'Internet : enjeux et perspectives

Catherine Lupovici

Directrice du Département de la Bibliothèque Numérique
Direction des Services et des Réseaux
Bibliothèque nationale de France

1

Internet est devenu un nouveau moyen de communication

- ◆ Croissance continue de l'Internet dans les pays développés
 - 22 habitants/serveur domaine national en France, 24 en Allemagne, 16 au Royaume Uni, 11 en Italie, 10 au Canada, 6 en Suède, 5 aux Pays Bas, 7 en Belgique, 8 en Australie, 4 pays scandinaves
 - + 831 % de serveurs du domaine .fr de 1998 à 2004
- ◆ Le web est devenu en 10 ans **la** plate-forme applicative de l'Internet
- ◆ OCLC Web Characterization study
 - Juin 2002, 3 080 000 sites publics
 - 35% du web, 1,4 milliards pages, moyenne de 441 pages par site

2

Les contenus du web

◆ Typologie des contenus

- Publications classiques (dont l'auto-publication) & littérature grise
- Données brutes d'institutions savantes et de sociétés commerciales
- Communication interpersonnelle
- Programmes de radio & TV, diffusion ou à la demande
- Services

◆ Caractéristiques et défis techniques

- Masse et fréquence de modification
- Liens à l'intérieur du web mais aussi avec les ressources classiques
- Web de surface et web profond (non accessible au robots)

3

Garder la mémoire de l'Internet

- ◆ Réflexion entamée mi-décennie 90. Très peu de pré-web conservé. Les « incunables » d'Internet
- ◆ Acteurs : organismes de recherche (exemple Internet Archive) et institutions de mémoire (exemple Bibliothèques nationales)
- ◆ Principes de l'archivage du web à grande échelle
 - Les unités documentaires sont les sites. Ils peuvent contenir des documents numériques encore publiés en parallèle de manière classique, mais ce ne sont plus les unités documentaires de mémorisation
 - La technique de base est proche de celle des moteurs de recherche (Google, Voila, Yahoo!..) : robot parcourant les liens internes et externes des sites

4

Approche des bibliothèques nationales

- ◆ Historiquement deux approches opposées dans les premières expérimentations
 - Canada (1994) et Australie(1996) dépôt des ressources uniquement numériques (ex revues électroniques)
 - Suède (1997), collecte périodique automatique du domaine suédois par robot
 - Library of Congress, USA, collecte thématique par robot (Elections présidentielles 2000, 11 septembre 2001)
- ◆ Approches complémentaires
 - Collecte par robot permet de conserver les liens et la capacité à naviguer dans l'archive. Des zooms peuvent permettre un suivi rapproché pour des événements
 - Le dépôt permet de conserver le web profond et d'effectuer des actions préventives de conservation, si besoin en relation avec les producteurs

5

International Internet Preservation Consortium

- ◆ Programme de R&D de trois ans (août 2003- juillet 2006)
- ◆ 11 bibliothèques nationales (Australie, Canada, Danemark, Finlande, France, Islande, Italie, Norvège, Suède, UK, US) et Internet Archive
- ◆ Programme de travail: architecture cadre pour le développement d'outils, besoins des chercheurs, mesures statistiques et d'évaluation, définition et traitement du web profond, gestion des contenus collectés, outils d'accès, développement d'un crawler complexe adapté
- ◆ Tous les outils développés seront en open source

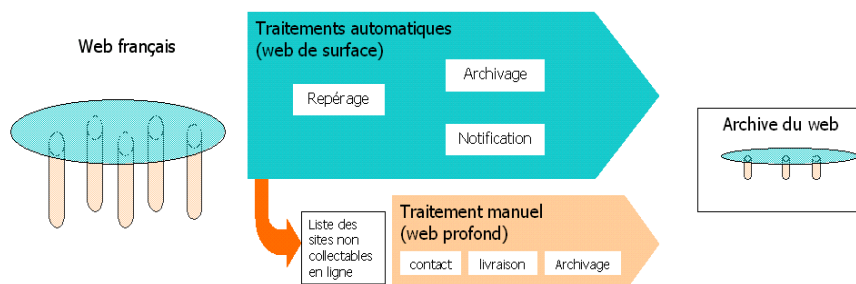
6

Les expérimentations à la BnF

- ◆ Préparer l'évolution de la législation sur le DL
 - titre IV du projet de loi droit d'auteur et droits voisins
 - décret d'application fixera les modalités pratiques
- ◆ Collecte thématique des sites électoraux
 - élections présidentielles et législatives 2002 (500 Go, 12 M de fichiers)
 - élections régionales et européennes 2004
- ◆ Test grandeur réelle du robot suédois sur .fr.
- ◆ Etude sur des outils de repérage de contenus: projet WATSON
- ◆ Etude de procédés de dépôts avec une centaine de producteurs

7

Modèle fonctionnel intégré défini à la suite des expérimentations



8

Conclusion

- ◆ La mémorisation de l'Internet en est à ses débuts
- ◆ Importance du suivi et de la contribution à l'évolution des technologies de traitement automatique de grandes masses d'informations liées à l'évolution du web
 - pour le recueil des informations à conserver
 - pour l'organisation et la conservation à long terme des informations recueillies
 - pour offrir l'accès à des entrepôts interopérables aux chercheurs de demain