World Library and Information Congress:
71th IFLA General Conference and Council

"Libraries - A voyage of discovery"

August 14th - 18th 2005, Oslo, Norway

Conference Programme:
http://www.ifla.org/IV/ifla71/Programme.htm

septembre 28, 2005

Code Number: **194-F**
Meeting: 150 SI - ICABS (IFLA/CDNL Alliance for Bibliographic Standards)

Accès à long terme et interopérabilité pour les archives du web: l'activité du Consortium International pour la Préservation d'Internet

Catherine Lupovici
IIPC Program Officer
Bibliothèque nationale de France
catherine.lupovici@bnf.fr

Résumé: *Le Consortium international pour la préservation d'Internet a été créé en juillet 2003 par douze institutions pionnières dans l'archivage du web dont onze bibliothèques nationales. Les objectifs du consortium sont d'offrir un lieu d'échanges des connaissances sur la préservation des contenus de l'Internet, de développer des outils interopérables, des méthodologies et des standards nécessaires pour collecter, archiver et donner accès aux contenus de l'Internet. Un ensemble d'outils pour la chaîne complète de traitement seront publiés en logiciels libres à la fin des trois premières années d'existence du consortium. L'activité de normalisation d'IIPC couvre les aspects de format pour les archives du web, les métadonnées de préservation et l'identification permanente des ressources concernant plus particulièrement les particularités des très larges archives.*

Je voudrais tout d'abord remercier ICABS pour son invitation à présenter l'activité du Consortium International pour la Préservation d'Internet qui a, dans son domaine spécifique, des objectifs très convergents avec ceux d'ICABS. Cette présentation donne tout d'abord des informations générales sur l'organisation, les objectifs et les résultats du consortium puis elle détaillera le travail de normalisation en cours pour l'accès à long terme et l'interopérabilité des archives du web.

1. IIPC (http://netpreserve.org)

L'IIPC a été créé par douze institutions déjà engagées dans l'archivage du web afin:

- d'offrir un lieu d'échanges des connaissances sur la préservation des contenus d'Internet entre les membres du consortium mais également plus largement
- de développer et recommander des standards
- de développer des outils intéropérables et des techniques pour collecter, préserver et donner accès aux contenus archivés
- de sensibiliser sur les questions et les initiatives relatives à la préservation de l'Internet au travers de conférences, d'ateliers, de formations, de publications etc.…

1.1. Organisation de l'IIPC

Le consortium a été créé en juillet 2003, pour une durée initiale de 3 ans par les membres suivants :

- Bibliothèque nationale de France, responsable du programme
- National Library of Australia
- Bibliothèque et archives Canada
- Bibliothèque nationale du Danemark
- Bibliothèque nationale de Finlande
- Bibliothèque nationale d'Islande
- Bibliothèque nationale d'Italie, Florence
- Bibliothèque nationale de Norvège
- Bibliothèque nationale de Suède
- The British Library (UK)
- Library of Congress (USA)
- Internet Archive

L'idée de créer ce consortium est née à l'occasion des discussions entre des bibliothèques nationales et des institutions de recherche déjà engagées dans des opérations d'archivage du web, et qui avaient pu échanger leurs expériences au cours d'ateliers organisés par la BnF lors des conférences ECDL (European Conference on Research and Advanced Studies on Digital Libraries) en 2001 et 2002. Ces échanges ont permis de mettre en évidence le besoin d'un robot intelligent et la BnF a proposé en janvier 2003 aux membres de COBRA et de Bibliotheca Universalis la création du consortium. Douze partenaires ont accepté de consacrer un budget à un tel programme et le consortium a été officiellement créé en juillet 2003 pour une durée initiale de trois ans. Les partenaires ont décidé de limiter la participation aux membres fondateurs pendant cette première période afin de construire un ensemble minimum d'outils avec des partenaires pionniers.
A la fin de l'année 2005, un appel à de nouvelles adhésions sera lancé pour une extension du consortium, non seulement pour d'autres bibliothèques nationales, mais aussi pour d'autres institutions comme les archives ou des institutions de recherche.

1.2. Programme de travail d'IIPC

Le consortium a adopté une démarche pragmatique en définissant deux niveaux de travail. Six groupes de travail ont été créés, pilotés chacun par un des membres, chaque partenaire contribuant en fonction de son expertise. De plus des projets spécifiques sont acceptés et soutenus par le consortium avec la possibilité d'une subvention partielle. Ces projets doivent impliquer au moins deux membres du consortium et bien entendu contribuer à ses objectifs.
Les livrables prévus comprennent des outils, qui seront publiés en open source avec licence gratuite de type GPL (General Public License), des recommandations sur les méthodologies, les traitements et des standards.

Le Comité de pilotage approuve les objectifs des groupes de travail, le soutient à des projets ainsi que le niveau de financement qui leur est accordé. Il suit l'état d'avancement des travaux des groupes de travail et des projets soutenus.

1.2.1. Groupes de travail sur les méthodes et les standards

- **Le groupe de travail « Framework »**

Le groupe de travail *Framework* a pour objectifs de définir le cadre général pour toutes les activités techniques du consortium de manière à garantir une interopérabilité future entre toutes les archives et de préparer les standards internationaux qui seront nécessaires. Ce groupe de travail a défini dès le tout début du projet l'architecture de haut niveau sur laquelle tous les autres groupes de travail appuient leur activité. Les modules fonctionnels ainsi identifiés concernent uniquement l'archivage du web et doivent pouvoir échanger via des APIs (Application Programming Interface) avec les systèmes de chaque institution en fonction des implémentations locales.

- **Le groupe de travail « Metrics and Test Bed »**

Le groupe de travail *Metrics & Test Bed* a pour objectifs de définir et de mettre en place un banc de test de robots pour évaluer un robot de collecte particulier par rapport aux différents problèmes de collecte qui peuvent être rencontrés, en raison soit des formats des contenus soit de la présence de passerelles d'accès à ces contenus à l'intérieur des sites web. Ces limitations techniques ont été identifiées à partir de travaux antérieurs et d'une étude conduite par la Library of Congress et sont publiées sur le site web du consortium[1].

1.2.2. Groupes de travail pour le développement d'outils

- **Le groupe de travail « Access tools »**

Le groupe de travail *Access tools* a commencé par la définition de *use cases* pour dresser la liste des différents besoins et type d'accès. On a distingué l'accès pour les conservateurs pendant l'étape de sélection, l'accès pour les besoins de gestion des contenus au niveau intellectuel, l'accès pour la gestion de l'archive et bien entendu l'accès pour l'utilisateur final. Il a été décidé pour la première étape de se limiter à l'accès pour l'utilisateur final et de développer un outil pour l'indexation et la navigation dans de très larges archives.

- **Le groupe de travail « Deep Web »**

Le groupe de travail *Deep Web* se consacre à l'identification du web profond inaccessible aux robots et au développement d'outils pour le dépôt et l'accès aux contenus auxquels on accède par une passerelle documentaire gérée dans une base de données. Il est en effet très important pour les institutions qui veulent gérer le dépôt de contenus du web profond de pouvoir conserver les descriptions stockées dans des bases de données avec les objets qui y sont décrits. Ce groupe de travail a déjà terminé ce programme initial et se consacre actuellement à la veille et à l'évaluation des évolutions technologiques des moteurs de recherche génériques pour l'indexation du web profond.

1.2.3. Les groupes de travail pour les politiques documentaires et les définitions de besoins

- **Le groupe de travail « Content management »**

Le groupe de travail *Content Management* a été créé pour permettre aux partenaires de construire une vision commune de la couverture documentaire et pour permettre à chaque entrepôt de pouvoir compléter les autres. La couverture documentaire peut être comprise de manière différente en fonction de la politique d'acquisition retenue. Cela peut correspondre à une collecte intensive en profondeur d'un petit nombre de sites ou à une collecte extensive transversale d'un très grand nombre de sites. La collecte peut également être thématique (comme par exemple des élections) ou limitée à un domaine comme par exemple un domaine géographique national tel que .se.

[1] http://netpreserve.org/publications/reports.php

Le travail a été recentré sur une partie des besoins pour faciliter le développement d'un ensemble d'outils centraux communs à toutes les approches possibles.

– **Le groupe de travail « Researchers »**

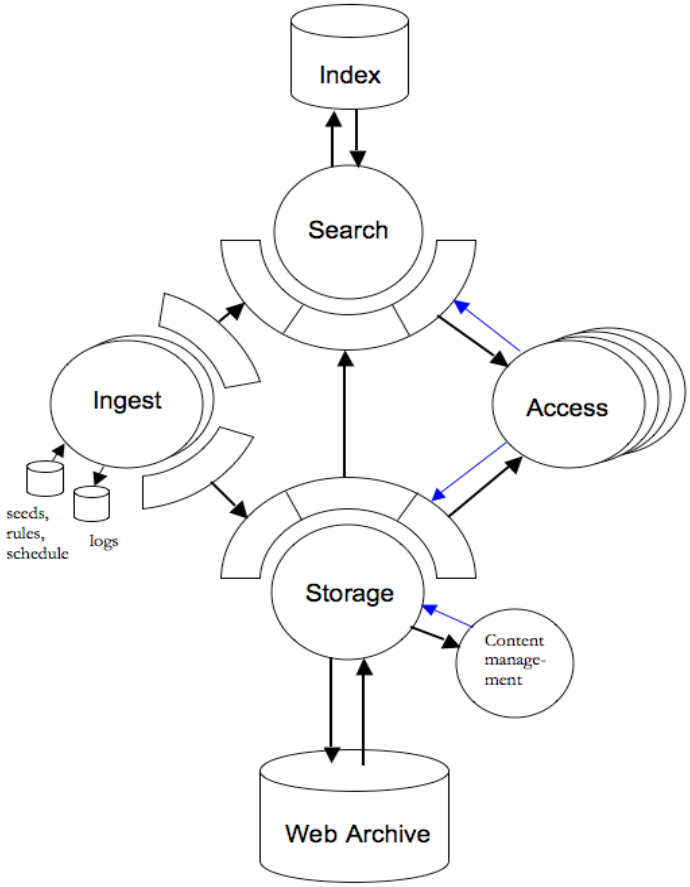
Le groupe de travail *Researchers* a été constitué pour donner des avis et des réactions sur ce qu'il convient d'inclure dans des archives du web et sur les modes d'indexation. Par exemple ne serait-il pas intéressant de conserver les liens entrants d'un site avec le site lui-même comme une information de contexte ? Les membres du groupe de travail seront également invités à évaluer les outils du consortium destinés à l'utilisateur final des archives. Le groupe est composé de 40 chercheurs représentant les différents secteurs représentés sur Internet, tels que la littérature, les sciences politiques, l'histoire, l'informatique. Certain d'entre eux sont membres de l'Association of Internet Researchers ce qui a conduit à coordonner les travaux du groupe de travail avec les conférences de l'AoIR.

1.3. Les outils de l'IIPC

Une fois l'architecture globale définie et stabilisée, le comité de pilotage a décidé, à mi-parcours du programme, de mettre l'accent sur la finalisation d'un ensemble complet d'outils pour les rendre disponibles à la fin de la première phase du programme. L'objectif est de couvrir les fonctions centrales de l'archivage du Web sur les quatre points suivants :

– Sélection et vérification pour alimenter les traitements d'acquisition
– Acquisition par collecte automatique ou par dépôt
– Stockage et maintenance des collections
– Accès

L'architecture de haut niveau définie pour le développement des outils est la suivante :



Les outils d'IIPC concernent toute la chaîne de production et seront finalisés avant mi 2006. Ils permettront les fonctions de base et devront être robustes et capables de fonctionner à l'échelle du web global. Ils implémenteront les standards d'IIPC pour le format d'archivage, les métadonnées et les APIs. Ils devront être faciles à installer et à utiliser pour un utilisateur avancé et sont destinés à un ingénieur informatique d'archivage du web. Ils seront en open source et utilisables gratuitement par la communauté d'archivage du web. Certains d'entre eux sont déjà réalisés et en cours de test et de validation. Les outils d'acquisition qui sont déjà disponibles et permettent l'utilisation des standards d'IIPC pour la préservation à long terme et l'interopérabilité sont les suivants :

- **Heritrix** est un robot de collecte à grande échelle développé conjointement par Internet Archive et les bibliothèques scandinaves sur la base de spécifications écrites dès 2003. Heritrix permet à la communauté de l'archivage du web la capacité d'améliorer la découverte de contenus au-delà de l'encodage des liens en HTML, tels que via CSS ou Flash. Sa configuration par défaut est destinée à la réalisation de collectes larges, mais il est également possible d'ajuster les paramètres de configuration pour focaliser la collectes automatique de manière différente et par exemple de gérer des priorités de capture au niveau des sites[2]. Heritrix génère directement des fichiers dans le format Arc adopté par l'IIPC. Des outils facilitant la gestion des fichiers stockés en format Arc sont aussi disponibles sur le site wiki d'Heritrix[3].

- Parmi les contenus qui ne peuvent être collectés automatiquement en raison des limitations des robots, le groupe de travail *Deep web* a identifié les bases de données qui servent de passerelles vers des contenus comme des métadonnées importantes à archiver en même temps que les contenus auxquels elles donnent accès. Afin d'assurer la conservation et l'accès à long terme à des bases de données, il est nécessaire de les migrer dans un format ouvert avant de les verser dans l'entrepôt numérique. La BnF a développé l'outil **DeepArc** en 2003 à l'occasion de son expérimentation pour le dépôt volontaire. Cet outil a été accepté par le consortium comme un élément de l'ensemble des outils. Il est disponible en test comme logiciel libre gratuit sur Source Forge[4]. L'outil offre une interface graphique permettant d'écrire le mapping entre la base de données initiale et le Schéma XML cible et de réaliser la migration de la base de données dans un fichier XML plat propre à sa conservation et sa réutilisation dans le cadre de l'accès à l'archive.

2. Activité de standardisation

L'activité de standardisation porte sur :
- Des APIs standards entre les modules fonctionnels
- Le format pour l'archivage du Web et les échanges entre les archives
- Les métadonnées pour la conservation et l'accès à long terme
- L'identification permanente

2.1. Les APIs standards

L'IIPC souhaite normaliser les APIs entre les différents modules fonctionnels pour construire une architecture ouverte et modulaire mais également pour permettre l'interopérabilité avec les systèmes en place dans chaque institution, évidemment si ce système est lui-même suffisamment techniquement ouvert pour permettre cette interopérabilité.
Un des aspects important déjà identifié comme nécessitant une interconnexion concerne les relations avec le système de gestion de bibliothèque et plus particulièrement avec les modules d'acquisition de catalogage et l'OPAC. Mais avec l'extension probable du consortium à d'autres types d'institutions que les bibliothèques, comme les archives et les institutions de recherche, l'interopérabilité avec d'autres types de système devra être examinée.

[2] http://sourceforge.net/projects/archive-crawler
[3] http://crawler.archive.org/cgi-bin/wiki.pl?BnfArcTools
[4] http://deeparc.sourceforge.net

Les APIs standards sont également très importantes dans une période d'évolution technologique très rapide, afin de permettre de remplacer les modules de manière indépendante au fur et à mesure de la disponibilité d'outils plus sophistiqués et performants.
Les APIs standards seront publiées sur le site du consortium. Les premières concerneront les fonctions d'accès et les interactions entre les modules de recherche et de stockage avec l'objectif de fournir à l'utilisateur final la possibilité de faire de la recherche plein texte et du feuilletage.

2.2. Le format IIPC pour l'archivage et l'échange de données

Le robot Heritrix produit directement des résultats de capture dans le format Arc adopté par l'IIPC comme format de stockage et d'échange. Les outils d'accès en cours de développement et de test par le consortium permettent de réaliser l'indexation directement à partir des fichiers Arc.

Le format Arc est utilisé par Internet Archive pour le stockage des données web depuis le début en 1996[5] et les membres d'IIPC qui utilisent Heritrix pour leur collecte stockent également leurs données en format Arc.

Le format Arc consiste en une séquence d'enregistrements d'URL. Chaque enregistrement commence par un en-tête contenant des métadonnées relatives au contexte technique de la capture et qui sont extraites du protocole d'échange entre le robot et le serveur, suivi du fichier correspondant à l'URL capturé. La taille d'un fichier Arc varie de 100 à 600 méga octets et permet, dans le contexte des archives du web, de minimiser les accès disques et de gérer spécifiquement un très grand nombre de très petits fichiers.
Un fichier DAT est associé à chaque Arc dans lequel sont enregistrées des métadonnées (dates, checksum, etc.) et surtout la liste des URLs des enregistrements qui sont contenus dans l'ARC
Un ficher CDX est associé avec chaque Arc fournissant l'index des URLs de l'Arc avec la position du début de l'enregistrement et sa longueur. Ce dispositif permet de retrouver directement un URL particulier et de passer l'information directement à l'interface d'accès pour l'affichage, en conservant les URLs réels sans avoir à les transformer pour une intégration dans un système de fichier qui ne saurait gérer certains caractères qui ont pu être utilisés dans les URLs.
Les caractéristiques du format actuel le rendent très efficace pour le stockage et l'extraction à partir d'une très grosse archive. Par exemple un snapshot du domaine .fr effectué en 2004 par la BnF est composé d'environ 25 000 fichiers Arc contenant environ 500 000 serveurs différents et environ 121 millions d'URLs.

Le consortium travaille avec d'autres communautés intéressées à l'extension du format Arc pour qu'il puisse répondre à davantage de besoins. Le consortium souhaite présenter ce nouveau format appelé Warc à l'ISO TC46/SC4 comme format normalisé pour les archives web. Les institutions de conservation souhaitent en effet un format international standard sur lequel elles puissent construire des entrepôts et des échanges de données fiables comme elles l'ont toujours fait par le passé pour tous les types de données.

Les besoins pris en compte par le format Warc sont tout d'abord les résultats du travail effectué au sein du consortium à l'intérieur du cadre de développement défini au départ :

- Une extension des capacités actuelles pour inclure des sites déposés et des sites collectés dans la même application de stockage et d'accès de manière à gérer dans un même format des collectes automatiques très large, des collectes automatiques ciblées et des dépôts de sites
- La possibilité de gérer à la fois des métadonnées et des contenus dans le même format pour faciliter le stockage et l'accès à des collectes incrémentales par exemple. Cette approche conduit à gérer davantage de relations entre des objets qui se complètent les uns les autres. Ce besoin résulte de l'instabilité des contenus de web sur la durée. L'outil de consultation offert par le consortium offre également la navigation dans une ligne de temps et souhaiterait pouvoir offrir des possibilités de comparaison entre les états d'un même objet à différents moments, ce qui a un impact sur les formats

[5] http://www.archive.org/web/researcher/ArcFileFormat.php

- L'extension du format doit également permettre de segmenter un objet très volumineux dans plusieurs Arc

Le projet du format Warc prend également en compte la conservation en incluant des métadonnées de préservation qui peuvent être générées indépendamment de l'enregistrement de l'objet lors de sa capture, ainsi que la possibilité de créer des enregistrements convertis par exemple après une opération de migration.

Les institutions associées à la discussion sur le format, en dehors des membres du consortium, sont le Los Alamos National Laboratory qui a réalisé une étude pour la Bibliothèque du Congrès sur l'extension du format actuel et la California Digital Library. Le format actuel est également utilisé par la NARA (Archives nationales américaines) qui est par ailleurs très impliquée dans la réflexion sur la conservation des ressources numériques et qui a contribué à la standardisation du modèle OAIS.

2.3. Le jeu de métadonnées IIPC pour l'archivage du Web

L'IIPC s'intéresse aux métadonnées pour la conservation à long terme et pour l'instant n'a pas considéré les métadonnées descriptives. Le travail est également focalisé sur la documentation des dépendances techniques en considérant principalement les caractéristiques de très grosses archives contenant un grand nombre de fichiers interdépendants via des liens. Les archives qui contiennent un nombre restreint de sites ont des spécificités techniques plus proches des objets électroniques discrets et qui sont déjà traités dans d'autres groupes.
Sur le web différents formats sont très souvent combinés dans une même page et tous les formats possibles sont potentiellement présents. Cependant 90 % des objets utilisent uniquement quatre formats principaux qui sont documentés publiquement (HTML, JPEG, GIF et PDF) et la plupart des autres ne sont pas dépendants de systèmes d'exploitation spécifiques. Dans le contexte du web l'Information de Représentation (au sens du modèle de référence OAIS)la plus importante à mémoriser est le format de fichier. Malheureusement on ne peut s'appuyer sur le type MIME fournit par le serveur lors de l'échange HTTP entre le robot et le serveur pour construire cette métadonnée de conservation.

Archiver le web, même par des collectes extensives, signifie échantillonner le web réel qui peut lui-même être modifié très rapidement. De plus un robot travaille en exécutant une politique de collecte qui peut dépendre de nombreux paramètres allant de l'appartenance du serveur à un domaine générique jusqu'à l'analyse des liens ou du contenu textuel ou la détection des changements etc.… Toutes ces informations de paramétrage font partie de l'information de contexte relative aux contenus collectés et qui doit être enregistrée pour pouvoir aider le chercheur du futur à comprendre ce que le contenu collecté représente par rapport au contenu original. Le jeu de métadonnées d'IIPC comprend donc le contexte de sélection, les éléments du contexte d'interaction et de transaction entre le client et le serveur (par exemple les identifiants, les cookies etc.). Ces informations de contexte constituent une extension de l'Information de Pérennisation du modèle de référence OAIS adapté à l'archivage du Web.

Enfin le niveau de granularité auquel s'appliquent les différentes métadonnées doit encore être finalisé. Les objets physiques définis dans le format Warc sont le niveau du fichier Warc et le niveau du fichier enregistré pour chaque URL. Cependant le groupe de travail *Metrics & Test Bed* est encore en train de définir les entités logiques et intellectuelles à considérer telles que le site, la page, le document ou des entités plus techniques telles que le serveur. La discussion est donc encore en cours et l'IIPC a déjà reconnu la nécessité de partager cette discussion avec les experts de PREMIS qui ont abordé la granularité intellectuelle du site Web davantage au travers du niveau des métadonnées descriptives.

La décision finale du niveau auquel appliquer chacune des métadonnées de préservation est également liée au cycle de génération des métadonnées et à ce qui peut être automatisé lors de la phase d'acquisition et en particulier lors du processus de collecte par robot. Cette phase d'acquisition est examinée par le groupe de travail *Content management* comme une phase de pré-versement permettant de générer un SIP OAIS (Paquet d'Information à Verser). La prochaine étape du travail devra considérer plus particulièrement la phase de versement et la génération des AIP (Paquet d'Information Archivé).

2.4. Les identifiants

L'identification est également objet de débats au sein du consortium en relation étroite avec les travaux sur le Framework et le format Warc.

Au niveau de granularité du ficher Warc, les membres du consortium ont accepté le principe d'un cadre commun de nommage des fichiers Warc de manière à faciliter l'organisation du miroring et de la redondance qui sont nécessaires pour minimiser le risque en cas de catastrophe et pour faciliter la coopération de la conservation dans le consortium. L'attribution de l'identifiant d'un fichier Warc sera de la responsabilité de l'institution qui l'a créé. La structure de l'identifiant n'est pas encore finalisée.

Au niveau de l'enregistrement de chaque URL, la discussion porte sur le choix d'un identifiant globalement unique et sur les besoins de cohérence avec l'identification des autres ressources dans les institutions.

Conclusion

Les trois premières années du consortium on été consacrées à la définition de l'infrastructure générale et à l'adoption des standards nécessaires pour le développement d'un ensemble complet d'outils spécifiques pour la totalité de la chaîne de traitement des contenus du web, depuis la sélection jusqu'à l'acquisition, l'indexation et l'accès par l'utilisateur final.

Un des axes possibles pour la prochaine phase et l'extension du consortium à d'autres types d'institutions est la préservation à long terme et l'entrepôt fiable. Cet aspect va être exploré par le groupe de travail *Content management* à l'occasion du travail concret sur le cycle de génération des métadonnées d'archivage à long terme.