Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Département fédéral de l'intérieur DFI
Office fédéral de la culture OFC
Bibliothèque nationale suisse BN

Version 1.7, 19.04.2024

Archives Web Suisse

Sites web représentatifs sur la Suisse

Une collection commune de bibliothèques cantonales suisses, de bibliothèques et archives spécialisées et de la Bibliothèque nationale suisse

Notice Archivage

Modifications dans le document

Version	Date	Remarque
1.0	22.02.2006	Création
1.1	01.04.2008	Actualisation
1.2	01.05.2009	Actualisation
1.3	15.07.2010	Actualisation du chapitre 8
1.4	15.01.2011	Actualisation
1.5	01.10.2013	Actualisation des chapitres 7 et 8
1.6	30.01.2015	Actualisation
1.7	19.04.2024	Petites corrections chapitres 5.1 et 6.3

1 Table des matières

2 Introduction

La notice Archivage décrit comment les sites web annoncés par les bibliothèques cantonales et d’autres bibliothèques spécialisées sont déposés et conservés dans le système de la Bibliothèque nationale suisse.

Deux composantes participent au processus d'archivage. D'une part, il s'agit du système Ingest, qui prépare les données pour l'archivage et qui garantit que les informations descriptives qui les accompagnent (métadonnées) soient également à disposition. D'autre part, il s'agit du système d'archivage lui-même (mémoire), dans lequel les publications numériques sont stockées avec leurs métadonnées. Il est important que les métadonnées soient aussi enregistrées dans un catalogue qui soit à disposition des utilisateurs/trices.

3 Principes

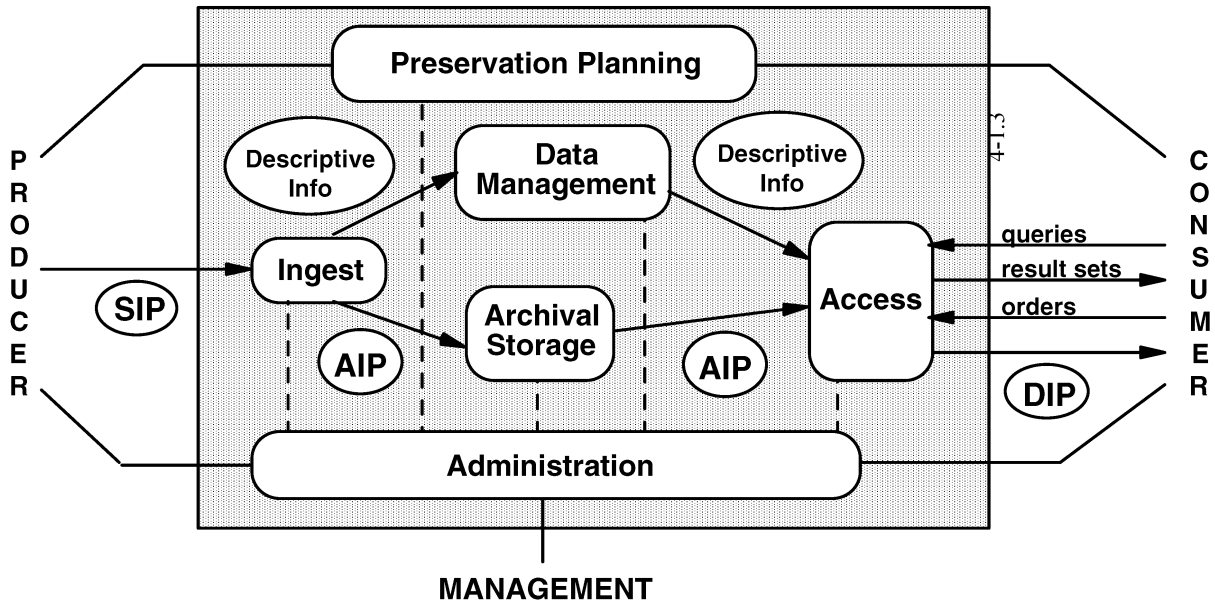
Les principes suivants ont guidé la Bibliothèque nationale suisse lors de l’élaboration du système Ingest:

- Automatisation la plus large possible
- Offrir aux producteurs de données différentes interfaces pour la prise en charge des données
- Procéder à la standardisation de l'information, en particulier des métadonnées, le plus tôt possible dans le processus
- Utiliser des outils librement combinables pour le traitement des données et métadonnées
- Obtenir les publications numériques dans leur format d’origine

Au niveau du système d'archives, l'information stockée ne doit pouvoir être ni écrasée, ni effacée. Si des informations devaient être modifiées (migrées) afin de rester lisibles pour l'utilisateur/trice, il faudrait créer une nouvelle version du paquet d'archives en question, tout en conservant toutes les versions antérieures de ce paquet d'archives.

4 OAIS

Pour réaliser son système d'archivage d'informations électroniques, la Bibliothèque nationale suisse se conforme au modèle de référence pour un système d'informations d'archives ouvert (OAIS) du Consultative Committee for Space Data Systems. Pour comprendre les développements exposés plus loin dans ce document, une connaissance sommaire du modèle OAIS est indispensable. En voici donc une brève explication.
OAIS s'est imposé dans le monde entier auprès des bibliothèques et des archives comme modèle de référence pour l'archivage numérique. C'est un modèle strictement logique et donc indépendant de toute implémentation. Il contribue largement à une compréhension commune de l'archivage numérique et à un langage commun dans ce domaine.



L'on distingue six fonctions principales:

Ingest (prise en charge des données)
- Prise en charge des SIP (Submission Information Package) créés par le producteur
- Contrôle de l'intégralité et de l'intégrité
- Transformation du SIP en AIP (Archival Information Package)
- Extraction de l'information descriptive pour la base de données de recherche
- Transmission de l'AIP à la mémoire des archives
- Communication au Data Management

Archival Storage (mémoire des archives)
- Conservation et réception des AIP
- Création de backups
- Contrôle régulier de l'intégrité des données
- Mécanismes de recréation en cas d'urgence
- Transmission des AIP à e-Helvetica Access pour l'utilisation

Access (consultation)
- Interface utilisateur
- Rendre possible la recherche et générer des réponses contenant la description des AIP et des informations quant à leur disponibilité
- Réception de demandes (requests) et livraison des DIP (Dissemination Information Packages)
- Garantie du respect des droits d'accès

Administration

- Contrôle des processus globaux dans OAIS et de ses relations extérieures
- Configuration du matériel et du logiciel
- Contrôle des droits d'accès
- Génération de DIP et leur transmission aux utilisateurs/trices

***Data Management (gestion des données*)**

- Gère les informations descriptives (base de données) qui identifient les fonds d'archives et les documents, ainsi que d'autres données nécessaires à l'utilisation du matériau d'archives
- Réception et traitement des demandes (queries) émanant du domaine de l'utilisation

***Preservation Planning (planification de l'archivage*)**

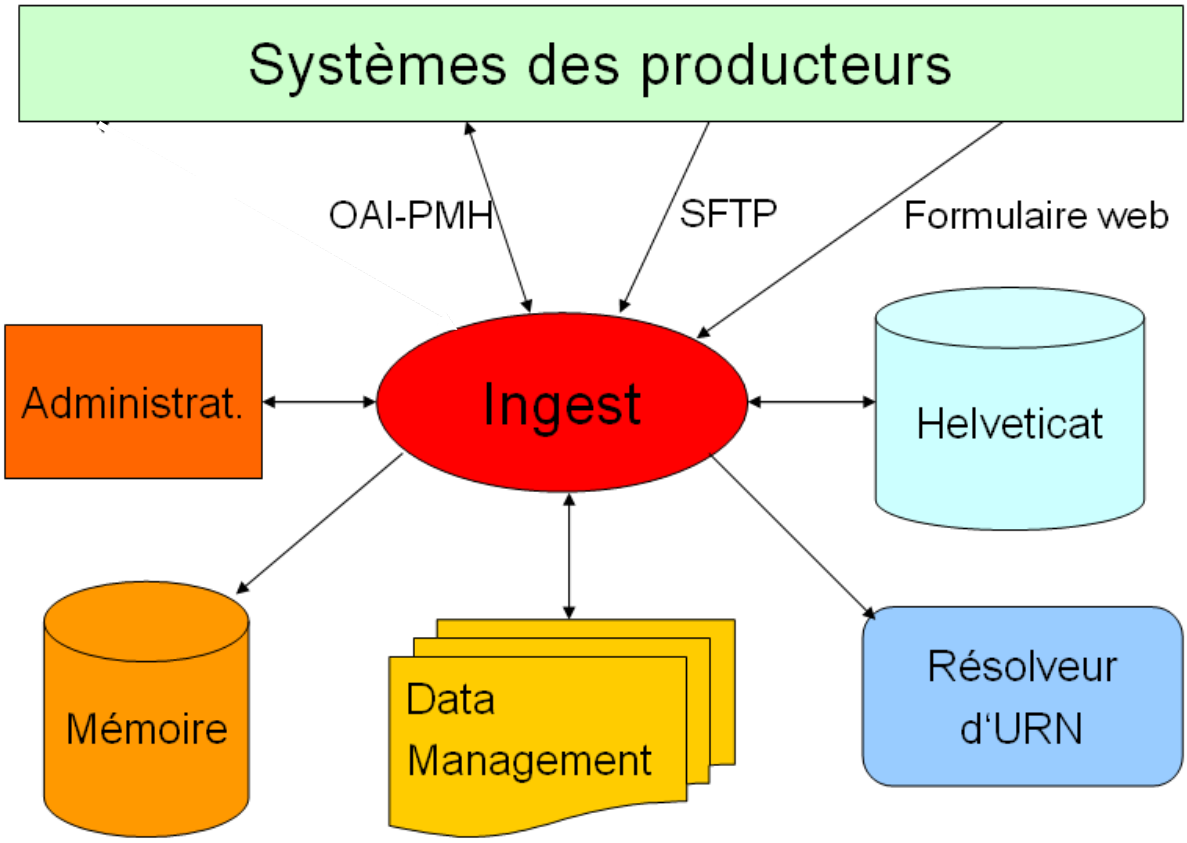
- Suivre les développements technologiques et formuler des recommandations par rapport aux standards et à la politique d'archivage
- Surveillance des efforts en matière d'archivage
- Formulation de recommandations pour le maintien de la lisibilité de l'information stockée
- Planification de migrations de données et de processus de copie

5 Ingest

On appelle processus Ingest tout le processus allant de la prise en charge des données du fournisseur ou de la source de données accessible par Internet jusqu'au stockage dans le système d'archives.

5.1 Environnement système de la Bibliothèque nationale suisse

En vue d'un bon fonctionnement, il est capital que le processus Ingest soit intégré à l'environnement système. Le graphique ci-dessous montre les composantes qui jouent un rôle. Pour l'explication des différents systèmes, prière de se référer au modèle OAIS. Les persistent identifiers sont transformés en liens par un résolveur d'URN.

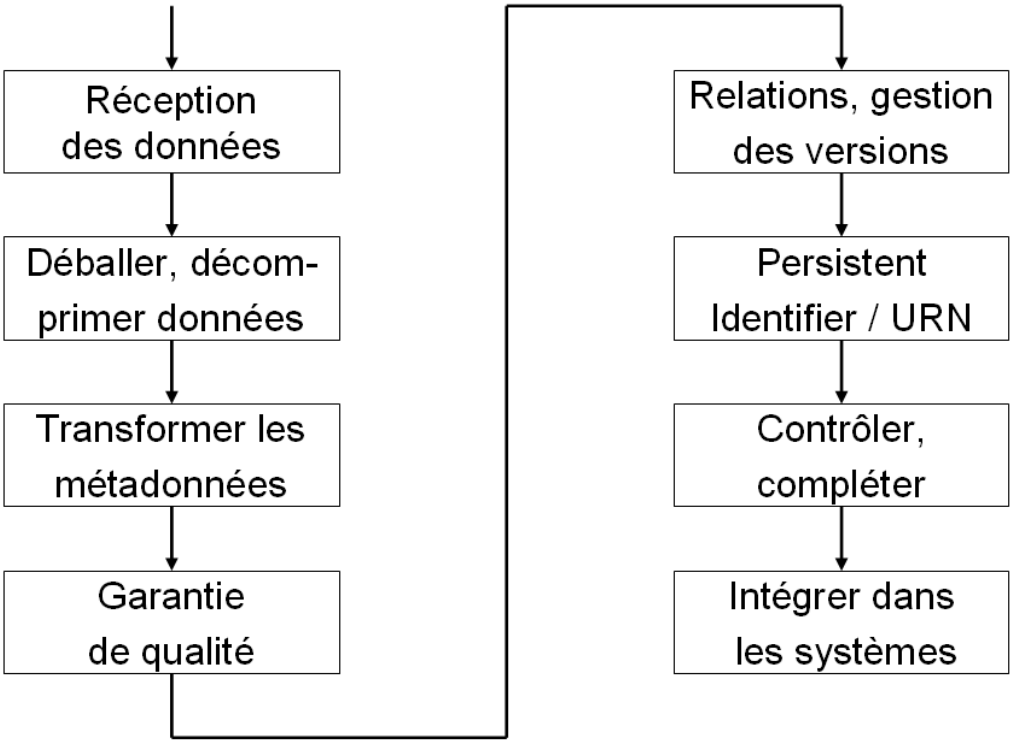


Le processus Ingest met à disposition plusieurs interfaces pour la reprise des données depuis les systèmes des producteurs.

- OAI-PMH: les plus grandes bibliothèques universitaires autorisent un harvesting OAI-PMH de leur catalogue. C'est par ce biais que la Bibliothèque nationale suisse se procure les métadonnées des thèses et des habilitations recensées récemment.
- SFTP: périodiquement, la Bibliothèque nationale suisse accède à un système du producteur via SFTP et va y chercher les nouveaux paquets de données qui y sont déposés. Pour l'instant, cette possibilité n'est utilisée par aucun producteur.
- Formulaire web: les fournisseurs de données peuvent annoncer les métadonnées des publications prévues pour l'archivage à long terme via des formulaires web. Cette annonce arrive sous forme d'e-mail avec un attachement XML dans des boîtes postales définies auxquelles accède le processus Ingest.

5.2 Le processus Ingest

Le processus Ingest est composé de toute une série d'étapes de travail.



Après la prise en charge des données et la décompression des fichiers qui arrivent p.ex. sous la forme de fichiers ZIP, il est nécessaire de transformer les métadonnées des formes les plus diverses en une structure interne de la Bibliothèque nationale suisse, afin que le traitement ultérieur de toutes les données entrantes puisse se faire de façon uniforme.

Dans le cadre du contrôle de qualité, les données entrantes sont examinées:

- Lisibilité
- Authenticité (concordance de la somme de contrôle)
- Conformité au format (un fichier se terminant par .pdf est-il vraiment un fichier PDF?)
- Absence de virus
- Intégralité des données
- Intégralité des métadonnées
- Livraisons à double (les données livrées n'existent-elles pas déjà à la Bibliothèque nationale suisse?)

L'attribution à des entrées supérieures doit être garantie. De nouveaux fascicules doivent p.ex. être attribués au bon titre de revue. Pour les sites web, il s'agira de collecter périodiquement de nouvelles versions du même site web et de les archiver.

Tous les paquets de données à archiver sont munis d'un identificateur univoque (Persistent Identifier).

Les métadonnées manquantes doivent être insérées et les données techniques et administratives doivent être complétées.

Ensuite, plusieurs systèmes reçoivent des données du processus Ingest:

- Helveticat: enregistrement des publications électroniques dans le catalogue
- Data Management: gestion des données dans le système d'archivage (données techniques et administratives avant tout)
- Mémoire (système d'archivage): maintien des paquets de données créés pour l'archivage

Des outils Java supplémentaires sont utilisés en plus d’Ingest pour traiter les cas spéciaux, comme par exemple les paquets très volumineux d’archives web.

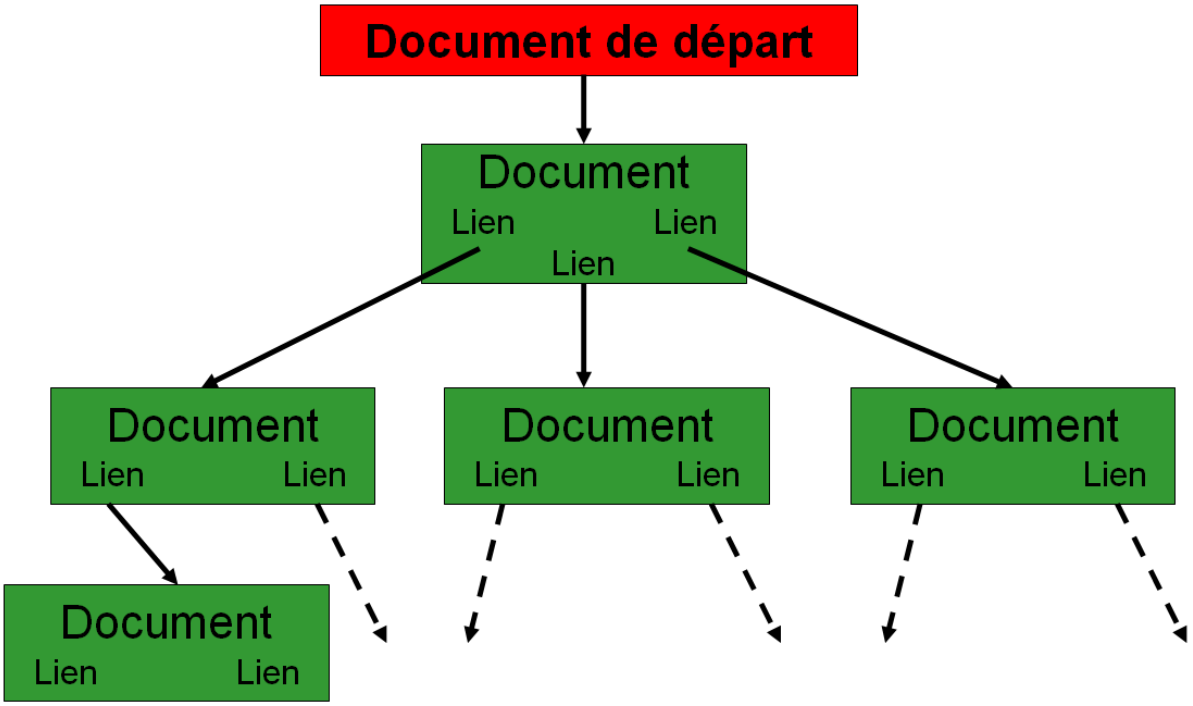
6 Harvesting

6.1 Mode de fonctionnement du harvester

On appelle harvesting la collecte de sites web de l'Internet. Lors du harvesting, des programmes spéciaux font en sorte que depuis une page de départ, tous les liens puissent être suivis et que les fichiers qui se trouvent dans le domaine de collecte défini soient téléchargés.

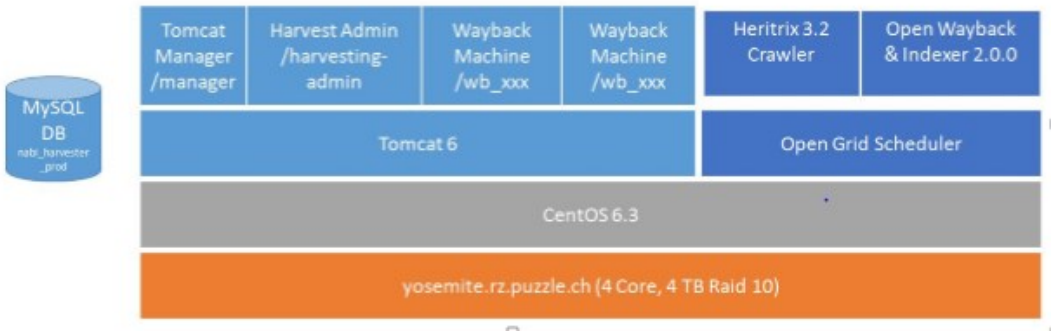
Les bibliothèques cantonales/spécialisées peuvent se permettre une flexibilité dans la définition des domaines de collecte. En principe, il y a deux possibilités différentes:

- Collecte d'un domaine comme http://www.rorschach.ch
  Collecte de tous les documents de tous les répertoires et sous-répertoires qui se trouvent dans le domaine www.rorschach.ch.
- Collecte de documents de certains répertoires comme http://www.swiss-world.org/de/geschichte
  Seuls les documents du répertoire /geschichte et de ses sous-répertoires sont collectés. Les documents qui se trouvent p.ex. sous http://www.swissworld.org/de/kultur/ ne sont pas inclus dans le harvesting.



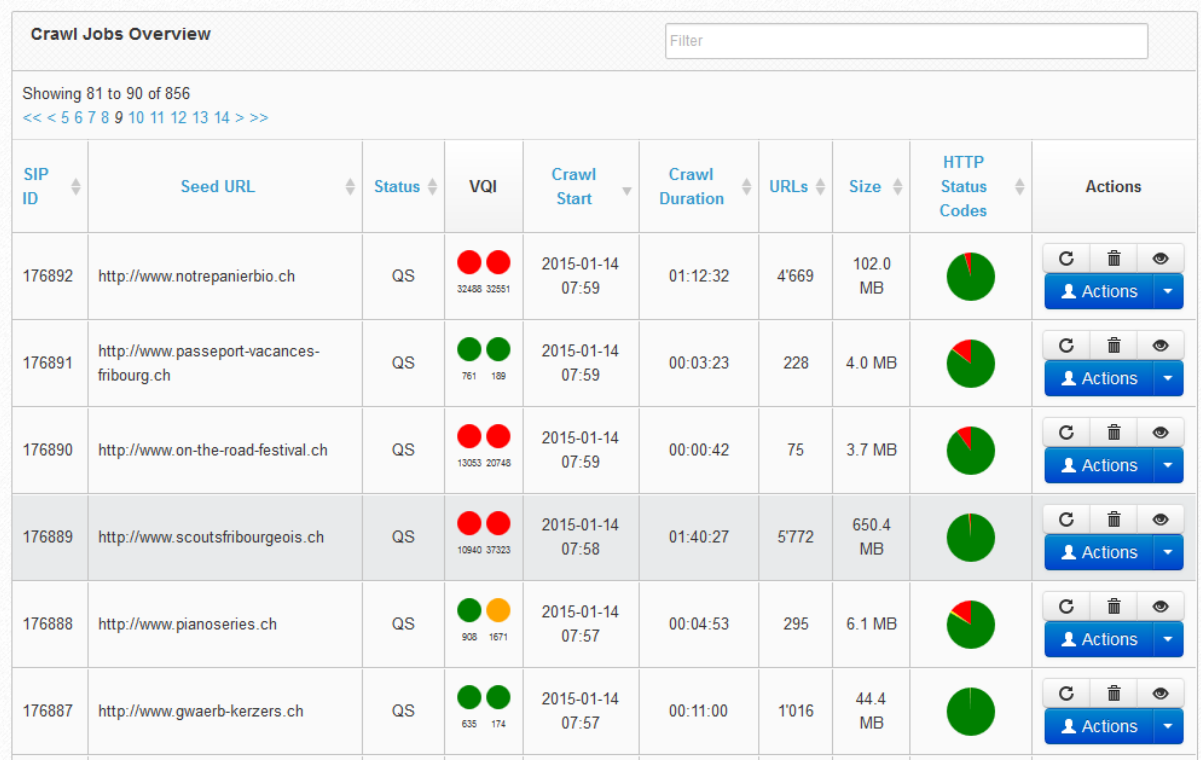
6.2 Structure du harvester

Le harvesting doit se dérouler dans un domaine de réseau qui soit si possible soumis à un minimum de restrictions, sans quoi l'on risque de ne pas pouvoir collecter tous les documents souhaités. Pour cette raison, l'infrastructure de harvesting se trouve chez un fournisseur externe.



Le robot d'indexation Heritrix collecte les sites web; une fois indexé, le site web moissonné peut être ouvert dans la Wayback Machine pour le contrôle de qualité.

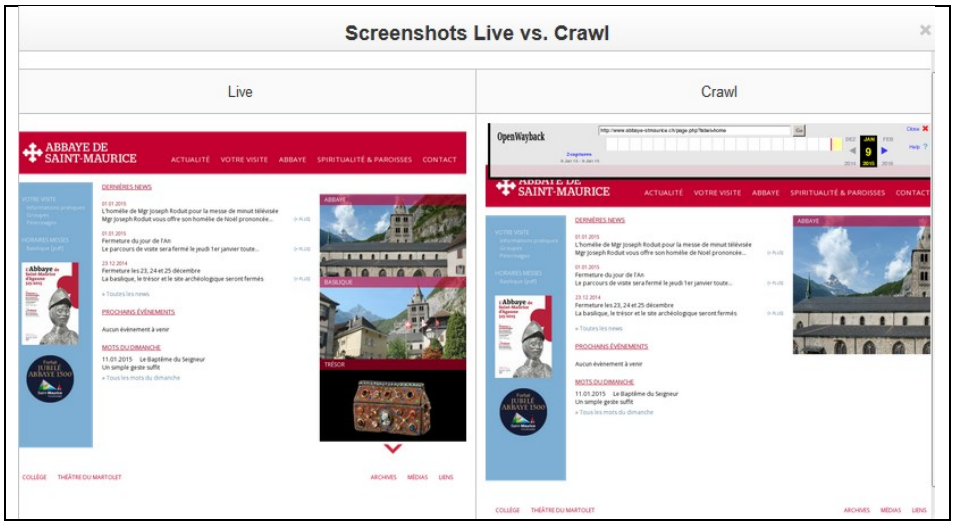
Le processus de harvesting peut être piloté et surveillé directement via une interface d'utilisation basée sur le web. La saisie du numéro SIP et de l'URL permet de démarrer les jobs; environ 20 mandats de harvesting peuvent tourner en même temps.



Les crawls sont surveillés en affichant plusieurs types d'informations comme le nombre d'URL détectées, la répartition des codes de status HTTP ou l'analyse du crawl log détaillé. La fonction pause permet de contrôler le résultat intermédiaire dans la Wayback Machine.

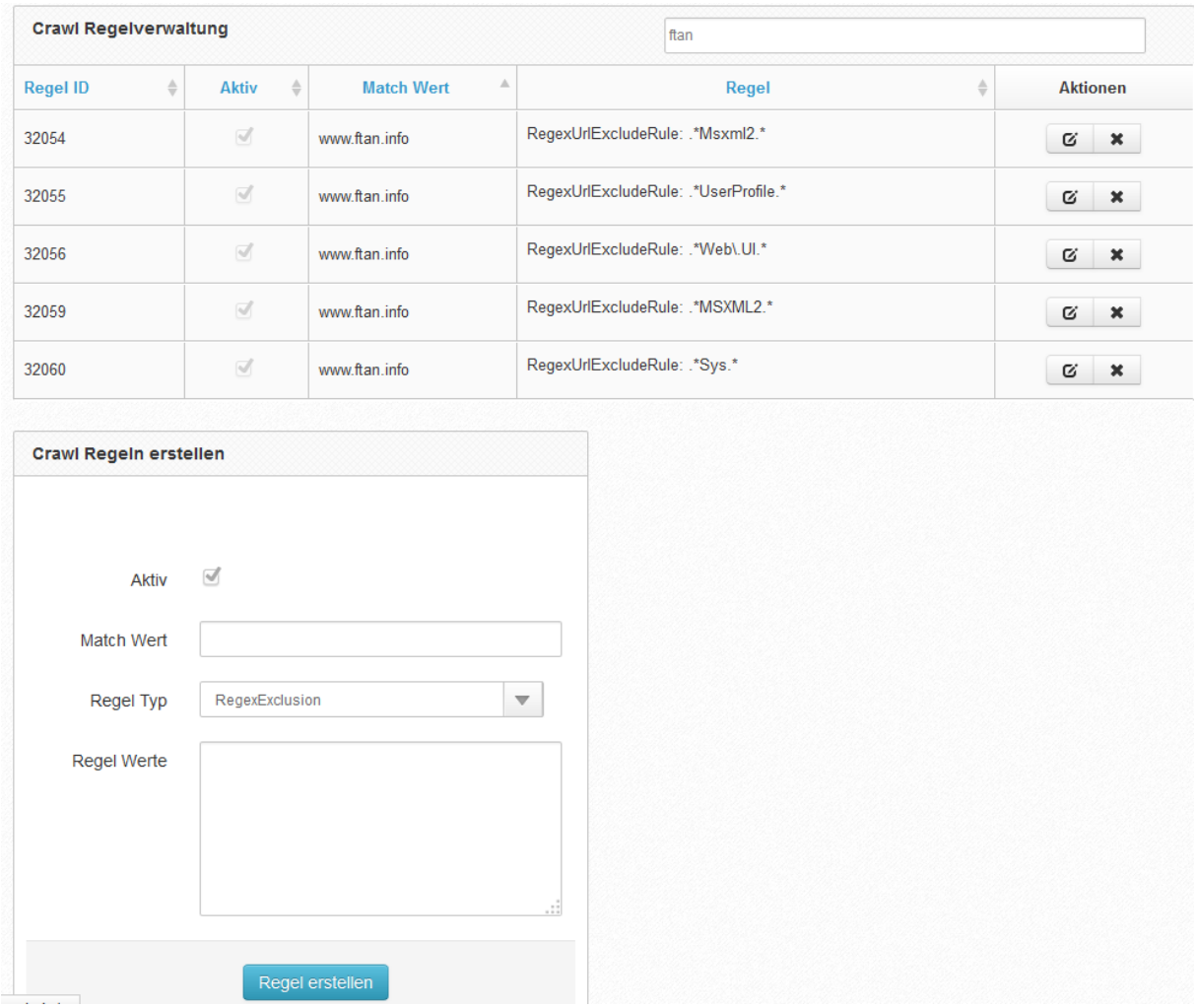
De plus, pour soutenir le contrôle de qualité, la Bibliothèque nationale suisse a implémenté un Visual Quality Index au sein de l'infrastructure de harvesting. Grâce aux outils PhantomJS et CasperJS, deux comparaisons de captures d'écran sont créées automatiquement pour tous les crawls.

1. La comparaison du site web live avec le crawl actuel:



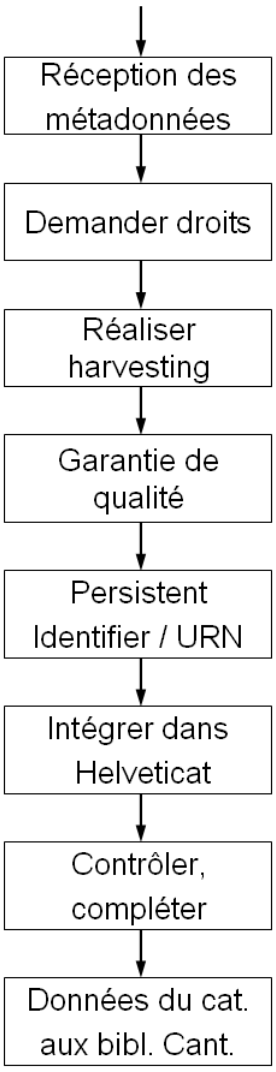
2. La comparaison de la dernière version archivée dans e-Helvetica Access avec le crawl actuel:

Au besoin, les crawls peuvent être améliorés en ajoutant des URL de base supplémentaires ou en excluant des domaines d'URL au moyen de Regular Expressions.


Crawl Regelverwaltung
ftan
Regel ID
Aktiv
Match Wert
Regel
Aktionen
32054
www.ftan.info
RegexUrlExcludeRule: .*Msxml2.*
32055
www.ftan.info
RegexUrlExcludeRule: .*UserProfile.*
32056
www.ftan.info
RegexUrlExcludeRule: .*Web\.UI.*
32059
www.ftan.info
RegexUrlExcludeRule: .*MSXML2.*
32060
www.ftan.info
RegexUrlExcludeRule: .*Sys.*
Crawl Regeln erstellen
Aktiv
Match Wert
Regel Typ
RegexExclusion
Regel Werte
Regel erstellen

6.3 Processus de traitement

Le processus de travail implémenté à l'heure actuelle pour le traitement de sites web est légèrement différent du chemin de traitement prévu normalement dans le processus Ingest.



Les métadonnées sont livrées via un formulaire Web protégé par mot de passe. Une fois les données arrivées à la Bibliothèque nationale suisse, l'on demande les droits pour récolter et conserver les sites web. La Bibliothèque nationale suisse envoie au propriétaire (exploitant) du site web un courriel contenant des informations sur Archives Web Suisse et lui offrant la possibilité de refuser l'archivage du site web. Si la Bibliothèque nationale suisse n'obtient pas de réaction à ce courriel, elle procède au harvesting du site web dans le sens du „fair use".

Les métadonnées ne sont contrôlées et complétées qu'après l'importation dans Helveticat, ce qui permet aux bibliothécaires d'utiliser leur interface d'utilisation habituelle (Alma, Ex Libris dans le cas de la Bibliothèque nationale suisse) pour traiter les métadonnées.

A la fin du processus, les métadonnées bibliographiques complètes peuvent être transmises aux bibliothèques cantonales/spécialisées pour être intégrées à leurs propres catalogues (voir la notice Mise à disposition).

Pour la collecte de sites web, la Bibliothèque nationale suisse se limite pour l'instant à un harvesting sélectif. Un harvesting de tout le domaine .ch ne doit toutefois pas être exclu d'emblée. Pour l'instant, les capacités pour le faire manquent.

7 Contrôle de qualité

Un contrôle de qualité fiable des sites web ne peut être réalisé qu'à l'aide d'un instrument technique qui fait une analyse détaillée des documents collectés et qui signale d'éventuelles erreurs. Il est prévu qu'un tel instrument soit développé dans le cadre de IIPC (International Internet Preservation Consortium).

Jusqu'à ce que cet instrument puisse être utilisé, le contrôle de qualité, par la force des choses, doit se faire manuellement et ne peut donc être que rudimentaire.
Le contrôle de qualité ne vise pas à contrôler la qualité d'un site web dans l'Internet, mais bien la qualité du processus de collecte.

7.1 Schéma de contrôle

La Wayback Machine permet d'accéder aux sites web collectés via une interface de navigation courante (p.ex. Firefox, Internet Explorer). Les tests suivants sont effectués:

1. Impression générale
   La cohérence de l'impression générale du site web collecté est comparée avec le site web original.
2. Nombre de documents présents
   Si le nombre de documents collectés pour un site web est inférieur à 100, on vérifie sur le site web original s'il est vraiment aussi petit.
3. Présentation
   La présentation au moyen des feuilles de style et des graphiques correspond-elle au site web original? Y a-t-il des galeries photo sur le site web et si oui, s'affichent-elles correctement?
4. Structure des répertoires
   A partir de l'URL annoncée, on contrôle si des fichiers de répertoires subordonnés ont aussi été livrés. Le crawl contient-il aussi des documents comme des PDF?
5. Contrôle de l'intégrité
   Le contrôle de l'intégrité d'un site web ne se fait que par échantillons. Le contrôle se limite aux éléments suivants:
   - Appeler la page d'accueil et contrôler si tous les éléments sont présents.
   - S'il y a plusieurs langues, contrôler les différentes versions linguistiques.
   - Suivre chacun des liens sur la page d'accueil et contrôler si tous les éléments sont présents dans les documents suivants.
   - Suivre une chaîne de liens depuis un document suivant sur 5 niveaux, en contrôlant à chaque fois si tous les éléments sont présents.
6. Contrôle des fonctions
   Si le site web contient des éléments dynamiques, il s'agit de contrôler leur fonction. Si l'on constate qu'une fonction ne peut pas être utilisée, on contrôle ce qui cause cette restriction.

Le résultat du contrôle de qualité est documenté comme OK ou insuffisant. Un site web téléchargé n'est rejeté que lorsqu'un deuxième et un troisième harvesting avec une configuration adaptée n'ont pas donné de meilleur résultat.

Raisons pouvant amener à rejeter un site web téléchargé:

- Présentation tout à fait différente ou fausse
  Feuilles de style manquantes (ou mal interprétées), police (caractères spéciaux), graphiques, composantes placées de façon erronée
- Contenu manquant
  Des parties du site web ou des documents intégrés au contenu important manquent (p.ex. PDF, DOC, XLS, ...)

- Fonctions du menu manquantes
  S'il n'y a pas d'autre moyen de consulter le contenu (p.ex. via un plan du site)
- Out of scope
  En dehors du domaine de collecte

Les problèmes suivants sont connus et ne génèrent pas automatiquement le rejet d'un site web téléchargé:

- La fonction de recherche manque ou ne fonctionne pas
- La fonction d'impression manque ou ne fonctionne pas
- Le calendrier manque ou ne fonctionne pas
- L'horloge système manque ou ne fonctionne pas
- Le compteur manque ou ne fonctionne pas
- La webcam manque ou ne fonctionne pas
- Le formulaire web / le champ de saisie manque ou ne fonctionne pas
- Le forum/wiki/blog manque ou ne fonctionne pas
- Une carte géographique manque ou ne fonctionne pas
- Une vidéo manque ou ne fonctionne pas
- Un document audio manque ou ne fonctionne pas
- Le diaporama multimédia, le jeu par navigateur etc. manque ou ne fonctionne pas
- Liens en direct (acceptables uniquement si le contenu archivé reste accessible)
- Une fonction pilotée par script manque ou ne fonctionne pas
- Une fonction pilotée par serveur manque ou ne fonctionne pas

Globalement, il n'y a que peu de critères durs qui conduisent forcément au rejet d'un site web téléchargé. Les „collection managers" (bibliothécaires) décident si un site web téléchargé est acceptable ou non.

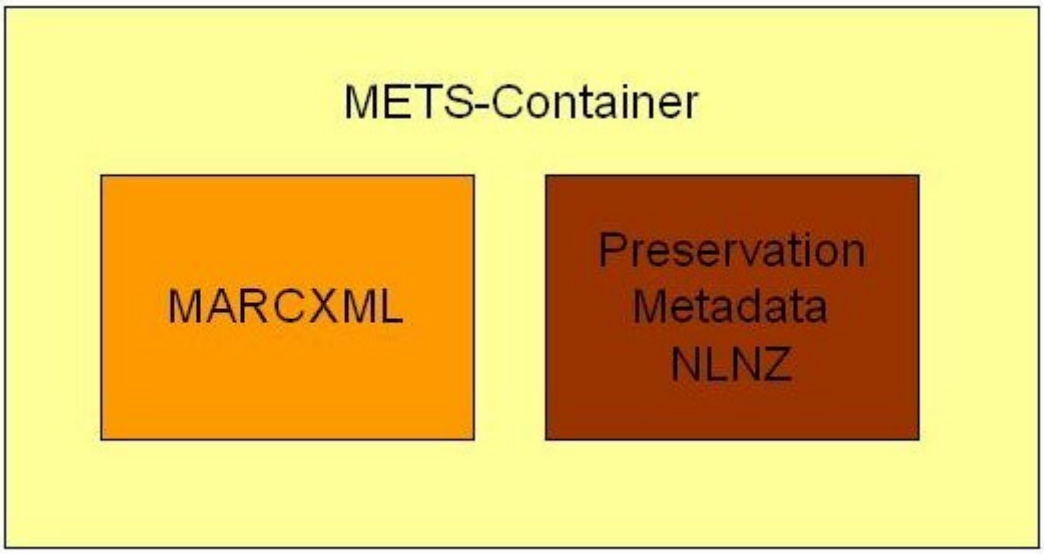
Les questions suivantes accompagnent la décision:

- Quelles sont les parties importantes du site web („significant properties"), et ces dernières sont-elles représentées dans les archives web?
- Un nouveau harvesting du site web peut-il améliorer le résultat?

8 Métadonnées

La Bibliothèque nationale suisse n'a pas développé sa propre structure de métadonnées. Elle profite des formats existant en format XML, ce qui lui évite également de s'investir dans le développement ultérieur de la structure de métadonnées.

Pour la structure interne des métadonnées, la Bibliothèque nationale suisse utilise le Container METS tenu à jour par la Library of Congress. MARCxml y est intégré pour les données bibliographiques. MARCxml est également tenu à jour par la Library of Congress ; il est compatible avec MARC21, la structure de métadonnées de Helveticat. Dans le schéma pour «Preservation Metadata» développé par la National Library of New Zealand, les métadonnées non bibliographiques sont également intégrées dans le Container METS.



9 Persistent Identifiers

9.1 Le système choisi par la Bibliothèque nationale suisse

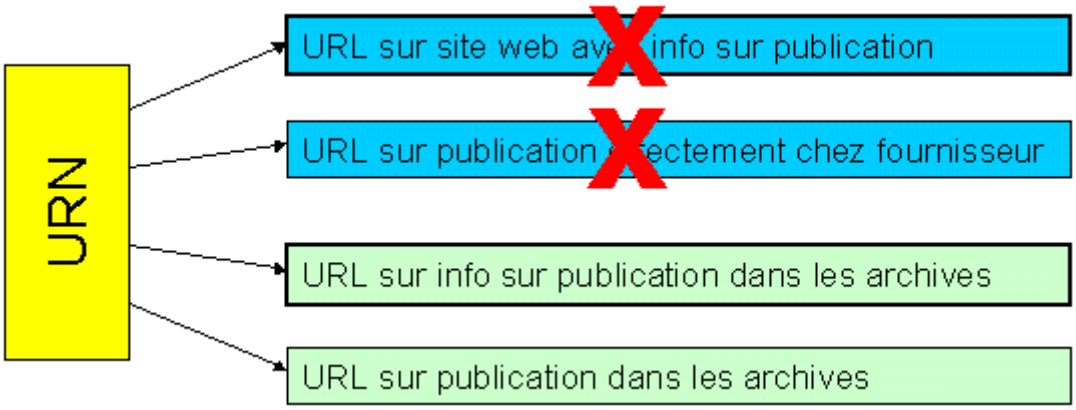
Un Persistent Identifier (identificateur univoque) doit couvrir deux besoins:

- Identification univoque des documents d'archives
- Renvoi stable à une source de données disponible en ligne (les liens se sont avérés très inconstants.)

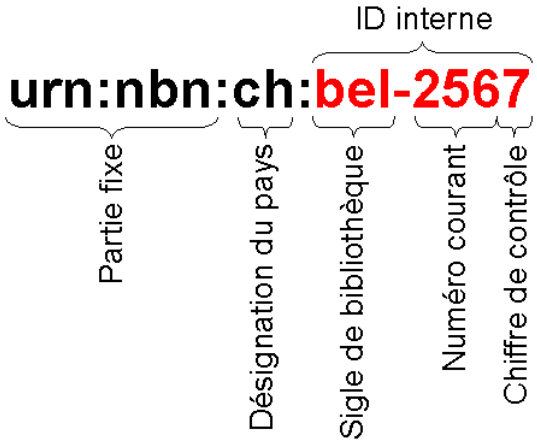
La Bibliothèque nationale suisse a décidé d'utiliser des Uniform Resource Names (URN) sous la forme de National Bibliographic Numbers (NBN), car l'URN répond aux besoins énumérés ci-dessus. Pour transformer les URN en liens, la Bibliothèque nationale suisse peut utiliser le résolveur d'URN de la Deutsche Nationalbibliothek.

Par un navigateur Web, tout/e utilisateur/trice peut enregistrer un URN sous la forme https://admin.nbn-resolving (adresse Web du résolveur d'URN et URN). Le résolveur d'URN renvoie un lien valable sur l'information souhaitée, à l'aide duquel cette information est automatiquement appelée dans l'Internet et affichée dans le navigateur.

Le résolveur d'URN apporte toute une série de fonctionnalités intéressantes pour la Bibliothèque nationale suisse. Il est ainsi possible de déposer toute une série de liens (Uniform Resource Locator, ou URL) derrière un URN. En priorisant, on peut définir dans quel ordre le résolveur d'URN livre les liens déposés aux utilisateurs/trices. Si le lien ayant la plus haute priorité est invalide, c'est le prochain qui est transmis automatiquement. Ainsi par exemple, on peut déposer via l'URN des liens vers un site web auprès du producteur avec des informations sur la publication souhaitée, vers la publication électronique elle-même qui est stockée chez le producteur, et également vers les informations qui se trouvent dans les archives de la Bibliothèque nationale suisse. Si le producteur retire la publication de l'offre de son site web, p.ex. parce qu'elle ne présente plus d'intérêt commercial, l'URN reste quand même valide et pointe alors directement sur la publication archivée à la Bibliothèque nationale suisse.



9.2 La structure du Persistent Identifier



L'identificateur univoque de la Bibliothèque nationale suisse correspond à la norme en vigueur pour les URN et contient tout d'abord une partie fixe qui indique qu'il s'agit d'un URN sous la forme d'un National Bibliographic Number (NBN).

La désignation du pays indique qu'un URN vient de Suisse.

La partie variable de l'URN contient d'abord une identification du service d'attribution. Lorsque les bibliothèques font office de service d'attribution, la Bibliothèque nationale suisse a décidé d'utiliser le sigle de la bibliothèque comme identification. Rien n'empêche de passer ultérieurement du sigle de bibliothèque à la nouvelle norme ISO ISIL (International Standard Identifier for Library Related Organizations). L'identification du service d'attribution est suivie d'un numéro courant. Le nombre de chiffres de ce numéro courant n'est pas limité. Le dernier chiffre est toujours un chiffre de contrôle qui est calculé à l'aide d'un algorithme et sur la base des indications précédentes.

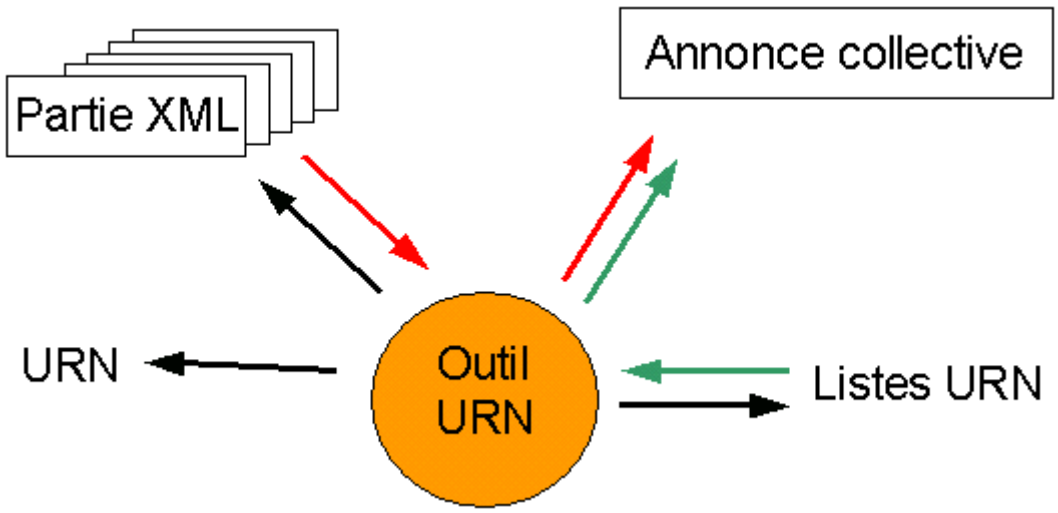
La Bibliothèque nationale suisse a renoncé délibérément à intégrer dans l'URN un numéro intelligent qui reproduit certaines structures ou systématiques. L'expérience a montré que de telles structures ont tendance à se modifier au fil du temps, ce qui enlèverait la plus-value d'un numéro intelligent. De plus, pour une attribution automatisée d'URN, c'est le numéro courant qui est le plus facile à manipuler.

Pour les publications électroniques qui ne sont pas accessibles par l'Internet, la Bibliothèque nationale suisse attribue les identificateurs univoques selon le même principe que pour les URN. Dans ce cas, elle utilise seulement la partie variable de l'URN comme identificateur interne pour une œuvre électronique.

9.3 Un outil simple pour l’attribution de Persistent Identifiers

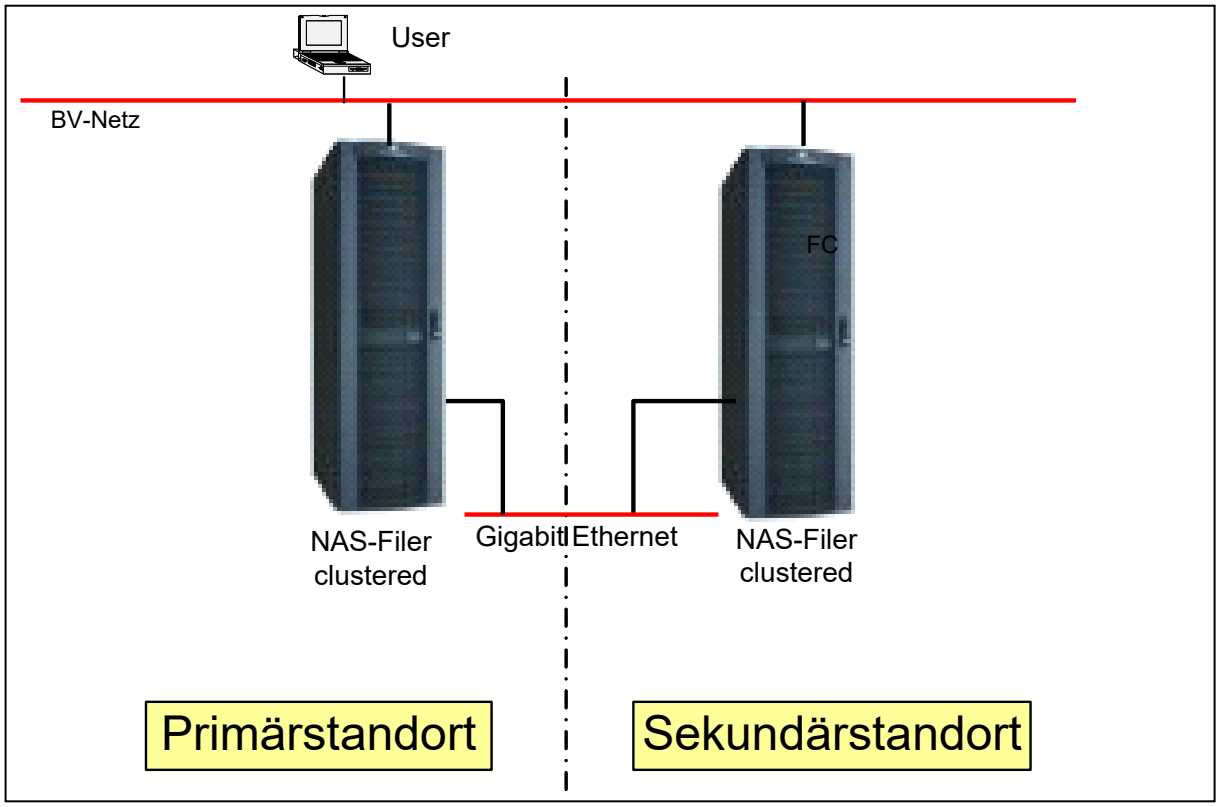
La Bibliothèque nationale suisse a créé un petit outil basé sur Excel pour l'attribution automatisée des URN. Dans le cadre de l'automatisation du processus Ingest, cet outil provisoire a été remplacé par un produit plus développé. Le travail quotidien a permis de mettre en évidence rapidement ce qu'on est en droit d’attendre d’un outil URN:

- Livraison d'un URN isolé qui peut aussi être utilisé comme identificateur interne au besoin.
- Génération de l'annonce XML pour le résolveur d'URN de la Deutsche Nationalbibliothek afin d'activer les URN nouvellement attribués. Chaque nouvelle annonce d'URN ne doit pas être transmise séparément au résolveur, mais les annonces doivent pouvoir être cumulées et transmises comme une annonce collective.
- Génération de listes électroniques (p. ex. sous forme de fichiers Excel) des nouveaux URN. Les liens correspondants sont enregistrés manuellement dans ces listes.
- Création d'une annonce collective pour le résolveur d'URN à partir des listes complétées avec les liens.



10 Stockage des données

10.1 La mémoire à long terme Ninive



La mémoire à long terme Ninive se compose, pour l'essentiel, d'un système NAS redondant (Network Attached Storage) de l'entreprise NetWork Appliance. Les deux composantes du système, d'une capacité de 9 TB de mémoire chacune, se trouvent à deux emplacements à Berne distants d'environ 4,5 km. Une synchronisation automatisée des données entre les deux composantes du système fait en sorte que les données stockées aux deux emplacements soient complètes. A l'emplacement secondaire, une troisième copie des données est faite sur bande magnétique via un lecteur à bandes IBM. Cette troisième copie est conservée séparément.
L'exploitation du système d'archivage Ninive est assurée par l'Office fédéral de l'informatique et de la télécommunication (OFIT).

10.2 Archivage des données

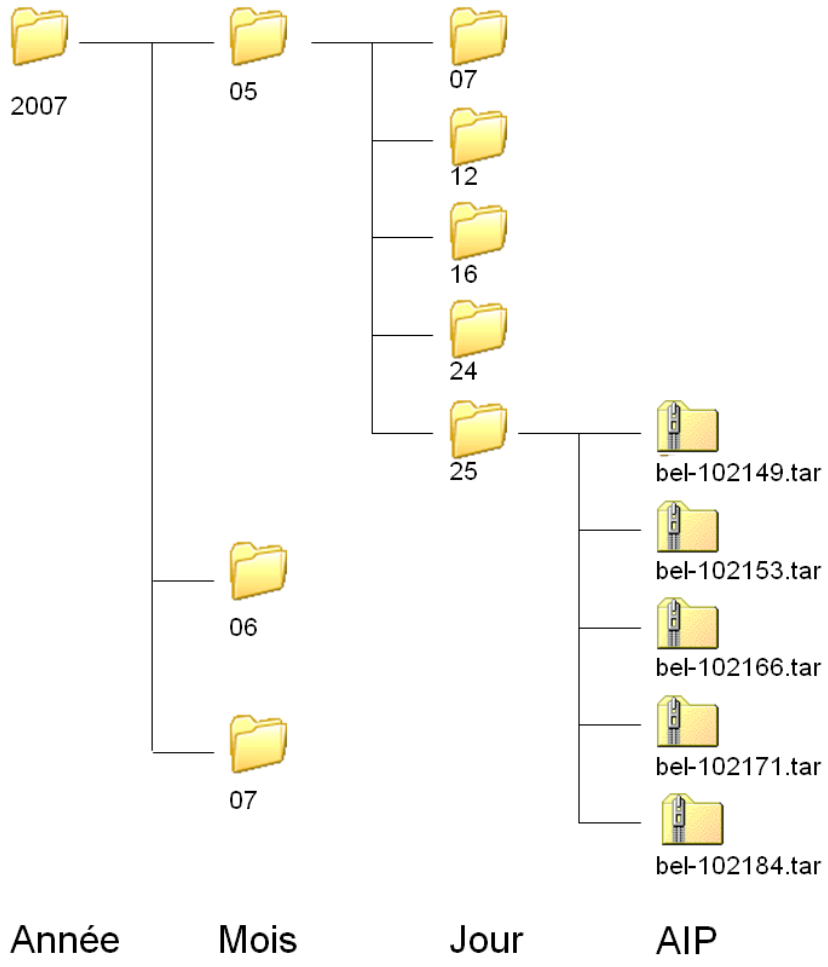
Pour pouvoir retrouver facilement les données archivées, il faut les archiver systématiquement. C'est pourquoi la Bibliothèque nationale suisse a défini l'organisation de l'archivage des données et établi des directives concernant les structures des répertoires, la dénomination et la structure des Archival Information Packages (AIP).

10.2.1 Structure des répertoires

Les données sont archivées dans une structure de répertoires composée de plusieurs niveaux. Ceci permet d'éviter qu'un trop grand nombre de paquets d'archives se trouve dans le même répertoire, car cela peut ralentir énormément l'accès aux données.

Au niveau supérieur, un répertoire est créé par année. Au niveau directement inférieur, il y a un répertoire par mois. Enfin, au niveau inférieur des répertoires, là où les paquets d'archives sont effectivement déposés, un répertoire séparé est créé pour chaque jour où des paquets d'archives entrent dans le système de stockage. Le chemin au sein du système d'archives reflète donc en même temps la

date de stockage. Dans le répertoire 2007\07\15 par exemple, on trouve les données qui ont été déposées dans le système de stockage le 15 juillet 2007.

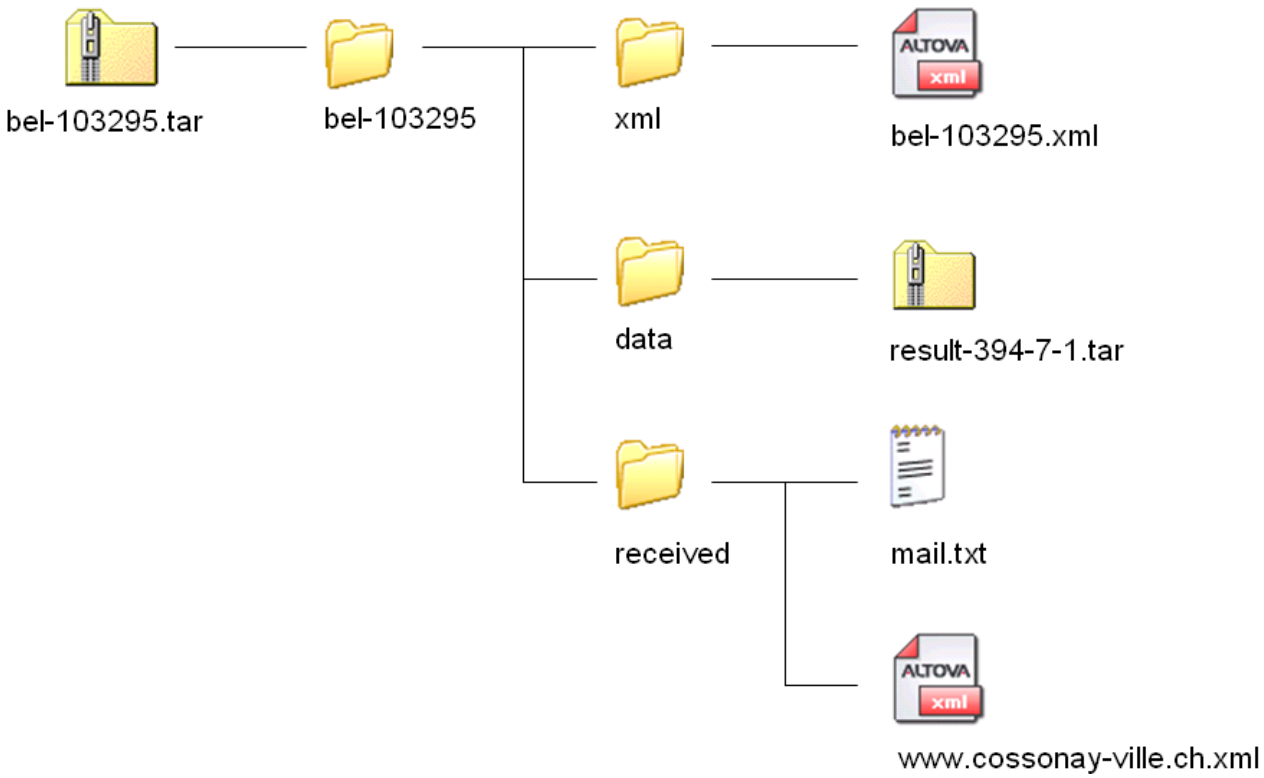


10.2.2 Dénomination

Chaque paquet d'archives reçoit un persistent identifier sous la forme d'un URN ou d'un ID interne si l'AIP n'est pas mis à disposition dans l'Internet. Cet ID interne et univoque est utilisé en même temps comme nom de fichier pour l'AIP (ex. d'un ID interne = bel-102149, nom de l'AIP = bel-102149.tar).

10.2.3 Structure des données

Les AIP sont archivés en tant que fichiers TAR. Ces fichiers TAR ne contiennent pas seulement les objets numériques, mais aussi les métadonnées.

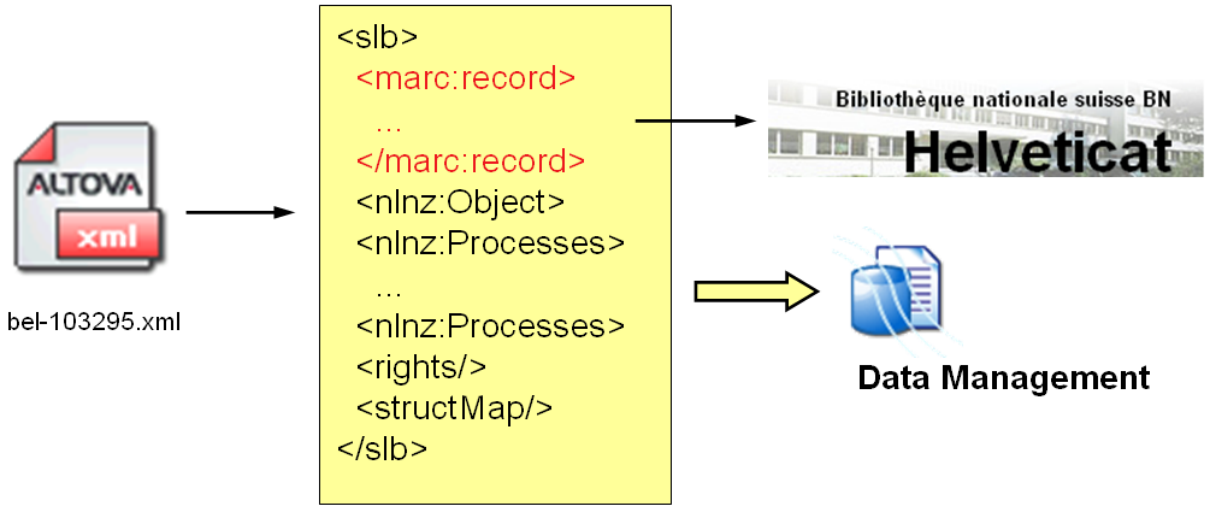


L'AIP est toujours archivé comme fichier TAR. Dans le fichier TAR se trouve un classeur qui est nommé d'après l'ID interne. Ceci devrait garantir qu'au moment du déballage des fichiers TAR, des données ne soient pas copiées par erreur dans le répertoire d'un autre AIP déballé.

Dans ce répertoire, il y a trois sous-répertoires.

- Le répertoire „received" contient les métadonnées reçues du service qui dépose ses publications. Dans le cas d'Archives Web Suisse, il s'agit du mail (mail.txt) et de l'annexe au mail en format XML (www.cossonay-ville.ch.xml). L'annexe a le même nom que le site web annoncé.
- Dans le répertoire „xml" se trouvent les métadonnées complètes de www.cossonay-ville.ch.xml complétées durant le processus Ingest, dans la structure de métadonnées interne de la Bibliothèque nationale suisse. Encore une fois, le fichier XML utilise l'ID interne pour le nom.
- Enfin, le répertoire „data" contient un fichier TAR qui contient l'ensemble du site web avec tous ses documents et répertoires. A moyen terme, le format TAR sera remplacé par un format WARC (Web ARChive) qui va être développé pour la conservation de sites web et qui va devenir un standard ISO.

10.2.4 Stockage des métadonnées



En plus de l'archivage direct dans l'AIP, les métadonnées sont encore enregistrées à deux autres endroits pour l'accès futur à l'AIP.

- Dans le Data Management, une base de données utilisée pour le processus Ingest, mais plus tard également pour l'accès aux données, se trouvent les métadonnées complètes dans le format interne de la Bibliothèque nationale suisse.
- Dans Helveticat, le catalogue de la Bibliothèque nationale suisse, les métadonnées bibliographiques sont également enregistrées. Ceci garantit l'accès par ce catalogue aux informations sur tous les documents archivés à la Bibliothèque nationale suisse.