Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement des Innern EDI
Bundesamt für Kultur BAK
Schweizerische Nationalbibliothek NB

Version 1.7, 19. April 2024

Webarchiv Schweiz

Repräsentative Websites zur Schweiz

Eine gemeinsame Sammlung von Kantonsbibliotheken, Fachbibliotheken und -archiven und der Schweizerischen Nationalbibliothek

Merkblatt Archivieren

Änderungen im Dokument

Version	Datum	Bemerkung
1.0	22.02.2006	Ersterstellung
1.1	01.04.2008	Aktualisierung
1.2	01.05.2009	Aktualisierung
1.3	15.07.2010	Aktualisierung Kapitel 8
1.4	15.01.2011	Aktualisierung
1.5	01.10.2013	Aktualisierung Kapitel 7 und 8
1.6	30.01.2015	Aktualisierung
1.7	19.04.2024	Kleine Korrekturen Kapitel 5.1 und 6.3

1 Inhaltsverzeichnis

2 Einleitung

Das Merkblatt Archivieren zeigt auf, wie die von den Kantonsbibliotheken und weiteren Spezialbibliotheken gemeldeten Websites im System der Schweizerischen Nationalbibliothek abgelegt und aufbewahrt werden.

Am Prozess Archivieren sind im Prinzip zwei Komponenten beteiligt. Zum einen ist dies das Ingest-System, welches die Daten für die Archivierung aufbereitet und sicherstellt, dass die dazugehörenden beschreibenden Informationen (Metadaten) ebenfalls zur Verfügung stehen. Zum anderen ist es das Archivierungssystem (Speicher) selber, auf welchem die digitalen Publikationen samt Metadaten abgelegt werden. Wichtig ist dabei, dass die Metadaten auch in einem Katalog verzeichnet werden, der den Benutzenden zur Verfügung stehen muss.

3 Grundsätze

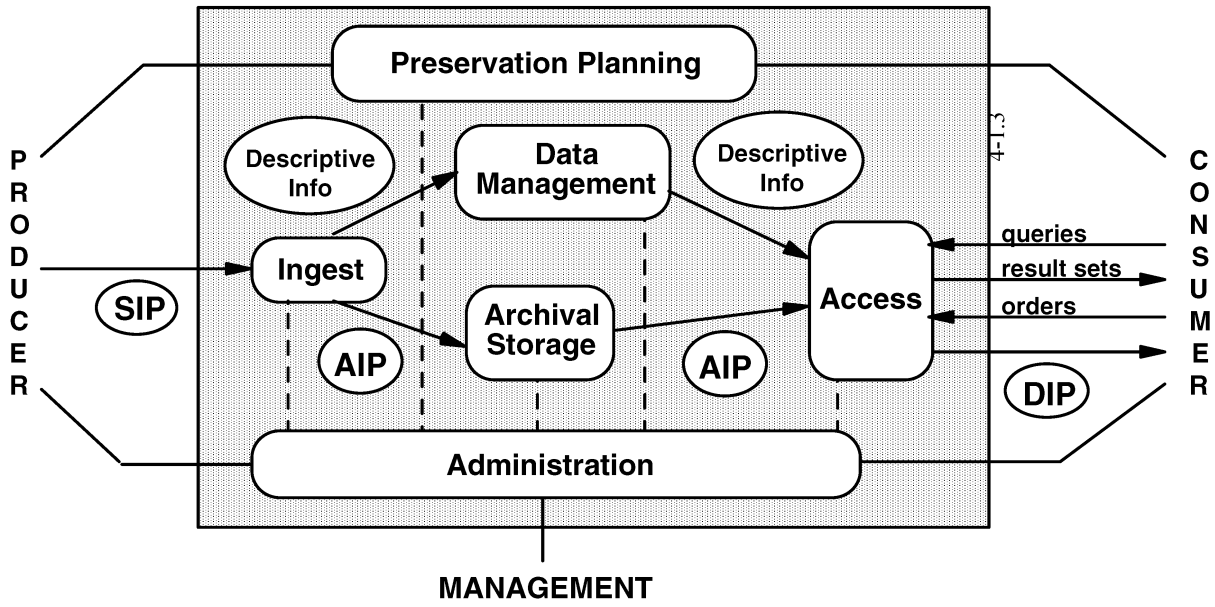
Die Schweizerische Nationalbibliothek hat sich bei der Ausarbeitung des Ingest-Systems von den folgenden Grundsätzen leiten lassen:

- Möglichst weitgehende Automation
- Den Datenproduzenten verschiedene Schnittstellen zur Datenübernahme anbieten
- Normierung der Information, insbesondere der Metadaten möglichst früh im Prozess vornehmen
- Frei kombinierbare Werkzeuge für die Daten- und Metadatenaufbereitung einsetzen
- Digitale Publikationen im Originalformat erhalten

Für das Archivsystem gilt, dass die eingelagerte Information nicht überschrieben respektive gelöscht werden kann. Sollten Informationen verändert (migriert) werden müssen, damit sie für den Benutzenden lesbar bleiben, wird eine neue Version des betroffenen Archivpakets erstellt. Alle bisherigen Versionen dieses Archivpakets bleiben dabei erhalten.

4 OAIS

Bei der Realisierung eines Systems für die Archivierung von elektronischen Informationen hält sich die Schweizerische Nationalbibliothek an das Referenzmodell für ein Offenes Archiv-Informationssystem (OAIS) des Consultative Committee for Space Data Systems. Zum Verständnis der weiteren Ausführungen in diesem Dokument ist eine grobe Kenntnis des OAIS-Modells unabdingbar. Es wird deshalb



kurz erklärt.

OAIS hat sich als Referenzmodell für die digitale Archivierung bei Bibliotheken und Archiven weltweit durchgesetzt. Es ist ein strikt logisches Modell und damit unabhängig von jeder Implementation. Es leistet einen grossen Beitrag zu einem gemeinsamen Verständnis bezüglich der digitalen Archivierung und einer gemeinsamen Sprache in diesem Bereich.

Es werden sechs Hauptfunktionen unterschieden:

Ingest (Datenübernahme)

- Übernahme der vom Produzenten erzeugten SIPs (Submission Information Package)
- Überprüfung auf Vollständigkeit und Unversehrtheit
- Umwandlung des SIP in ein AIP (Archival Information Package)
- Extraktion der beschreibenden Information für die Findmitteldatenbank
- Übermittlung des AIP an den Archivspeicher
- Mitteilung an das Data Management

Archival Storage (Archivspeicher)

- Aufbewahrung und Erhaltung der AIPs
- Erstellen von Backups
- Regelmässige Prüfung der Datenintegrität
- Wiederherstellungsmechanismen für Notfälle
- Weitergabe von AIPs an e-Helvetica Access für die Nutzung

Access (Abfrage)

- Benutzerinterface
- Ermöglichen von Recherche und generieren von Antworten mit Beschreibung der AIPs und Angaben zu deren Verfügbarkeit

- Empfangen von Anfragen (requests) und Ausliefern der DIPs (Dissemination Information Package)
- Sicherstellen der Einhaltung von Zugriffsberechtigungen

Administration
- Steuerung der Gesamtabläufe im OAIS und seiner Aussenbeziehungen
- Konfiguration von Hard- und Software
- Überprüfung von Zugriffsrechten
- Erzeugung von DIPs und deren Übergabe an die Benutzenden

***Data Management (Datenverwaltung*)**
- Verwaltet die beschreibenden Informationen (Datenbank), die Archivbestände und Dokumente identifizieren, sowie weitere Daten, die für den Umgang mit dem Archivgut notwendig sind
- Entgegennahme und Bearbeitung von Anfragen (queries) aus dem Nutzungsbereich

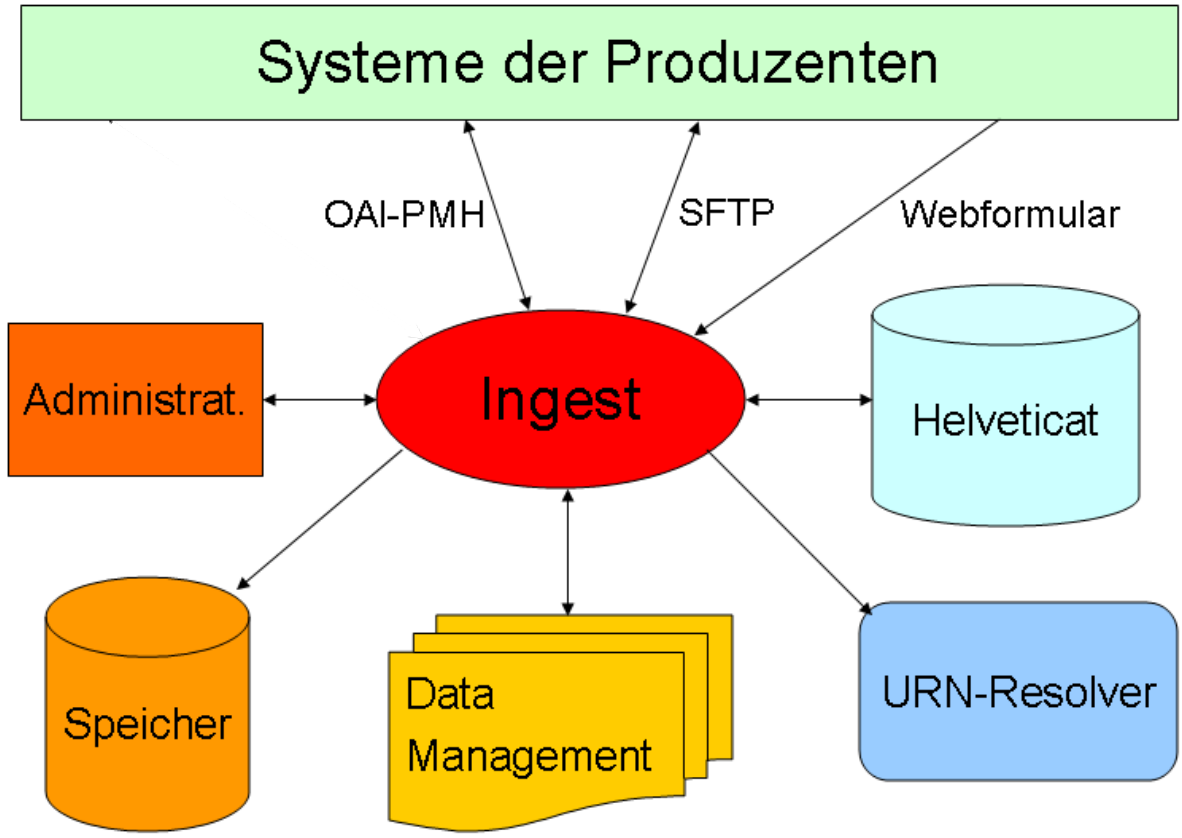
***Preservation Planning (Archivierungsplanung*)**
- Verfolgen der Technologieentwicklung und entwickeln von Empfehlungen in Bezug auf Archivierungsstandards und -politik
- Überwachen der Archivierungsbemühungen
- Ausarbeiten von Empfehlungen für die Erhaltung der Lesbarkeit der gespeicherten Information
- Planen von Datenmigrationen und Kopiervorgängen

5 Ingest

Der ganze Prozess der Datenübernahme vom Lieferanten oder von der über das Internet zugänglichen Datenquelle bis hin zur Einlagerung in das Archivsystem wird als Ingest-Prozess bezeichnet.

5.1 Systemumgebung der Schweizerischen Nationalbibliothek

Entscheidend für das gute Funktionieren des Ingest-Prozesses ist seine Einbettung in die Systemumgebung. Die folgende Grafik zeigt die Komponenten, die dabei eine Rolle spielen. Für die Erläuterung der einzelnen Systeme sei auf das OAIS-Modell hingewiesen. Über einen URN-Resolver werden Persistent Identifiers in Links umgewandelt.

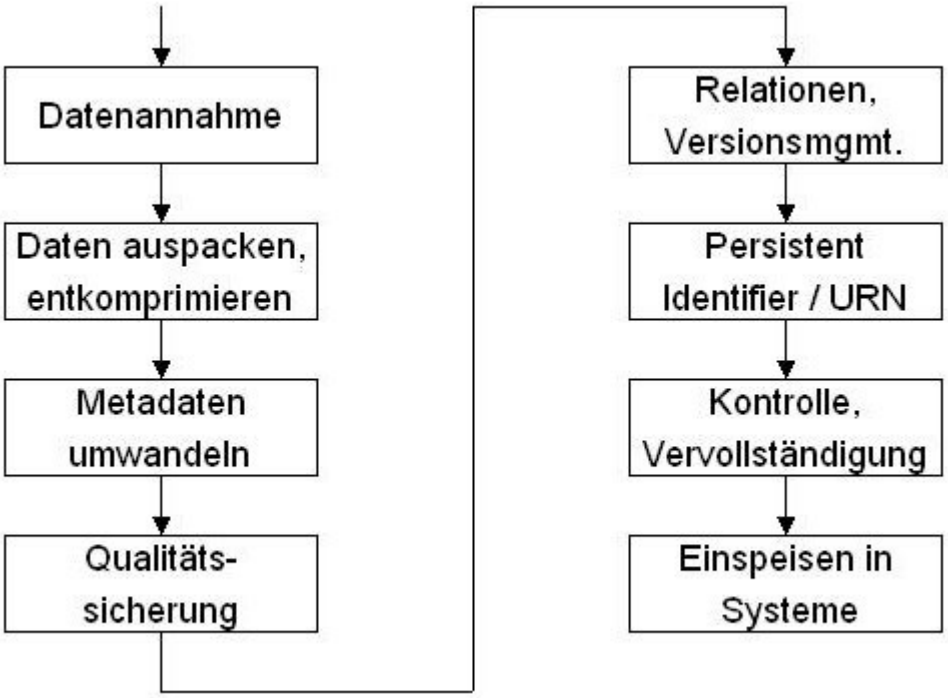


Für die Übernahme der Daten aus den Systemen der Produzenten stellt der Ingest-Prozess verschiedene Schnittstellen zur Verfügung.

- OAI-PMH: Die grösseren Universitätsbibliotheken erlauben ein OAI-PMH-Harvesting ihres Bibliothekskatalogs. Auf diesem Weg beschafft sich die Schweizerische Nationalbibliothek die Metadaten der neu verzeichneten Dissertationen und Habilitationen.
- SFTP: Die Schweizerische Nationalbibliothek greift mit SFTP periodisch auf ein System des Produzenten zu und holt die dort zur Verfügung gestellten neuen Datenpakete ab. Diese Möglichkeit wird bisher von keinem Produzenten genutzt.
- Webformular: Über Webformulare können Datenlieferanten die Metadaten der für die Langzeitarchivierung vorgesehenen Publikationen anmelden. Diese Anmeldung trifft als E-Mail mit einem XML-Attachment in bestimmten Postfächern ein, auf die der Ingest-Prozess zugreift.

5.2 Der Ingest-Prozess

Der Ingest-Prozess besteht aus einer ganzen Reihe von einzelnen Arbeitsschritten.



Nach der Datenübernahme und dem entkomprimieren von Dateien, die z.B. als ZIP-Files eintreffen, ist es nötig, die in den verschiedensten Formen vorliegenden Metadaten in eine interne Struktur der Schweizerischen Nationalbibliothek umzuwandeln, damit die weitere Verarbeitung für alle eintreffenden Daten einheitlich erfolgen kann.

Im Rahmen der Qualitätssicherung werden die eintreffenden Daten überprüft:

- Lesbarkeit
- Authentizität (Übereinstimmen der Checksumme)
- Formattreue (Ist eine Datei mit der Endung .pdf auch wirklich ein PDF-File?)
- Virenfreiheit
- Vollständigkeit der Daten
- Vollständigkeit der Metadaten
- Doppellieferungen (Sind die angelieferten Daten nicht schon in der Schweizerischen Nationalbibliothek vorhanden?)

Die Zuordnung zu übergeordneten Einträgen muss sichergestellt werden. Neue Hefte sind z.B. dem richtigen Zeitschriftentitel zuzuordnen. Bei Websites geht es darum, periodisch neue Versionen der gleichen Website zu sammeln und abzulegen.

Alle zu archivierenden Datenpakete werden mit einem eindeutigen Identifikator (Persistent Identifier) versehen.

Fehlende Metadaten müssen eingefügt und die technischen und administrativen Angaben vervollständigt werden.

Verschiedene Systeme werden anschliessend mit Daten aus dem Ingest-Prozess versorgt:

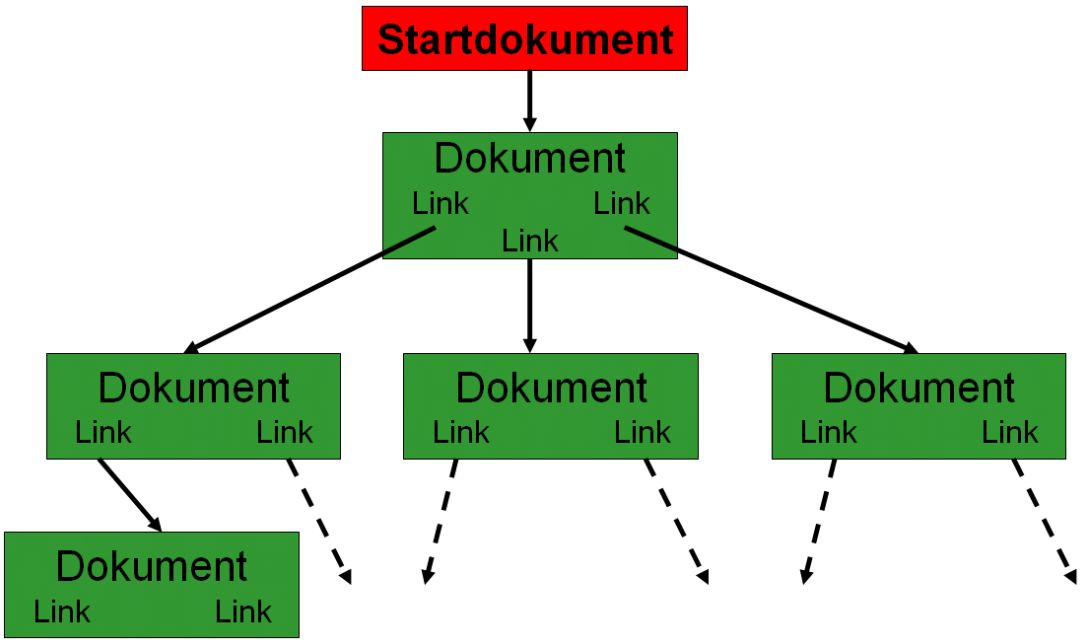
- Helveticat: Verzeichnung der elektronischen Publikationen im Katalog
- Data Management: Datenverwaltung auf dem Archivierungssystem (vor allem auch technische und administrative Daten)
- Speicher (Archivierungssystem): Haltung der für die Archivierung hergestellten Datenpakete

Zur Verarbeitung von Spezialfällen wie beispielsweise von sehr grossen Webarchiv-Paketen sind neben Ingest zusätzliche Java-Tools im Einsatz.

6 Harvesting

6.1 Funktionsweise des Harvesters

Das Einsammeln von Websites aus dem Internet wird als Harvesting bezeichnet. Spezielle Programme sorgen beim Harvesting dafür, dass ausgehend von einer Startseite alle Links weiterverfolgt werden und die Dateien, die innerhalb des definierten Sammelgebiets liegen, heruntergeladen werden.

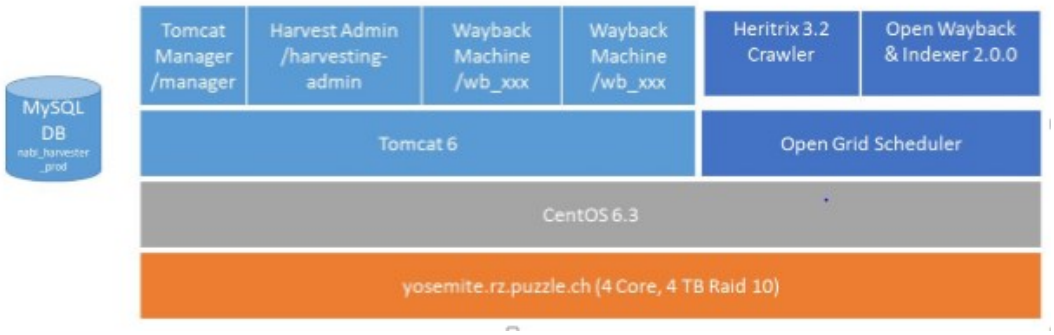


Die Sammelgebiete können durch die Kantonsbibliotheken/Spezialbibliotheken flexibel definiert werden. In der Hauptsache gibt es zwei verschiedene Möglichkeiten:

- Sammlung einer Domäne wie http://www.rorschach.ch
  Damit werden alle Dokumente in sämtlichen Verzeichnissen und Unterverzeichnissen gesammelt, die in der Domäne www.rorschach.ch zu finden sind.
- Sammlung von Dokumenten aus bestimmten Verzeichnissen heraus wie http://www.swissworld.org/de/geschichte
  Nur die Dokumente im Verzeichnis /geschichte und seinen Unterverzeichnissen werden gesammelt. Dokumente, die sich beispielsweise unter http://www.swissworld.org/de/kultur/ befinden, werden nicht ins Harvesting einbezogen.

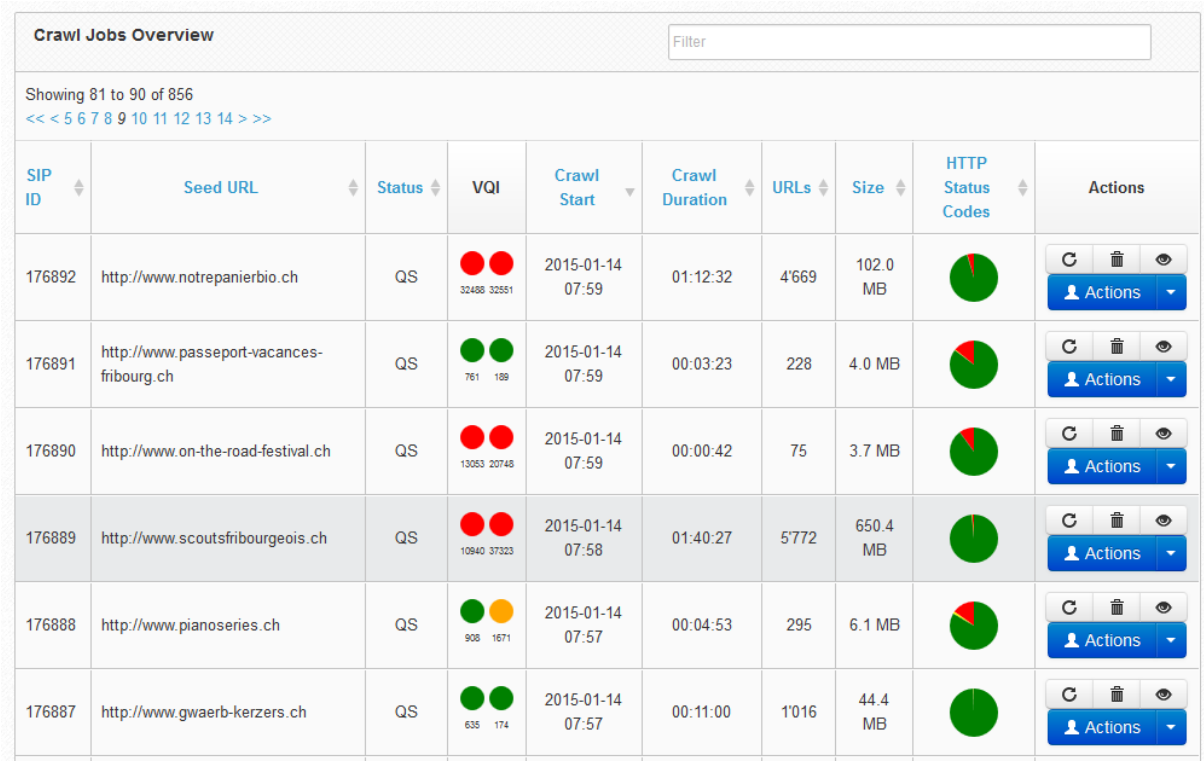
6.2 Aufbau des Harvesters

Das Harvesting muss in einem Netzbereich erfolgen, der möglichst wenig Einschränkungen unterworfen ist, sonst ist die Gefahr gross, dass nicht alle gewünschten Dokumente eingesammelt werden können. Aus diesem Grund befindet sich die Harvesting-Infrastruktur bei einem externen Provider.



Der Crawler Heritrix sammelt die Websites ein, danach erfolgt die Indexierung und dann kann die geharvestete Website zur Qualitätskontrolle in der Wayback Machine geöffnet werden.

Über ein webbasiertes Benutzerinterface kann der Harvesting-Prozess direkt gesteuert und überwacht werden. Die Jobs werden durch die Eingabe von SIP Nummer und URL gestartet und es können um die 20 Harvesting-Aufträge gleichzeitig laufen.



Crawl Jobs Overview

Filter

Showing 81 to 90 of 856

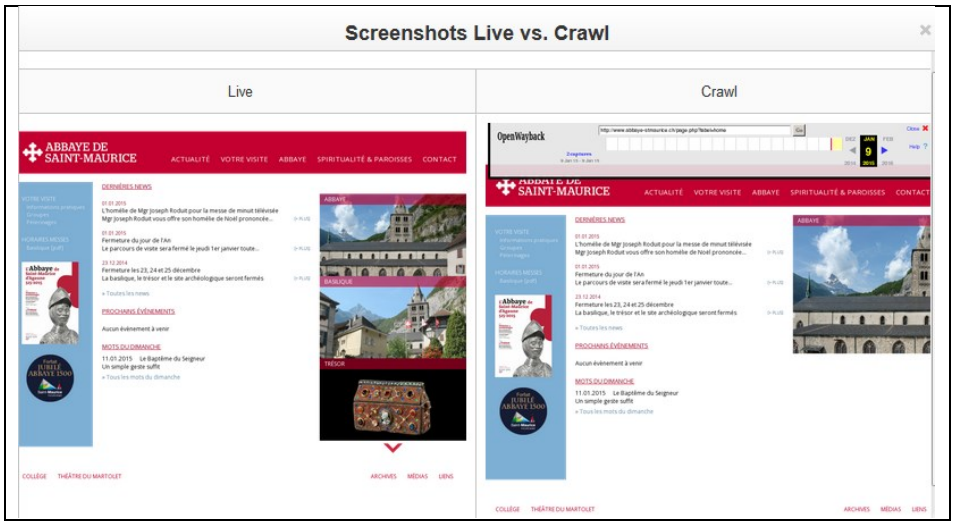
<< < 5 6 7 8 9 10 11 12 13 14 > >>

SIP ID	Seed URL	Status	VQI	Crawl Start	Crawl Duration	URLs	Size	HTTP Status Codes	Actions
176892	http://www.notrepanierbio.ch	QS	32488 32551	2015-01-14 07:59	01:12:32	4'669	102.0 MB		Actions
176891	http://www.passeport-vacances-fribourg.ch	QS	761 189	2015-01-14 07:59	00:03:23	228	4.0 MB		Actions
176890	http://www.on-the-road-festival.ch	QS	13053 20748	2015-01-14 07:59	00:00:42	75	3.7 MB		Actions
176889	http://www.scoutsfribourgeois.ch	QS	10940 37323	2015-01-14 07:58	01:40:27	5'772	650.4 MB		Actions
176888	http://www.pianoseries.ch	QS	908 1671	2015-01-14 07:57	00:04:53	295	6.1 MB		Actions
176887	http://www.gwaerb-kerzers.ch	QS	635 174	2015-01-14 07:57	00:11:00	1'016	44.4 MB		Actions

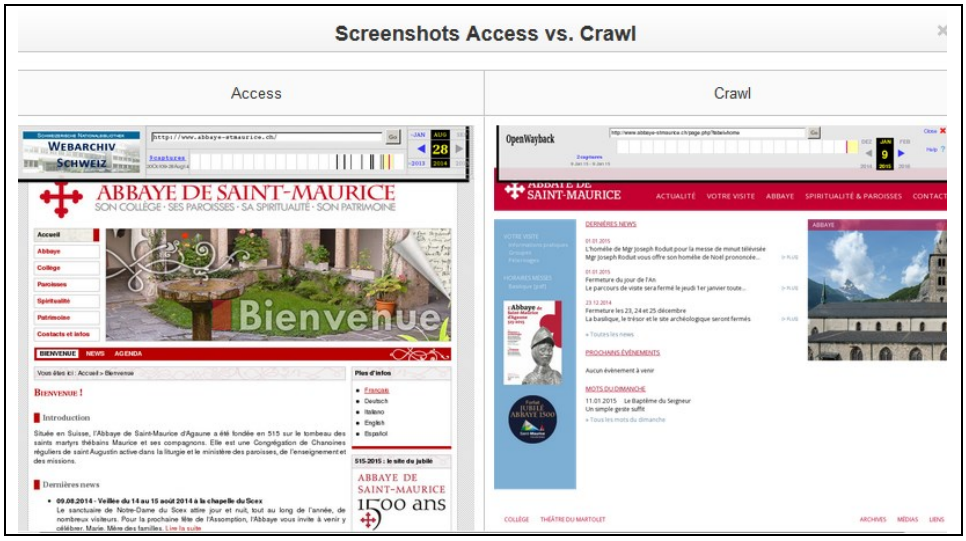
Die Crawls werden überwacht durch verschiedene Anzeigemöglichkeiten wie die Anzahl detektierter URLs, die Verteilung der HTTP Status Codes oder die Analyse des detaillierten Crawl Logs. Durch die Funktion „Pausieren“ kann bereits das Zwischenresultat in der Wayback Machine überprüft werden.

Zur Unterstützung der Qualitätskontrolle hat die Schweizerische Nationalbibliothek zudem innerhalb der Harvesting-Infrastruktur einen Visual Quality Index implementiert. Von allen Crawls werden mittels der Zusatztools PhantomJS und CasperJS automatisch zwei Screenshot-Vergleiche erstellt.

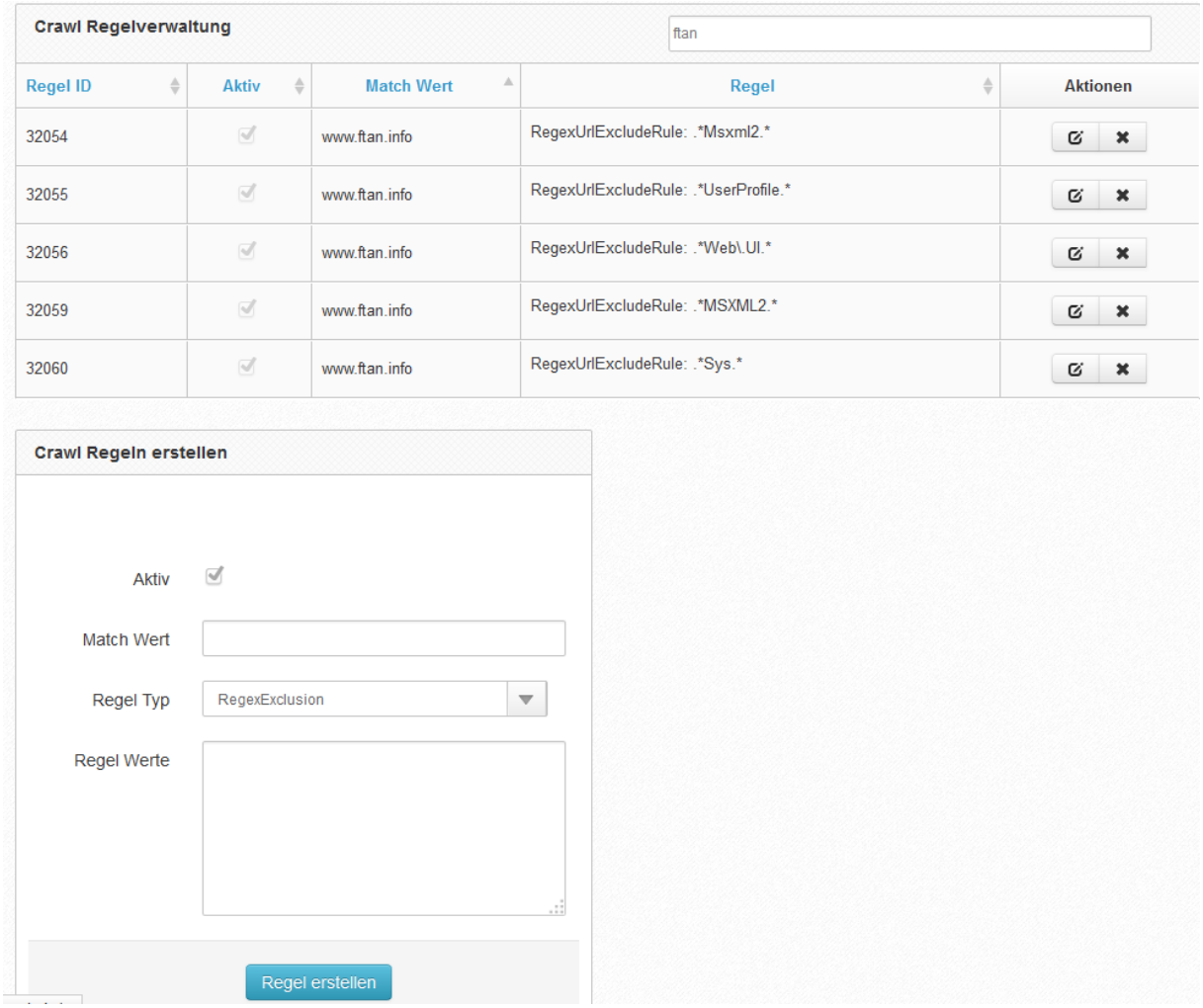
1. Der Vergleich der Live Website mit dem aktuellen Crawl:



2. Der Vergleich der letzten archivierten Version in e-Helvetica Access mit dem aktuellen Crawl:

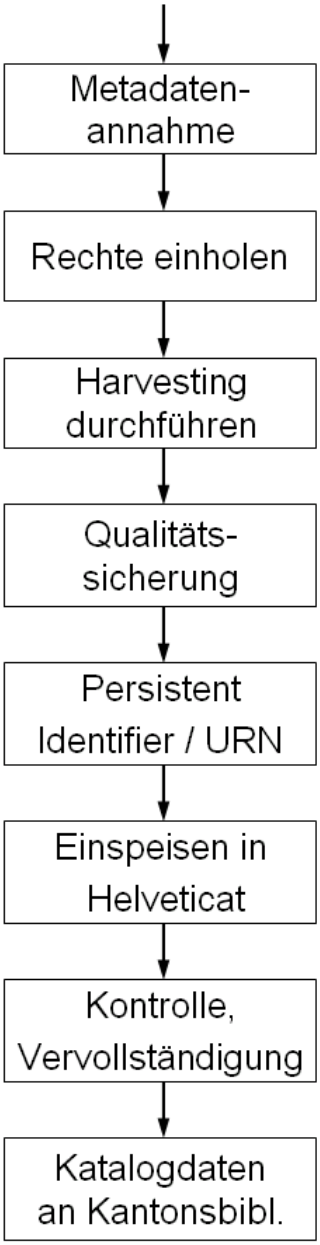


Bei Bedarf lassen sich die Crawls verbessern durch das Hinzufügen zusätzlicher Seed URLs oder durch das Ausschliessen von URL-Bereichen mittels Regular Expressions.



6.3 Verarbeitungsprozess

Der gegenwärtig implementierte Arbeitsablauf für die Behandlung von Websites unterscheidet sich leicht vom im Ingest-Prozess sonst vorgesehenen üblichen Bearbeitungsweg.



Die Lieferung der Metadaten erfolgt über ein passwortgeschütztes Webformular. Nach dem Eintreffen der Daten in der Schweizerischen Nationalbibliothek werden die Rechte für das Einsammeln und das Aufbewahren der Websites durchgeführt. Die Schweizerische Nationalbibliothek versendet eine E-Mail an den Eigentümer der Website (Website-Betreiber) mit Informationen zu Webarchiv Schweiz und der Möglichkeit, die Archivierung der Website abzulehnen. Erhält die Schweizerische Nationalbibliothek daraufhin keine Rückmeldung erfolgt das eigentliche Harvesting der Website im Sinne von „Fair Use“.

Die Überprüfung und Vervollständigung der Metadaten erfolgt erst nach dem Import in Helveticat, weil die Bibliothekar/innen dadurch ihre gewohnte Benutzeroberfläche (im Falle der Schweizerischen Nationalbibliothek Alma, Ex Libris) bei der Bearbeitung der Metadaten verwenden können.

Am Schluss des Prozesses können die vollständigen bibliographischen Metadaten den Kantonsbibliotheken/Spezialbibliotheken zur Integration in ihre eigenen Kataloge weitergegeben werden (siehe Merkblatt Bereitstellen).

Die Schweizerische Nationalbibliothek beschränkt sich beim Einsammeln von Websites gegenwärtig auf ein selektives Harvesting. Ein Harvesting der ganzen .ch-Domäne soll aber nicht von vornherein ausgeschlossen werden. Im Moment sind die Kapazitäten dazu aber nicht vorhanden.

7 Qualitätskontrolle

Eine zuverlässige Qualitätssicherung von Websites kann erst mit einem technischen Instrument vorgenommen werden, das die eingesammelten Dokumente genauer analysiert und allfällige Fehler aufzeigt. Ein solches Instrument soll im Rahmen von IIPC (International Internet Preservation Consortium) entwickelt werden.

Bis dieses Instrument eingesetzt werden kann, muss die Qualitätssicherung zwangsläufig von Hand durchgeführt werden und kann deshalb nur rudimentär vorgenommen werden. Bei dieser Qualitätskontrolle geht es nicht darum, die Qualität der Website im Internet zu überprüfen sondern die Qualität des Sammelvorgangs zu kontrollieren.

8.1 Überprüfungsraster

Die Wayback Machine ermöglicht den Zugriff auf die gesammelten Websites durch ein gängiges Browser-Interface (z.B. Firefox, Internet Explorer). Dabei werden folgende Überprüfungen vorgenommen:

1. Gesamteindruck
   Die Stimmigkeit des Gesamteindrucks der eingesammelten Website wird mit der Originalwebsite verglichen.

2. Anzahl der vorhandenen Dokumente
   Liegt die Anzahl der für eine Website eingesammelten Dokumente unter 100 wird auf der Originalwebsite nachgeschaut, ob diese wirklich so klein ist.

3. Darstellung
   Entspricht die Darstellung mittels Stylesheets und Graphiken der Originalwebsite? Hat es Photogalerien auf der Website und werden diese korrekt angezeigt?

4. Verzeichnisstruktur
   Ausgehend von der angemeldeten URL wird überprüft, ob Dateien aus untergeordneten Verzeichnissen mitgeliefert worden sind. Sind auch Dokumente wie PDFs im Crawl enthalten?

5. Vollständigkeitsüberprüfung
   Die Vollständigkeitsüberprüfung erfolgt innerhalb der jeweiligen Website nur anhand von Stichproben. Diese beschränkt sich auf die folgenden Elemente:
   - Aufruf der Ausgangsseite und Überprüfung ob alle Elemente vorhanden sind.
   - Prüfen der verschiedenen Sprachversionen, falls mehrere Sprachen vorhanden sind.
   - Verfolgen jedes auf der Ausgangsseite vorhandenen Links und Überprüfung ob auf den Folgedokumenten alle Elemente vorhanden sind.
   - Verfolgen einer Kette von Links von einem Folgedokument aus über weitere 5 Ebenen mit der jeweiligen Kontrolle auf das Vorhandensein aller Elemente.

6. Funktionsprüfung
   Verfügt die Website über dynamische Elemente, so wird deren Funktion überprüft. Wird festgestellt, dass eine Funktion nicht abgerufen werden kann, wird geprüft wodurch die Beschränkung ausgelöst wird.

Das Ergebnis der Qualitätssicherung wird dokumentiert als OK oder als ungenügend. Eine eingeholte Website wird erst abgewiesen, wenn ein zweites und drittes Harvesting mit angepassten Einstellungen kein besseres Ergebnis gebracht hat.

Gründe, aus denen eine eingeholte Website abgewiesen werden kann:

- Wesentlich andere oder falsche Darstellung
  Fehlende (oder falsch interpretierte) Stylesheets, Fonts (Sonderzeichen), Grafiken, falsch platzierte Komponenten
- Fehlender Inhalt
  Teile der Website oder integrierte Dokumente, die inhaltlich relevant sind, fehlen (z.B. PDF, DOC, XLS, ...)
- Fehlende Menufunktionen
  Wenn es keinen anderen Weg gibt, die Inhalte abzurufen (z.B. durch eine vorhandene Sitemap)
- Out of scope
  Ausserhalb des Sammelgebiets

Nachfolgende Probleme sind bekannt und führen nicht automatisch zur Abweisung einer eingeholten Website:

- Suchfunktion fehlt oder funktioniert nicht
- Druckfunktion fehlt oder funktioniert nicht
- Kalender fehlt oder funktioniert nicht
- Systemzeit fehlt oder funktioniert nicht
- Zähler fehlt oder funktioniert nicht
- Webcam fehlt oder funktioniert nicht
- Webformular/Eingabefeld fehlt oder funktioniert nicht
- Forum/Wiki/Blog fehlt oder funktioniert nicht
- Geografische Karte fehlt oder funktioniert nicht
- Video fehlt oder funktioniert nicht
- Audio fehlt oder funktioniert nicht
- Multimedia Diashow, Browserspiel usw. fehlt oder funktioniert nicht
- Live-Links (nur akzeptabel, wenn der archivierte Inhalt zugänglich bleibt)
- Scriptgesteuerte Funktion fehlt oder funktioniert nicht
- Servergesteuerte Funktion fehlt oder funktioniert nicht

Global betrachtet gibt es nur wenige harte Kriterien, die zwingend dazu führen, eine eingeholte Website abzuweisen. Die „Collection Manager“ (Bibliothekar/innen) entscheiden, ob eine eingeholte Website akzeptabel ist oder nicht.

Folgende Fragestellungen begleiten die Entscheidung:

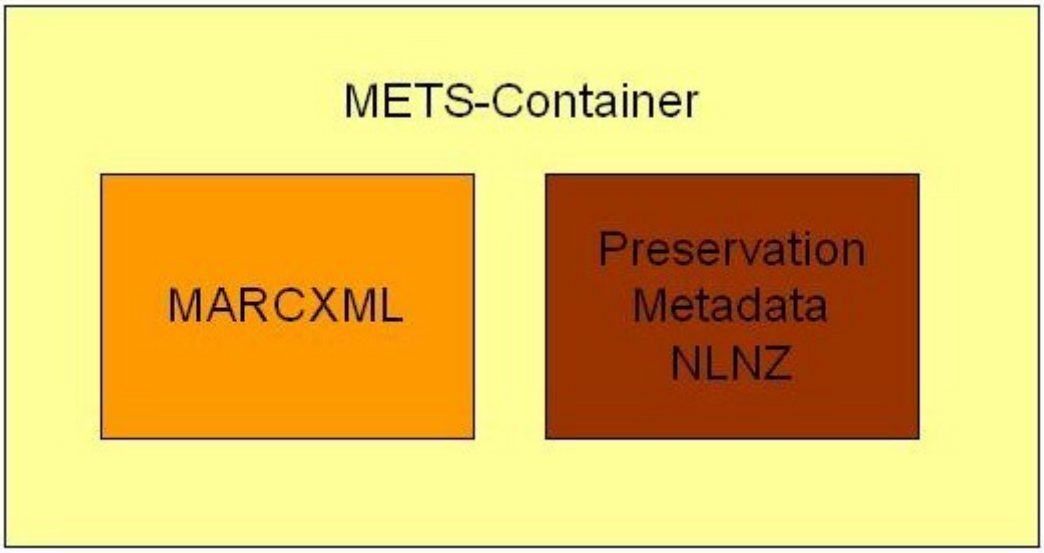
- Was sind die relevanten Teile der Website („Significant properties“) und werden diese im Webarchiv abgebildet?
- Kann mit einem erneuten Harvesting der Website das Ergebnis verbessert werden?

8 Metadaten

Die Schweizerische Nationalbibliothek hat keine eigene Metadatenstruktur entwickelt. Sie profitiert von bestehenden Formaten im XML-Format. Damit entfällt auch der Aufwand für die Weiterentwicklung der Metadatenstruktur.

Für die interne Metadatenstruktur verwendet die Schweizerische Nationalbibliothek den von der Library of Congress gepflegten METS-Container. In diesen wird für die bibliographischen Daten MARCxml eingebettet. MARCxml wird ebenfalls von der Library of Congress unterhalten und ist kompatibel mit MARC21, der Metadatenstruktur von Helveticat. Die nicht bibliographischen Metadaten

werden im von der National Library of New Zealand entwickelten Schema für «Preservation Metadata» ebenfalls in den METS-Container integriert.



9 Persistent Identifiers

9.1 Das gewählte System der Schweizerischen Nationalbibliothek

Ein Persistent Identifier (eindeutiger Identifikator) soll zwei Bedürfnisse abdecken:

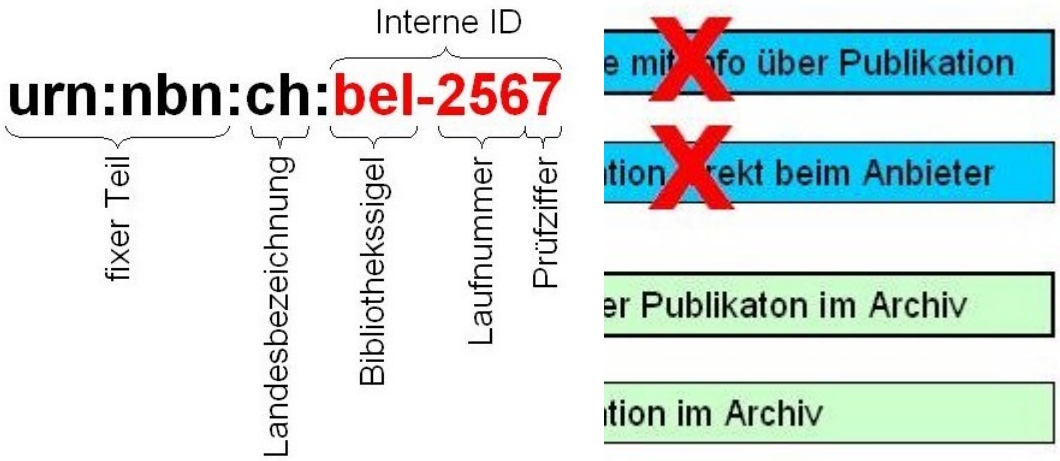
- Eindeutige Identifikation des Archivguts
- Stabiler Verweis auf eine online verfügbare Datenquelle (Links haben sich als sehr unbeständig erwiesen.)

Die Schweizerische Nationalbibliothek hat sich entschieden, Uniform Resource Names (URN) in der Form von National Bibliographic Numbers (NBN) zu verwenden, denn die URN erfüllt die oben aufgeführten Bedürfnisse. Um die URNs in Links umzuwandeln, kann die Schweizerische Nationalbibliothek den URN-Resolver der Deutschen Nationalbibliothek mitbenutzen.

Über den Webbrowser können beliebige Benutzer/innen eine URN in der Form https://admin.nbn-resolving (Webadresse des URN-Resolvers und URN) aufgeben. Der URN-Resolver liefert dann einen gültigen Link auf die gewünschte Information zurück, mit dem diese Information im Internet automatisch aufgerufen und im Browser angezeigt wird.

Der URN-Resolver bringt für die Schweizerische Nationalbibliothek eine ganze Reihe interessanter Funktionalitäten. So ist es möglich, eine ganze Reihe von Links (Uniform Resource Locator, resp. URL) hinter einem URN abzulegen. Mit Hilfe einer Priorisierung kann festgelegt werden, in welcher Reihenfolge der URN-Resolver die hinterlegten Links den Benutzenden liefert. Ist der am höchsten priorisierte Link ungültig, wird automatisch der nächste weitergegeben. Damit können zum Beispiel über eine URN Links auf eine Website beim Produzenten mit Informationen über die gewünschte Publikation, auf die beim Produzenten gespeicherte elektronische Publikation selber und ebenso auf die Informationen im Archiv der Schweizerischen Nationalbibliothek gelegt werden. Entfernt der Produzent die Publikation aus dem Angebot auf seiner Website, weil z.B. kein kommerzielles Interesse mehr an ihr besteht, bleibt die URN trotzdem gültig und zeigt dann direkt auf die bei der Schweizerischen Nationalbibliothek archivierten Publikation.

9.2 Der Aufbau des Persistent Identifiers



Der eindeutige Identifikator der Schweizerischen Nationalbibliothek entspricht der für URN vorgegebenen Norm und enthält zuerst einen fixen Teil an dem abzulesen ist, dass es sich um eine URN in der Form einer National Bibliographic Number (NBN) handelt.

Mit der Landesbezeichnung wird mitgeteilt, dass eine URN aus der Schweiz stammt.

Der variable Teil der URN enthält zuerst eine Identifikation der Vergabestelle. Die Schweizerische Nationalbibliothek hat sich entschlossen bei Bibliotheken als Vergabestelle das Bibliothekssigel als Identifikation zu verwenden. Einem späteren Wechsel vom Bibliothekssigel auf die neue ISO-Norm ISIL (International Standard Identifier for Library Related Organizations) steht nichts im Weg. Die Identifikation der Vergabestelle wird gefolgt von einer Laufnummer. Die Anzahl Ziffern für diese Laufnummer ist nicht begrenzt. Die letzte Ziffer ist immer eine Prüfziffer, die aufgrund der vorangehenden Angaben mit Hilfe eines Algorythmus berechnet wird.

Die Schweizerische Nationalbibliothek hat bewusst darauf verzichtet, eine intelligente Nummer in die URN einzubinden, die gewisse Strukturen oder Systematiken wiedergibt. Im Verlauf der Zeit pflegen sich solche Strukturen erfahrungsgemäss zu verändern und damit wäre der Mehrwert einer intelligenten Nummer nicht mehr vorhanden. Für eine automatisierte URN-Vergabe ist eine Laufnummer zudem am einfachsten zu handhaben.

Für elektronische Publikationen die nicht über das Internet zugreifbar sind, vergibt die Schweizerische Nationalbibliothek die eindeutigen Identifikatoren nach dem gleichen Prinzip wie die URNs. Sie verwendet in diesem Fall einfach nur den variablen Teil der URN als internen Identifikator für ein elektronisches Werk.

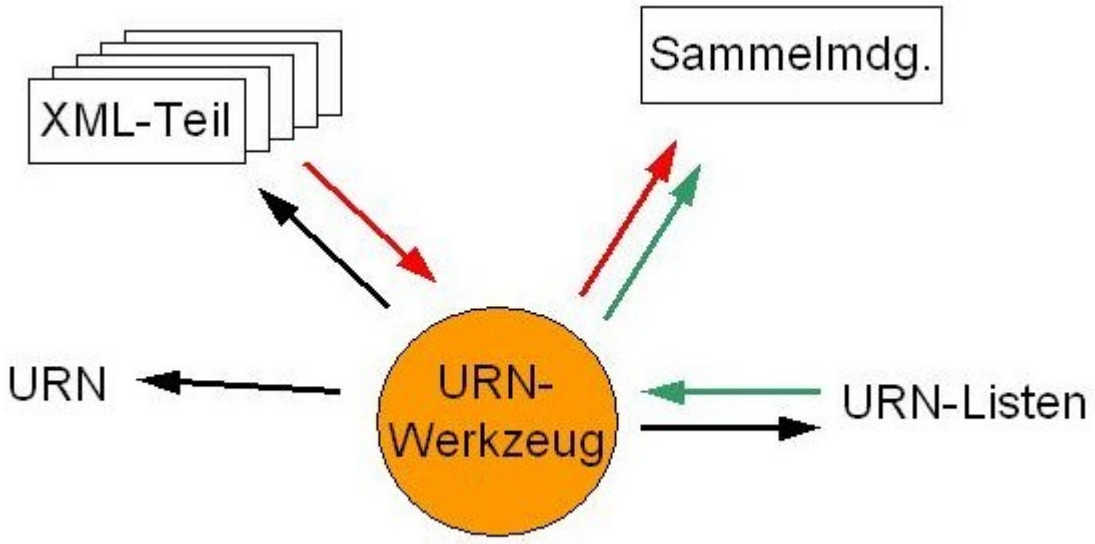
9.3 Ein einfaches Tool zum Vergeben von Persistent Identifiers

Die Schweizerische Nationalbibliothek hat zur automatisierten Vergabe der URNs ein kleines Tool auf Excel-Basis erstellt. Im Rahmen der Automatisierung des Ingest-Prozesses wurde dieses provisorische Tool durch ein ausgereifteres Nachfolgeprodukt ersetzt. In der täglichen Arbeit hat sich die Anforderung an ein URN-Tool rasch herauskristallisiert:

- Lieferung einer isolierten URN, die bei Bedarf auch als interner Identifikator eingesetzt werden kann.
- Generieren der XML-Meldung für den URN-Resolver der Deutschen Nationalbibliothek, um die neu vergebenen URNs zu aktivieren. Dabei darf nicht jede neue URN-Meldung einzeln an

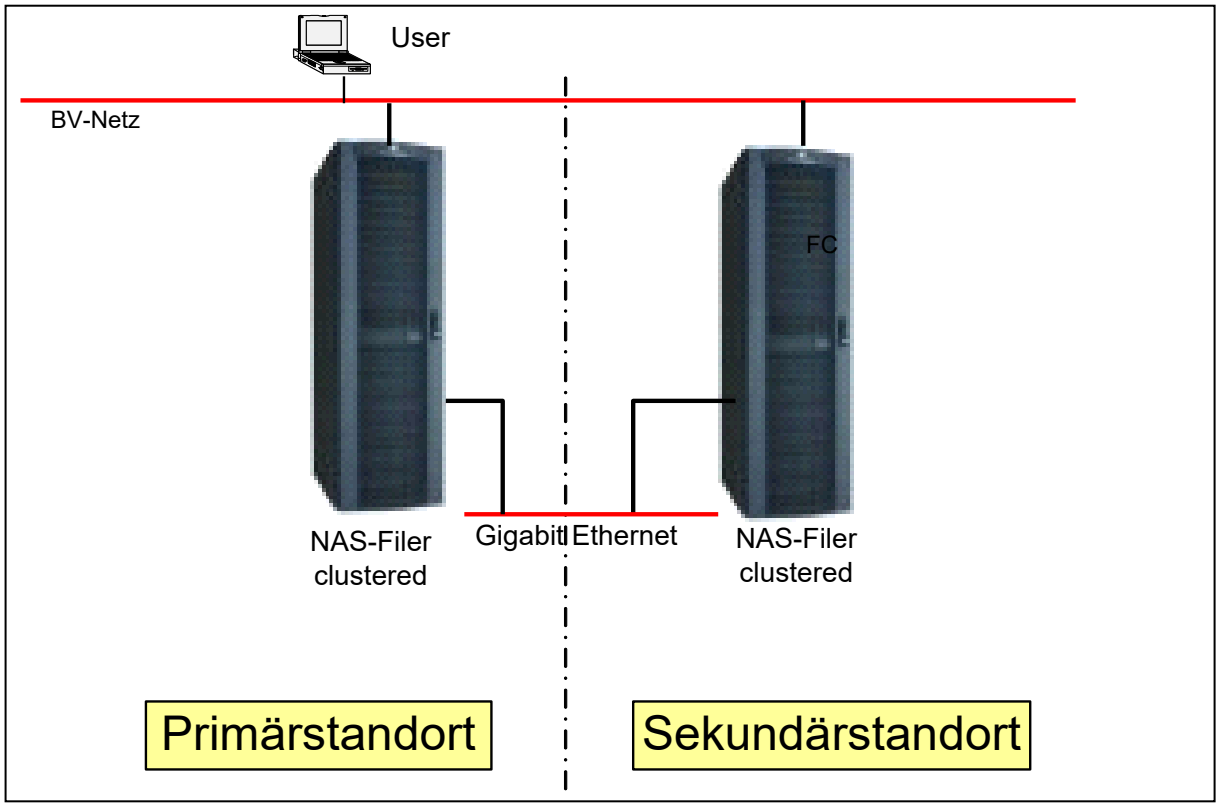
den Resolver übermittelt werden, sondern die anfallenden Meldungen sollen kumuliert und als Sammelmeldung übermittelt werden können.

- Generieren von elektronischen Listen (z.B. in Form von Excel-Files) mit neuen URNs. In diese Listen werden die dazugehörenden Links manuell eingetragen.
- Erstellen einer Sammelmeldung für den URN-Resolver aus den mit den Links ergänzten Listen.



10 Datenspeicherung

10.1 Der Langzeitspeicher Ninive



Der Langzeitspeicher Ninive besteht im Wesentlichen aus einem redundanten NAS-System (Network Attached Storage) der Firma NetWork Appliance. Die beiden Systemkomponenten mit je 9 TB Speicherkapazität stehen an zwei Standorten in Bern, welche rund 4,5 km voneinander entfernt sind. Ein automatisierter Datenabgleich zwischen diesen beiden Systemkomponenten sorgt dafür, dass die gespeicherten Daten an beiden Standorten vollständig vorhanden sind. Am Sekundärstandort wird zusätzliche über ein IBM-Bandlaufwerk eine dritte Kopie der Daten auf Magnetband erstellt. Diese dritte Kopie wird separat aufbewahrt.
Der Betrieb des Archivierungssystems Ninive wird durch das Bundesamt für Informatik und Telekommunikation (BIT) sichergestellt.

10.2 Datenablage

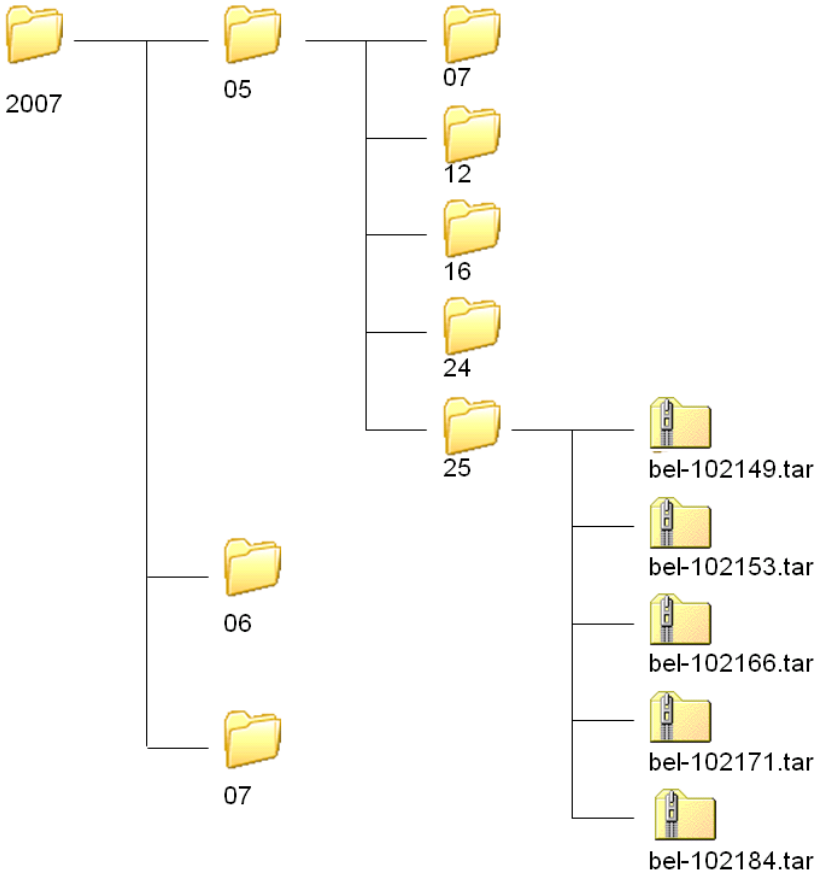
Damit die archivierten Daten einfach aufgefunden werden können, ist eine systematische Ablage nötig. Die Schweizerische Nationalbibliothek hat deshalb die Organisation der Datenablage festgelegt und Richtlinien bezüglich Verzeichnisstrukturen, Namensgebung und Struktur der Archival Informaton Packages (AIP) erstellt.

10.2.1 Verzeichnisstruktur

Die Ablage der Daten erfolgt in einer Verzeichnisstruktur, die mehrere Ebenen aufweist. Damit soll vermieden werden, dass sich zu viele Archivpakete im gleichen Verzeichnis befinden, weil das den Zugriff auf die Daten enorm verlangsamen kann.

Auf der obersten Ebene wird pro Jahr ein Verzeichnis angelegt. In der darunterliegenden Ebene gibt es pro Monat ein weiteres Verzeichnis. Auf der untersten Verzeichnisebene schliesslich, auf der die Archivpakete dann tatsächlich abgelegt werden, wird für jeden Tag, an dem Archivpakete ins Speichersystem eingespiesen werden ein separates Verzeichnis angelegt. Der Pfad im Archivsystem gibt

also gleichzeitig das Speicherdatum wieder. Im Verzeichnis 2007\07\15 befinden sich beispielsweise



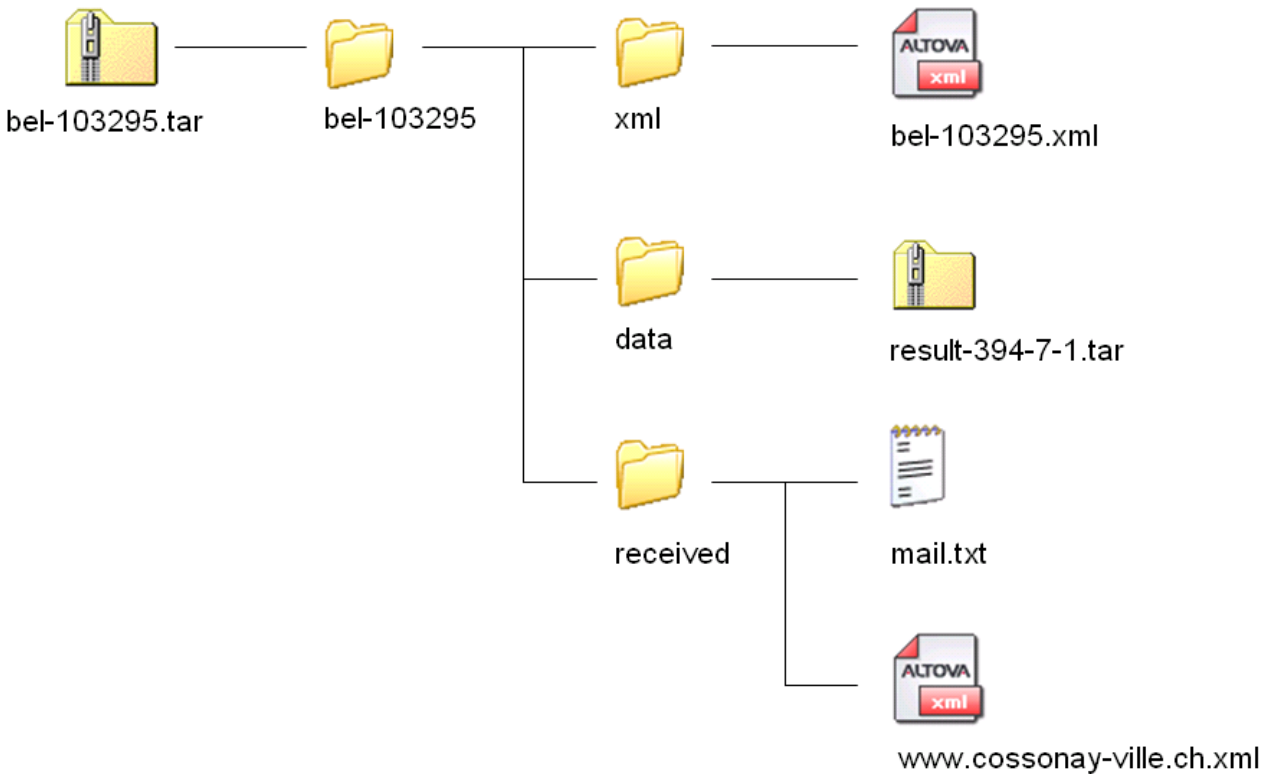
die Daten, welche am 15. Juli 2007 dem Speichersystem übergeben wurden.

10.2.2 Namensgebung

Jedes Archivpaket erhält einen Persistent Identifier in Form einer URN oder einer internen ID, wenn das AIP nicht im Internet zur Verfügung gestellt wird. Diese eindeutige interne ID wird gleichzeitig als Dateinamen für das AIP verwendet (Bsp. interne ID = bel-102149, Namen des AIP = bel-102149.tar).

10.2.3 Datenstruktur

Die AIPs werden als TAR-Files abgelegt. In diesen TAR-Files sind nicht nur die digitalen Objekte, sondern auch die Metadaten enthalten.

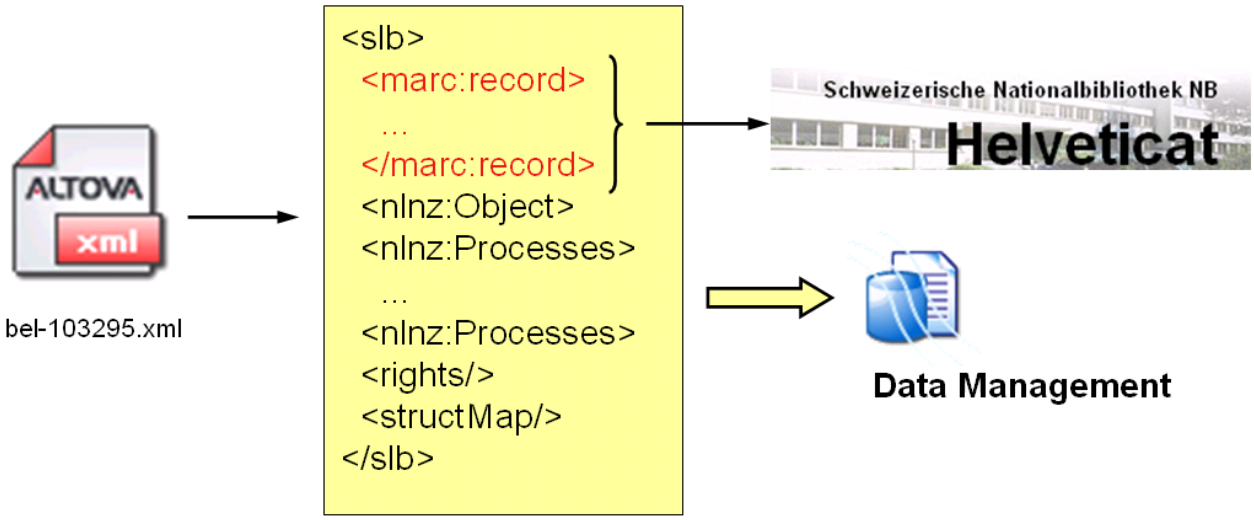


Das AIP wird immer als TAR-File abgelegt. Innerhalb des TAR-Files findet sich ein Ordner der nach der internen ID benannt ist. Damit sollte sichergestellt sein, dass beim Auspacken von TAR-Files nicht irrtümlicherweise Daten in das Verzeichnis eines anderen ausgepackten AIPs kopiert werden.

Innerhalb dieses Verzeichnisses befinden sich drei Unterverzeichnisse.

- Im Verzeichnis „received“ sind die von der abliefernden Stelle erhaltenen Metadaten zu finden. Im Falle von Webarchiv Schweiz sind dies die E-Mail (mail.txt) und das E-Mail-Attachment im XML-Format (www.cossonay-ville.ch.xml). Das Attachment trägt denselben Namen wie die angemeldete Website.
- Im Verzeichnis „xml“ sind die im Verlauf des Ingest-Prozesses ergänzten vollständigen Metadaten aus www.cossonay-ville.ch.xml in der internen Metadatenstruktur der Schweizerischen Nationalbibliothek zu finden. Das XML-File verwendet für den Namen wiederum die interne ID.
- Im Verzeichnis „data“ schliesslich ist ein TAR-File zu finden, das die gesamte Website mit all ihren Dokumenten und Verzeichnissen enthält. Mittelfristig wird das TAR-Format hier durch ein WARC-Format (Web ARChive) abgelöst, das für die Aufbewahrung von Websites neu entwickelt und nun zum ISO-Standard erhoben werden soll.

10.2.4 Ablage der Metadaten



Die Metadaten werden neben der Ablage direkt im AIP für den späteren Zugriff auf das AIP noch an zwei weiteren Orten verzeichnet.

- Im Data Management, einer Datenbank die für den Ingest-Prozess aber später auch für den Zugriff auf die Daten verwendet wird, sind die vollständigen im internen Format der Schweizerischen Nationalbibliothek vorliegenden Metadaten zu finden.
- In Helveticat, dem Bibliothekskatalog der Schweizerischen Nationalbibliothek, werden die bibliographischen Metadaten ebenfalls verzeichnet. Damit ist gewährleistet, dass über diesen Katalog auf Informationen über sämtliche in der Nationalbibliothek archivierten Dokumente zugegriffen werden kann.