Zeitschrift: Geschichte und Informatik = Histoire et informatique
Herausgeber: Verein Geschichte und Informatik
Band: 13-14 (2002-2003)

Rubrik: Projekte = Projets

Le projet SIRANAU: Système Intégré Radio pour l'Archivage Numérique Audio

Ralf Dahler[1]

Zusammenfassung

Das Radio Suisse Romande (RSR) hat 2001 beschlossen, ein Spezialsystem zur Aufbewahrung und Verwaltung ihrer Audio- und Multimedia-Produktionen einzuführen. Mit dem System SIRANAU soll dieses Ziel erreicht werden. Seine wichtigsten Funktionalitäten sind:
- Übernahme von zur Archivierung vorgesehenen digitalen Audio- oder Multimediadokumenten inkl. Metadaten aus dem Produktionssystem
- Verwaltung des Archivierungs-Workflows
- Ausführliche Katalogisierung aller akzeptierten Dokumententypen
- Dokumentationsrecherche aus einer einzigen Quelle
- Abfragen digitaler Dokumente von den RSR-Arbeitsplätzen aus
- Möglichkeit, Audiodokumente wieder ins Produktionssystem zurückzuladen.

Résumé

La Radio Suisse Romande (RSR) a décidé en 2001 de se doter d'un système spécifique pour stocker et gérer ses productions audio et multimédia. C'est cet objectif que vise SIRANAU, dont les principales fonctionnalités sont:
- Prise en charge, à partir des systèmes de production, des documents numériques audio ou multimédia «candidats à l'archivage», accompagnés de leurs métadonnées;
- Gestion du workflow de l'archivage;
- Catalogage détaillé de tous les types de documents acceptés;
- Recherche documentaire dans une source unique;
- Consultation des documents numériques à partir des postes bureautiques de la RSR;
- Commande de rechargement des documents audio (réutilisation en production).

1 Les notes ont été rajoutées par la rédaction.

La Radio Suisse Romande (RSR) a décidé en 2001 de se doter d'un système spécifique pour stocker et gérer ses productions audio et multimédia.[2]

Entre 1993 et 2000, l'ensemble des moyens de production traditionnels ont été remplacés par des outils de production numérique. Au sein de chaque chaîne et de l'Information, les sons sont stockés dans des serveurs informatiques et peuvent être écoutés, édités et diffusés à partir de stations de travail. L'archivage de ces productions se fait aujourd'hui en gravant des CD enregistrables (CD-R), ou en transférant les fichiers sur des disques magnéto-optiques (MOD), gérés manuellement. Ces solutions sont lentes et peu fiables. Une rationalisation de la chaîne de production passe par l'implantation d'un outil destiné au stockage à long terme et à la réutilisation des contenus audio, sans manipulation de supports physiques.

	type de support	nombre	heures de son
1935 – 1951	disques	85'000	1'300
1951 – 2002	bandes magnétiques	170'000	85'000
1992 – 2002	CDR	7'000	8'000
1998 – 2002	MOD	350	3'500

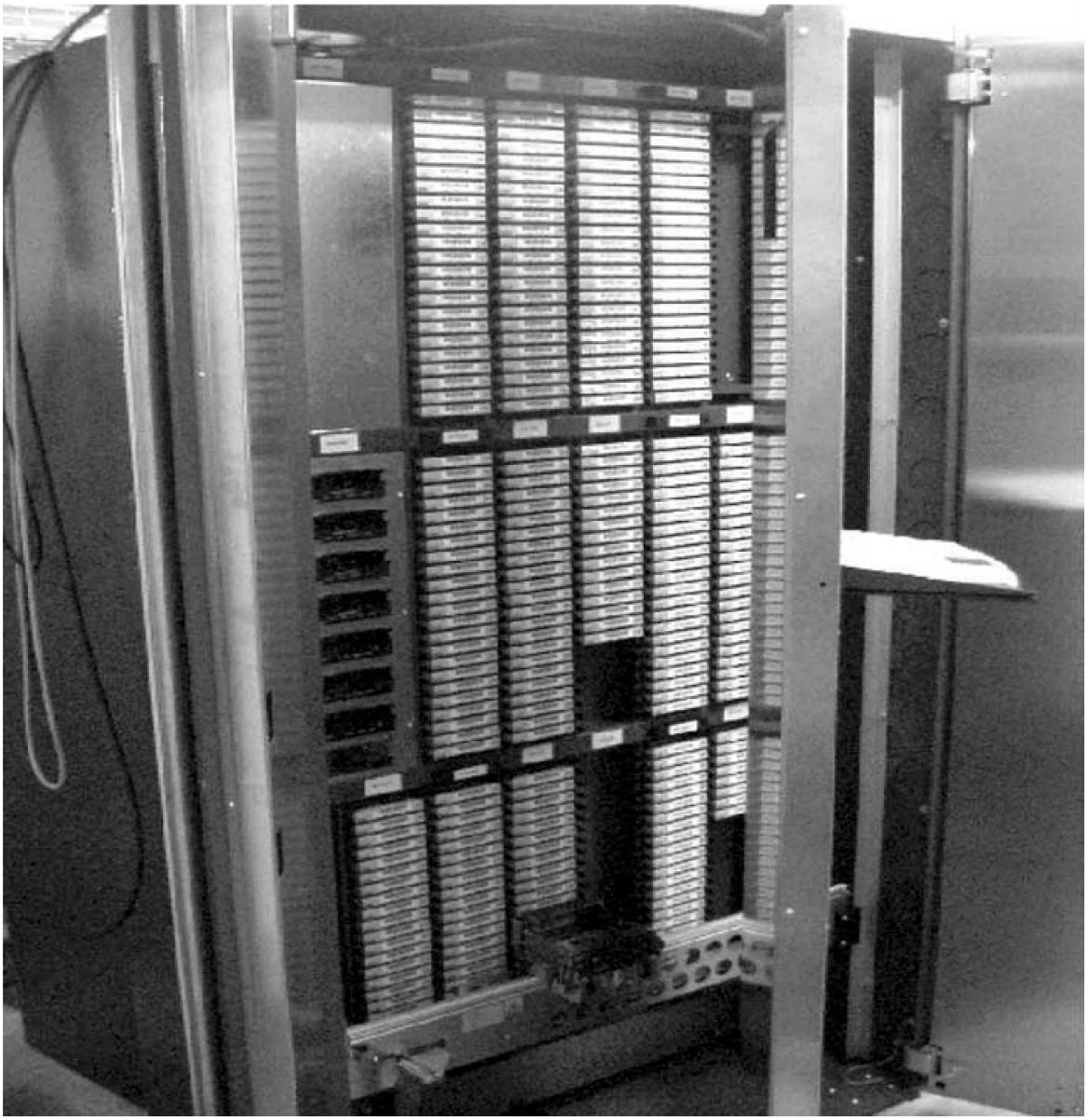
Tableau 1: Situation avant l'introduction de SIRANAU (sans 2003)

C'est cet objectif que vise SIRANAU, qui a été conçu et développé à partir de fin 2001. Ses principales fonctionnalités sont les suivantes:

- Prise en charge, à partir des systèmes de production, des documents numériques audio ou multimédia «candidats à l'archivage», accompagnés de leurs métadonnées.
- Gestion du workflow de l'archivage (identification, acceptation ou refus des candidats, compléments d'information, transmission au catalogage) avec la possibilité de consulter les contenus: son, image, etc.
- Catalogage détaillé de tous les types de documents acceptés (sons parlés ou musicaux, bruitages, images fixes, vidéos, textes)
- Recherche documentaire, au moyen de nombreux critères détaillés, sur l'ensemble des documents stockés, ainsi que sur les supports physiques déjà à disposition (par import du contenu des anciennes bases de données Basis): SIRANAU doit être la source unique pour toutes les recherches documentaires.

2 En Suisse alémanique un projet analogue est en cours. Il s'agit du projet *Digitale Archiv-Speicherung* (DAS). Le système devrait être opérationnel dès mi 2004. Dans les 10 ans à venir, un quart des archives sonores de la Radio Suisse alémanique (SR DRS), soit 50'000 contributions, sera numérisé et intégré au système. Le projet est soutenu par l'association MEMORIAV.

– Consultation des documents numériques, en basse résolution, à partir de l'ensemble des postes bureautiques de la RSR.
– Commande de rechargement des documents audio pour réutilisation dans les systèmes de production.



Ill. 1: Les documents stockés sur casettes sont rapidement accessibles grâce a un robot (accès «near line»).

En ce qui concerne la production audio, SIRANAU devra pouvoir archiver, chaque année, environ 4'500 heures de son. Les fichiers «sources», issus des systèmes de production, à 256 kbit/s., sont accompagnés d'un fichier

XML contenant les métadonnées. Ils sont stockés tels quels pour réutilisation directe, et déclinés en format RealAudio (pour préécoute) et en format BWF linéaire (pour stockage sécurisé à long terme).[3] Un format type MP3 pour la distribution Internet pourra venir s'y ajouter par la suite.

La solution technique s'articule autour des composants suivants:
- Divers serveurs Compaq sous Windows 2000 Server,
- Baie de disques durs Clariion de EMC^2 (1 TB en première étape),
- Robot de stockage Infinistore de Grau, avec cassettes Sony AIT2 (10 TB en première étape)
- Base de données Oracle,
- Application documentaire Thesaurus Rex 2 de Question d'Image,
- Modules d'interfaçage avec les systèmes de production développés par Dalet,
- Systèmes clients sur les postes des utilisateurs professionnels (documentalistes) développés par Cap Gemini Ernst & Young
- Interface Web de recherche et consultation pour l'ensemble des collaborateurs RSR.

L'infrastructure réseau constitue le noyau d'un SAN (storage area network), qui pourra être étendu par la suite pour accueillir le stockage des divers systèmes de production.

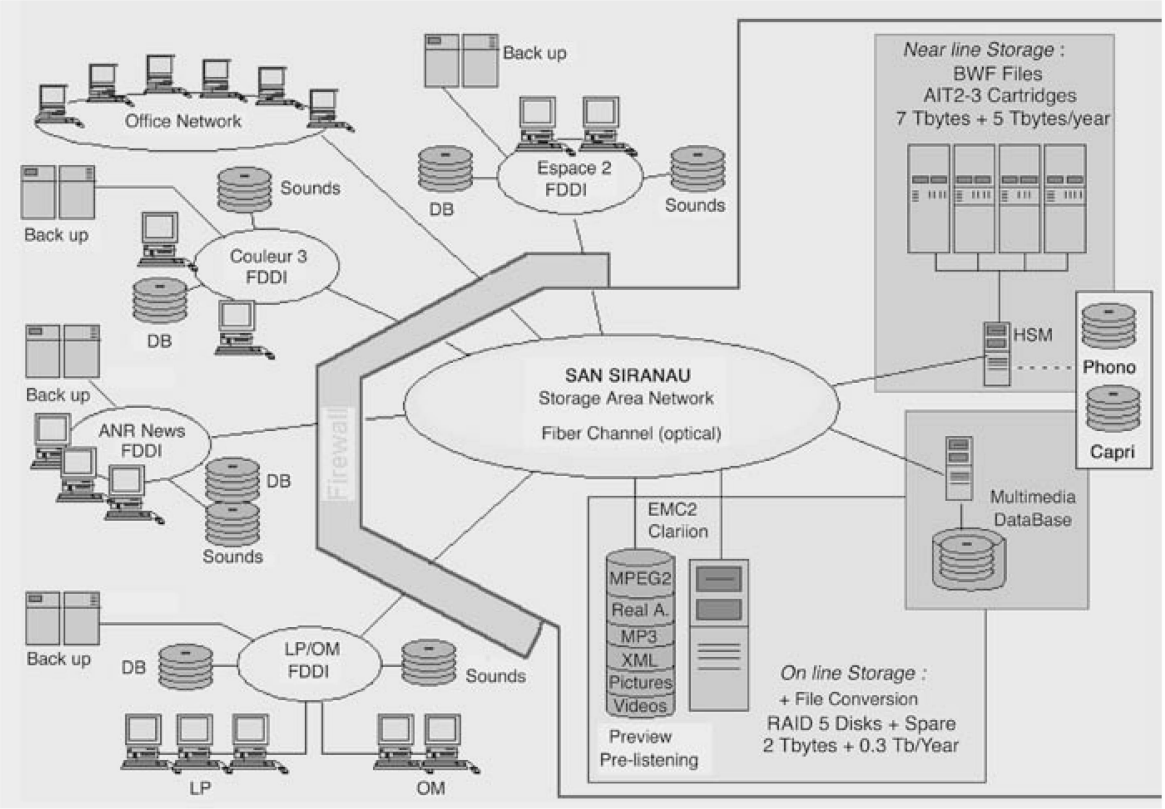
Pour l'intégration générale du système SIRANAU, la RSR a choisi la société Cap Gemini Ernst & Young, qui avait déjà procédé à une étude d'architecture générale des systèmes de bases de données.

Dans une phase ultérieure, SIRANAU se substituera complètement aux bases de données documentaires *Basis* actuelles, et deviendra l'outil de recherche et de gestion central pour tous les contenus, sur supports physiques aussi bien que numériques. Il pourra également s'ouvrir vers Internet, afin de délivrer sélectivement certains documents numériques, soit gratuitement en basse résolution, soit contre paiement. Un accès préférentiel à des institutions partenaires intéressées par la conservation et la mise en valeur du patrimoine audiovisuel est également prévu.

---

3 Le Broadcast Wave Format (BWF) est un standard de l'Union Européenne de Radio-Télévision (UER) basé sur le format Wave de Microsoft. Chalmers, R[ichard]: «The Broadcast Wave Format – an introduction». In: *EBU Technical Review* No. 274, Winter 1997 <http://www.ebu.ch/trev_274-chalmers.pdf> EBU Technical Recommendation R97-1999 <http://www.ebu.ch/tech_texts/tech_text_r97-1999.pdf>

Outre son rôle dans une gestion plus rationnelle des contenus sonores et multimédias, SIRANAU permettra de créer un véritable outil d'*Assets Management,* notamment en incluant toutes les données sur les droits qui leur sont attachés.



Ill. 2: Vue d'ensemble du système et de ses liaisons.
AIT: Advanced Antelligent Tape. Casettes produites par Sony, capacité entre 36 (AIT-2) et 100 (AIT-2) GB
ANR: Audio News Releases.
DB: Database. Base de données.
FDDI: Fiber Distributed Data Interface. Standard de réseau local à haut débit (100 Mbps) utilisant la fibre optique.
HSM: Hierarchical Storage Management. Logiciel de gestion hiérarchique du stockage.

Enfin, comme outil de conservation numérique et de mise en valeur, SIRANAU va pouvoir accueillir les documents audio issus de la numérisation des très riches archives de la Radio Suisse Romande.[4] Cette opération est d'ores et déjà entreprise avec le soutien de MEMORIAV (Association

4 Cirio, Yves: «Quadriga: un nouvel outil de numérisation des archives sonores». In: *Les Inouïs*, Bulletin no. 45, Sept. 2002, <www.memoriav.ch/fr/home/son/pdf/Inouis.pdf>

pour la sauvegarde du patrimoine audiovisuel suisse)[5]. Il contribuera donc à garantir la conservation des contenus – aujourd'hui tributaires de supports menacés de dégradation – et la mise en valeur de documents irremplaçables.

5 Cosandier, Jean-François: «SIRANAU: archiver le son à l'ère numérique». In: *bulletin MEMORIAV*, 3/1999, p. 11-12 (deutsch S. 7) <http://www.memoriav.ch/fr/home/memoriav/services/pdf/Memoriav_3_1998.pdf>
Description du projet VOCS et bibliographie: «Voix de la culture suisse». <http://www.memoriav.ch/de/home/son/projets/d-proj-vocs.htm>

ARCHIMED: une base de données historique?

Jean-Daniel Zeller

Zusammenfassung

Seit 1995 bilden die Genfer Universitätsspitäler einen Spitalverbund, der jährlich 45'000 Aufnahmen bewältigt. Ab 1976 entwickelte die Spitalinformatik ein patientenorientiertes System. Der Artikel beschreibt die Fortentwicklung des Systems und die Verwirklichung der integrierten Datenbank ARCHIMED. Das Schlüsselkonzept ist das Umschreiben von an Patienten gebundenen Daten in «Elementarfakten», welche anschliessend eine äusserst leistungsfähige statistische Auswertung ermöglichen. Konkret werden täglich vorselektierte Daten aus den verschiedenen operationellen Informatiksystemen in eine Applikation gesandt, welche diese zu Elementarfakten verarbeitet und danach in die Datenbank ARCHIMED leitet.

Résumé

Les Hôpitaux universitaires de Genève forment depuis 1995 un réseau d'établissements hospitaliers offrant annuellement 45'000 hospitalisations. Dès 1976, le centre d'informatique hospitalière a développé un système orienté patient. L'article décrit l'évolution du système et la réalisation de la base de données intégrée ARCHIMED. Le concept clé de cette application réside dans la transcription des données liées aux patients en une série de «faits élémentaires», qui peuvent être manipulés ultérieurement à des fins statistiques de manière extrêmement performante. Pratiquement, les données préalablement sélectionnées des différents systèmes informatiques opérationnels sont déversées quotidiennement dans une application qui les transforme en faits élémentaires puis les déverse dans la base de données ARCHIMED.

1. Introduction

A Genève, l'informatique médicale a développé depuis plus de 20 ans un système «orienté patient». Une application a été développée à des fins d'exploitation statistique des données ainsi rassemblées. La base de données ARCHIMED regroupant les variables exploitables annualisées doit-elle être considérée comme base de données historique et à ce titre digne d'archivage à long terme?

2. Contexte institutionnel et informatique

Les Hôpitaux universitaires de Genève forment depuis 1995 un réseau d'établissements hospitaliers répartis sur 5 sites géographiques, qui emploie près de 7900 collaborateurs, compte 2185 lits et totalise annuellement environ 175'000 hospitalisations (HC: 52'000) et 620'000 consultations ambulatoires (chiffres de l'année 2001). Il est issu du regroupement des établissements suivants:

Etablissement	*Créé en**	*Type de soins*	*Lits*
Hôpital cantonal (HC)	1856	Soins aigus (i.c. pédiatrie et maternité)	1176
Hôpital psychiatrique	1900	Psychiatrie	343
Hôpital de Loëx	1900	Soins de longue durée	268
Hôpital gériatrique	1970	Gériatrie	293
Centre de soins continus	1980	Soins palliatifs	105
* sur leur site actuel			2185

Héritage historique, cette répartition tend à être remplacée par une organisation structurée en fonction des spécialités médicales, regroupées actuellement en 11 départements correspondant également aux activités académiques que se doit d'assumer un hôpital universitaire.

L'hôpital cantonal de Genève a été un des pionniers de l'informatique médicale en Suisse voire en Europe. Dès 1976, l'équipe d'informaticiens-médecins, emmenée par le professeur Jean-Raoul Scherrer, concevait le système DIOGENE qui allait connaître de nombreux développements, dont les principales étapes sont résumées ci-dessous (applications patients uniquement):

Année	*Application*	*Remarques*
1978	Démarrage de l'application hospitalière DIOGENE	Au départ, uniquement pour les patients hospitaliers
1982	Intégration des policliniques dans le système DIOGENE	Bases de données en parallèle. L'intégration sera achevée seulement en 1993.
1988-98	Déploiement progressif des applications UNI-LAB	Gestion des examens de laboratoire
1994	Mise en application d'UNI-IMAGE	Gestion des examens d'imagerie médicale
1993-96	Mise en application d'UNI-DOC	Production de documents à partir des données DIOGENE
1995	Migration de DIOGENE sur des machines et applications standards. Les stations de travail sont remplacé par des PC	Restructuration des bases de données. Changement d'OS et de DBMS
1998	Création du concept DOMED (médical), DOSSI (soins infirmier) puis DPI (Dossier du Patient Intégré)	Navigateur médical permettant l'accès aux données/documents des patients issus de diverses applications
2003	Début de l'analyse de la fusion des applications de gestion administratives des patients (DIOGENE+PHILOS)	Intégration des systèmes de l'hôpital cantonal avec ceux des autres sites hospitaliers

Actuellement, la base de données patients de DIOGENE rassemble des données pour un peu plus d'un million de patients, avec un accroissement annuel de l'ordre de 45'000 nouveaux patients.

3. Problématique …

Dès le départ, le système DIOGENE a été conçu comme un système «orienté patient» et non pas comme un système de gestion administratif. Cela est particulièrement bien illustré par la représentation constamment affirmée de la «roue» DIOGENE (voir figure 1).

Cela apparaît également dans la structure fonctionnelle des applications représentée dans le schéma des tables de données de DIOGENE dont le point d'entrée est également le patient (voir figure 2).

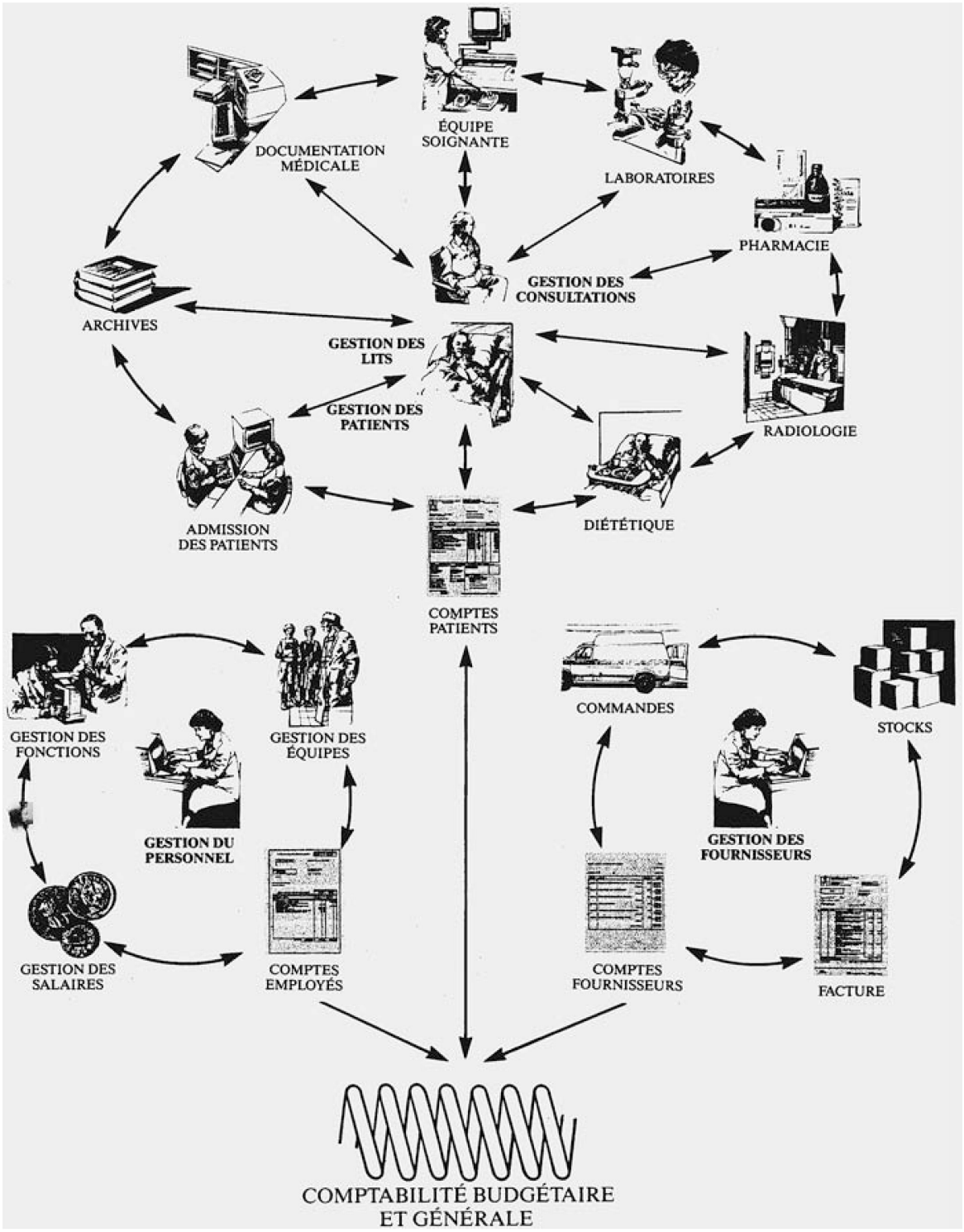


Figure 1: Structure globale des applications informatiques des HUG

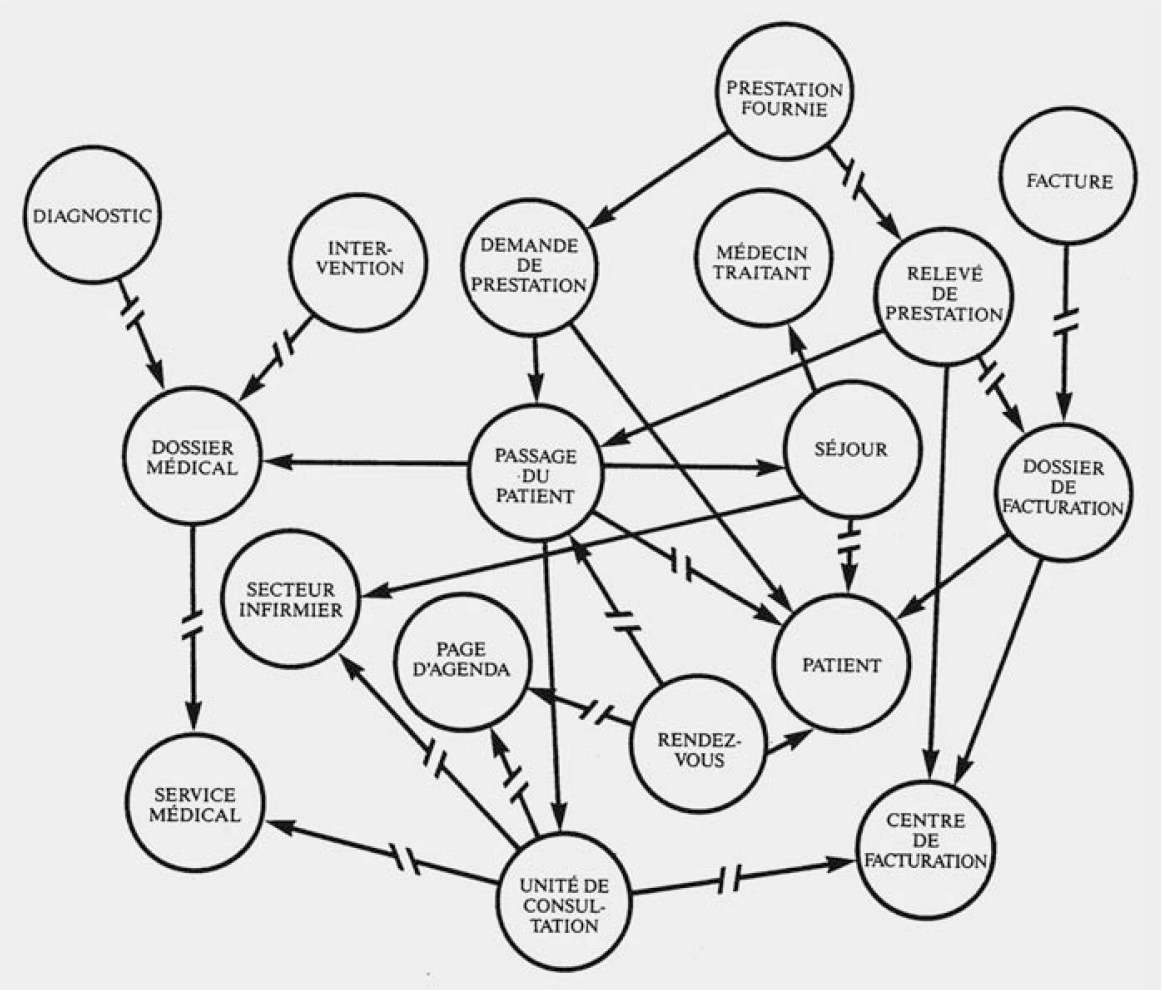


Figure 2: Schéma des relations des données dans le système DIOGENE; les flèches simples indiquent une relation fonctionnelle, les flèche barrées indiquent une application.

Cependant, une telle base de données ne pouvait pas ne pas être exploitée du point de vue statistique, tant pour des interrogations de type médical (charge des services, épidémiologie) qu'administratif (statistiques annuelles, coût par patient, etc.).

Durant les premières années d'utilisation de DIOGENE, ce type d'interrogation s'effectuait en mode batch, durant les heures creuses de l'exploitation, entre 22 heures et 4 heures du matin. Cependant, ce type d'exploitation s'est progressivement heurté à plusieurs obstacles:

a) L'accroissement continuel de la base et les besoins de sécurité élevés nécessitaient de plus en plus d'opérations de maintenance devant s'effectuer pendant ces mêmes heures creuses;

b) L'accroissement de la base rendait également les analyses statistiques de plus en plus intéressantes pour les gestionnaires, qui prirent l'habitude d'en solliciter de plus en plus;
c) L'organisation hiérarchique de la base ne facilitait pas le lancement de requêtes statistiques, dont les critères n'étaient pas directement liés au patients mais plutôt à des groupes d'événements, qui pouvaient se situer à des niveaux hiérarchiques très divers dans la base et nécessiter la constitution de fichiers temporaires très volumineux, perturbant d'autant le fonctionnement opérationnel courant de la base de données.

Pour toutes ces raisons, le responsable de ces éditions statistiques envisagea dès les années 90 la possibilité de créer une base de données normalisée unique, qui permettrait:

a) Le lancement de requêtes statistiques sophistiquées et paramétrables sans mobiliser les ressources de la base de données opérationnelle.
b) La possibilité d'interroger une base de données anonymisée, directement par les utilisateurs finaux au moyen de navigateurs (en limitant la réalisation à la demande de requêtes SQL par les informaticiens).
c) La possibilité d'intégrer des données relatives aux patients issues de systèmes d'information divers. Cette option allait s'avérer cruciale quelques années plus tard lorsqu'il s'est agit de préparer la fusion des HUG, dont les établissements possédaient des systèmes informatiques divers et non intégrés.

Ce projet vit le jour en 1993 par la réalisation de la base de données intégrée ARCHIMED. Le concept clé de cette application réside dans la transcription des données liées aux patients en une série de «faits élémentaires» qui peuvent être manipulés ultérieurement de manière extrêmement performante. Avec l'ouverture du service intranet à l'Hôpital cantonal en 1995, cette version «dynamique» des atlas statistiques a remplacé définitivement la version papier produite entre 1983 et 1993.

4. Le concept d'ARCHIMED[1]

Si le concept se rapproche des notions actuellement bien connues des «entrepôts de données» (Data Warehouses), il s'en distingue dans la mesu-

1 On trouvera une description historique de ces travaux dans G. Thurler, F. Borst, C. Bréant, D. Campi, J. Jenc, B. Lehner-Godinho, P. Maricot, J.R. Scherrer: «ARCHIMED: A Network of Integrated Information systems». In: *Methods on Information in Medecine*, vol. 39, no 1, 2000, pp. 36-43.

re ou les systèmes de «data warehousing» s'occupent plus d'offrir une présentation comparable des données sélectionnées issues de différents système opérationnels en vue de leur consultation par les gestionnaires, tandis qu'ARCHIMED effectue de surcroît un travail d'harmonisation et d'homogénéisation des données, aux trois niveaux syntaxique, sémantique et ontologique (voir plus bas) afin de constituer une base de données intégrée. Cette base de données constitue le noyau d'un système complété par une série d'outils d'aide à la décision.

Pratiquement, les données préalablement sélectionnées des différents systèmes informatiques opérationnels sont filtrées quotidiennement par une série d'interfaces (8 à ce jour), qui les transforment en faits élémentaires puis les déversent dans la base de données ARCHIMED, sur une base annualisée. L'annualisation a été imposée au départ par des contraintes techniques de place mémoire et de gestion des applications. Avec les moyens actuels elle ne serait plus nécessaire et les tables annuelles pourraient être regroupées en une seule grande table. Notons, du point de vue de l'archiviste, que cette organisation, si elle implique certaines contraintes lorsque l'on effectue des requêtes sur des données pluri-annuelles, représente en fait un mode de stockage intéressant sur le long terme. En effet, il permet d'une part une maîtrise prévisible des volumes de données et d'autre part il offre une rupture temporelle claire et explicite pour le cas où l'on devrait mettre certaines données off-line.

4.1. Le modèle ontologique d'ARCHIMED

Le modèle ontologique conçu pour construire l'application repose sur une structure de domaines se superposant en couches d'abstraction successives, telle qu'illustrée sur le schéma ci-dessous (figure 3).

Ce schéma montre que s'il s'agit bien sur le plan technique d'extraire des données des bases de données opérationnelles, la cohérence de ces données doit être assurée en amont; d'une part en fonction des définitions adoptées lors de la constitution des applications opérationnelles et d'autre part en fonction des «découpages» des unités opérationnelles, définies indépendamment des applications informatiques mais dont celles-ci doivent tenir compte. Il existe par ailleurs d'autres découpages liés à la gestion budgétaire ou à la localisation physique par exemple.

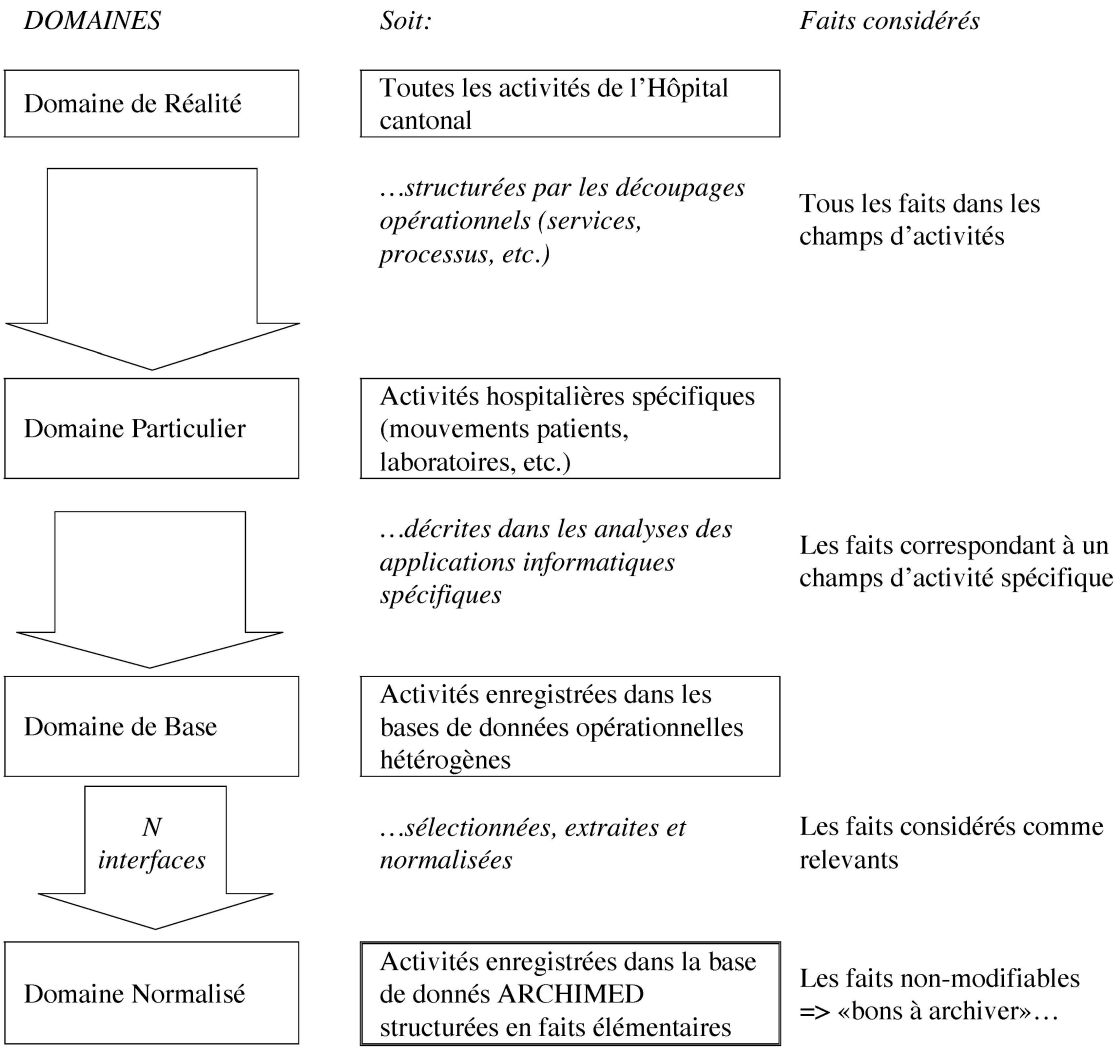


Figure 3: modèle ontologique d'ARCHIMED

A partir de ce modèle, sont définis des *liens de base* qui ont été déduits en examinant les relations ou les similarités entre les faits contenus dans les différents domaines de base (par ex: patient, service médical, centre budgétaire, etc.)

A ces liens de base sont attachés des *propriétés*, caractérisées par un type, une valeur, un temps. La connexion entre un lien de base et une propriétés forme un *fait élémentaire*. On a alors l'équation suivante:

Lien de base + propriété = fait élémentaire

A ce niveau, il faut noter que chaque fait élémentaire est estampillé d'un temps. Celui-ci peut être un temps ponctuel, quand le fait n'a lieu qu'une

fois (date d'examen par ex.), ou une durée avec ses deux limites de début et de fin (durée de séjour par ex.). Celle-ci est alors représentée par un processus avec un fait de début et un fait de fin, ce qui formalise la notion de trajectoire hospitalière, mais permet, plus généralement, de définir n'importe quelle chronologie de manière stricte.

Il faut également remarquer que les faits ne sont sélectionnés que lorsqu'ils ne sont plus modifiables. Cela correspond partiellement à la notion de clôture de dossier en archivistique classique, avec une nuance de taille cependant: c'est le «fait» qui est clos et cette clôture s'effectue indépendamment d'une notion de dossier, bien que celle-ci existe par ailleurs dans certains domaines de base. Ceci est rendu possible sans perte d'information et de contexte parce que chacun des faits élémentaires est d'une part daté pour lui-même et d'autre part «lié» à un domaine de base.

4.2. Pratique quotidienne

Dans la pratique, une fois les domaines de base définis, des routines d'extraction sont appliquées aux différentes bases opérationnelles hétérogènes, en général sur un rythme quotidien pendant les heures creuses nocturnes. Ces tables, issues pour la plupart de bases INGRESS, sont extraites en fichiers ASCII et envoyées par protocole FTP vers l'application ARCHIMED. Elles sont alors distribuées dans des «paniers», en fonction de leur domaine de provenance. Ces paniers sont traités pour transformer les données en faits élémentaires normalisés, qui sont déversés dans les tables de la base de données ARCHIMED. Par ailleurs, un traitement parallèle sur les faits élémentaires permet de produire des «données réduites» qui sont en fait des indicateurs dérivés. (voir schéma de la constitution de la base de données ARCHIMED en fin d'article, figure 4).

4.3. L'accès aux données ARCHIMED

Dans un premier temps, les gestionnaires d'ARCHIMED ont développé des requêtes d'interrogation similaires à celles préalablement utilisées pour produire les statistiques à partir des applications opérationnelles. Celles-ci étaient éditées sous la forme d'atlas statistiques annuels ou mensuels, ainsi que d'autres atlas particuliers en fonction des demandes. Ces éditions ont été remplacées par un navigateur disponible sur l'Intranet des HUG, qui permet aux collaborateurs de consulter ces données communes de manière rétrospective (ce qui était difficile avec les atlas publiés sur papier) et avec une actualisation quotidienne (ce qui était impossible avec des éditions annuelles et mensuelles). Ces statistiques d'activités sont affichables à

différentes échelles temporelles (années, mois, jour en général) et à différents niveaux d'agrégation, des HUG en entier jusqu'à l'unité de soins. Les tableaux sont exportables vers un tableur Excel pour un traitement localisé.

La véritable valeur de la base de données intégrée d'ARCHIMED n'apparaît cependant qu'avec les différents outils d'aide à la décision développés ultérieurement par l'Unité d'information médico-économique (UIME), qui permettent diverses interrogations sophistiquées.

La plupart se présentent sous forme de navigateur, nécessitant un droit d'accès. Les outils disponibles sont, par exemple:
- le calcul d'indicateurs (DRG: diagnosis related groups, ré-admissions);
- les archives des faits des patients (recherche de cas similaires, recherche des faits d'un patients, données de laboratoires);
- les statistiques médicales (code diagnostic, interventions, statistique selon normes OFS, etc.);
- les services Archimed (analyse des mouvements d'urgence, calcul des scores de gravité, suivi des patients sur plusieurs années, etc.).[2]

5. Etat actuel de la base ARCHIMED

A fin 1998 la base de données ARCHIMED était constituée de:
- 5 MB de données (faits) incorporés quotidiennement dans la Base de Données Intégrée (BDI);
- des données distribuées dans 750 tables, incluant 50 millions d'enregistrements pour un total de 8 GB.
- des données couvrant les activités hospitalières depuis 1990, la plupart des activités de laboratoire depuis 1993 et les activités ambulatoires depuis 1996.

Après dix ans d'activité (1993-2003), les évolutions annuelles sont les suivantes:
- accroissement d'environ 1,5 Gigabytes par an (5MB/j * 365j = 1,8 GB);
- 70 millions d'enregistrements (faits élémentaires);
- ajout de 200 tables annuelles.

---

2 On trouvera des descriptions plus détaillées de ces outils dans les articles suivants: Lehner B., Thurler G., Bréant C:, Tahintzi P. Borst F: «Retrieval of Similar Cases using the ARCHIMED Navigator», In: MIE 2003 (Medical Informatics Europe 2003, St-Malo), et Thurler G., Bréant C., Lehner B., Bunge M., Samii K., Hochstrasser D., Nendaz M., Gaspoz J.M., Tahintzi P., Borst F.: «Toward a Systemic Approch to Disease», In: *ComPlexUs*, 2003.

6. Une base de données historique(s)?

Comme expliqué plus haut, le concept qui a présidé à la naissance d'ARCHIMED a été d'ordre médical et opérationnel. Cependant, la nécessité d'harmoniser les données provenant de plusieurs systèmes opérationnels différents a forcé ses concepteurs à une réflexion ontologique qui a mené à une structuration qui rencontre les préoccupations d'une conservation des données à long terme.

Actuellement, ARCHIMED représente non seulement un outils d'aide à la décision, raison pour laquelle il a été conçu, mais également une source de données historiques sans égale, car les données consolidées dans cette base unique ne pourraient être rassemblées autrement, chaque système opérationnel (et il y en a environs une centaine au sein des HUG) ayant sa propre structure et ses propres dictionnaires.

Cependant, ARCHIMED n'est pas une base de données historique «idéale» pour les raisons suivantes.

a) Un outil plutôt qu'un système:
ARCHIMED a été initialement conçu comme un outil technique, dont l'objectif était de simplifier l'accès à des données provenant de bases de données hétérogènes. Bien que ses concepteurs aient très rapidement appréhendés ses possibilités d'outil d'aide à la décision, ce n'est qu'après la migration des systèmes informatiques en 1995 et la mise en place d'un plan d'investissement de grande envergure en 1998 qu'ARCHIMED apparaît pour la première fois comme une application identifiée au sein de la division informatique et de l'institution.

b) Un manque de vision institutionnelle
Conséquence du point précédent, la hiérarchie tant informatique qu'administrative n'a réalisé l'intérêt de ce système d'information que lorsqu'il a été rendu visible sur l'intranet institutionnel. N'ayant pas participé directement à sa conception, les décideurs ont mis longtemps à reconnaître sa valeur.

c) Une normalisation à posteriori
Bien que certains dictionnaires de données soient communs à toutes les applications opérationnelles, comme les découpages, certaines harmonisations ne sont effectuées que lors du transfert des données dans ARCHIMED. Il manque à l'institution une instance qui définirait certains référen-

tiels communs de manière univoque et uniforme, permettant une normalisation en amont.

d) Une absence de politique de conservation à long terme
Le système ARCHIMED représente surtout aux yeux de ses concepteurs et de ses utilisateurs un outil d'aide à la décision basé sur des périodes longues (10 ans) mais qui restent relativement courtes en terme d'archivistique. Jusqu'à présent, la place mémoire n'ayant pas fait défaut, la question de la conservation à long terme ne s'est pas posée, l'intérêt étant d'offrir en ligne le plus de données possible. Ce contexte est également valable pour les données des bases opérationnelles. La nature même de l'activité hospitalière portant sur la totalité de la vie des patients, cette tendance restera une constante. De plus en plus de données se trouvant nativement dans les systèmes informatiques des HUG, nous entamons seulement maintenant, secteur par secteur, des discussions sur la conservation ou la non-conservation à long terme de ces données.

Nonobstant ces réserves, ARCHIMED représente un intérêt historique considérable, pour les raisons suivantes:

a) Une validation préalable des données
L'analyse ontologique et sémantique des données déversées dans ARCHIMED implique que les données ainsi conservées ont été considérées comme pertinentes à la base. Ceci évitera un travail d'évaluation supplémentaire lors d'un futur archivage historique. On devra cependant veiller à ce que ces critères de validation soient explicitement documentés.

b) Une structure simple et documentée
La gestion des données en faits élémentaires et en tables annuelles rendent leur manipulation très indépendante des logiciels de gestion de base de données. Chaque type de relation étant documenté dans des dictionnaires, il ne serait par exemple pas difficile de conserver ces tables sous forme XML. Dans la perspective d'une conservation à très long terme, on doit cependant se poser la question de la conservation des navigateurs, qui donnent une image de l'usage actuel de ces données, et la possibilité de construire à long terme d'autres navigateurs, répondant à des questions d'ordre historique plutôt que médico-économique. Les HUG ne se sont pas encore prononcés à cet égard.

c) Une structure indépendante du temps
Comme chaque fait élémentaire est daté et qu'il n'est transféré dans la base de donnée intégrée qu'une fois qu'il n'est plus susceptible de changement, on évite un problème récurent dans les entrepôts de données courants, qui est celui de la mise à jour des données. De ce fait, la base est parfaitement cohérente dans le temps. La structuration en tables annuelles permet potentiellement une mise off-line par tranche chronologique sans aucune manipulation supplémentaire (les trajectoires de soins qui «passent» d'une année sur l'autre sont signalées dans les tables par un drapeau, ce qui permet leur identification et la concaténation des données entre les tables annuelles).

Conclusion

Dans un article de 2001, Edward Atkinson défend la proposition selon laquelle les «Data Warehouses» - ARCHIMED peut y être assimilé - sont des «records» et à ce titre dignes d'être conservés.[3] La justification de sa position n'est pas très étayée et l'exemple de la base de données ARCHIMED peut fournir au moins un argument de taille: si les entrepôts de données sont strictement documentés chronologiquement au niveau du fait élémentaire, ils représentent une source historique de première qualité.

En conclusion on peut affirmer que les entrepôts de données sont des records (historiques) si:
– ils sont strictement documentés (ontologies)
– ils sont strictement délimités (faits élémentaires)
– ils sont strictement datés (attributs temporels)
– ils permettent la construction de nouveaux critères de navigation

3 Atkinson, Edward: «Data warehousing - a boat records managers should not miss». In: *Records Management Journal*, vol. 11, no. 1, avril 2001, pp 35-43.

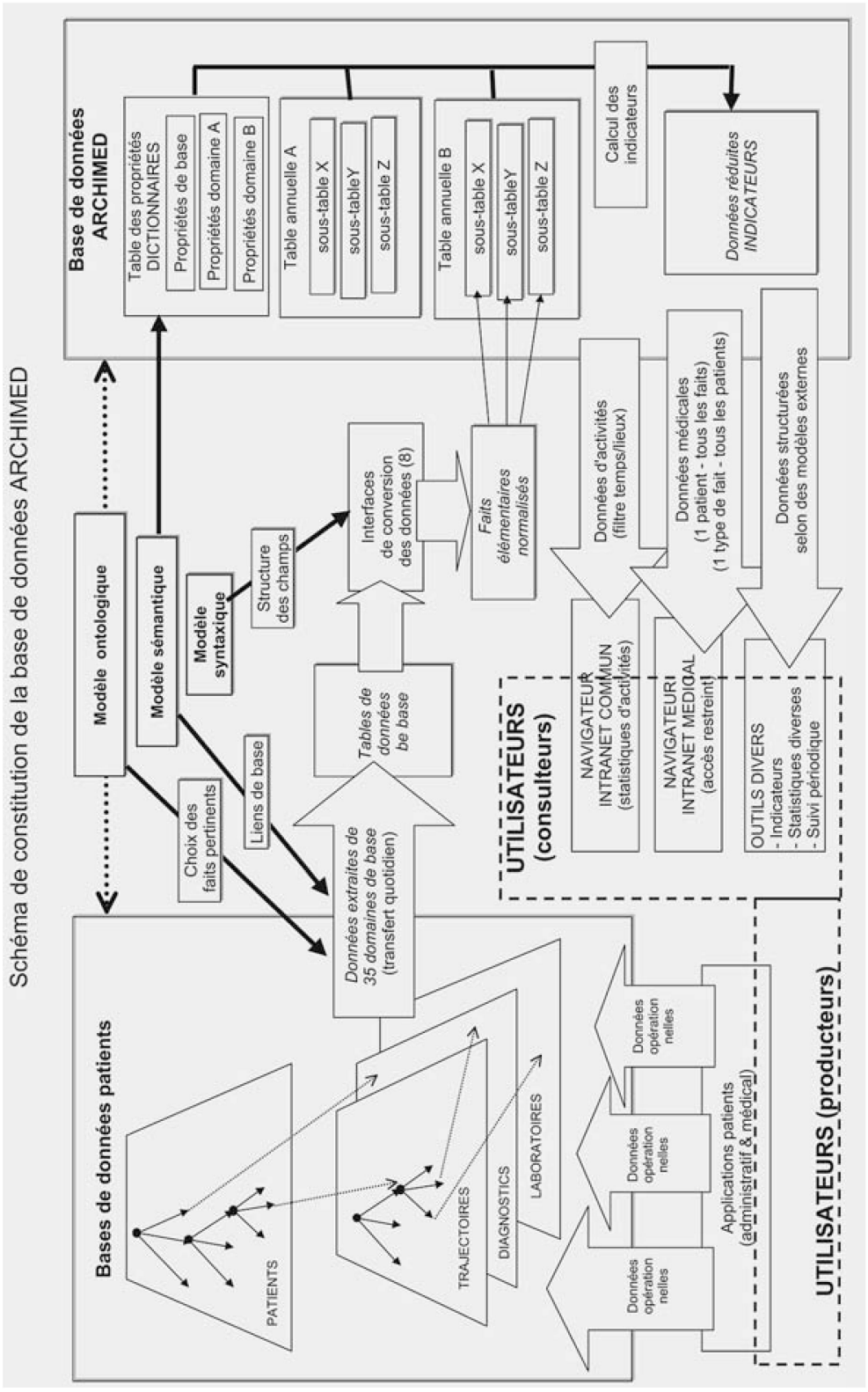


Figure 4: Schéma de constitution de la base de données ARCHIMED

Kopieren – oder verlieren: Grenzen der Rettung digitaler Medien aus der Sicht einer Praktikerin

Barbara Signori

Résumé

Le projet e-Helvetica a pour but de créer les bases nécessaires à la collection, l'indexation, la mise à disposition et la conservation à long terme des Helvetica électroniques.

La Bibliothèque nationale suisse (BNS) a fait un premier pas en direction de la conservation à long terme en copiant ses disquettes sur disque dur. Le vieillissement des disquettes entraîne une dégradation massive (démagnétisation) et le problème est aggravé par des formats de support ou de fichier devenus inutilisables. D'octobre 2002 à mars 2003, toutes les disquettes des fonds de la BNS ont été copiées dans un système de fichiers sur disque dur. La lisibilité et le fonctionnement correct de l'installation ont été contrôlés lors du transfert et les mesures de conservation ont été inscrites dans les notices bibliographiques respectives.

Zusammenfassung

Ziel des Projekts e-Helvetica ist es, Voraussetzungen für die Sammlung, Erschliessung, Bereitstellung und Langzeiterhaltung elektronischer Helvetica zu schaffen.

Der erste Schritt in Richtung Langzeiterhaltung hat die Schweizerische Landesbibliothek (SLB) mit dem Umkopieren ihrer Disketten getan. Disketten sind mit zunehmendem Alter massiv vom Verfall (Entmagnetisierung) bedroht. Sie lassen sich unter Umständen schon frühzeitig gar nicht mehr einlesen. Dabei verschärft sich das Problem durch unbrauchbar gewordene Datenträger- bzw. Dateiformate zusätzlich. Von Oktober 2002 bis März 2003 wurden sämtliche Disketten, die sich in den Beständen der SLB befinden, in ein Filesystem auf Harddisk umkopiert. Beim Umkopieren wurden die Disketten auf ihre Les- und Installierbarkeit geprüft. In den jeweiligen bibliographischen Titelaufnahmen wurden die vorgenommenen Bestandserhaltungsmassnahmen vermerkt.

Die Schweizerische Landesbibliothek und ihr Projekt e-Helvetica

Schon früh verfügte die Schweizerische Landesbibliothek über den gesetzlich festgeschriebenen Auftrag, das schweizerische Schrifttum, die sogenannten «Helvetica»[1], nicht nur zu sammeln und zu erschliessen, sondern auch zu erhalten und zu vermitteln, also dem Publikum zugänglich zu machen. Dieser Auftrag wurde im neuen Gesetz über die Schweizerische Landesbibliothek 1992 modernisiert und bezieht sich heute nicht mehr nur auf gedruckte Publikationen, sondern umfasst auch «auf andern Informationsträgern gespeicherte Informationen».[2] Somit gehören sowohl Offline-Publikationen - elektronische Publikationen, die auf physischen Datenträgern publiziert und verbreitet werden (z.B. Diskette, CD-ROM) - als auch Online-Publikationen - elektronische Publikationen, die ohne die Bindung an physische Datenträger im Internet publiziert und verbreitet werden (z.B. e-Journal, Datenbank) - zum Auftrag mit dazu.

Es ist die Aufgabe der Nationalbibliotheken in ihrer Rolle als nationales Gedächtnis, dem Verlust des geistigen und kulturellen Erbes aktiv entgegen zu wirken. Die elektronischen Medien machen immer mehr einen wichtigen Teil dieses geistigen und kulturellen Erbes aus. Zusätzliche und umfangreichere Aufgaben kommen auf die Nationalbibliotheken zu. So müssen unter anderem die Sammeltätigkeiten der Nationalbibliotheken neu auf elektronische Medien ausgeweitet werden. Dies bedingt wiederum eine Neudefinition der Sammelrichtlinien. Auch die Suche nach geeigneten Archivierungstechnologien und die Anpassung der Arbeitsprozesse sind neue Aufgabenbereiche, die es nicht zu unterschätzen gilt.

All diesen Herausforderungen stellt sich seit 2001 das Projekt e-Helvetica (http://www.e-helvetica.admin.ch/) der Schweizerischen Landesbibliothek. Ziel des Projekts e-Helvetica ist es, bis Ende 2006 Voraussetzungen für die Sammlung, Erschliessung, Bereitstellung und Langzeiterhaltung elektronischer Helvetica zu schaffen. Dies beinhaltet auch den Aufbau eines digitalen Archivs zur Langzeiterhaltung elektronischer Medien. Fünf Personen (210 Stellenprozente) beschäftigen sich mit diversen Hauptthemen dieses neuen Tätigkeitsfeldes.

---

1 Als «Helvetica» gelten die gesamte literarische Produktion des Landes, alle weiteren in der Schweiz publizierten Informationsträger, alle im Ausland erschienenen Werke, die sich auf in der Schweiz lebende Personen und auf schweizerische Sachverhalte beziehen, sowie Werke von Schweizer Autorinnen und Autoren.

2 «Bundesgesetz über die Schweizerische Landesbibliothek». SR 432.21. <http://www.snl.ch/d/download/gesetz92.pdf>, Art. 2

Die aktuellen Hauptthemen des Projekts e-Helvetica sind:[3]

– die Archivierung
  Die grösste Herausforderung beim Aufbau eines Archivs für elektronische Publikationen liegt in der Entwicklung eines Archivierungssystems. Nicht nur Inhalte, sondern auch Darstellung und Funktionalität von elektronischen Medien müssen langfristig erhalten werden. Die Schweizerische Landesbibliothek kooperiert beim Aufbau ihres Archivierungssystems mit dem Schweizerischen Bundesarchiv. Bis zur Inbetriebnahme eines definitiven Archivierungssystems werden bereits gesammelte Publikationen auf einem Testserver gespeichert.
– die Dissertationen
  Doktorandinnen und Doktoranden von einigen Schweizer Universitäten liefern ihre Dissertationen seit kurzer Zeit in elektronischer Form ab. Deren Langzeitarchivierung erfolgt in der Schweizerischen Landesbibliothek in elektronischer Form. Die Meldung über ihr Erscheinen erfolgt mittels eines interaktiven Formulars.
– die Sammelrichtlinien
  Die Schweizerische Landesbibliothek baut die Sammlung von elektronischen Publikationen selektiv auf. Dies bedeutet, dass die Sammlung aufgrund einer intellektuellen und kontrollierten Auswahl von elektronischen Publikationen erfolgt. Dabei wird eng mit den Produzentinnen und Produzenten von elektronischen Medien zusammengearbeitet. Die Publikationen werden im Bibliothekskatalog Helveticat der Schweizerischen Landesbibliothek erfasst und beschrieben. Als begleitende Massnahme plant die Schweizerische Landesbibliothek in regelmässigen Abständen eine Momentaufnahme (Harvesting) der Internet-Domäne .ch zu erstellen. Diese würde zu Dokumentationszwecken verwendet werden. Für das Harvesting würde keine beschreibende Erschliessung im Bibliothekskatalog Helveticat gemacht werden.
  Die Sammlung von Online-Publikationen kann nur kooperativ bewältigt werden. Die Schweizerische Landesbibliothek arbeitet mit Der Deutschen Bibliothek und mit der Österreichischen Nationalbibliothek im Bereich der Abstimmung der Sammelgebiete eng zusammen. Potentielle Kooperationspartnerinnen in der Schweiz sind die Kantonsbibliotheken mit Archivierungsauftrag. Im November 2003 fand in der Schweizeri-

3 Siehe auch: Balzardi, Elena: «Das Projekt e-Helvetica: eine Momentaufnahme». In: *Jahresbericht der Schweizerischen Landesbibliothek*, Jg. 89, 2002, S. 38-40.

schen Landesbibliothek diesbezüglich eine Tagung «Langzeitverfügbarkeit digitaler Publikationen in Schweizer Archivbibliotheken – eine gemeinsame Herausforderung» statt.

– die Langzeiterhaltung von Objekten auf Disketten
  Im Magazin der Schweizerischen Landesbibliothek befinden sich knapp 900 Disketten. Disketten sind massiv vom Zerfall (Entmagnetisierung) bedroht. Teilweise können sie nicht mehr eingelesen werden, weil das Datenträger- oder Dateiformat nicht mehr gebräuchlich ist. Zum Beispiel musste für die 5 1/4-Zoll-Disketten – vor einigen Jahren noch ein gebräuchliches Format – erst lange nach einem Lesegerät gesucht werden. Zur Erhaltung dieser Publikationen wurden die Disketten der Schweizerischen Landesbibliothek auf ein File-System umkopiert. Die Dateien und die dazugehörenden technischen Angaben wurden vorderhand auf dem Testarchivierungssystem der Schweizerischen Landesbibliothek gelagert. Nach der Inbetriebnahme des definitiven Archivierungssystems werden sie umgelagert werden.

Das Projekt e-Helvetica und die Langzeiterhaltung von Objekten auf Disketten

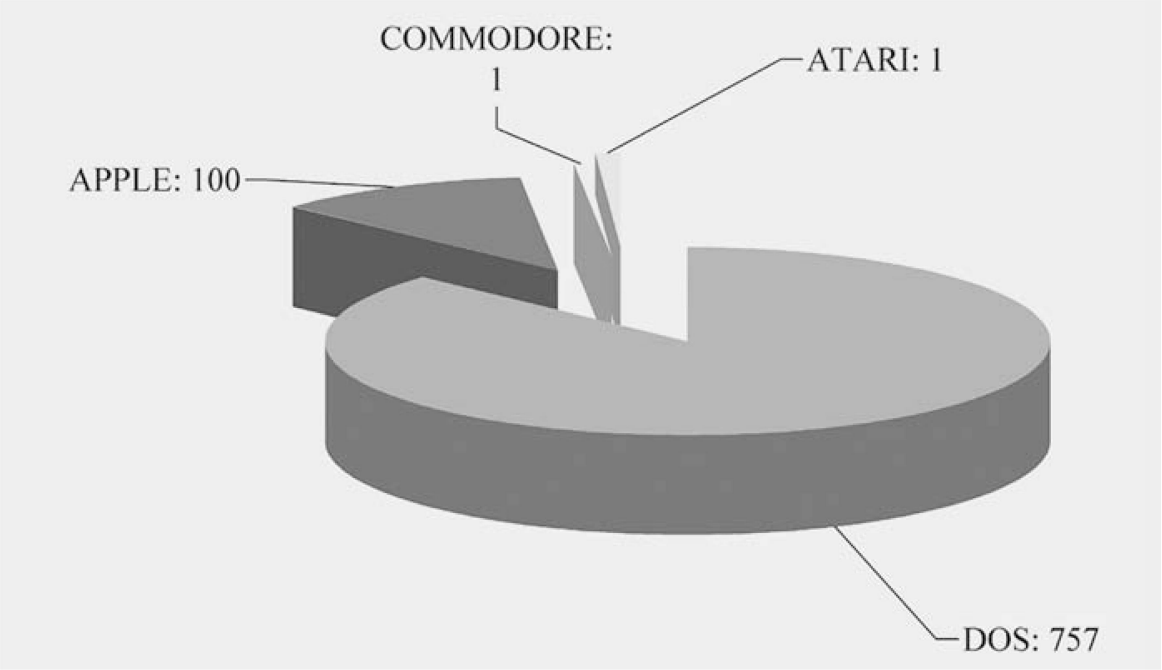
Viele der heute auf Disketten gespeicherten Daten werden in einigen Jahrzehnten nicht mehr lesbar sein. Digitale Informationen sind anfälliger als Papier. Papier muss man im schlimmsten Fall entsäuern. Lücken auf einer Seite oder eine ausgeblichene Schrift kann ohne allzu grosse Schwierigkeiten ausgebessert werden. Ein einziges fehlerhaftes Bit auf einem elektronischen Datenträger aber kann einen kompletten Datensatz unleserlich machen. Einige Wissenschaftler sprechen darum schon von einem neuen «dunklen Zeitalter» in der Menschheitsgeschichte. Während sich die Informationsmenge auf der Welt alle sechs Monate vervierfacht, wird mehr als die Hälfte dieser Daten nur noch digital bearbeitet. Eine Kopie auf Papier existiert nicht. Die Wissenschaftler vergleichen den drohenden Datenverlust mit dem Brand in der Bibliothek von Alexandria im Jahr 47 vor Christus. Nach zwanzig Jahren könnten sämtliche auf elektronischen Datenträgern gespeicherten Informationen durch Erosion, nicht mehr gebräuchliche Datenträger oder veraltete Hardware unlesbar geworden sein. Das kulturelle Erbe aus den Anfängen des digitalen Zeitalters könnte verschwinden und der heutigen Informationsgesellschaft droht ein Gedächt-

nisverlust.[4] Diesen Gedächtnisverlust so gering wie möglich zu halten ist eine zentrale Aufgabe des Projekts e-Helvetica der Schweizerischen Landesbibliothek.

Den ersten Schritt in Richtung Langzeiterhaltung ihrer elektronischen Helvetica hat die Schweizerische Landesbibliothek mit dem Umkopieren ihrer Disketten getan. Die Arbeiten begannen Anfang Oktober 2002 und endeten Ende März 2003. Ein zusätzlicher Informatiker wurde während dieses Zeitraums eingestellt.

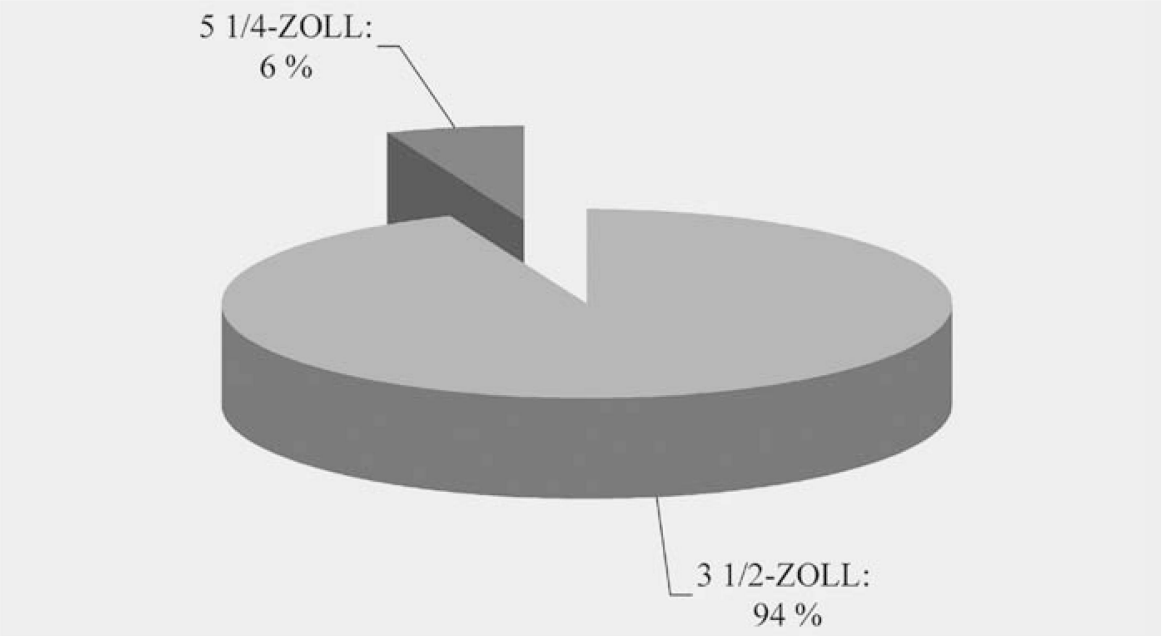
Insgesamt wurden aus dem Magazin der Schweizerischen Landesbibliothek 859 physische Disketten bearbeitet.[5] Dies entspricht 517 bearbeiteten bibliographischen Aufnahmen im Bibliothekskatalog Helveticat.

Von diesen 859 Disketten sind 757 DOS-Disketten, 100 APPLE-Disketten, 1 COMMODORE-Diskette und 1 ATARI-Diskette. Die beiden Letzteren wurden im Austausch von Der Deutschen Bibliothek bearbeitet.

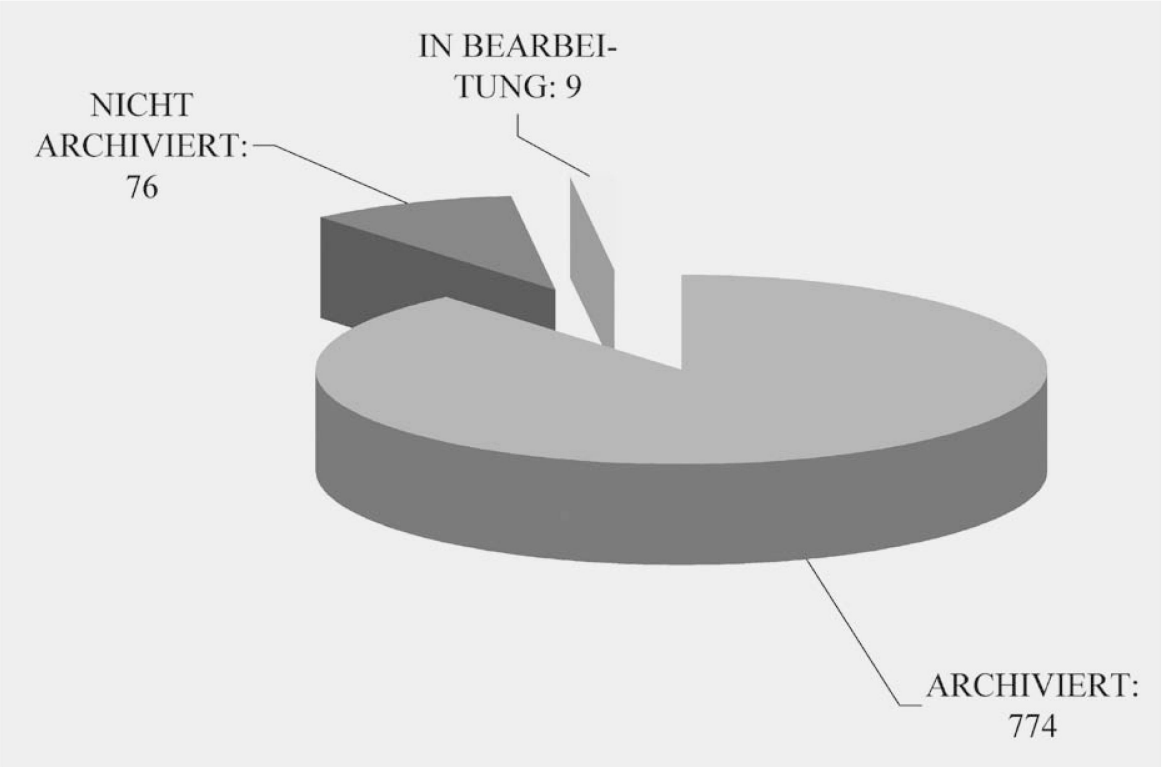


4 Siehe auch: Karsack, Hendrik: «Eine digitale Zeitbombe: Ein Kampf gegen das Vergessen: auch elektronische Datenträger sind nicht für die Ewigkeit gemacht». In: *Frankfurter Allgemeine Zeitung,* Montag, 30. Juli 2001, Nr. 174, S. 9.

5 Die Zahlen basieren auf dem Stand von Ende März 2003.

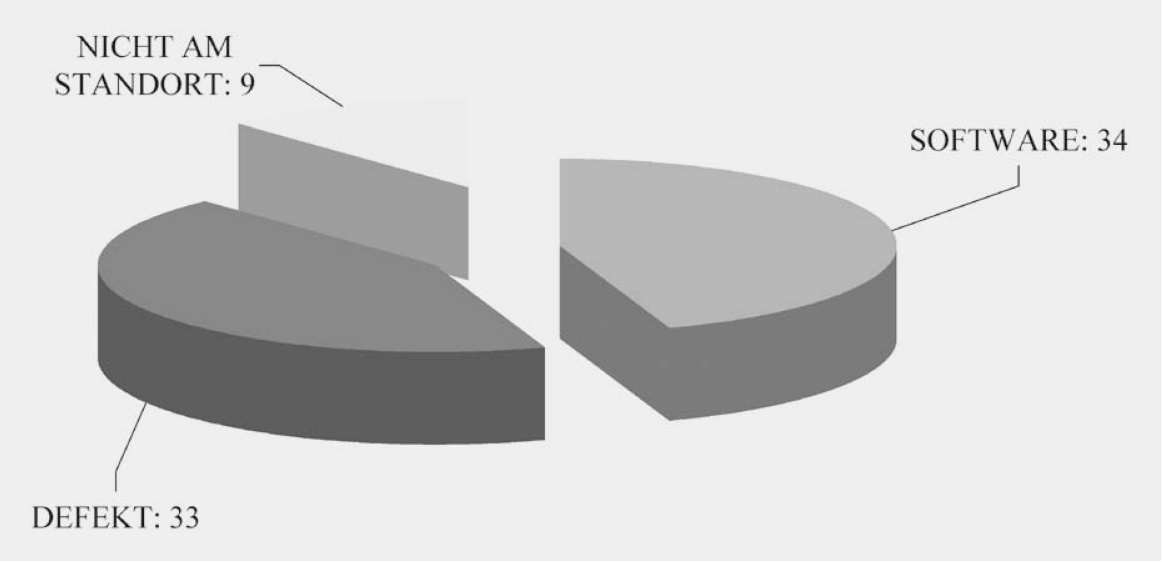


Die Mehrzahl, d.h. 94 % der Disketten sind 3 1/2-Zoll-Disketten. Die restlichen 6 % sind 5 1/4-Zoll-Disketten.



774 Disketten konnten erfolgreich auf dem Testserver der Schweizerischen Landesbibliothek archiviert werden. Diese Disketten wurden auf ihre Les-

und Installierbarkeit überprüft. Einige Disketten benötigen für die Installation spezielle Schritte. Diese sind im Bibliothekskatalog Helveticat in der entsprechenden bibliographischen Aufnahme im MARC21-Feld 922 (Installation) beschrieben. 76 Disketten konnten nicht archiviert werden. Ende März 2003 befanden sich noch 9 Disketten in Bearbeitung.



Von den 76 nicht archivierten Disketten bestehen 34 Disketten aus reiner Anwender-Software. Diese konnten sogleich als «nicht archivierungswürdig» eingestuft werden, zumal die Sammelrichtlinien der Schweizerischen Landesbibliothek das Sammeln von reiner Anwender-Software ausschliessen. 33 Disketten sind effektiv nicht mehr les- und installierbar. Entweder ist der Träger selber beschädigt oder die Diskette ist leer oder Dateien auf der Diskette sind fehlerhaft usw. 9 Disketten sind trotz bibliographischer Aufnahme im Bibliothekskatalog Helveticat nicht an ihrem Standort im Magazin und gelten somit als verloren. Der Erwerbungsdienst der Schweizerischen Landesbibliothek hat versucht, defekte und verlorene Disketten neu zu beschaffen. In diesen Fällen war eine Neubeschaffung leider nicht mehr möglich.[6]

6 Für vertiefte Informationen siehe: Signori, Barbara: «Langzeiterhaltung von Objekten auf Disketten: Schlussbericht», <http://www.e-helvetica.admin.ch/pdf/ger/tp-org/Langzeiterhaltung%20von%20Objekten%20auf%20Disketten_Schlussbericht.pdf>, 6. Mai 2003.

Informationen über e-Helvetica

Das Projekt e-Helvetica hat sich das Ziel gesetzt, breit und umfassend über seine Aufgaben und Tätigkeiten zu informieren. Die Website www.e-helvetica.ch richtet sich an das allgemeine Publikum, an Verlage und an Bibliotheken. Zahlreiche Informationen, Fachberichte und weiterführende Links stehen Interessierten zur Verfügung.

Quellenhinweise

Balzardi, Elena: «Das Projekt e-Helvetica: eine Momentaufnahme». In: *Jahresbericht der Schweizerischen Landesbibliothek*, Jg. 89, 2002, S. 38-40.

Bilfinger, Monica: *Die Schweizerische Landesbibliothek in Bern.* Bern 2001.

Karsack, Hendrik: «Eine digitale Zeitbombe: Ein Kampf gegen das Vergessen: auch elektronische Datenträger sind nicht für die Ewigkeit gemacht». In: *Frankfurter Allgemeine Zeitung,* Montag, 30. Juli 2001, Nr. 174, S. 9.

Signori, Barbara; Locher, Hansueli: «Langzeiterhaltung von Objekten auf Disketten», <http://www.e-helvetica.admin.ch/pdf/ger/tp-org/Langzeiterhaltung%20von%20Objekten%20auf%20Disketten3.pdf>, 10. April 2002.

Signori, Barbara; Locher, Hansueli: «Langzeiterhaltung von Objekten auf Disketten: Arbeitsablauf», <http://www.e-helvetica.admin.ch/pdf/ger/tp-org/Langzeiterhaltung%20von%20Objekten%20auf%20Disketten_Arbeitsablauf2.pdf>, 8. Oktober 2002.

Signori, Barbara: «Langzeiterhaltung von Objekten auf Disketten: Schlussbericht», <http://www.e-helvetica.admin.ch/pdf/ger/tp-org/Langzeiterhaltung%20von%20Objekten%20auf%20Disketten_Schlussbericht.pdf>, 6. Mai 2003.

Website der Schweizerischen Landesbibliothek: <http://www.snl.ch>.

Website des Projekts e-Helvetica: <http://www.e-helvetica.ch>.

Archivierung von Internetseiten – eine Standortbestimmung

Hansueli Locher

Résumé

Les informations disponibles via Internet changent rapidement. La durée d'existence moyenne d'une page Web ne dépasse pas les 100 jours. Une telle fugacité de l'information va à l'encontre des efforts d'archivage.

Cette contribution évoque les expériences faites par «l'Internet Archive» avec sa «Wayback Machine», explique une stratégie intéressante choisie par la Bibliothèque nationale de France et donne une vue d'ensemble des tests que la Bibliothèque nationale Suisse (BNS)a effectués avec «Web Harvester». Elle présente finalement un aperçu des projets de la BNS concernant l'archivage de pages Internet.

Zusammenfassung

Informationen im Internet ändern rasch. Die durchschnittliche Lebensdauer einer Website beträgt knapp 100 Tage. Diese Flüchtigkeit der Information steht den Archivierungsbestrebungen entgegen.

Im Beitrag wird auf die Archivierungsbemühungen von «Internet Archive» mit ihrer «Wayback Machine» eingegangen, eine interessante Strategie der Bibliothèque nationale de France erläutert und ein Einblick in die aktuellen Tests der Schweizerischen Landesbibliothek (SLB) mit einem «Web Harvester» geboten. Zum Schluss erfolgt ein kurzer Ausblick auf die weiteren Pläne der SLB bezüglich der Archivierung von Internetseiten.

1. Einleitung

Lesen Sie schnell, denn nichts ist beständiger als der Wandel im Internet!
(Anita Berres, deutsche Publizistin)

Das Zitat von Anita Berres lässt sich durch Zahlen untermauern: Die mittlere Lebensdauer einer Website beträgt rund 19 Monate und die durchschnittliche Lebensdauer eines HTML-Dokuments gerade mal 100 Tage.[1]

Für die Archivierungsbemühungen hat das grosse Auswirkungen. Wenn wir die im World Wide Web gegenwärtig verfügbaren Dokumente nicht jetzt sammeln, sind sie für immer verloren.

Das ist aber einfacher gesagt als getan. Verschiedene weitere Eigenschaften des World Wide Web stellen sich dieser Absicht entgegen.

- Riesiger Datenumfang
  Für das Jahr 2002 wird der Datenumfang auf dem öffentlich zugänglichen Web ohne datenbankbasierte Webdokumente (Surface Web) auf 167 Terabytes geschätzt.[2] Das entspricht etwa dem Speicherplatz von 120 Millionen 3 1/2-Zoll-Disketten.
- Rasches Wachstum
  Schätzungen aus dem Jahr 2000 gehen von täglich 7,3 Millionen neuen Webdokumenten aus, die Speicherplatz in der Grösse von 0,1 Terabytes (100 Gigabytes) belegen.[3]
- Unübersichtlichkeit
  Vorgaben zur Datenorganisation oder zu inhaltlichen Strukturen fehlen gänzlich. Auf gegenwärtig 172 Millionen im Internet registrierten Servern sind irgendwelche Informationen zu finden, auf die mit verschiedensten Instrumenten zugegriffen werden kann. Das Angebot reicht dabei von Websites über Mail-Server, Datenbanken, FTP-Server bis hin zu Diskussionsgruppen und Chat-Foren.[4]

---

1 Stata, Raymie: Saving the Web (Vortrag an der European Conference on Digital Libraries 2002 in Rom), <http://webapp.bnf.fr/bibnum/ecdl/2002/ia/ia.html>

2 Lyman, Peter; Varian, Hal R. et al.: «How Much Information 2003», <http://www.sims.berkeley.edu/research/projects/how-much-info-2003/printable_report.pdf>, 27. Oktober 2003.

3 Murray, Brian H; Moore, Alvin: «Sizing the Internet: A White Paper», Cyveillance, <http://www.cyveillance.com/web/corporate/white_papers.htm>, 10. Juli 2000, S. 2.

4 Quelle: Internet Software Consortium, <http://www.isc.org/ds/>.

Die wesentliche Herausforderung beim Archivieren von Websites liegt darin, die Dokumente zeitgerecht und mit vernünftigem Aufwand im Internet abzuholen und auf einem Archivsystem abzulegen. Das Aufbewahren an und für sich bietet genau die gleichen Probleme wie bei allen übrigen elektronischen Daten auch.

2. Verschiedene Ansätze zur Archivierung von Internetseiten

Verschiedene Stellen, insbesondere aber auch die Nationalbibliotheken, bemühen sich heute um die Archivierung von Websites. Ihre Strategien sind zum Teil sehr unterschiedlich. In den nachfolgenden Unterkapiteln soll auf zwei wichtige Initiativen in diesem Bereich eingegangen werden.

2.1 Internet Archive[5]

Internet Archive ist eine Nonprofit-Organisation, die 1996 mit dem Ziel gegründet wurde, eine «Internet-Bibliothek» zu erstellen, die sowohl Wissenschaftler/innen wie Forscher/innen und Historiker/innen als auch dem breiten Publikum den ständigen Zugriff zu historischen Sammlungen von digitaler Information bieten soll.

Seit der Gründung wurden rund 30 Milliarden Web-Dokumente archiviert und erschlossen. Das Sammeln geschieht in automatisierter Form durch WebCrawler von Alexa.[6] Gesammelt werden dabei im Prinzip alle öffentlich zugänglichen Files im World Wide Web.

Die Datenmenge im Internet Archive beträgt gegenwärtig rund 250 Terabytes. Das entspricht ungefähr 250 Millionen Büchern à je 200 Seiten. Das sind mehr Bücher, als seit der Erfindung der Buchdruckerkunst weltweit produziert worden sind.[7]

Pro Monat nimmt die Datenmenge um rund 10-12 Terabytes zu. Die einzelnen Dateien aus dem Internet werden dabei zu ARC-Files von je 100 Megabytes Grösse zusammengefasst und auf DLT-Tapes abgelegt.

Die Lebensdauer dieser DLT-Tapes beträgt etwa 30 Jahre. Internet Archive plant aber öfter als alle 10 Jahre ihre Archivdaten umzukopieren.[8]

---

5 Siehe Website Internet Archive, <http://www.archive.org>.

6 Informationen zu diesem WebCrawler: <http://pages.alexa.com/company/technology.html>. Die Firma Alexa gehört Amazon.com.

7 «Auffrischung des «nationalen Gedächtnisses». Grundzüge einer schweizerischen Memopolitik». In: *Neue Zürcher Zeitung*, 13. Mai 2003.

8 «About the Internet Archive», <http://www.archive.org/about/about.php>.

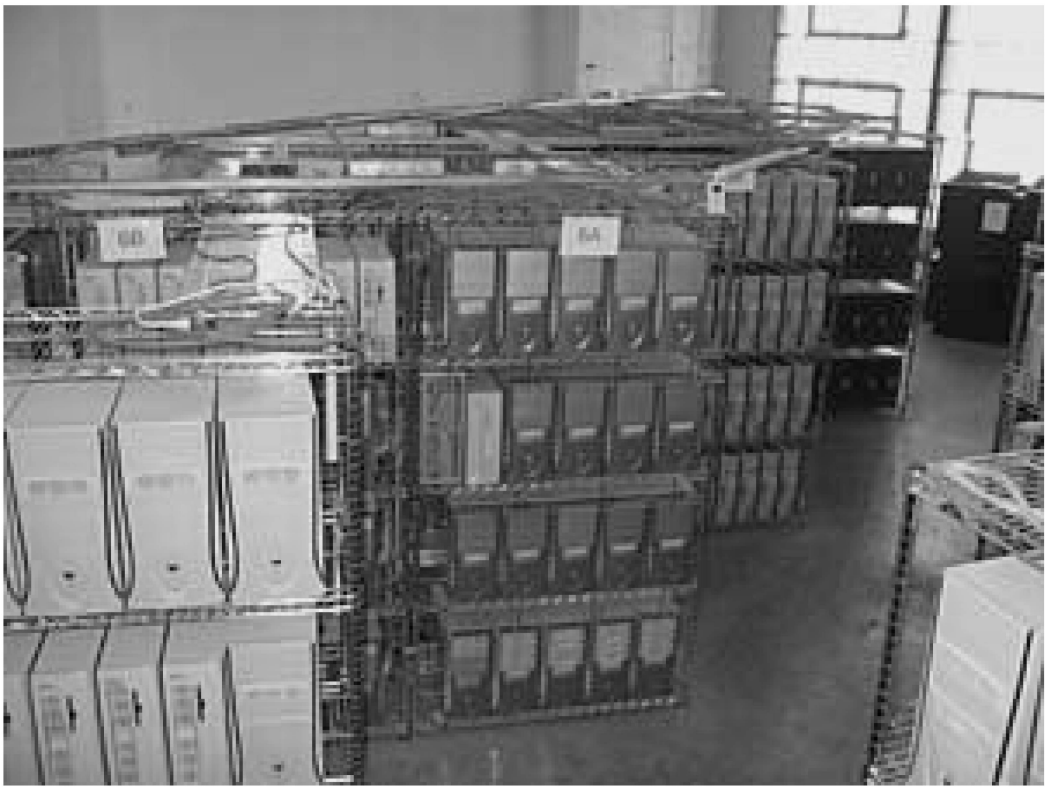


Abb. 1: Teil der Wayback Machine

Mit Hilfe der sogenannten Wayback Machine kann der Benutzer auf die archivierten Webseiten zugreifen, die ihm zu diesem Zweck auf den Harddisks von einigen 100 Servern zur Verfügung gestellt werden. Es genügt dabei die Adresse einer Website oder eines Webdokuments einzutippen, um einen Überblick über die archivierten Versionen zu erhalten, die via Link nun aufgerufen werden können.

2.2 Bibliothèque nationale de France (BnF)

Bei den Nationalbibliotheken stehen zwei Strategien zum Sammeln von Websites im Vordergrund. Die Nordländer führen ein Domain-Harvesting durch, das heisst, sie versuchen alle Websites, deren Domänen-Namen die Kennzeichnung ihres Landes führen, zu sammeln.[9] Andere wie beispielsweise die amerikanische Nationalbibliothek, die Library of Congress, setzen auf ein selektives Harvesting, bei dem ausgewählte Websites oder auch

9 Die nordischen Nationalbibliotheken (Dänemark, Finnland, Island, Norwegen und Schweden) arbeiten im Bereich des Web-Harvesting eng zusammen. Als Forum für Koordination und Erfahrungsaustausch dient das Nordic Web Archive (NWA).

möglichst alle Websites zu einem bestimmten Thema (Präsidentenwahlen 2000 in den USA, 11. September 2001) gesammelt werden.

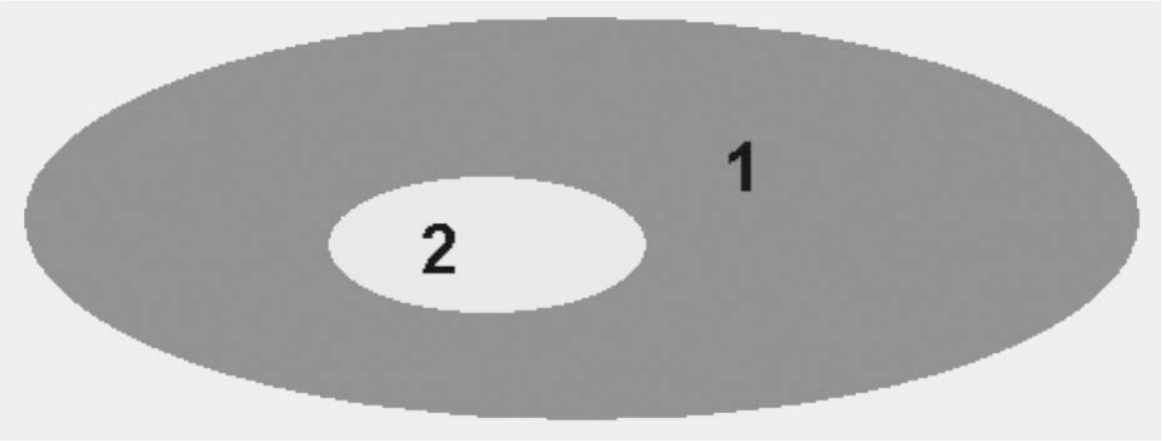


Abb. 2: Selektives Harvesting
1 Gesamte französische Webdomäne (.fr)
2 Ausgewählte Websites

Die Bibliothèque nationale de France (BnF) mischt diese beiden Ansätze. Die gesamte französische Webdomäne wird in einem Harvestig-Durchgang gesammelt. Dieses generelle Harvesting dient dazu, eine Link-Topologie zu erstellen. Damit lässt sich feststellen, auf welche Websites am häufigsten mittels Links referenziert wird. Diese Websites bilden dann zusammen mit weiteren Websites, die durch die Referenzbibliothekarinnen und -bibliothekare bestimmt werden, eine Auswahl. Diese wird genauer analysiert. Je nach Bedarf werden dann für diese Websites, die zur Auswahl gehören, spezielle Massnahmen vorgesehen.

– Bei Sites, die sich rasch verändern (z.B. Online-Zeitungen oder -Zeitschriften), wird der Harvesting-Prozess mit hoher Periodizität durchgeführt.
– Bei dynamischen Websites wird mit der verantwortlichen Stelle vereinbart, dass sie die Inhalte in regelmässigen Abständen auf einem Datenträger der BnF zur Verfügung stellt.

3. Tests der SLB

Die Schweizerische Landesbibliothek (SLB) hat sich im Rahmen des Projekts Networked European Deposit Library (NEDLIB)[10] an der Entwick-

10 Website NEDLIB <http://www.kb.nl/coop/nedlib>.

lung einer Software zum Sammeln von Internet-Dokumenten beteiligt. Dieser NEDLIB-Harvester wurde Ende 2002 mit Hilfe eines externen Dienstleisters auf einem Testsystem in der SLB installiert, um erste Erfahrungen im Umgang mit Websites sammeln zu können.

3.1 Funktionsweise des NEDLIB-Harvesters

Der NEDLIB-Harvester kann auf einer Linux- oder Unix-Plattform eingesetzt werden. Er benötigt eine MySQL-Datenbank. Diese wird sowohl für das Speichern von Informationen zu den archivierten Dokumenten als auch zur Konfiguration der Software verwendet.

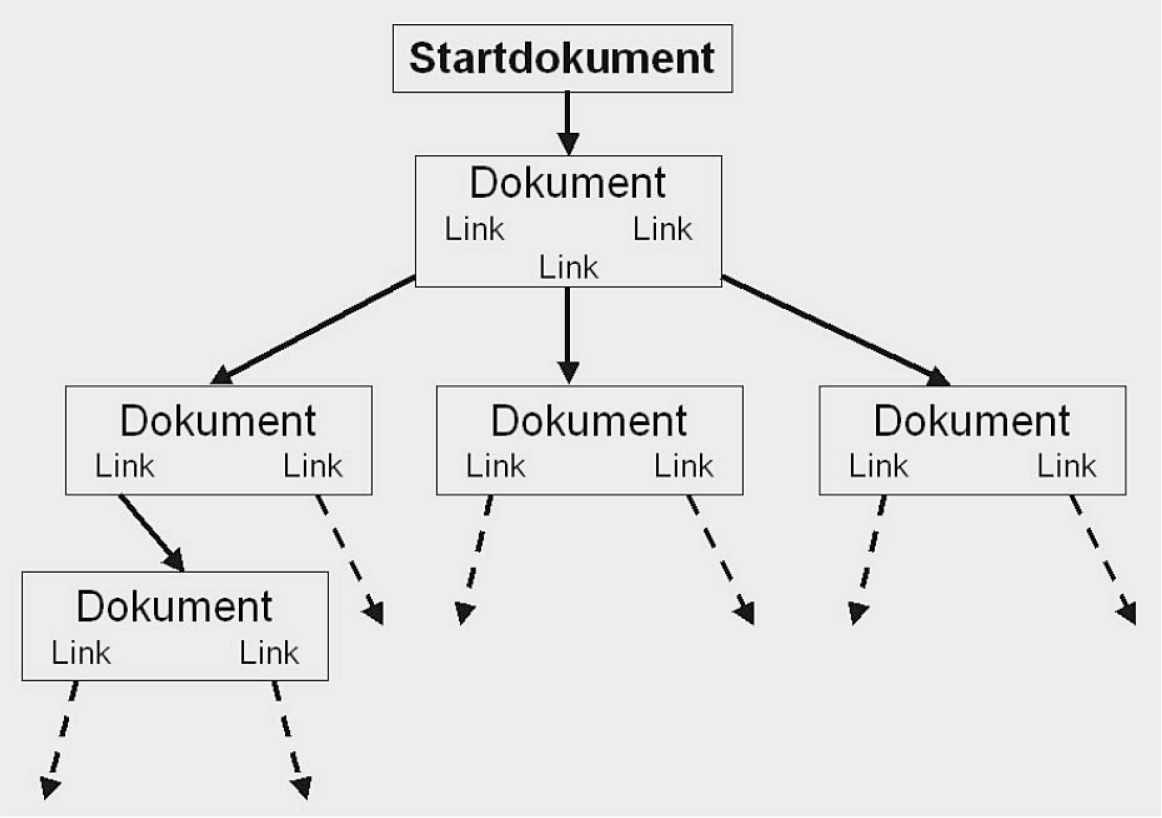


Abb. 3: Harvesting-Prozess

Die Startadresse (Link auf ein Dokument im Internet), von der aus der Sammelprozess durchgeführt werden soll, muss manuell konfiguriert werden. Ausgehend von dieser Adresse beginnt dann das Harvesting, indem den im Startdokument vorhandenen Links gefolgt wird, die zu weiteren Dokumenten führen. In diesen werden erneut die Links ausgewertet.

Dieser Prozess kann mit Hilfe einer vorgängigen Konfiguration auf bestimmte Websites oder Domänen beschränkt werden.

3.2 Die Tests

Der NEDLIB-Harvester wurde in der SLB auf einem normalen Pentium 4 (1,7 GHz Prozessor, 256 MB RAM) getestet, der auch als Arbeitsplatzgerät bei den Mitarbeitenden eingesetzt wird.

In einer ersten Testphase war der Harvester nach rund 40 Minuten und etwa 5000 archivierten Dokumenten überlastet. Es wurden kaum noch Dokumente archiviert. Der Grund dafür lag bei der MySQL-Datenbank, die in der Folge für grosse Datenmengen optimiert wurde.

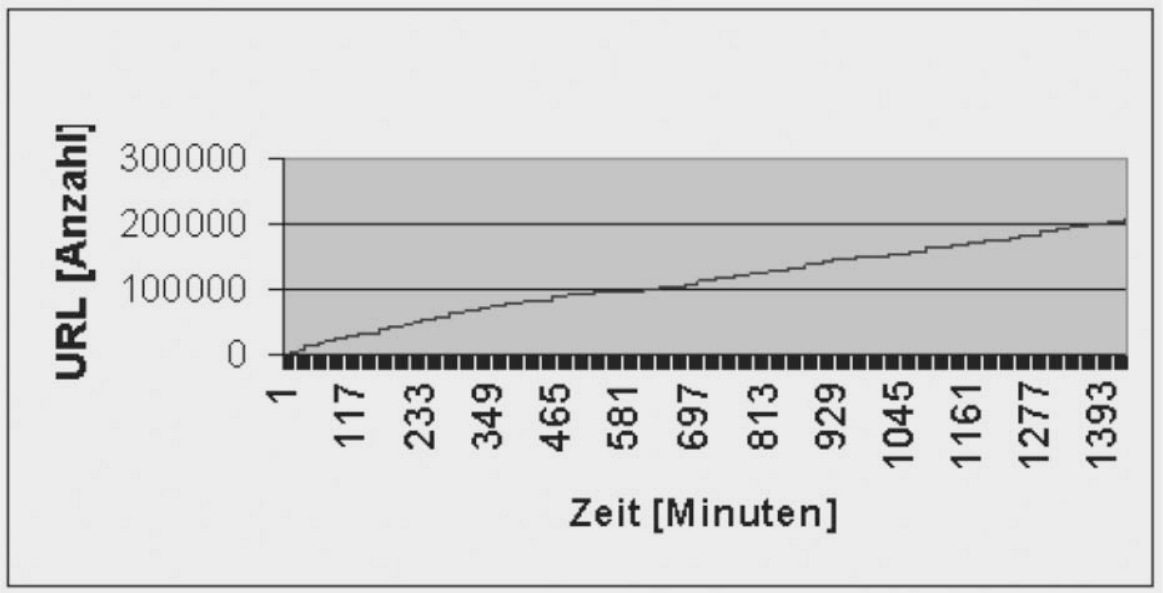


Abb. 4: Leistungsmessung beim Harvesting

Nun war es möglich innerhalb von 24 Stunden 206'000 Dateien mit einem Umfang von gut 8 GB zu sammeln. Die Zuwachskurve verlief linear (vgl. Abbildung). Die Leistungsfähigkeit des Systems blieb konstant hoch.

Als entscheidend für die Geschwindigkeit des Sammelprozesses stellten sich in dieser Testphase verschiedene Faktoren heraus:

– Leistungsfähigkeit der Datenbank
– Datendurchsatz von und zu der Harddisk
– Geschwindigkeit der Harddisk

Der Einsatz mehrerer Harddisks bringt einen weiteren Geschwindigkeitsgewinn.

3.3 Erkenntnisse

Durch den Einsatz eines leistungsfähigen Servers für den Sammel-Prozess und einer Konfiguration, die alle Möglichkeiten zur Optimierung des Pro-

zesses ausschöpft, müsste es möglich sein, rund 25 Gigabytes Daten pro Tag hereinholen zu können. Da die Datenmenge im .ch-Domain im Internet durch die SLB auf gut 2 Terabytes geschätzt wird, heisst das konkret, dass das Sammeln der ganzen Domäne mit dem NEDLIB-Harvester mindestens 80 Tage beanspruchen würde.

Die praktische Arbeit hat gezeigt, dass es dabei eine ganze Reihe von Punkten zu beachten gilt:

- Gewisse Datentypen müssen vom Sammelprozess ausgeschlossen werden können. Die SLB hat beispielsweise kein Interesse an Programm-Files, die in öffentlichen Software-Bibliotheken angeboten werden und mit ihrem Datenumfang den Prozess deutlich verlangsamen.
- Grosse Websites mit zum Teil gegen 100'000 Dokumenten müssen ganz am Anfang des Sammelprozesses bereits besucht werden, weil diese sonst die Dauer des Prozesses wesentlich verlängern können.
- Die gesammelten Datenfiles müssen mit einer eindeutigen, immer gleich bleibenden Identifikation (Persistent Identifier) versehen und verzeichnet werden, damit sie auffindbar bleiben.
- Sobald periodische Sammelprozesse vorgesehen werden, braucht es ein Versionenmanagement.
- Der NEDLIB-Harvester – wie übrigens praktisch alle anderen Harvester auch – kann keine Informationen sammeln, die nur über ein Abfrageformular zu erreichen oder durch Passwörter geschützt sind. Er beschränkt sich hauptsächlich auf statische Websites.

Der Zugriff auf die durch den NEDLIB-Harvester gesammelten Dokumente dürfte kein grosses Problem darstellen, da das Nordic Web Archive mit dem NWA Toolset[11] ein Instrument entwickelt hat, das den Benutzerinnen und Benutzern dafür zur Verfügung gestellt werden könnte.

4. Ausblick

Momentan ist die SLB daran, die Voraussetzungen dafür zu schaffen, die in Zukunft ein Web-Harvesting erlauben sollen. So wird ein Konzept für die Vergabe von Persistent Identifiers erarbeitet, die als Zugriffsadressen auch im Internet verwendet werden können. Die SLB arbeitet dabei mit Der Deutschen Bibliothek zusammen und wird Unified Resource Names (URN) auf der Basis von National Bibliographic Numbers (NBN) vergeben.

11 «About the NWA Toolset», <http://nwa.nb.no/aboutNwaT.php>, 5. September 2002.

Zusammen mit dem Bundesarchiv wird Speicherplatz von vorerst 30 Terabytes beschafft, um unter anderem auch die gesammelten Informationen aus dem Web ablegen zu können.

Voraussichtlich wird die SLB beim Sammeln und Archivieren von Websites zwar gelegentlich ein Domain-Harvesting betreiben, vor allem aber ausgewählte Websites archivieren.

Dies durchaus im Bewusstsein, dass es unmöglich ist, voraus zu sehen, was die Menschen von morgen interessieren wird.

Es gibt trotzdem verschiedene Gründe, die für den selektiven Ansatz sprechen:

- Genau wie im Printbereich sollen auch im Web-Bereich Sammelrichtlinien angewendet und diejenigen Websites aufbewahrt werden, die ihnen entsprechen.
- Die vorhandenen Mittel lassen sich bei einem selektiven Sammeln von Websites zielgerichteter einsetzen, als das mit einem regelmässigen Harvesting der .ch-Domäne möglich wäre. Dieses ist aus technischen Gründen (Zugriff auf Datenbanken, passwortgeschützte Information) ebenfalls nicht vollständig.

Neben der selektiven Sammlung von Websites wird aber auch eine enge Zusammenarbeit mit Verlagen oder anderen Produzenten von elektronischen Publikationen, die online oder auf Datenträgern verfügbar gemacht werden, angestrebt.