KOST-Newsletter Quartal 1, 2015

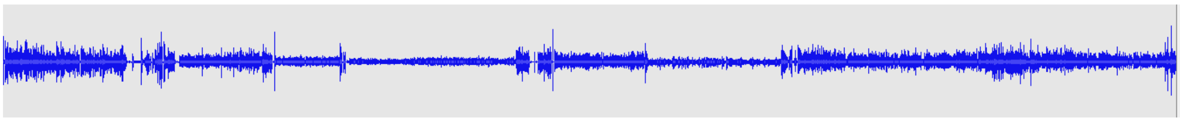
AudioVault

Im Projekt *AudioVault* haben das Liechtensteinische Landesarchiv, die Staatsarchive Bern, Uri, Jura sowie die Stadtarchive Bern und St. Gallen konkrete Problemstellungen bei der Archivierung von Audiounterlagen und -dokumenten untersucht und Lösungsansätze dazu erarbeitet. Drei Kategorien von Audiodokumenten standen für die beteiligten Archive im Vordergrund: Aufnahmen von Ratsdebatten, dokumentarische Aufzeichnungen und künstlerische Darbietungen. Die erste Kategorie ist für die beteiligten Archive eindeutig prioritär, weswegen sie im Projekt auch als Beispiel dient. Als Ergänzung und Weiterentwicklung des KOST-Projektes *Sauvez les CD* bot *AudioVault* die Gelegenheit, auf zwei Themenkomplexe im Bereich der Audioarchivierung vertieft einzugehen: die Erschliessung von Audiomaterialien und die Frage der Metadaten.

Erschliessung

Bei Aufnahmen von Ratsdebatten klaffen die Granularitäten der Primär- und Metadaten sichtbar auseinander. In der Regel sind detaillierte deskriptive Metadaten verfügbar (Granularität Redebeitrag), aber sie ermöglichen keinen präzisen Zugriff auf die Primärdaten, da die Aufzeichnungen in der Regel eine ganze Ratssitzung umfasst. Für dieses Problem sind grundsätzlich verschiedene Lösungen denkbar:

1. Audiostream in die Granularität der Verzeichnung schneiden
2. Eine Timecode-Referenz auf den Audiostream in den Metadaten ablegen
3. Mit einer externen Schnittanweisung für das Abspielprogramm (*Cue Sheet)* direkt auf einen Redebeitrag zugreifen
4. Eine *Playlist* direkt in den Audiocontainer einbetten



Aus archivischer Sicht ist vermutlich Lösung 3 mit einem externen *Cue Sheet* zur Audiodatei die beste Lösung. Es vermeidet Eingriffe in die Primärdatei und ermöglicht dennoch einen gewissen Komfort beim Zugang.

Ebenfalls Teil der Erschliessung ist die Frage, ob im Findmittel als „Vorschau“ ein *Sample* zur Verfügung gestellt werden soll. Falls es sich bei den Audioarchivalien um urheberrechtlich geschütztes Material handelt, stellt sich die Frage, ob dennoch mit Berufung auf das Zitatrecht ein kurzer Ausschnitt verwendet werden kann. Da das Zitatrecht in der juristischen Literatur widersprüchlich diskutiert wird und Gerichtspraxis weitgehend fehlt, kann diese Frage nicht abschliessend beantwortet werden. Mit der Aufnahme von *Samples* ins Findmittel würde ein Archiv ein gewisses, unter Umständen jedoch vertretbares rechtliches Risiko eingehen.

Metadaten

Neben der angesprochenen rechtlichen Problematik bei *Samples* darf nicht übersehen werden, dass diese im Gegensatz zu einem *Thumbnail* für Bilder oder einer *Key Frame Extraction* für Videos keine besonders aussagekräftige oder automatisch erstellbare Kurzansicht oder Vorschau sind. Umso wichtiger ist für Audioarchivalien die präzise Verzeichnung von Metadaten. Leider ist die Standardvielfalt für Audiometadaten sehr hoch. Für eingebettete Metadaten fällt die Wahl sinnvollerweise auf BWF Metadaten und RIFF INFO tags bei WAV bzw. IDv3 bei MP3. Für externe deskriptive Metadaten schlägt *AudioVault* aufgrund der Analyse diverser Schemata ein Minimalset von Metadaten vor, das den meisten Ansprüchen genügen sollte und in verschiedenen Metadatenschemata gespeichert werden kann, nämlich MPEG-7, PBCore oder EBUCore.

Weitere Aktivitäten der KOST

KOST-Projekte

Die priorisierten KOST-Projekte 14-001 ViaCar/CARI (Strassenverkehrsamt), 14-017 TAXAR (Steuerdaten und -dossiers), 14-025 EDat (Einwohnerregister) und 14-026 AIS (Referenzmodell) haben die Arbeit aufgenommen. Weitere Interessierte an diesen Themen sind herzlich eingeladen, sich bei der Geschäftsstelle zu melden. Eine Übersicht über weitere potentielle KOST-Projekte gibt wie üblich die letzte Seite dieses Newsletters.

JBIG2-Komprimierung

Die fehlerhafte Implementierung der JBIG2-Komprimierung (die beispielsweise in PDF/A-Dokumenten enthalten sein kann) durch verschiedene Scanner gewinnt immer grössere Aufmerksamkeit. Aktuell hat das deutsche Bundesamt für Sicherheit in der Informationstechnik BSI dieses Verfahren für das Ersetzende Scannen verboten (BSI TR-03138). Die bereits im Herbst 2013 erarbeitete und im August 2014 publizierte Studie der KOST wird in der öffentlichen Diskussion dabei regelmässig zitiert.

eCH-0165, SIARD-Format, Addendum

Bei der ersten Spezifikation des SIARD-Formates war der beim Erstellen eines ZIP-Archivs gängige verlustfreie Komprimierungsalgorithmus *Deflate* noch mit einem Patent belastet. Weil andere Public-Domain-Komprimierungsalgorithmen damals noch keine grosse Verbreitung hatten und in der Archivwelt generell ein Vorbehalt gegenüber Komprimierungsalgorithmen bestand, wurde im SIARD-Format bislang auf die Komprimierung verzichtet. Inzwischen ist aber diese patentrechtliche Einschränkung schon seit einiger Zeit weggefallen.

Die Anwendung von SIARD zur Archivierung von grossen Datenbanken und CSV-Sammlungen hat gezeigt, dass hier durch Datenkomprimierung für die Archive ein enormes Sparpotential brach liegt. SIARD-Dateien komprimieren wie alle XML-basierten Formate unglaublich gut, das heisst in der Regel um den Faktor 9 bis 10. Zudem wird das Handling von SIARD-Dateien im Archivsystem wesentlich vereinfacht, wenn diese um den Faktor 10 kleiner sind: Es werden dann in der Regel keine Dateigrössen-Einschränkungen von Seiten des Betriebs- oder Dateisystem verletzt.

Ein Addendum zu eCH-0165, SIARD-Format, Version 1.0, hebt das Komprimierungsverbot nun auf und ermöglicht es, die Vorteile der verlustfreien Komprimierung mit dem *Deflate*-Algorithmus für SIARD zu nutzen.

csv2siard, Version 1.9

Das KOST-Tool zur Migration von CSV-Dateien ins SIARD-Format gemäss eCH-0165 liegt neu in der Version 1.9 vor. Die neue Version beseitigt die bisherige Grössenbeschränkung und kann nun CSV-Dateien beliebiger Grösse in SIARD umwandeln.

OAIS-Terminologie in TERMDAT

2012 beteiligte sich die KOST an einer nestor-Arbeitsgruppe, die eine deutsche Übersetzung des OAIS (ISO 14721) erarbeitete und publizierte. Diese Arbeit hat die Bundeskanzlei zum Anlass genommen, die OAIS-Terminologie in Deutsch, Französisch, Italienisch und Englisch in die Datenbank TERMDAT einzuarbeiten. Die entsprechenden Einträge sind unter http://www.termdat.ch zugänglich. Bitte verwenden Sie dazu im Suchfeld den speziellen Abfrageparameter ty:oai13.

Newsletter CECO du 1er trimestre 2015

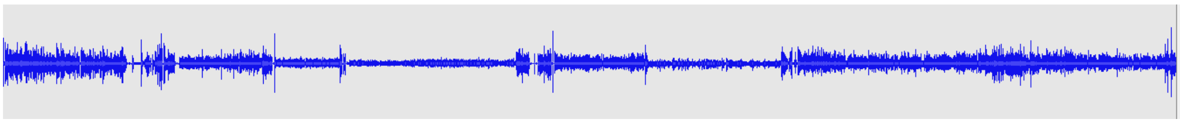
AudioVault

Dans le cadre du projet *AudioVault*, les Archives nationales du Liechtenstein, les Archives d'État de Berne, d'Uri et du Jura ainsi que les Archives des villes de Berne et St-Gall ont examiné les problèmes concrets posés par l'archivage de documents audio et élaboré des approches pour les résoudre. Les archives participantes ont rangé au premier plan trois catégories de documents audio : enregistrements des débats des conseils, enregistrements documentaires et prestations artistiques. La première catégorie leur est clairement apparue prioritaire, raison pour laquelle elle sert d'exemple dans le projet. En complément et en prolongement du projet du CECO *Sauvez les CD*, *AudioVault* offre l'opportunité d'approfondir deux thématiques de l'archivage audio, à savoir le catalogage de matériel audio et la question des métadonnées.

Catalogage

Dans les enregistrements des débats des conseils, les granularités des données primaires et des métadonnées divergent totalement. En général, des métadonnées descriptives détaillées sont disponibles (granularité intervention), mais ne permettent pas d'accès précis aux données primaires, puisque les enregistrements contiennent en principe une séance de conseil entière. Pour y remédier, différentes solutions sont envisageables :

1. Découper le flux audio dans la granularité de l'enregistrement
2. Sauvegarder une référence temporelle (*timecode*) sur le flux audio dans les métadonnées
3. Accéder directement à une intervention à l'aide d'une indication de découpage externe pour le programme de lecture (fichier *cue sheet*)
4. Intégrer une liste de lecture directement dans le conteneur audio



D'un point de vue archivistique la meilleure solution est probablement la troisième avec le fichier *cue sheet* externe accompagnant le fichier audio. Cela évite d'intervenir dans le fichier primaire et permet cependant un certain confort d'accès.

Tout catalogage soulève la question de la mise à disposition en aperçu d'un échantillon dans le catalogue des archives. S'il s'agit d'archives audio protégées par des droits d'auteur, il faut se demander si un bref extrait peut néanmoins être utilisé en vertu du droit de citation. Ce dernier donnant lieu à des discussions contradictoires dans la littérature juridique et une jurisprudence faisant grandement défaut, cette question ne peut être réglée définitivement. L'intégration d'échantillons dans le catalogue ferait courir aux archives un certain risque juridique qui peut toutefois être considéré comme acceptable.

Métadonnées

Outre la problématique juridique évoquée au sujet des échantillons, il ne faut pas oublier que ceux-ci, contrairement à une miniature ou vignette (*thumbnail*) pour des images ou à une extraction d'images clé pour des vidéos, ne constituent pas des résumés ou des aperçus très éloquents et qu'ils ne peuvent être générés automatiquement. La précision de l'indexation des métadonnées est d'autant plus importante pour les archives audio. Hélas, la diversité des standards pour les métadonnées audio est très élevée. Pour des métadonnées intégrées, le choix se porte judicieusement sur des métadonnées BWF et des balises RIFF INFO avec WAV ou IDv3 avec MP3. Pour des métadonnées descriptives externes, *AudioVault* propose, sur la base de l'analyse de différents schémas, un set de métadonnées minimal censé répondre à la plupart des exigences et qui peut être stocké dans différents schémas de métadonnées, notamment MPEG-7, PBCore ou EBUCore.

Autres activités du CECO

Projets du CECO

Les travaux des projets prioritaires du CECO 14-001 ViaCar/CARI (services des automobiles), 14-017 TAXAR (dossiers fiscaux électroniques), 14-025 EDat (registres des habitants) et 14-026 AIS (modèle de référence) ont démarré. Toute autre personne intéressée à ces thèmes est cordialement invitée à s'annoncer auprès du Bureau. Un aperçu des autres projets potentiels du CECO se trouve comme d'habitude à la fin de cette newsletter.

Compression JBIG2

L'implémentation défectueuse de la compression JBIG2 (qui peut être contenue par exemple dans les documents PDF/A) par différents scanners retient de plus en plus l'attention. Actuellement, l'Office fédéral allemand pour la sécurité des techniques de l'information (BSI) a interdit ce procédé pour la numérisation (BSI TR-03138). À ce propos, l'étude du CECO, élaborée en automne 2013 déjà et publiée en automne 2014, est régulièrement citée dans le discours public.

eCH-0165, format SIARD, addenda

Lors de la première spécification du format SIARD, l'algorithme de compression courant et sans pertes *Deflate* utilisé lors de la création d'une archive ZIP était encore protégé par une patente. Comme d'autres algorithmes de compression du domaine public n'étaient à l'époque pas très répandus et que le milieu des archives était en général plutôt réservé à l'égard des algorithmes de compression, on a jusqu'ici renoncé à la compression dans le format SIARD. Cette contrainte juridique liée à la patente a cependant été supprimée entretemps.

L'utilisation de SIARD pour l'archivage de grandes bases de données et de collections CSV a montré que la compression des données représente pour les archives un énorme potentiel d'économie encore inutilisé. À l'instar de tous les formats basés sur XML, les fichiers SIARD compriment incroyablement bien, soit en général à un facteur de 9 à 10. De plus, l'utilisation de fichiers SIARD dans le système d'archivage est très simplifiée s'ils sont plus petits d'un facteur 10. Le procédé n'enfreint alors généralement aucune restriction de taille de fichier émise par le système de fichiers ou d'exploitation.

Un addenda à eCH-0165, format SIARD version 1.0, lève désormais l'interdiction de compression et permet d'utiliser les avantages de la compression sans pertes avec l'algorithme *Deflate* pour SIARD.

csv2siard, version 1.9

L'outil du CECO pour la migration de fichiers CSV en format SIARD conformément à eCH-0165 est maintenant disponible en version 1.9. La nouvelle version élimine la restriction de taille en vigueur jusqu'ici et peut dorénavant convertir des fichiers CSV de n'importe quelle taille dans le format SIARD.

La terminologie OAIS dans TERMDAT

Le CECO a collaboré en 2012 au groupe de travail nestor qui a réalisé et publié une traduction allemande d'OAIS (ISO 14721). La Chancellerie fédérale a profité de ces travaux pour récupérer la terminologie OAIS en allemand, français, italien et anglais dans la base de données TERMDAT. Les entrées correspondantes sont accessibles sur http://www.termdat.ch. Veuillez saisir dans le champ de recherche le paramètre spécifique ty:oai13.

Potentielle KOST-Projekte / Projets CECO potentiels

Die Liste der potentiellen KOST-Projekte listet Projektvorschläge auf, die für eine Betreuung als KOST-Projekt in Frage kommen, und für die weitere Partner gesucht werden (* priorisierte Projektvorschläge).
La liste des projets CECO potentiels répertorie les propositions de projets des archives membres qui entrent en ligne de compte pour devenir un projet CECO et pour lesquels des partenaires additionnels sont recherchés (propositions de projets prioritaires).*

GIGA StAGR	Archivierung der Daten und Dossiers von industriellen Firmen
GIS StAZH	Spezifikation zur Archivierung von GIS Daten *Spécifications pour l'archivage des données GIS*
Kompass3 StAZH	Archivierung der Daten der Berufsbildungsämter *Archivage des données des services de la formation professionnelle*
JURIS offen	Archivierung von Unterlagen der Rechtspflege

Archive, die an der Mitarbeit an einem dieser Projekte interessiert sind, werden gebeten, dies der Geschäftsstelle mitzuteilen (info@kost-ceco.ch).
Les archives souhaitant collaborer à l'un de ces projets sont priées de le communiquer au Bureau (mailto:info@kost-ceco.ch).