Metadaten und Primärdaten

Langzeiterhaltung digitaler Inhalte für das CultLib-Projekt IV

Wir sichern Kulturgut in digitaler Form
und machen es frei verfügbar,
heute und für zukünftige Generationen.

CultLib-ID: d300b0c6-b8f5-49bd-83f5-a36ff56e5896
Version: 1.2
Datum: 24. April 2014
Verfasser: Dr. sc. math. Hartwig Thomas

Übersicht

Das Projekt *CultLib* des Vereins *Digitale Allmend* bezweckt die Gründung einer schweizerischen Stiftung *Pro Cultura Libera* welche ein digitales Repositorium freier, kultureller Werke aufbaut und betreibt.

Die Berichte über Aspekte der Langzeiterhaltung[1] erklären, wie CultLib technisch konzipiert ist und dienen als Anforderungen an die Umsetzung. Ausserdem können sie für Entscheidungsträger und IT-Architekten von Memo-Institutionen von Interesse sein.

Eigentliche Metadaten sind zu unterscheiden von technischen Metadaten und Primärdaten. Diese eigentlichen Metadaten dürfen im Gegensatz zu den Primärdaten während der Archivierung ergänzt und korrigiert werden.

Archivunabhängige Metadaten müssen im Archivgut eingebettet sein und bei der Denormalisierung erhalten bleiben. Pfad und Dateiname des Archivguts gehören nicht zu den Metadaten und können für die Regulierung des Arbeitsablaufs verwendet werden.

Die Gefahren des Semantic Web sollen bei der Festlegung des Metadatensatzes für CultLib vermieden werden.

Die CultLib-Metadaten basieren auf den "Dublin Core" Metadaten und enthalten die UUID des Werks und den Digest der Werkausprägung.

Die Felder der CultLib-Metadaten werden einzeln aufgezählt und beschrieben.

[1] Der Begriff Nachhaltigkeit wird im Zusammenhang mit der Langzeiterhaltung bewusst vermieden, weil die virtuelle Allmend in ihrer Ausdehnung nicht beschränkt ist, und deshalb die mit dem Begriff Sustainability adressierte Übernutzung, die traditionelle Tragödie der Allmend, im Bereich der digitalen Inhalte keine Gefahr darstellt. Das Ziel der Tradierung kultureller Inhalte ist würdig und gewichtig genug und hat nicht nötig, sich auf die Leere übernutzter Warmluft-Begriffe zu stützen.

Inhaltsverzeichnis

1 Was sind Metadaten?

Auf der Titelseite dieses Kurzberichts und in der Kopf- und Fusszeile findet man viele Metadaten: Autor, Publikationsdatum, Titel. In einem Brief steht oft der Absender, der Empfänger, der Ort und das Datum. Viele heute in Archiven[2] angewendete Metadaten-Sätze enthalten Inhaltsverzeichnisse, Kurzbeschreibungen und Schlagwörter, sodass der Eindruck entstehen könnte, das ganze Dokument bestehe eigentlich nur aus Metadaten.

1.1 Metadaten und Primärdaten

Dieser Eindruck täuscht: Im Gegensatz zu Primärdaten sind Metadaten strukturierte Texte, die Angaben enthalten, die nicht in den Primärdaten enthalten zu sein brauchen. Die Tatsache, dass es sich um Texte handelt, ist besonders wichtig für Dokumente, die selber keinen Textcharakter haben: Metadaten können wie andere Texte indexiert und als Basis für Suchabfragen genutzt werden. Ohne Metadaten wäre es schwierig bis unmöglich nach Bildern, Audiodaten oder Videodaten zu suchen. Die Tatsache, dass es sich um strukturierte, beziehungsweise halbstrukturierte Texte handelt, ermöglicht die systematische Erfassung der Metadaten in Datenbanken, welche Abfrage und Suche unterstützen.

In einem gewissen Sinne gehören die Metadaten dem Archiv, bzw. der Nachwelt. In Metadaten werden einige – möglichst für alle Dokumente und Dokumentarten sinnvolle – Eigenschaften des Archivguts in strukturierter Weise festgehalten, die nicht immer in den Primärdaten korrekt verzeichnet sind. Während das Archiv verpflichtet ist, die Primärdaten des Archivguts unangetastet zu lassen, damit die Authentizität garantiert werden kann, darf es die Metadaten eines archivierten Dokuments verändern.

Generell sollten Metadaten einen kleineren Umfang haben als die Primärdaten und keine langen unstrukturierten Texte oder Unterdokumente enthalten.

1.2 Eigentliche Metadaten und technische Metadaten

Der Begriff „Metadaten“ ist in neuerer Zeit überstrapaziert worden. So ist etwa im Zusammenhang mit der internationalen Wirtschaftsspionage durch die Geheimdienste der Welt immer wieder von Metadaten oder Randdaten der Telefonüberwachung die Rede. Im Zusammenhang mit der Langzeiterhaltung unterscheiden wir die eigentlichen Metadaten von sogenannten technischen Metadaten. Technische Metadaten sind Eigenschaften der Primärdaten wie etwa die Anzahl Seiten des Berichts, die Länge einer Tonaufnahme, die Breite und Höhe eines Bildes, die Schriftarten eines Dokuments, der Farbraum eines Bildes etc. Diese Eigenschaften können ohne Änderung der Primärdaten nicht verändert werden und gehören somit eigentlich zu den Primärdaten. Normalerweise sind sie nur für einen einzigen Dokumenttyp sinnvoll.

Es ist durchaus denkbar, dass einige dieser technischen Metadaten auch in die Datenbank des Archivinformationssystems aufgenommen werden und die Suche nach Dokumenten unterstützen. Im Folgenden meinen wir mit dem

[2] Wo im Folgenden von Archiven die Rede ist, sind ganz allgemein Memo-Institutionen wie Bibliotheken, Museen und eigentliche Archive gemeint; das Wort Archivierung wird synonym für Erhaltung oder Aufbewahrung benutzt.

Wort Metadaten aber nur eigentliche Metadaten und zählen die technischen Metadaten zu den Primärdaten. Die eigentlichen Metadaten beziehen sich mehrheitlich auf das Werk und nicht auf die Werkausprägung, während die technischen Metadaten je nach Ausprägung verschieden sein können.

1.3 Metadaten und Verzeichnisdaten

Viele Dateiverzeichnisse enthalten neben den Dateinamen eine ganze Reihe von Verzeichnisdaten wie Erzeugungszeitpunkt, Zeit der letzten Änderung, ID des Benutzers, der für die letzte Änderung verantwortlich ist, Länge der Datei in Bytes und vieles Andere mehr.

Diese automatischen Daten enthalten oft zweifelhafte, beziehungsweise missverständliche Informationen. Wenn ich eine Datei kopiere, um deren Format zu übernehmen, alles lösche, und ein neues Werk schaffe, enthält das automatische Erzeugungsdatum immer noch den Zeitpunkt der Erzeugung der ursprünglichen Datei. Wenn ich einen Brief ausdrucke, erhält er ein neues Datum der letzten Änderung, obwohl nichts verändert wurde.

Aus diesem Grund sind alle hier beschriebenen eigentlichen Metadaten nur manuell änderbar.

Ein Dateiname kann geändert werden, obwohl der Inhalt unverändert bleibt. Dateiname und Pfad dienen also eigentlich der Organisation im Arbeitsablauf und gehören weder zu den Primärdaten noch zu den Metadaten des Archivguts. Sie stellen eine momentane Adresse der Datei dar, die in der Datenbank des Archivinformationssystems (AIS) mit der ID des Werk und dem Digest der Werkausprägung verknüpft wird.

Im weiteren Sinn sind alle auf das Archiv bezogenen Metadaten (zum Beispiel der Name des Servers, des Archivs, die Einordnung in die Archivstruktur) Verzeichnisdaten und haben in den eigentlichen Metadaten nichts verloren, da sie bei der Integration in ein anderes Archiv den Sinn verlieren. Solche Metadaten gehören wie der Dateiname in die Datenbank des Archivinformationssystem (AIS), wo sie nützliche Angaben zur Organisation des Archivs darstellen.

2 Eingebettete Metadaten

In heutigen Archiven ist es üblich, die Metadaten zum Archivgut in einer separaten Datenbank des Archivinformationssystems (dem AIS in der OAIS-Empfehlung) zu speichern. Dieser Ansatz hat verschiedene Nachteile: Einerseits ist es sehr schwierig, die Verbindung zwischen Metadaten und Archivgut über lange Zeit gesichert aufrechtzuerhalten. Immer wieder kommt es vor, dass der Bezug in der Tabelle um eines „verrutscht“ und auf das falsche Dokument zeigt. Andererseits wird der Austausch mit anderen Archiven in einem weltweiten verteilten Repositorium stark erschwert, da jede Datenlieferung von einer synchronen Metadatenlieferung begleitet werden muss. Schliesslich sieht man denormalisierten Ausprägungen des Archivguts nicht mehr an, woher es kommt und welches seine Metadaten sind.

Glücklicherweise erlauben fast alle moderneren Dokumentformate, Metadaten im Dokument einzubetten. Dieser Ansatz wird im Zusammenhang mit Langzeiterhaltung oft empfohlen, aber selten realisiert. CultLib-Metadaten sind immer im Archivgut einzubetten und werden bei jeder Konversion mitgeführt. Sollte die Datenbank des Archivsystems einmal verloren gehen, kann diese jederzeit aus den eingebetteten Metadaten der archivierten Dokumente wieder hergestellt werden. Erhält ein Archiv eine Sendung mit Archivgut von einer anderen Institution, kann es die Datenbank aus den eingebetteten Metadaten ergänzen.

Bei der Normalisierung und Denormalisierung werden die eingebetteten Metadaten transferiert. So stehen sie jedem unmittelbaren oder mittelbaren Empfänger eines Dokuments aus dem Archivgut zur Verfügung und ermöglichen eine klare Referenzierung.

2.1 Normalisierung kopiert eingebettete Metadaten

Falls ein neu angeliefertes Dokument schon CultLib-Metadaten enthält, werden diese bei der Normalisierung in das Archivformat kopiert. Alle anderen Metadaten (z.B.: EXIF-Metadaten) werden im Beschreibungsfeld der CultLib-Metadaten erhalten. Falls das angelieferte Dokument keine CultLib-Metadaten enthält, werden diese im Normalisierungsschritt automatisch angelegt und mit Hilfe vorhandener anderer Metadaten so gut wie möglich initial gefüllt.

2.2 Archiv bearbeitet eingebettete Metadaten

Unmittelbar nach der Übernahme (OAIS: Ingest) eines Dokuments werden die Metadaten ergänzt. Dieser Prozess wurde im ersten Bericht zur Langzeiterhaltung digitaler Objekte für das CultLib-Projekt beschrieben. Nach der Übernahme werden die Metadaten sporadisch korrigiert. Denn während der Aufbewahrungszeit können sich die für ein Werk relevanten Metadaten ändern. Die Lizenz kann wegfallen. Ein Autor eines anonymen oder pseudonymen Werks kann bekannt werden. Weit verbreitete Titel in anderen Sprachen können hinzutreten.

2.3 Denormalisierung kopiert eingebettete Metadaten

Auch bei der Denormalisierung werden die eingebetteten Metadaten in den Byte-Strom integriert, der an den Endnutzer abgeliefert wird. Denn auch die Denormalisierungsformate unterstützen in aller Regel das Einbetten von Metadaten.

2.4 Werkzeuge zur Bearbeitung von eingebetteten Metadaten

Vorläufig können CultLib-Metadaten in normalisierten Dateien nur mit den Werkzeugen von CultLib bearbeitet werden. Der CultLib-Standard ist aber öffentlich publiziert und die CultLib-Metadaten im XML-Format ist vielen anderen Werkzeugen zugänglich.

3 Warum ein weiterer Metadaten-Standard?

Es gibt eine Unmenge von Metadaten-Standards für die Langzeiterhaltung. Man könnte den Eindruck gewinnen, dass Dokumentalisten sich in ihren Standards zu verwirklichen trachten und mehr Aufwand in die Standardisierung stecken als in die Langzeiterhaltung des Archivguts.

Wir haben Metadatenstandards wie PREMIS oder METS verworfen, weil sie uns einerseits zu komplex und andererseits zu flexibel schienen. Wenn ein Standard zu flexibel ist, degeneriert er zu einem Metadaten-Behälter-Standard wie etwa XMP. Auf einem solchen Standard kann keine Datenbank mit Metadaten aufgebaut werden und die Suche unterstützen. Wenn ein Standard zu komplex ist, d.h. zu viele Felder definiert, werden diese Felder nur sporadisch befüllt und es ist oft zweifelhaft, welche Angaben in welches Feld gehören.

3.1 Die Gefahren des Semantic Web

Wir basieren den CultLib-Metadaten-Standard nicht auf dem Resource Description Format (RDF) des World Wide Web Consortium, weil wir nicht an die Segnungen des „Semantic Web“ oder des „Web 3.0“ glauben.

Das „semantische Gewebe“ scheitert an mehreren Fronten. Zum einen sind die Subjekte, Objekte und Prädikate nie genügend genau definiert, dass sich mehrere „Ontologien“ verknüpfen lassen. Zum anderen ist die Beschränkung auf semantische Tripel (Subjekt, Prädikat, Objekt) zwar ungemein beliebt, weil sie sich so schön an der Tafel oder im PowerPoint als beschriftete Striche darstellen lassen, die von einem Punkt zu einem anderen führen. Aber damit werden alle Sätze mit mehr als einem Objekt und somit alle komplexeren Relationen von vornherein ausgeschlossen. So beschränkt die Notation unbewusst, was denkbar ist. Diese für Geisteswissenschaften typische Fixierung auf zweistellige Relationen wird zwar von Weiterentwicklungen wie der Web Ontology Language (OWL) des World Wide Web Consortium behoben. Die Ontologien ufern aber in schöner Regelmässigkeit in – mit religiösem Eifer verteidigte – komplexe Systeme aus. Diese nützen einem einfachen Benutzer nichts, da er sie erst verstehen müsste, um sie bei seiner Suche nach Archivgut nutzbringend anzuwenden. Ausserdem bindet ihre Erzeugung, Pflege und Durchsetzung personelle und maschinelle Ressourcen, die der eigentlichen Archivierung der Dokumente entzogen werden.

Bei Herstellen von Ontologien ist es sehr schwierig, Einigkeit auch nur unter Experten zu erzielen. Selbst sehr eingeschränkte, wissenschaftliche Ontologien – etwa die Taxonomie der Arten der Flora und Fauna – scheitern an der klaren Definition der Begriffe. Und ein „weltweites semantisches Gewebe“ liesse sich ja nur herstellen, wenn zum Beispiel ein scheinbar so klar und eindeutig definierter Begriff wie das Geschlecht eines Menschen in jeder Ontologie gleich oder vergleichbar definiert wäre.

3.2 Hauptzweck von korrekten Metadaten

Der Hauptzweck korrekter Metadaten besteht in der Hilfe für den Benutzer des Archivs bei der Suche nach einem speziellen Dokument.

Wenn es gelingen soll, einigermassen korrekte Metadaten für das Archivgut zu erhalten, dürfen diese nicht zu präzis definiert werden. Ausserdem ist eine relativ kleine Anzahl von Feldern, die dafür auch mehrheitlich ausgefüllt sind, einer grossen Zahl von Hunderten von Feldern vorzuziehen, deren Definition man gar nicht gleichzeitig im Kopf haben kann, und deren Datenqualität somit in keiner Weise erstellt oder überprüft werden kann.

Schliesslich sollten sich korrekte Metadaten auf das abstrakte Werk und nicht auf die Werkausprägung beziehen und einen kleinen Satz von semistrukturierten Texten zur Charakterisierung des Inhalts (der Primärdaten) enthalten.

Mindestens ein Feld muss eine unstrukturierte Beschreibung ermöglichen, wo etwa die Namen von in einem Bild dargestellten Personen oder Zeit und Ort der Aufnahme festgehalten werden können. Diese Beschreibung kann mit Hilfe von Volltextsuche durchmustert werden. Sie sollte nicht zu lang werden und keine Abhandlung über das Werk darstellen.

CultLib-Metadaten haben nie Werkcharakter, sind immer frei zugänglich. Ihre Urheber und Änderungsdaten sind normalerweise nicht in den Metadaten selber verzeichnet. Selbstverständlich steht es einem Archiv frei, Metadaten-Änderungen in einer Datenbank zu verzeichnen.

3.3 Neuer Standard benutzt Werk-UUID

Diesen Anforderungen genügt der ursprüngliche „Dublin Core“ Metadaten-Standard weitgehend. Der einzige Grund, dass wir trotzdem einen eigenen CultLib-Metadaten-Standard definieren, besteht darin, dass wir einerseits Werk-UUID und Ausprägungs-Digest als Metadaten-Felder hinzufügen wollen, und andrerseits eine konkrete XML Schema Definition (XSD) festzulegen wünschen, welche die formale Struktur der CultLib-Metadaten eindeutig festlegt.

4 Die Felder der CultLib-Metadaten

Die CultLib-Metadaten bestehen aus folgenden Feldern:

- Werk-UUID: id
- Ausprägungs-Digest: digest
- Publikationsdatum: publication
- Lizenz: license
- Typ: type
- Titel (mehrspachig): title
- Thema (mehrsprachig): content/subject
- Beschreibung (mehrsprachig): content/description
- Urheber (mehrsprachig): creator
- Mitwirkende (mehrsprachig): contributor
- Original (Ausprägung von): origin/id und origin/digest
- Quelle (Abwandlung von): source/id und source/digest
- Herkunft (mehrsprachig): provenance
- Sprache: language

4.1 Werk-UUID: id

Die UUID des Werks, dessen Ausprägung das Dokument ist, wie sie im dritten Bericht zur Langzeiterhaltung digitaler Objekte für das CultLib-Projekt ausführlich vorgestellt wurde.

4.2 Ausprägungs-Digest: digest

Identifikation der Primärdaten wie sie ebenfalls im dritten Bericht zur Langzeiterhaltung digitaler Objekte für das CultLib-Projekt ausführlich vorgestellt wurde. Streng genommen ist dies ein technisches Metadatum, denn es handelt sich beim Digest um eine Funktion der Primärdaten, der nicht frei manuell eingegeben werden kann. Er kann höchstens neu berechnet, bzw. überprüft werden. Trotzdem behalten wir ihn in den Metadaten bei, weil es zu aufwendig ist, ihn jedesmal neu zu berechnen. Wenn der Wert in diesem Feld mit dem aus den Primärdaten berechneten Wert übereinstimmt, ist dies kein hundertprozentiger Beweis für die Authentizität der Primärdaten. Wer die Primärdaten ändert, kann auch dieses Metadatenfeld geändert haben. Um die Authentizität sicherzustellen, benötigt man einen angemessenen Arbeitsablauf, in welchem der Digest extern (im AIS) gespeichert wird, und seine Änderung durch Unbefugte mit kryptographischen Methoden verhindert wird.

Die vorliegende Metadaten-Definition adressiert keine Fragen der Sicherheit oder des Schutzes vor unbefugten Änderungen. Diese müssen anderweitig behandelt werden.

4.3 Publikationsdatum: publication

In diesem Feld sollte das früheste gesicherte Datum nach der Publikation des Dokuments eingetragen werden. Wenn es fehlt, bedeutet dies, dass das Dokument nie publiziert wurde. Wenn man weiss, dass es publiziert ist, aber nicht weiss wann, kann man wenigstens das aktuelle Datum eingeben. Denn dieses ist sicher nach dem Publikationsdatum. Wenn man nur das Jahr kennt, kann man den 31.12. diese Jahres als das früheste gesicherte Datum nach der Publikation verwenden. Für „irgendwann im 20. Jahrhundert“ ist entsprechend „31.12.1999“ einzugeben.

4.4 Lizenz: license

In diesem Feld sollte eine allfällige von den Urhebern erteilte Lizenz für das allgemeine Publikum eingetragen werden. Zum Beispiel die GNU Public Documentation License (GPDL), eine Creative Commons (CC) Lizenz oder die Feststellung der Gemeinfreiheit (Public Domain, PD). Wenn es fehlt, bedeutet dies, dass die Urheber oder Rechteinhaber keine Lizenz für das allgemeine Publikum gewährt haben und der normale Urheberrechtsschutz uneingeschränkt gilt. Im CultLib-Archiv werden nur Werke aufgenommen, die gemeinfrei sind (PD) oder mit einer der erwähnten Lizenzen dem allgemeinen Publikums freien Zugang gewähren.

4.5 Typ: type

Auch dieses Feld ist streng genommen ein technisches Metadatum, da man den Dokumenttyp – etwa mit Hilfe von JHOVE – aus den Primärdaten bestimmen kann. Für CultLib sind vorläufig nur folgende Typenwerte erlaubt:

picture für Bilddaten

audio für Tondaten

video für Bewegtbilddaten

ooxmla für MS Office Dokumente im OOXML/A Format

odfa für OpenOffice/LibreOffice Dokumente im ODF/A Format

4.6 Titel (mehrsprachig): content/title

Dies ist ein Freitextfeld, das normalerweise ziemlich kurz sein sollte. Es sind Titel in mehreren Sprachen möglich, aber nur einer pro Sprache. Denn es wird möglicherweise in mehreren Sprachen nach Titeln gesucht.

4.7 Thema (mehrsprachig): content/subject

Ein Untertitel, ein Betreff, eine Kurzcharakterisierung. Auch dieses Feld ist kurz und mehrsprachig.

4.8 Beschreibung (mehrsprachig): content/description

Hier darf ein längerer Freitext eingegeben werden, der aber Metadatencharakter haben sollte. Etwa eine Kurzbeschreibung eines Bildes für Sehbehinderte, der Tonaufnahme für Hörbehinderte. Ausserdem zum Beispiel: die Namen aller dargestellten Personen; Ort und Zeit der Aufnahme; Referenzen auf andere Werke, die dem Werk zugrunde liegen, sowie deren Urheber.

Dieses Feld sollte keine Abhandlung über das Werk enthalten. Eine solche ist ein eigenständiges Werk. Metadaten haben keinen Werkcharakter!

4.9 Urheber (mehrsprachig): creator

Dieses Feld wird so oft wiederholt, wie es Urheber gibt. Es enthält den Namen oder das Pseudonym des Urhebers. Das Feld ist mehrsprachig, weil es international verschiedene Schreibweisen eines Namens gibt. Zum Beispiel „Aristophanes“ und „Ἀριστοφάνης“.

4.10 Mitwirkende (mehrsprachig): contributor

Dieses Feld wird so oft wiederholt, wie es bekannte Mitwirkende gibt, die nicht schon in einem Urheberfeld erwähnt wurden. Zusätzlich zum Namen sollte hier auch die Funktion (z.B.: Chorsängerin, Kameramann etc.) erwähnt werden.

4.11 Original (Ausprägung von): origin/id und origin/digest

Original, das der Ausprägung zugrunde liegt. Also beim Archivgut ein Verweis auf die noch nicht normalisierten, übernommenen Rohdaten, bei der Auslieferung einer denormalisierten Ausprägung ein Verweis auf die Masterversion im Archiv.

4.12 Quelle (Abwandlung von): source/id und source/digest

Fast jedes Werk beruht auf einem oder mehreren anderen Werken. Für jedes dieser Werke kann ein Quelleneintrag eingefügt werden.

4.13 Herkunft (mehrsprachig): provenance

Viele Memo-Institutionen tun sich schwer mit der völlig freien Herausgabe ihres Archivguts. Zum Teil beanspruchen sie illegalerweise ein Urheberrecht auf den Inhalten, die sie zur Verfügung stellen. Zum Teil erzwingen sie mit den absurdesten technischen Massnahmen den Besuch ihrer eigenen Website und eine Registrierung, wo man äusserst restriktiven Bedingungen zustimmen muss, welche die freie Publikation – etwa in wissenschaftlichen Untersuchungen zum Archivgut – verbieten. Zum Teil verschandeln sie das Archivgut, das sie öffentlich zur Verfügung stellen mit digitalen Wasserzeichen.

Diese Einschränkungen sind dem eigentlichen Zweck einer Memo-Institution entgegengesetzt und verfolgen eigentlich nur ein Ziel: Dritten, die ein Objekt nicht direkt von der Institution erhalten haben, soll bewusst sein, wem sie es zu verdanken haben. Diese Sichtbarkeit der Institution ist für viele Archive lebenswichtig, weil ihr Budget davon abhängt, ob ihre Leistung öffentlich wahrgenommen wird.

Aus diesem Grund haben wir das Herkunftsfeld hinzugefügt. Dieses kann mehrmals aufgeführt werden, weil ja eine ganze Kette von Institutionen beteiligt sein kann. In den eingebetteten Metadaten wandert es bei fast allen Konversionen mit und ist für den Endbenutzer leicht einsehbar.

Das Feld ist ein Freitextfeld, welches eher kurz die Leistung einer Institution beschreibt. In CultLib ist mindestens das CultLib-Archiv selber hier ver-

zeichnet. In verteilten Repositorien findet man hier auch die ursprüngliche Institution, die das Dokument zugänglich gemacht hat.

4.14 Sprache: language

Dieses Attribut eines Werks ist fakultativ, da es nicht auf alle Werke anwendbar ist. So ist etwa das Musikalische Opfer von Johann Sebastian Bach in keiner Sprache, beziehungsweise in der internationalen Sprache der Musik verfasst. Wenn ein Werk aber in einer konkreten Sprache vorliegt, ist diese Eigenschaft für die Suche von grosser Bedeutung, da es mir wenig nützt, ein Buch, ein Hörspiel oder einen Film in einer Sprache zu beziehen, die ich nicht verstehe.

Es gibt auch mehrsprachige Werke wie etwa ein Film mit Untertiteln in vielen Sprachen. Für solche Werke ist hier die „Hauptsprache“ festzuhalten. Die Mehrsprachigkeit drückt sich dann eher darin aus, dass Inhaltsangaben (Titel, Thema, Beschreibung) in mehreren Sprachen vorhanden sind.

4.15 Warum keine Schlagworte?

Schlagworte oder Tags sind aus folgenden Gründen bewusst nicht in die CultLib-Metadaten aufgenommen.

- Wenn 1900 jemand den Bestand der Zentralbibliothek Zürich beschlagwortet hätte, hätte er kaum je erwähnt, ob Frauen oder Männer in den Büchern vorkommen. Ab 1970 ist dies für die Feministinnen von höchster Bedeutung.
- Mittels Volltextsuche à la Google kommt man schneller und besser zum Ziel. Das Dokument enthält seine eigene Beschlagwortung.
- Wenn im Dokument gewisse wichtige Schlagworte zur Charakterisierung fehlen, kann man sie in der (mehrsprachigen!) Beschreibung eingeben und diese mittels Volltextsuche nutzen.
- Beschlagwortung kostet menschliche Ressourcen, automatische Indexierung – auch eine Form der Beschlagwortung – kostet keine.

Beschlagwortung ist also tendenziell ein Versuch zur Gängelung des Erkenntnisinteresses der zukünftigen Nutzer. Beschlagwortung ist teuer. Indexierung zwecks Volltextsuche ist besser. In CultLib überlassen wir den traditionellen Suchmaschinen das Indexieren.

5 Anhang: Formale XML Schema Definition

Die formale XML Schema Definition der CultLib Metadaten findet man online unter: http://www.cultlib.ch/xmlns/metadata/1.0/cultlib-metadata.xsd

Dies ist ihre initiale Version:

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- cultlib-metadata.xsd ============================================
XML Schema Definition of CultLib metadata.
Version     : $Id: cultlib-metadata.xsd 3811 2014-03-28 15:59:47Z hartwig $
Application: CultLib
Platform    : XML 1.0, XSD 1.0
Description: XML Schema Definition of CultLib metadata.
=====================================================================
Created     : Feb 03, 2014, Hartwig Thomas
Copyright   : 2014, Enter AG, Zurich, Switzerland
License     : http://creativecommons.org/licenses/by-sa/3.0/ch/deed
===================================================================== -->
<xs:schema
  xmlns:xs="http://www.w3.org/2001/XMLSchema"
  xmlns="http://www.cultlib.ch/xmlns/metadata/1"
  targetNamespace="http://www.cultlib.ch/xmlns/metadata/1"
  version="1.0">

  <!--  Simple type definitions -->
  <xs:simpleType name="UuidType">
    <xs:annotation><xs:documentation>
      UUID with optional braces and suitable hyphens inserted in hexadecimal representation.
    </xs:documentation></xs:annotation>
    <xs:restriction base="xs:string">
      <xs:pattern value="([0-9a-fA-F]{8}-[0-9a-fA-F]{4}-[0-9a-fA-F]{4}-[0-9a-fA-F]{4}-[0-9a-fA-F]{12})|(\{[0-9a-fA-F]{8}-
[0-9a-fA-F]{4}-[0-9a-fA-F]{4}-[0-9a-fA-F]{4}-[0-9a-fA-F]{12}\})"/>
    </xs:restriction>
  </xs:simpleType>

  <xs:simpleType name="DigestType">
    <xs:annotation><xs:documentation>
      Message digests with prefix "MD5:", "SHA-1:" or "SHA-3:" followed by hexadecimal or - for the SHA variants - Base64
code.
    </xs:documentation></xs:annotation>
    <xs:restriction base="xs:string">
      <xs:pattern value="(MD5:[0-9a-fA-F]{32})|(SHA-1:[0-9a-fA-F]{40})|(SHA-1:[A-Za-z0-9\+/]{27}=)|(SHA-3:[0-9a-fA-F]
{40})|(SHA-3:[A-Za-z0-9\+/]{27}=)"/>
    </xs:restriction>
  </xs:simpleType>

  <xs:simpleType name="NonEmptyString">
    <xs:annotation><xs:documentation>
      Definition of a non-empty string. It doesn't test for white spaces, though.
    </xs:documentation></xs:annotation>
    <xs:restriction base="xs:string">
      <xs:minLength value="1"/>
    </xs:restriction>
  </xs:simpleType>

  <!--  Enumerations -->
  <xs:simpleType name="DocumentType">
    <xs:annotation><xs:documentation>
      Accepted CultLib document types:

      picture: a PNG file.
      audio: a WAV file in RIFF format.
      video: a video file.
      ooxmla: a document in OOXML/A format.
      odfa: a document in ODF/A format.
    </xs:documentation></xs:annotation>
    <xs:restriction base="xs:string">
      <xs:enumeration value="picture"/>
      <xs:enumeration value="audio"/>
      <xs:enumeration value="video"/>
      <xs:enumeration value="ooxmla"/>
      <xs:enumeration value="odfa"/>
    </xs:restriction>
  </xs:simpleType>

  <!--  Complex type definitions -->
  <xs:complexType name="ContentType">
    <xs:annotation><xs:documentation><![CDATA[
      The language-dependent content metadata of an object in the CultLib repository
      describing the abstract work, not the concrete representation.

      <title>: title of the work,
```

```
      <subject>: subject of the work,
      <description>: free text describing the work,
      @language: ISO language of the content metadata - required.

      Generally: Empty strings within an element are not allowed. Instead, drop the element.
    ]]></xs:documentation></xs:annotation>
    <xs:sequence>
      <xs:element name="title" type="NonEmptyString"/>
      <xs:element name="subject" type="NonEmptyString" minOccurs="0" maxOccurs="1"/>
      <xs:element name="description" type="NonEmptyString" minOccurs="0" maxOccurs="1"/>
    </xs:sequence>
    <xs:attribute name="language" type="xs:language" use="required"/>
  </xs:complexType>

  <xs:complexType name="ReferenceType">
    <xs:annotation><xs:documentation><![CDATA[
      Describes the complete reference of a representation of a concrete document.

      <id>: id of the work (may not be known),
      <digest>: digest of the document (exists always),
    ]]></xs:documentation></xs:annotation>
    <xs:sequence>
      <xs:element name="id" type="UuidType" minOccurs="0" maxOccurs="1"/>
      <xs:element name="digest" type="DigestType"/>
    </xs:sequence>
  </xs:complexType>

  <!--  Root element -->
  <xs:element name="cultlib-metadata">
    <xs:complexType>
      <xs:annotation><xs:documentation><![CDATA[
        The metadata of an object in the CultLib repository.

        <id>: Id of the abstract work - it is not (yet?) a work, if this element is missing,
        <digest>: Id of the concrete document (digest over primary data),
        <publication>: publication date - it is not published, if this is missing,
        <license>: publication license   it is not freely usable, if this is missing.
        <type>: type of document (picture, audio, video, document),
        <content>: list of at least one mandatory title, optional subject, optional description with (ISO) language as
key,
        <creator>: possibly empty list of creators,
        <contributor>: possibly empty list of contributors,
        <provenance>: possibly empty list of persons or institutions that financed or published a work or made it
accessible to the public.
        <source>: optional id and digest of source,
        @language: Language of the abstract work, if applicable.
                   If XSD 1.1 were supported by the XML validator, one should like to add 'inheritable="true"' to it.

        Generally: Empty strings within an element are not allowed. Instead, drop the element.
      ]]></xs:documentation></xs:annotation>
      <xs:sequence>
        <xs:element name="id" type="UuidType" minOccurs="0" maxOccurs="1"/>
        <xs:element name="digest" type="DigestType"/>
        <xs:element name="publication" type="xs:date" minOccurs="0" maxOccurs="1"/>
        <xs:element name="license" type="NonEmptyString" minOccurs="0" maxOccurs="1"/>
        <xs:element name="type" type="DocumentType"/>
        <xs:element name="content" type="ContentType" minOccurs="1" maxOccurs="unbounded">
          <xs:key name="contentKey">
            <xs:selector xpath="."/>
            <xs:field xpath="@language"/>
          </xs:key>
        </xs:element>
        <xs:element name="creator" type="NonEmptyString" minOccurs="0" maxOccurs="unbounded"/>
        <xs:element name="contributor" type="NonEmptyString" minOccurs="0" maxOccurs="unbounded"/>
        <xs:element name="origin" type="ReferenceType" minOccurs="0" maxOccurs="1"/>
        <xs:element name="source" type="ReferenceType" minOccurs="0" maxOccurs="unbounded"/>
        <xs:element name="provenance" type="NonEmptyString" minOccurs="0" maxOccurs="unbounded"/>
      </xs:sequence>
      <xs:attribute name="language" type="xs:language" use="optional"/>
      <xs:attribute name="version" type="xs:token"/>
    </xs:complexType>
  </xs:element>

</xs:schema>
```

6 Anhang: Beispiel

Ein einfaches Beispiel für eine CultLib Metadaten-Datei illustriert das intendierte Versionierungskonzept:

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<cm:cultlib-metadata
  xmlns:cm="http://www.cultlib.ch/xmlns/metadata/1"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.cultlib.ch/xmlns/metadata/1
                      http://www.cultlib.ch/xmlns/metadata/1.0/cultlib-metadata.xsd"
  version="1.0"
  language="de">
  <id>{550e8400-e29b-41d4-a716-446655440000}</id>
  <digest>MD5:57edf4a22be3c955ac49da2e2107b67a</digest>
  <publication>2014-02-03</publication>
  <license>http://creativecommons.org/licenses/by-sa/3.0/ch/deed</license>
  <type>picture</type>
  <content language="en">
    <title>Metadata XML example</title>
    <subject>CultLib</subject>
    <description>This XML does not belong to a real object in the CultLib repository but serves as an example documenting
the metadata structure.</description>
  </content>
  <content language="de">
    <title>Metadata XML Beispiel</title>
    <subject>CultLib</subject>
    <description>Diese XML-Datei beschreibt nicht die Metadaten eines Objekts im CultLib Repositorium, sondern dient der
Illustration der Metadaten-Struktur.</description>
  </content>
  <creator>Hartwig Thomas, hartwig.thomas@enterag.ch</creator>
  <creator>Sigrun Zink</creator>
  <contributor>Hansjörg Zürcher</contributor>
  <provenance>http://www.cultlib.ch</provenance>
  <!-- no "source" -->
</cm:cultlib-metadata>
```