Magglinger Rechtsformationsseminar 2010

CHDocML

Data Factory AG, 8057 Zürich

Hubert Münst
01.06.2010

CHDocML: Zweck, Abgrenzungen

Einleitung

CHDocML ist ein Projekt des SVRI (Schweizerischer Verein für Rechtsinformatik). Es bezweckt eine strukturierte Aufbereitung und Darstellung juristischer Dokumente in Form eines XML-Schemas. CHDocML ist auf dokumentarische Bedürfnisse ausgerichtet. Es definiert die Struktur juristischer Texte (ohne Normen [werden abgedeckt durch CHLexML] und Entscheide [werden abgedeckt durch CHDecML]): Lehrbücher, Monographien, Kommentare, Zeitschriften-Artikel, Zeitungsartikel, Beiträge in Festschriften oder Sammelbänden, Online-Publikationen usw. Dabei geht es nur (aber immerhin) darum, den Inhalt eines solchen Dokuments zum Zweck der besseren Erschliessung zu gliedern und eine brauchbare Darstellung am Bildschirm zu ermöglichen. Eine solche Wiedergabe auf elektronischem Weg wird nur Informationsbedürfnisse befriedigen, nie aber versuchen, das Aussehen einer Originaldarstellung (beispielsweise einer gedruckten Version) auch nur annähernd zu reproduzieren. Zudem soll CHDocML jene Angaben liefern, die für ein hinreichendes und eindeutiges Zitieren nötig sind. CHDocML deckt demnach nicht die Bedürfnisse von Bibliothekaren ab; es liefert zu wenig und zu wenig genaue Informationen für eine fachgerechte Katalogisierung. Es soll vielmehr dem Nutzer elektronischer Dokumente deren besseres Handling (Erschliessung, Weiterverarbeitung, verbesserte und automatisierte Recherchemöglichkeit…) und zudem einen Mehrwert im Umgang mit Papierdokumenten liefern. Und es soll so offen sein, dass es Erweiterungen zum Zweck der Drucklegung zulässt; damit wäre es auch als Informationsträger für Produzenten (Verlage, Herausgeber) verwendbar.

Die nachfolgenden Gedanken stellen nicht abschliessende Feststellungen, sondern Thesen dar, die der Diskussion dienen sollen. Wenn über den Zweck und das Ziel Einigkeit besteht, kann in einem nächsten Schritt eine entsprechende logische Struktur für die Beschreibung eines Dokuments vorgeschlagen werden.

Zweck und Ziele von CHDocML im Einzelnen

- (Automatisierte) Austauschbarkeit von Dokument-Beschreibungen, mitsamt Inhalt (so vorhanden), zwischen unterschiedlichen Software-Systemen
- Möglichkeit der systematischen Aufbewahrung/Archivierung von Dokumenten
- Suchmöglichkeit über ein oder mehrere Dokumente
- Wissensmanagement
- Möglichkeit, die Metadaten (und, sofern vorhanden, den Inhalt) des Dokuments auf beliebige Weise formatiert zu speichern und weiterzuverwenden, z.B. auf elektronischem Weg (am Bildschirm) sichtbar zu machen als reiner *Content*, ohne Formatierung oder Layout, unabhängig von vorgegebenen Dateiformaten; es soll also möglich sein, ein Dokument entweder nur zu beschreiben (Metadaten) oder zusätzlich auch den Inhalt des Dokuments zu erschliessen und zu speichern.
- eindeutige Benennung eines elektronischen Dokuments, um es in Informationssystemen adressierbar zu machen
- sprachunabhängige Beschreibung von Dokumenten via strukturierte Metadaten (standardisierte Angaben zu Typ, Thema, Publikation, usw.)
- für Dokumentationszwecke hinreichend detaillierte Strukturierung von Unterlagen, damit *Inhalt* automatisiert erschlossen und verarbeitet werden kann
- Möglichkeit, Tabellen, Medienobjekte und Formeln im Text darzustellen oder als Anhang zu referenzieren.
- Möglichkeit, folgende Schriftarten darzustellen (*kursiv*, **fett**, normal, <u>unterstrichen</u>, $^{hoch}/_{tief}$ - gestellt; sie können den Inhalt verdeutlichen) eindeutige Referenzierbarkeit eines *Dokuments* nach den in der Schweiz üblichen wissenschaftlichen Zitierregeln

- logische Verknüpfung mit anderen Dokumenten
- Verknüpfung mit Anhängen (attachments)
- Kompatibilität mit anderen eCH-Standards (CHLexML, CHDecML): nach Möglichkeit gleiche Benennung von Elementen bzw. Attributen, Übernahme vordefinierter Strukturen

Was gehört *nicht* zu den Zielen?

- Bibliographische Erfassung für die Bedürfnisse von Bibliotheken (Katalogisierung)
- eindeutige Adressierung eines bestimmten physischen (gedruckten) Dokuments
- Lokalisierung des physischen Standortes einer Unterlage (Bibliothek, Signatur)
- Erstellen einer Grundlage für den gepflegten Ausdruck (PDF oder Papier); dies ist natürlich möglich durch die nachträgliche Verwendung geeigneter Formatvorlagen oder Stylesheets.
- Beschreibung unterschiedlicher Ausgaben ein und desselben Dokuments (verschiedene Auflagen, verschiedene Sprachen, gebunden/broschiert usw.)
- Qualifikation (Bewertung) der Unterlage
- Aussagen zu Verfügbarkeit/Preis der Unterlage
- Aussagen zur physischen Beschaffenheit der Unterlage (Grösse in cm, Farbe, Zustand, Vollständigkeit)
- Verhinderung von Redundanz bei Autoren, Verlagen, Publikationsreihen
- Darstellung von Normen (siehe CHLexML) oder Entscheiden (siehe CHDecML)
- Unterstützung von Systemen zur Rechte-Verwaltung (DRM)
- Basis für Redaktionssysteme
- Wiederverwendung von *Teilen* von Dokumenten

→ Manche dieser Punkte können weiterführende Systeme aufgreifen und realisieren; sie gehören aber nicht zu den eigentlichen Zielen von CHDocML.

Was gehört nur bedingt zu den Zielen?

- Formal exakte Wiedergabe von Titel oder Inhalt (inkl. Ligaturen, Leerstellen, Schriftarten [Times, `Courier`, Arial usw.], Schriftgrössen, Kapitälchen, Seitenumbruch, Layout); Darstellung in einem Originalformat. Solche Angaben sind für die Dokumentation nicht interessant, *für den Hersteller bzw. Verlag hingegen vital*. Optimal wäre es, wenn sich das zu entwickelnde Schema für den gesamten Produktionsprozess von der Erfassung durch den Autor über die Drucklegung bis hin zur Einspeisung in ein Dokumentationssystem verwenden liesse. Dies kann möglicherweise erreicht werden mit der Erweiterbarkeit durch nicht vordefinierte Strukturen, die produktionsspezifische Informationen transportieren (sog. Custom- bzw. Blackbox-Elemente im `xs:anyType`-Format).

Zielpublikum

- inhaltlich: Redaktoren und Leser von juristischen Texten
- technisch: Architekten und Entwickler von juristischen Dokumentationssystemen, Verleger, Hersteller, Autorensysteme

Mögliche Unterlagen, die erschlossen werden können:

Art	*Metadaten*	*Wiedergabe des Inhalts möglich?*
Gedruckte Texte	ja	ja, nach Konvertierung z.B. in XML/PDF
Elektronische Dokumente	ja	ja
Bilder, Photographien	ja	ja
Geographische Karten	ja	ja
Bewegte Bilder	ja	nein
Multimedia	ja	nein
Tonfolgen	ja	ja
Musiknoten	ja	ja

Anstelle der Wiedergabe des Inhalts innerhalb einer CHDocML-Instanz kann der Verweis auf eine Fundstelle treten, an der die betreffende Unterlage gefunden werden kann (Beilage, URL).

Metadaten

Welche Angaben sind (zusätzlich zum ev. Inhalt) für die dokumentarische Erschliessung interessant?

- Autor(en)
- Herausgeber
- Titel
- Sprache (nur solche mit lateinischem Alphabet, also ohne kyrillisch, arabisch, hebräisch, chinesisch usw.)
- Klassifizierung (amtliche Geheimhaltung)
- Angaben zum Datenschutz (Dokument könnte schützenswerte Daten enthalten)
- Ausgabe
    - Typ (Einzelbuch, Artikel in Sammelband (z.B. Festschrift), Zeitschrift, Zeitungsartikel, URL [=Internetveröffentlichung] usw.)
    - Verlag, Name der Zeitschrift bzw. der Zeitung (Abkürzung möglich: SJZ, NZZ, ZSR), einsprachig, ohne Übersetzungen
    - vollständiges Zeitschriftenzitat in üblicher Form (Beispiel: ZBJV 144/2008 S.514)
    - Auflage (bei Büchern)
    - Publikationsort (geographisch)
    - Datum (bei Büchern: nur Jahr; bei URL exaktes (Stand-)Datum)
    - exakter Verweis: von Seite … bis Seite (bei Büchern, die mehrere Dokumente enthalten, z.B. Festschriften), und/oder Randnummer(n)
    - ISBN|ISSN-Nummer

Je nach Typ sind mehr oder weniger Editionsangaben möglich bzw. sinnvoll. Ein URL ist immer möglich, entweder als einzige Angabe zum Fundort oder als Ergänzung.

Welche Arten von Angaben werden *nicht* erfasst und abgelegt?

- Status des Dokuments (immer definitiv)
- Eigentümer/Unterzeichner

Dokument

Ein Dokument kann beispielsweise sein:

- ein Buch
- ein Artikel in einer Zeitung oder Zeitschrift
- ein Artikel in einem Buch (Sammelbände, Festschriften)
- amtliche Publikation (z.B. Botschaft des Bundesrates, publiziert im Bundesblatt)
- eine elektronische Publikation im Internet

Das Dokument kann in- oder ausländisch, neu oder alt, aktuell oder überholt sein. Inwieweit der *Inhalt* eines Dokuments elektronisch erschlossen wird, bleibt dem Entscheid des Redaktors bzw. Dokumentalisten überlassen.

Erschliessung des Inhalts

Hier geht es darum, unterschiedliche Textteile zu qualifizieren und zu bezeichnen. Juristische Texte haben keine allgemein definierbare Struktur; sie folgen dem Gedankengang des Autors. Hingegen finden sich häufig ein Inhaltsverzeichnis oder Zwischentitel, die eine Gliederung (und damit den Überblick und die Verwendbarkeit) erleichtern. Wichtig sind Zitate und Verweise im Fliesstext:

- auf andere Unterlagen
- auf Normen
- auf Entscheide
- auf andere Passagen im selben Text (Anker für XLink)

Welche Teile eines Dokuments werden nicht übernommen in CHDocML?

- Inhaltsübersicht [auf die obersten Titel-Ebenen reduziertes Inhaltsverzeichnis]: fällt ersatzlos weg, lässt sich aus dem Inhaltsverzeichnis generieren
- Inhaltsverzeichnis: wird automatisch generiert aufgrund der Zwischentitel; verweist nicht mehr auf Seitennummern, sondern auf Dokumentabschnitte
- Liste der Abbildungen: wird automatisch generiert aufgrund der im Text gefundenen Abbildungen
- Alphabetischer Index: wird ersetzt durch Online-Suche

Alle diese Originaldokument-Teile können natürlich in Form eines PDF dem Dokument (als Anhang) beigefügt werden. Manchmal wirkt ein Blick auf das Original-Inhaltsverzeichnis oder auf den Index erhellend; deshalb bietet amazon.com bei einzelnen Büchern eine solche Option im Internet ja auch an.

Standards

Auf diesem Gebiet sind folgende Standards interessant:

- DocBook[1]
- DITA[2]
- Dublin Core[3]
- MARCXML[4]
- MADS[5]
- MODS[6]
- URN-Handbuch der Schweizerischen Nationalbibliothek[7]

Zitierweise im deutschen Sprachraum Raum: Peter Moesgen, Wissenschaftliches Zitieren[8]

USA: Long Island University: MLA Citation Style[9].

Offene Fragen

1. Unklar ist bislang, wie die eindeutige Identifikation eines Dokuments aussehen könnte. Bei Büchern bietet sich die ISBN-Nummer an. Bei Zeitungs- und Zeitschriftenartikeln, bei einzelnen Beiträgen in Sammelbänden oder Festschriften ist die Frage schwieriger zu beantworten. Bei Dokumenten, die nur elektronisch verfügbar sind, bietet sich ein Hyperlink an, der aber im Laufe der Zeit erfahrungsgemäss ändern kann. Die Nationalbibliothek (NB) teilt bestimmten digitalen Dokumenten einen URN (Uniform Resource Name) zu. Inwieweit dieser Ansatz in den hier interessierenden Fällen funktioniert, muss abgeklärt werden; URNs werden nur vergeben für Objekte, die in der NB langzeitarchiviert oder auf Dokumentservern im Hochschulbereich Schweiz mit der Perspektive auf Langzeitarchivierung verwaltet werden.

2. Der erste Entwurf eines XML-Schemas in der herkömmlichen Art liegt vor. Er bildet eine Struktur ab, die die bisher zusammengestellten Informationen enthält. Daneben wird eine Variante untersucht, bei der statt eines spezifisch für die juristische Dokumentation geschaffenen XML-Schemas ODF[10] die Grundlage für einen fachspezifischen Standard liefern würde. Dies hätte den Vorteil, dass Inhalt, Formatierung und Metadaten in einem Dokument zusammengefasst und mit den heute üblichen Texteditoren bearbeitet werden könnten. Ob dies gelingt, steht noch nicht fest und ist Gegenstand weiterer Abklärungen.

[1] http://de.wikipedia.org/wiki/DocBook

[2] http://de.wikipedia.org/wiki/Darwin_Information_Typing_Architecture

[3] http://dublincore.org/

[4] http://www.loc.gov/standards/marcxml/

[5] http://www.loc.gov/standards/mads/

[6] http://www.loc.gov/standards/mods/

[7] http://www.nb.admin.ch/nb_professionnel/01693/01695/01706/index.html?lang=de#sprungmarke0_5

[8] http://www.moesgen.de/pmoezit.htm

[9] http://www.liu.edu/CWIS/CWP/library/workshop/citmla.htm

[10] ISO 26300. OASIS Open Document Format for Office Applications: international genormter quelloffener Standard für Dateiformate von Bürodokumenten (http://de.wikipedia.org/wiki/OpenDocument). Wird u.a. unterstützt von MS-Word, Open Office, TextEdit (Apple), Scribus, Lotus Notes, WordPad. ODF organisiert sich intern in Form mehrerer XML-Dateien, die zu einem gezippten File zusammengefasst werden.

Anhang

Regeln für (deutschsprachige) Quellenangaben[11] geben Anhaltspunkte für die zu definierenden Elemente:

a) Zitieren aus Büchern

1. Vorname des Verfassers, so dass keine Verwechslungen möglich sind
2. Familienname des Verfassers; ist kein Verfasser angegeben, dann „o.V." = *ohne Verfasserangabe*; bis zu drei Verfasser werden jeweils komplett ausgeschrieben, bei mehr als drei Verfassern sind nach dem Erstautor die Abkürzungen „u. a." oder „et al." üblich (z. B. „Theisen et al. 2004")
3. Titel des Buches
4. Auflage
5. Verlagsort; bei mehr als drei Verlagsorten wird – wie bei den Verfassern – zumeist abgekürzt
6. Verlagsjahr; ist kein Verlagsjahr angegeben, dann „o.J." = *ohne Jahresangabe*
7. Seitenangabe; erstreckt sich die zitierte Stelle über die folgende Seite, so ist dieses mit dem Zusatz „f." zu kennzeichnen. Erstreckt sie sich über mehrere folgende Seiten, so ist der Zusatz „ff." notwendig

b) Zitieren aus Zeitschriftenaufsätzen

1. Vorname des Verfassers, so dass keine Verwechslungen möglich sind
2. Familienname des Verfassers
3. Titel des Aufsatzes
4. Name der Zeitschrift = „in"
5. Nummer des Jahrgangs
6. Nummer des Bandes
7. Seitenangabe

c) Zitieren aus Zeitungsartikeln

1. Vorname des Verfassers, so dass keine Verwechslungen möglich sind
2. Familienname des Verfassers; fehlen 1. und 2., dann Signatur angeben; ansonsten wie bei b) bearbeiten

d) Zitieren aus Sammelwerken

1. Vorname des Verfassers, so dass keine Verwechslungen möglich sind
2. Familienname des Verfassers
3. Titel des Aufsatzes
4. Titel des Sammelwerkes = „in"
5. Name des Herausgebers = „Hrsg. …"
6. Auflage
7. Verlagsort
8. Verlagsjahr
9. Seitenangabe

Anmerkungen und Anregungen zu diesem Thema sind erwünscht und werden gerne diskutiert.
Kontaktadressen: info@datafactory.ch oder an den SVRI via urspaul.holenstein@bj.admin.ch

[11] http://de.wikipedia.org/wiki/Zitieren