



Werkzeuge zur Nutzung von RDF

Bei der Präsentation von Werkzeugen zum Umgang mit RDF geht es zunächst darum, exemplarisch die Probleme, die existierende Dienste zum Retrieval von mit Metadaten angereicherten Internetdokumenten aufwerfen, zu diskutieren. Eines der hier auftretenden Probleme ist die Benutzung unterschiedlicher Metadaten Formate. Wir konzentrieren uns hier auf den Bereich von Personendaten und stellen den Internetstandard vCard vor.

Abschließend wird ein Werkzeug vorgestellt, das die Erstellung von Metadaten in RDF unterstützt und die Personendaten im vCard Profil speichert.

TheO

Der Suchdienst TheO (Theses Online) ermöglicht die Recherche in elektronisch veröffentlichten Dissertationen in der Bundesrepublik Deutschland. Momentan wird das elektronische Angebot von 31 Bibliotheken indexiert, von denen 15 Dublin Core Metadaten anbieten.

Da die Kodierung der Metadaten mit Hilfe des HTML `<Meta>` Tags erfolgt, lassen sich nur sehr flache Datenstrukturen repräsentieren. Insbesondere stellt die unzureichende Darstellung von Personendaten für das Retrieval ein Problem dar. Zum Beispiel liefert die Anfrage `Autor="Husmann"` das folgende Resultat (gekürzt):

<i>Titel:</i>	Ein Triangulierungsverfahren zur Approximation mit Dahmen-Micchelli-Seidel-Splines
<i>Autor:</i>	Husmann, Markus
<i>Gutachter:</i>	Gonska, Heinz H.; Prof.Dr. Zhou, Xinlong; Priv.Doiz.Dr.

Es ist erst nach einiger Überlegung zu erkennen, wer von den Gutachtern Professor und wer Privatdozent ist. Zusätzliche Angaben wie Email Adresse oder Homepage lassen sich mit der HTML Kodierung nicht realisieren.

Dieses Beispiel macht deutlich, dass für das ein qualitativ hochwertiges Retrieval ein möglichst anerkannter Standard für die strukturierte Beschreibung von Personendaten benötigt wird.

Speicherung von Personendaten

Weltweit gibt es viele verschiedene Standards zur Speicherung von Personendaten. Ein Beispiel dafür liefern die verschiedenen Dialekte von MARC. Dies führt zu Problemen beim Zusammenführen verschiedener Datenpools oder bei der Speicherung von Personendaten, die aus dem Internet gewonnen werden.

Um dieses Problem zu lösen, gibt es einerseits die Möglichkeit, die Entwicklung von Krosskonkordanzen zwischen den einzelnen Formaten zu nutzen. Leider steigt die Anzahl der benötigten Konkordanzen quadratisch in Abhängigkeit von der Anzahl der

Formate. Jedes neue Format verlangt damit einen sehr großen Arbeitsaufwand durch den Bedarf an neuen Krosskonkordanzen.

Eine weitere Möglichkeit ist die Entwicklung eines Superformates, in das jedes Format übersetzt werden kann. Versuche zur Entwicklung eines solchen Superformates sind bisher allerdings immer gescheitert.

Position des W3C

Das World Wide Web Consortium (W3C) hat dieses Problem erkannt und durch die Request for comments (RFC) 2425 und 2426 den vielen Formaten für Personendaten ein weiteres hinzugefügt.

Dieses neue Format hat allerdings gegenüber vielen anderen Formaten einen sehr großen Vorteil: Es ist relativ unabhängig von kulturellen und sprachspezifischen Besonderheiten, die es auf der ganzen Welt gibt. Diese Besonderheiten, wie Buchstaben, die nicht aus dem 26er Alphabet stammen, oder Namen, die nicht in Vor- und Nachnamen aufgeteilt sind, führen in klassischen Formaten zu Transliterationen, die für Menschen aus anderen Sprachräumen (z.B. s-Laute aus Osteuropa) nicht verständlich sind.

Ziel von RFC2425 ist die Standardisierung von Adressenlisten. Aufbauend darauf definiert RFC2426 mit vCard3.0 ein Format für elektronische Visitenkarten von Personen. RFC2426 ist nach Definition nicht für die Speicherung von Organisationen anwendbar.

vCard (RFC 2426)

Hier eine Liste von Attributen, die mit vCard3.0 kodiert werden können. Einige dieser Attribute müssen in einer vCard vorkommen, andere sind optional.

- Name, der die Person eindeutig identifiziert (Prefix, Suffix, Titel)
- formatierte Version des Namens zur Präsentation
- strukturierte und präsentierbare Form der Adressen (Dienst- und Privatadresse)
- Telefon- und Faxnummern
- Information über Firmen- oder Organisationszugehörigkeit
- Geburtsdatum
- Bilder: Photo oder Logo der Organisation
- Audioinformation z.B. zur Präsentation der Aussprache des Namens
- Längen- und Breitengrad, die in Relation zu der Person stehen
- Emailadresse
- URL einer Homepage
- public key Information
- Zeit der letzten Änderung / Version
- Künstlername/Spitzname

Eine vollständige Liste der Attribute findet sich im RFC 2426 selbst. Der RFC baut auf anderen RFCs auf und garantiert dadurch gute Verträglichkeit zu anderen Internetkonventionen.

Die Kodierung von vCard3.0 erfolgt in ASCII Dateien, dies garantiert Portabilität und einfachen Transport per Email oder anderen Internetprotokollen.

Hier ein kurzes Beispiel einer vCard:

Beispiel

```
BEGIN:vCard
VERSION:3.0
FN:Judith Plümer
N: Plümer;Judith;;Dr.
SORT-STRING: Pluemer Judith
ORG: Universität Osnabrück
REV: 2000-06-01
TEL;TYPE=VOICE,MSG,WORK:+4-541-969-2526
EMAIL;TYPE=INTERNET,PREF:judith@work
EMAIL;TYPE=INTERNET:judith@home
END: vCard
```

Die Zeilen BEGIN, VERSION und END rahmen die vCard ein. Unter FN steht der Name, wie er z.B. auf einer Webseite präsentiert wird. Unter N steht eine detaillierte Liste der Namenskomponenten in der Reihenfolge Nachname, Vorname, zusätzlicher Name, Prefix, Suffix. Unter SORT-STRING kann festgelegt werden, wie der Name in einer alphabetischen Liste einsortiert werden soll. ORG kennzeichnet die Organisation, für die die Person arbeitet und REV das Datum, an dem die vCard zuletzt geändert wurde.

Hinter TEL verbirgt sich eine Telefonnummer (VOICE), die zum Arbeitsplatz gehört (WORK) und über die Möglichkeit zum hinterlassen einer Nachricht verfügt (MSG). Als letztes werden zwei Internet-Emailadressen angegeben, wobei die Person bevorzugt die Email an der ersten Adresse empfängt.

Weitere wichtige Elemente einer vCard sind binäre Daten, die nach festen Kodierungsmechanismen (RFC2047) in eine ASCII Form überführt werden. Dadurch wird der Transport von Bilder, Audio- und Videodaten möglich.

RDF vCard

Renato Iannella hat eine Kodierung von vCard in RDF vorgeschlagen. Sie stellt keine separate Definition von vCard dar, sondern ermöglicht den Transport von vCard im RDF Datenmodell. Da wir uns schon eingehend mit RDF beschäftigt haben, ist es nicht mehr nötig, die RDF Kodierung des obigen (nun etwas verkürzten) Beispiels zu kommentieren:

```
<? xml version="1.0" ?>
<RDF xmlns=http://w3.org/TR/WD-rdf-syntax#
      xmlns:vCard=http://imc.org/vCard/3.0#>
<Description>
  <vCard:FN>Judith Plümer</vCard:FN>
  <vCard:N rdf:parseType="Resource">
    <vCard:Family>Plümer</vCard:Family>
```

```

    <vCard:Given>Judith</vCard:Given>
  </vCard:N>
  <vCard:SORT-STRING>Pluemer Judith </vCard:SORT-STRING>
  <vCard:TEL rdf:parseType="Resource">
    <rdf:value>+4-541-969-2526</rdf:value>
    <rdf:type resource="http://imc.org/vCard/3.0#work">
    <rdf:type resource="http://imc.org/vCard/3.0#voice">
    <rdf:type resource="http://imc.org/vCard/3.0#msg">
  </vCard:TEL>
  <vCard:EMAIL>
    <rdf:Alt>
      <rdf:li>
        <value>judith@work</value>
        <type resource=".../vCard/3.0#internet">
        <type resource=".../vCard/3.0#pref">
      </rdf:li>
      <rdf:li>
        <value>judith@home</value>
        <type resource=".../vCard/3.0#internet">
      </rdf:li>
    </rdf:Alt>
  </vCard>
</Description>
</RDF>

```

Die Praxis

MMM3.0 dient zur Generierung von Dublin Core Metadaten für mathematische Preprints. Die Metadaten werden in RDF unter Nutzung von vCard kodiert. Der Nutzer von MMM3.0 gibt dabei die betreffenden Daten in ein Forminterface ein und muss sich um die konkrete, syntaktisch korrekte Kodierung nicht kümmern.

Das Werkzeug MMM3.0 ist in Perl5 unter intensiver Nutzung des CGI Perl-Moduls. MMM3.0 ist im Internet frei nutzbar, und die Programmquellen sind frei zugänglich. Ausgehend von MMM3.0 können für andere Dokumententypen Werkzeuge zur Generierung von RDF programmiert werden.

Literatur

1. MMM3.0: <http://www.mathematik.uni-osnabrueck.de/cgi-bin/MMM3.0.cgi>
2. MMM3.0 Quellen: <ftp://ftp.mathematik.uni-osnabrueck.de/pub/MMM/>
3. TheO: <http://www.iuk-initiative.org/iwi/TheO/>
4. F. Dawson, T. Howes: RFC 2425. <http://www.imc.org/rfc2425>
5. RFC 2426: <http://www.imc.org/rfc2426>
6. Renato Iannella: Representing vCard in RDF.
<http://www.dstc.edu.au/Research/Projects/rdf/draft-iannella-vcard-rdf-00.txt>