



## Auf der Suche nach Brücken zwischen den Webs

Anders sieht es im Moment noch auf dem Gebiet des Webs der Daten aus. Das Web der Daten, auch Semantisches Web genannt, bildet im Moment noch eine kleine Insel mit wenigen Brücken zum Web der Menschen. Vielleicht ist auch das ein Grund, warum Bibliotheken als benutzerorientierte Einrichtungen wenig in diesen Bereich investieren. Dies gilt aber nicht nur für Bibliotheken, die wohl genug Daten haben, um das Web der Daten der Lösung des Henne-Ei-Problems – der Frage, ob erst genügend Daten da sein müssen, damit das Semantic Web funktioniert und seine Vorteile zeigen kann oder ob die Daten in großem Maße erst geliefert werden, wenn das Semantic Web als technische Infrastruktur vorhanden ist – einen großen Schritt näher zu bringen, sondern auch für andere Einrichtungen, die die Vorteile eines Webs, das nicht primär zum Konsum durch Menschen geeignet ist, noch erkennen müssen.

Die Referenten des W3C Thementages 2009 versuchten unter dem diesjährigen Motto diesem Phänomen, dem Nichterreichen einer kritischen Masse von Daten, die durch RDF-Triple beschrieben werden, entgegenzutreten, indem sie die aktuellen Ergebnisse der Standardisierungsgremien des World Wide Web Consortiums vorstellten und, soweit dies möglich war, die Brücken zwischen den Webs, zum Teil sogar als Ziel des Standardisierungsprozesses, darstellten.

Nicht alle Vorträge konnten jedoch auf die drei zugrunde liegenden Fragen eingehen, weil sie von ihrem Ansatz eher aus der Organisationsperspektive (Vorstellung des Semantic Technology Institute (STI)<sup>8</sup> durch Prof. Dr.-Ing. Robert Tolksdorf) vorgetragen wurden oder rein dem Themengebiet des Webs der Menschen zuzuordnen waren (die deutsche Übersetzung der Web Content Accessibility Guidelines (WCAG) 2.0<sup>9</sup>). Dies soll die Qualität der beiden Vorträge nicht schmälern, trugen sie doch zu einem vollständigen Bild dessen bei, was das Umfeld und die Arbeit des W3C auszeichnet.

Für diejenigen im öffentlichen Dienst und somit auch einem Großteil derer, die sich in Bibliotheken mit dem Erstellen von Webseiten beschäftigen, dürfte es erfreulich sein, zu vernehmen, dass die Verordnung zur Schaffung barrierefreier Informationstechnik nach dem Behindertengleichstellungsgesetz (BITV)<sup>10</sup> nun wohl endgültig als veraltet angesehen werden dürfte. Die BITV basiert laut Thomas Caspers, der über die Standardisierungsarbeit in diesem Gremium berichtete, noch auf der alten Version der Web Content Accessibility Guidelines (WCAG) 1.0<sup>11</sup>, die mit ihrer vollständigen Überarbeitung nun den Anforderungen eines modernen Webdesigns für behinderte Menschen entsprechen soll und nicht, wie die ältere Version, zum Beispiel Applikationen mit Flash, Silverlight oder Javascript als nicht barrierefrei ausschließt. Eine Aktualisierung der BITV steht dennoch noch aus. Zumindest steht jetzt aber ein großer Teil der deutschen Übersetzung der WCAG 2.0<sup>12</sup> zur Verfügung.

---

<sup>8</sup> <http://www.sti2.org/>

<sup>9</sup> <http://www.w3.org/TR/WCAG20/>

<sup>10</sup> <http://bundesrecht.juris.de/bitv/>

<sup>11</sup> <http://www.w3.org/TR/WAI-WEBCONTENT/>

<sup>12</sup> <http://wcag2.0-blog.de/2009-07-28/>

## Klassifikation und Ontologie

Zurück zu dem Web der Daten. Dieses bekommt durch die Weiterentwicklung der Ontology Working Language (OWL 2)<sup>13</sup> ein im Gegensatz zu seiner Vorgängerversion in seiner Syntax vereinfachtes, aber nicht weniger mächtiges Werkzeug an die Hand, um Daten mittels einer Ontologie zu beschreiben, berichtete Prof. Dr. Adrian Paschke von der FU Berlin. Ein besonderes Augenmerk sei auf die OWL 2 Profile (EL, QL, RL) gelegt, welche Subsets der OWL 2 beschreiben, um eine verbesserte Effizienz der OWL 2 durch die Ausrichtung auf bestimmte Anwendungsszenarien zu erreichen.

Wer hier aber eine Brücke vom Web der Daten zum Web der Menschen sucht, wird erst einmal enttäuscht. Ähnlich ist es beim Rule Interchange Format (RIF)<sup>14</sup>, welches zum Austausch von Regeln zwischen zwei proprietären Regelsystemen erdacht worden ist. Ein interessantes Werkzeug, denkt man nur daran, dass es in Bibliotheken weltweit von Regel(werke)n nur so wimmelt.

Einen ähnlichen Ansatz wie die OWL 2 und eher dem Bibliothekar ein Begriff, nutzt das Simple Knowledge Organization System (SKOS)<sup>15</sup>, welches eine Art Thesaurus-Beschreibungssprache für das Web der Daten liefert und von Joachim Neubert von der ZBW Kiel/Hamburg vorgestellt worden ist. Mittels SKOS wurde der Standard-Thesaurus Wirtschaft<sup>16</sup> für das Semantische Web modelliert. SKOS bietet ähnliche Beschreibungsmöglichkeiten und Relationsarten der Begriffe wie sie die ISO 2788 bzw. DIN 1463-1 vorsieht. Ein Beispiel:

```
<http://zbw.eu/stw/descriptor/10343-6> a skos:Concept, zbwext:Descriptor ;
  rdfs:isDefinedBy <http://zbw.eu/stw/descriptor/10343-6/about> ;
  rdfs:seeAlso <http://zbw.eu/econis/search/descriptor/Wirtschaftskrise> ;
  skos:broader <http://zbw.eu/stw/thsys/70021> ;
  skos:inScheme <http://zbw.eu/stw> ;
  skos:narrower <http://zbw.eu/stw/descriptor/11792-0>, <http://zbw.eu/stw/descriptor/19664-4> ;
  skos:prefLabel "Economic crisis"@en, "Wirtschaftskrise"@de ;
  skos:related
    <http://zbw.eu/stw/descriptor/10338-6>,
    <http://zbw.eu/stw/descriptor/10342-1>,
    <http://zbw.eu/stw/descriptor/10348-3>,
    <http://zbw.eu/stw/descriptor/10412-6>,
    <http://zbw.eu/stw/descriptor/10413-4>,
    <http://zbw.eu/stw/descriptor/16295-3>,
    <http://zbw.eu/stw/descriptor/19285-5>,
    <http://zbw.eu/stw/descriptor/19683-0> ;
  skos:scopeNote "Schwerwiegende, strukturbedingte Einbrüche in der gesamtwirtschaftlichen Tätigkeit"@de ;
  zbwext:indexedItem <http://zbw.eu/econis/search/descriptor/Wirtschaftskrise> .

<http://zbw.eu/econis/search/descriptor/Wirtschaftskrise> dcterms:subject <http://zbw.eu/stw/descriptor/10343-6>
```

Abbildung 1: Triple-Beispielcode für den Deskriptor „Wirtschaftskrise“ in Turtle-Syntax

Wer dem Link auf die Website des Standard-Thesaurus Wirtschaft folgt, wird diese Art von Datenkodierung nicht so leicht finden, da diese nicht primär für den Konsum durch Menschen gedacht ist. Dank der ersten wirklichen Brückentechnologie für die beiden Webs, die an diesem Tag benannt wurde, nämlich RDFa<sup>17</sup>, konnte der Standard-Thesaurus Wirtschaft mit sämtlichen Relationen in die Webseiten eingebettet werden. Mittels RDFa-Parsern, wie dem

<sup>13</sup> <http://www.w3.org/TR/owl2-overview/>

<sup>14</sup> <http://www.w3.org/TR/rif-core/>

<sup>15</sup> <http://www.w3.org/TR/skos-reference/>

<sup>16</sup> <http://zbw.eu/stw/versions/8.04/about.de.html>

<sup>17</sup> <http://www.w3.org/TR/rdfa-syntax/>

W3C-Distiller<sup>18</sup> oder den Gleaning Resource Descriptions from Dialects of Languages (GRDDL)<sup>19</sup> können dann wieder RDF-Triple wie im obigen Beispiel extrahiert werden.

## Nur 4,13 % aller Webseiten sind valide

Diese Verketzung von W3C-Technologien lassen zum ersten Mal das Web der Daten und das Web der Menschen zusammenrücken. Zu früh gefreut jedoch, betrachtet man die neuesten Entwicklungen im Bereich HTML, wie sie Felix Sasaki vorstellte. Die Entwicklung von XHTML 2 als Weiterentwicklung des grundlegenden Datenschemas für RDFa und GRDDL wird zugunsten des einfacheren HTML 5 aufgegeben.

Was bedeutet das für die semantische Auszeichnung in HTML? Zuerst einmal wird XHTML, sowie RDFa und GRDDL fortbestehen und Browser werden wohl auch weiterhin XHTML interpretieren können. Die Weiterentwicklung der Technologien ist jedoch gestoppt und somit auch mittel- oder langfristig ihr Ende eingeleitet.

Warum sich das W3C zu diesem Schritt entschlossen hat, erklärte Felix Sasaki zum Teil mit dem Einfluss und den Argumenten der am Standardisierungsprozess beteiligten Organisationen. Es soll eine Vereinfachung und Konzentration auf ein einziges DOM (Document Object Model) geben, in dem ausschließlich der HTML-Namespaces gelten soll. In HTML 5 sollen damit alle anderen Namespaces von Browsern ignoriert werden. Dafür setzt HTML 5 durch die besondere Fehlerbehandlungsalgorithmen und native Einbindung von multimedialen Elementen auf eine langsame Evolution, anstatt des revolutionären Gedankens, dass alle Webdokumente in valider Weise durch XHTML ausgezeichnet werden könnten. Dies ist nämlich trotz aller Bemühungen des W3C im WWW nur bei 4,13 % aller Webseiten der Fall.

Bei den Teilnehmern des W3C-Tages kam erkenntlich Unmut über die Entscheidung des W3Cs auf. Sollte hier eine Brücke eingerissen worden sein? Und wie kommt das W3C darauf, seinen eigenen Empfehlungen (RDFa und GRDDL) die Grundlage zu entziehen? Felix Sasaki verwies auf den noch fortlaufenden Standardisierungsprozess von HTML 5, in dem noch nicht alle Einsprüche – wie dem, semantische Informationen in HTML 5 implementieren zu können – mit der Endgültigkeit beantwortet wurden, die im Standardisierungsprozess des W3C vorgesehen ist. Die Möglichkeiten zur Einbettung von Semantik sollen noch geprüft werden.

Vorerst bleibt jedoch nur das Konzept der Microdata in HTML 5, die ähnlich wie die eher bekannten Microformats<sup>20</sup> bestimmte Typen von Informationen wie Kontakt-, Adress- und Kalenderdaten in HTML semantisch beschreiben und somit eine Interoperabilität und Austauschmöglichkeit unter Anwendungen ermöglichen.

## Aktuelle Standardisierungsvorhaben des W3C

Über zwei weitere Vorhaben des W3Cs berichtete wieder Felix Sasaki. Die Media Annotations Working Group<sup>21</sup> arbeitet zurzeit an einem Mapping von Eigenschaften von verschiedenen Multimedia-Metadatenformaten<sup>22</sup> auf das Metadatenformat MAWG<sup>23</sup>. Das

---

<sup>18</sup> <http://www.w3.org/2007/08/pyRdfa/>

<sup>19</sup> <http://www.w3.org/TR/grddl/>

<sup>20</sup> <http://microformats.org/>

<sup>21</sup> <http://www.w3.org/2008/WebVideo/Annotations/>

<sup>22</sup> [http://www.w3.org/2008/WebVideo/Annotations/drafts/ontology10/WD/mapping\\_table.html](http://www.w3.org/2008/WebVideo/Annotations/drafts/ontology10/WD/mapping_table.html)

Ziel sei hierbei, die Erleichterung des cross community-Austauschs von Information über multimediale Objekte im Web wie Video, Audio und Images. Die in bisher schwer zu integrierenden Metadatenformaten wie XMP, ID3, EXIF, MIX et cetera beschriebenen Objekte sollen so standardisiert über MAWG recherchierbar sein.

Das zweite W3C-Vorhaben, die XML Pipeline Language (XProc)<sup>24</sup>, habe mittlerweile den Status des W3C Recommendation Candidate erreicht. XProc kann, ähnlich wie GRDDL, als Brückentechnologie zwischen den webs angesehen werden, da es zur Aufgabe hat, Schritte zu definieren, die auf einem System mittels eines XProc-Processors auf ein in XML kodiertes Dokument angewendet werden sollen. Somit könnten zum Beispiel in RDF/XML kodierte Daten vereinfacht in (X)HTML umgewandelt werden, oder umgekehrt. XProc selber besteht dabei aus einem XML-Dokument, welches Schritt für Schritt Anweisungen an den XProc-Processor gibt, wie dieser die Input-Dokumente bearbeiten soll.

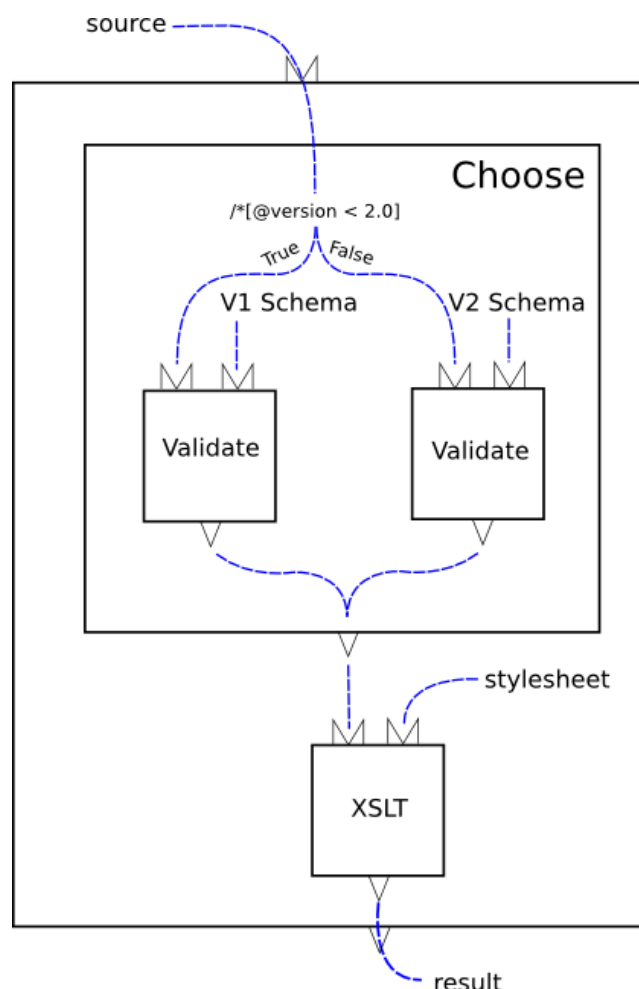


Abbildung 2: Xproc: Validation und Transformation von XML-Dokumenten

<sup>23</sup> <http://www.w3.org/TR/mediaont-10/>

<sup>24</sup> <http://www.w3.org/TR/xproc>

## Mehr Engagement von Bibliotheken

Zumindest konnten durch die Vorstellung der Standardisierungsarbeiten für MAWG und XProc den Teilnehmern des W3C-Tages noch einige Verbindungsmöglichkeiten zwischen dem Web der Daten und dem Web der Menschen aufgezeigt werden. Trotzdem blieb die Einstellung der Weiterentwicklung von XHTML 2 zugunsten von HTML 5 das beherrschende Thema.

Kein Grund, den Kopf in den Sand zu stecken. Besonders Bibliotheken müssen noch viel aufholen, wenn es um die webtaugliche „Semantifizierung“ ihrer Daten geht. Bei SKOS endet dieses Engagement nicht. Hierbei ist lediglich die „Sacherschließung“ abgedeckt und der formale Teil noch völlig außen vor gelassen. Aber auch in diesem Feld können Bibliotheken auf die Vorarbeit der Semantic-Web Community zurückgreifen. Hier sei nur kurz die Bibliographic Ontology<sup>25</sup> erwähnt, die auf eine Adaption ins und Mitentwicklung aus dem bibliothekarischen Umfeld wartet. Bibliotheken, bringt eure Daten ins Semantic Web!

---

<sup>25</sup> <http://bibliontology.com/>