

Dublin Core: Metadaten für WWW-Inhalte

Wer kennt das nicht: Auf der Web-basierten Suche nach Informationen oder Dokumenten zu einem bestimmten Thema gerät man leicht in die unangenehmen Untiefen der Volltextsuche. Wenn sich dann auf eine Suchanfrage hin mehrere zehntausend vermeintliche Treffer im Browserfenster stapeln, ist man oft nicht schlauer als ohne Suchmaschine, die dem entnervten Suchenden die ersehnte Nadel nur mit Heuhaufen präsentiert hat. Denn es versteht sich von selbst, dass die wenigsten der gefundenen Dokumente auch nur annähernd etwas mit dem gesuchten Thema zu tun haben.

Natürlich ist dieses Dilemma eine der meist strapazierten Binsenweisheiten des Internets. Und längst gibt es Versuche, vermittels statistischer und graphentheoretischer Methoden und Methoden der Künstlichen Intelligenz, Suchmaschinen zu einer Art Ranking bezüglich der Relevanz und zu einer vagen Klassifikation der gefundenen Dokumente zu befähigen¹.

Ein anderer Ansatz gründet sich darauf, dass Web-Autoren ihre Seiten selbst klassifizieren, indem sie ihnen bestimmte Informationen über ihre Eigenschaften hinzufügen – sogenannte *Metadaten*. Und natürlich ist auch das nichts neues: Der Inhalt des <TITLE>-Tags im Head-Bereich der HTML-Seite erleichtert Besuchern nicht nur das Erstellen von Lesezeichen. Viele Suchmaschinen messen einem gefundenen Schlagwort im Titel der Seite eine weit größere Bedeutung bei als einem Treffer irgendwo im Volltext.

Die Idee

Um die Erschließung von elektronischen Dokumenten zu verbessern und neben der einfachen Suche qualifizierte Retrieval-Funktionen zur Verfügung stellen zu können², wurde vor einigen Jahren ein weltweit standardisierter Metadatensatz entworfen. Dazu trafen sich 1995 neben Bibliothekaren vor allem Fachwissenschaftler aus aller Welt im amerikanischen Dublin (was dem Datensatz auch seinen Namen eingebracht hat: *Dublin Core*), um über einen Datensatz nachzudenken, der sich zur Beschreibung elektronischer Dokumente eignet.

Beim Entwurf ließen sich die Experten vor allem von den folgenden Prinzipien leiten:

- **Einfachheit:** Dublin Core soll insbesondere auch für Nicht-Bibliothekare benutzbar und verständlich sein.
- **Semantische Kompatibilität:** Die Metadaten sollen sich über Grenzen zwischen unterschiedlichen Fachbereichen hinweg semantisch entsprechen, um eine qualifizierte Suche im gesamten Internet zu ermöglichen.
- **Internationaler Konsens:** Die Initiative stützt sich auf die aktive Beteiligung von etwa 30 Ländern in Nordamerika, Europa, Australien und Asien.

- **Erweiterbarkeit:** Der vorliegende Metadatensatz zeichnet sich durch die Möglichkeit aus, feinere semantische Strukturen durch flexible Erweiterungen abzubilden.
- **Modularität von Metadaten im Web:** Das World Wide Web Consortium³ (W3C) hat eine Architektur für Metadaten im Internet implementiert, das sogenannte Resource Description Framework (RDF), das auch Dublin Core unterstützt.

Der Standard

Zum Standard von Dublin Core gehörten neben den eigentlichen Elementen natürlich vor allem die semantische Beschreibung derselben, was eine Recherche mit Hilfe dieser Angaben überhaupt erst sinnvoll macht. Jedes Element von Dublin Core ist durch die zehn Attribute gekennzeichnet, die in ISO/IEC 11179⁴ für die Beschreibung von Datenelementen vorgesehen sind: *Name*, *Identifier*, *Version*, *Registration Authority*, *Language*, *Definition*, *Obligation*, *Datatype* sowie *Maximum Occurrence* und *Comment*. Von denen besitzen allerdings sechs identische Werte für alle Elemente von Dublin Core:

Version	1.1
Registration Authority	Dublin Core Metadata Initiative
Language	en
Obligation	Optional
Datatype	Character String
Maximum Occurrence	Unlimited

Die wichtigsten der Elemente des Dublin Core sind in der nachfolgenden Tabelle zusammengefasst⁵.

Die Benutzung

Um eigene Internetseiten oder -dokumente mit Metadaten aus dem Dublin Core Standard auszustatten, sind lediglich ein paar Zeilen im ersten Teil der HTML-Seite, also zwischen <HEAD> und </HEAD> einzufügen. Das sieht dann etwa so aus:

```
<META NAME="DC.Title" CONTENT="Dublin Core:
  Metadaten für WWW-Inhalte">
<META NAME="DC.Creator" CONTENT="Uwe
  Müller">
<META NAME="DC.Publisher" CONTENT="HU Berlin,
  Rechenzentrum">
<META NAME="DC.Language" CONTENT="ger">
...
```

Um Autoren von Webseiten mühsame Tipparbeit zu ersparen, wurden so genannte Meta-Maker entwickelt, die die z. B. über ein Web-Formular abgefragten Daten in die angegebene Datei integrieren. Ein Beispiel dafür bietet unter anderem das Institut für Mathematik der Universität Osnabrück⁶ (siehe Abb. 1).

Name	Identifier	Definition
Title	Title	Name der Quelle.
Creator	Creator	Eine Körperschaft, die für das Dokument bzw. dessen Inhalt verantwortlich ist (z. B. Person, Institution etc.).
Subject and Keywords	Subject	Das Thema, mit dem sich die Quelle beschäftigt.
Description	Description	Ein Überblick über den Inhalt der Quelle (z. B. Abstract, Inhaltsverzeichnis etc.).
Publisher	Publisher	Eine Körperschaft, die für die Verfügbarkeit der Quelle verantwortlich ist (z. B. Person, Organisation, Service).
Date	Date	Ein Datum, das mit dem Lebenszyklus der Quelle assoziiert ist (z. B. Erstellungsdatum, Veröffentlichungsdatum).
Resource Type	Type	Die Art oder das Genre der Quelle (z. B. generelle Kategorien, Funktionen etc.).
Format	Format	Die physikalische oder digitale Manifestation der Quelle (z. B. Datenformat, Größe, Systemvoraussetzungen).
Resource Identifier	Identifier	Eine eindeutige Referenz der Quelle innerhalb eines gegebenen Kontextes (z. B. URI inkl. URL, DOI, ISBN).
Source	Source	Eine Referenz zu der Quelle, aus der die vorliegende Quelle hergeleitet ist (z. B. durch eine Nummer o. ä., die dem Ursprungsdokument in einem Identifikationssystem entspricht).
Language	Language	Die (bzw. eine) Sprache, in der der Inhalt der Quelle verfasst ist.
Relation	Relation	Eine Referenz auf eine verwandte Quelle.
Rights Management	Rights	Information über Rechte, die an der Quelle (oder deren Teilen) gehalten werden.

Tabelle 1: Elemente des Dublin Core

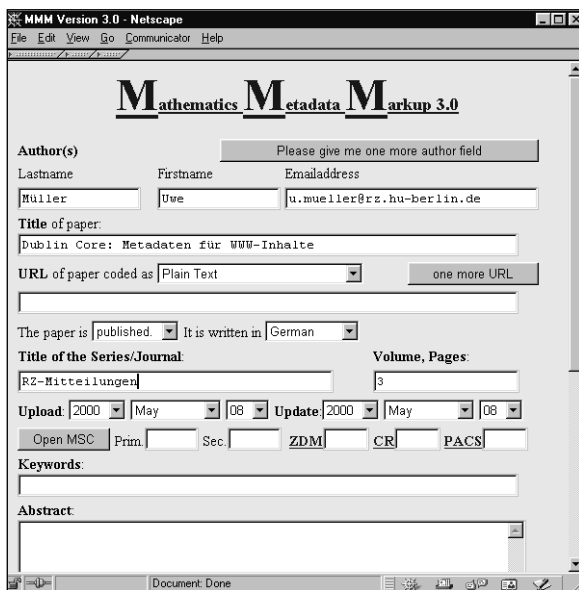


Abb. 1: Der Meta-Maker für mathematische Texte der Uni Osnabrück

Ausblick

Derzeit wird von der Dublin Core Initiative über eine Weiterentwicklung des Standards nachgedacht. Dabei kommen sogenannte *Qualifier* ins Spiel, mit Hilfe derer – rein optional natürlich – sich die Elemente weiter spezialisieren lassen. DC.Publisher ließe sich beispielsweise durch DC.Publisher.Name und DC.Publisher.Address weiter qualifizieren. Doch dieser erweiterte Datensatz ist bisher nicht beschlossen und somit im Sinne einer einheitlichen Verwendung auch nicht unbedingt für die Benutzung zu empfehlen.

Fazit

In vielerlei Hinsicht gleicht das Internet heute einem informationellen Chaos, das sich mit Begriffen wie „Digitale Bibliothek“ oder „Verteiltes Archiv“ kaum schmücken kann. Das liegt zum einen an der wenig hierarchischen Struktur des Internet aber auch daran, dass die meisten Dokumente „nur“ ihren Inhalt selbst enthalten, nicht aber Informationen *über* denselben. Die erforderliche Mehrarbeit, die überdies wenig Kenntnisse im bibliothekarischen Bereich voraussetzt, lohnt sich allemal – nicht nur für den Suchenden, auch für den Autor des Gefundenen.

Uwe Müller
uwe.mueller@rz.hu-berlin.de

1 z. B. <http://www.google.com> und <http://www.gerhard.de>
 2 Beispiele für Suchmaschinen, die den Metadatensatz von Dublin Core verwenden, stehen u. a. unter <http://www.luk-Initiative.org/iwi/TheO/> und http://www.MathGuide.de/advanced_search.html zur Verfügung
 3 siehe <http://www.w3c.org>
 4 siehe <ftp://sdct-sunsv1.ncsl.nist.gov/x318/11179/>
 5 siehe <http://purl.org/dc/>
 6 siehe <http://www.mathematik.uni-osnabrueck.de/cgi-bin/MMM3.0.cgi>