

Ontologien in den historischen Wissenschaften

*Leif Scheuermann**

Abstract: Databases are for historians a mighty tool, especially for the statistical analysis of huge quantities of data. But the tool can only work with a solid data structure that is defined on a meta-level. To guarantee this structure, computer scientists developed for most different scopes ontologies – the topic of the following elementary article. After the definition of ontologies as formal defined systems of concepts and relations with rules of inference and integrity, a special vocabulary for entities and relations and conditions of completeness and correctness of the terms, the following article will introduce the basic concepts of a specific ontology – CIDOC CRM (Comité International pour la Documentation Conceptual Reference Model). The fundamental terms Entities and Properties will be explained as well as their structure and their underlying rules.

Einleitung

In den letzten Jahren sind Datenbanken – besonders in Kombination mit einer Internetoberfläche – zu wichtigen Werkzeugen für die historischen Wissenschaften geworden. Große datenbankgestützte Quellensammlungen im Bereich der Hilfswissenschaften wurden erstellt, doch auch genuin historische Fragestellungen in Datenbanken umgesetzt. So kann heute mit Fug und Recht behauptet werden, dass dem Historiker eine große Datengrundlage in Form von Datenbanken zur Verfügung steht, welche er für seine spezifischen Fragestellungen auswerten kann. Doch ergeben sich hierbei einige Probleme. Zum einen gibt es erhebliche Redundanzen in den Daten, die er nutzen kann,¹ zum anderen

* Address all communications to: Leif Scheuermann, Historisches Institut, Abteilung Alte Geschichte, Keplerstr. 17/8a, 70174 Stuttgart, Germany;
e-mail: L.scheuermann@ike.uni-stuttgart.de.

¹ So gibt es im Bereich der Epigraphik beispielsweise alleine drei große Datenbanken, die größtenteils die gleichen geographischen Räume abdecken, wobei häufig dieselben Daten in jeweils anderem Kontext enthalten sind.

erfordern neue Analysen eine Kombination und Umstrukturierung bereits vorhandener Daten, da eine Datenbank nie unabhängig von der Fragestellung, welche der Analyse zugrunde liegt, aufgebaut ist. Es bedarf daher einer systemunabhängigen Beschreibung der Daten, durch welche das jeweilige aufgenommene Charakteristikum des Untersuchungsgegenstandes eindeutig und vollständig definiert wird. In den letzten Jahren wurden hierfür Ontologien entwickelt, welche im Folgenden einführend vorgestellt werden sollen.

Was sind Ontologien?

In seinem Ursprung ist Ontologie als philosophischer Fachterminus die Lehre vom Wesen des Seins. Dieser Begriff wurde in den 1990er Jahren von Informatikern aufgegriffen und für eine allgemein gültige, explizit definierte Terminologie² verwendet, welche die Struktur von Daten beschreibt.

Dieses formal definierte System von Konzepten und Relationen benötigt Inferenz- und Integritätsregeln, ein spezifisches Vokabular, Bedingungen zur Vollständigkeit und Richtigkeit der Begriffe sowie Beziehungen zwischen den Begriffen.

Aufgabe einer Ontologie ist die Schaffung einer Grundlage für die Strukturierung komplexer Daten, die Validierung bereits vorhandener Datenmodelle sowie der Datenaustausch zwischen Menschen, aber auch zwischen Mensch und Maschine und zwischen einzelnen Maschinen. Dadurch kann eine Aufteilung komplexer Probleme auf mehrere Bearbeiter oder eine Verknüpfung bereits bestehender Daten verwirklicht werden. Durch den gezielten Einsatz dieser Verknüpfung kann ontologisches Lernen verwirklicht werden. Hierbei soll der Computer als künstliche Intelligenz einzelne, durch die Ontologie eindeutig definierte Daten miteinander in Beziehung setzen und so den eigenen Datenbestand erweitern.

Um ontologisches Lernen zu ermöglichen, bedarf es allerdings der Erweiterung des World Wide Web um für Maschinen verarbeitbare Daten. Im Gegensatz zur heute eingesetzten HyperText Markup Language (HTML) muss nicht nur die Darstellung des Wortes, sondern auch die Bedeutung der Inhalte, also die Semantik, verzeichnet sein, welche durch die Ontologie definiert wird.

Ziel der Ontologie ist es also, die Welt in einem formalen inhaltsbezogenen System abzubilden um dadurch das Semantic Web³ zu verwirklichen. Sie ist das Modell eines spezifischen, für das jeweilige Themengebiet angepassten Ausschnittes der Welt und dient zur Spezifikation der Abbildung der Realität in einem Datenmodell (Abb. 1).

² Reches, R. / Fikes, R. / Finin T. / Gruber, T. / Patil, R. / Senator, T. / Swartoutt, W.R.: Enabling technology for knowledge sharing. In: AI Magazine Bd. 13 Nr. 3 1991.

³ Weitere Informationen hierzu: <<http://www.w3.org/2001/sw/>> (Mai 2006).

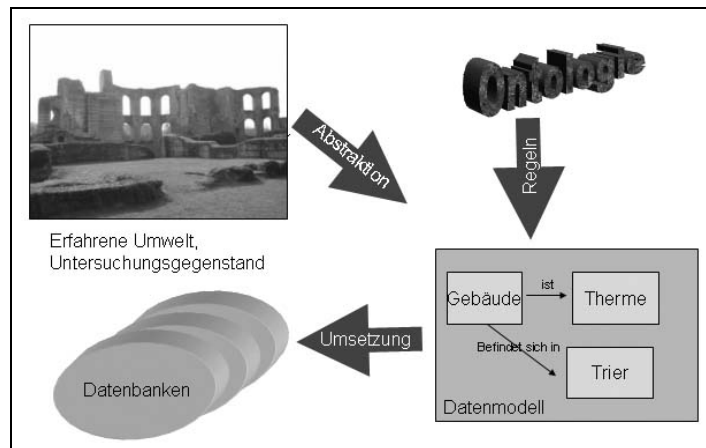


Abb. 1: Problemstellung beim Entwurf von Datenbanken

Im Folgenden soll nun CIDOC CRM – eine Ontologie für historische Fragestellungen – vorgestellt werden.

CIDOC CRM

Das CIDOC CRM (Comité International pour la Documentation – Conceptual Reference Model) ist ein Datenmodell, welches 1996 begründet und 1998 zum ersten Mal durch das CIDOC (Comité International pour la Documentation) des ICOM (International Council of Museums) publiziert wurde. Grundlagen des Modells sind die im Juni 1995 veröffentlichten „CIDOC Information Categories“ und das „CIDOC Relational Data Model“ aus dem Jahr 1994.

Im September 2000 wurde die Version 3.2.1 als ISO TC 46/SC4⁴ Working Draft von der ISO (International Organization for Standardization)⁵ verabschiedet; sie soll in Kürze als internationale Norm ISO/CD 21127 angenommen werden. Die Begründer des Modells bezeichnen es selbst als „*domain ontology for cultural heritage information*“⁶.

⁴ ISO TC 46/SC4: <<http://www.niso.org/international/SC4>> (Mai 2006).

⁵ Die ISO ist ein Netzwerk von nationalen Institutionen aus 147 Ländern zur Aufstellung weltweit gültiger Standards. Als NGO (non governmental organization) hat die ISO selbst keinen Einfluss auf die Annahme der Standards. Da jedoch alle führenden Industrienationen an der ISO beteiligt sind, werden die ISO-Normen als weltweit verbindlich angesehen. Nähere Informationen hierzu unter: <http://www.iso.com/about_iso/index.html> (Mai 2006).

⁶ Croft, Nick; Doerr, Martin; Gill Tony: The CIDOC Conceptual Reference Model. A Standard for Communicating Cultural Contents. In: Cultivate Interactive 9, 2003. <<http://www.cultivate-int.org/issue9/chios>> (Mai 2006).

Was ist nun unter kulturellem Erbe zu verstehen, oder besser, was ist kein kulturelles Erbe? Mit diesem Thema haben sich bereits Generationen von Kulturschaffenden auseinandergesetzt, ohne sich auf eine einheitliche Definition einigen zu können. Der Anwendungsbereich des CIDOC CRM beschränkt sich deshalb auf museale Gegenstände und deren Umfeld. Da dieser jedoch immer noch – besonders, wenn es um die Verwaltung der Objekte, deren Wert, Versicherungen oder Besitzstand etc. geht – einen zu großen Rahmen für das Modell darstellt, muss nochmals differenziert werden:

The CRM is specifically intended to cover contextual information: the historical, geographical and theoretical background in which individual items are placed and which gives them much of their significance and value.⁷

Entities und deren Hierarchie

Es geht also um den übergeordneten Kontext und den Bedeutungsgehalt der Objekte. Die aufgenommenen Grundkategorien sind, konzeptionelle und physische Objekte, Perioden, Teilnehmer und Orte (Abb. 2). Im Zentrum steht ein „Event“, also ein spezifisches historisches, das bedeutet zeitlich gebundenes Ereignis. Dieses findet an einem Ort statt und wird von Teilnehmern (Actors) begangen. Bei diesem Ereignis entsteht etwas, es verändert sich oder wird zerstört. Durch diesen Vorgang werden wiederum geistige oder physikalische Entitäten geschaffen.

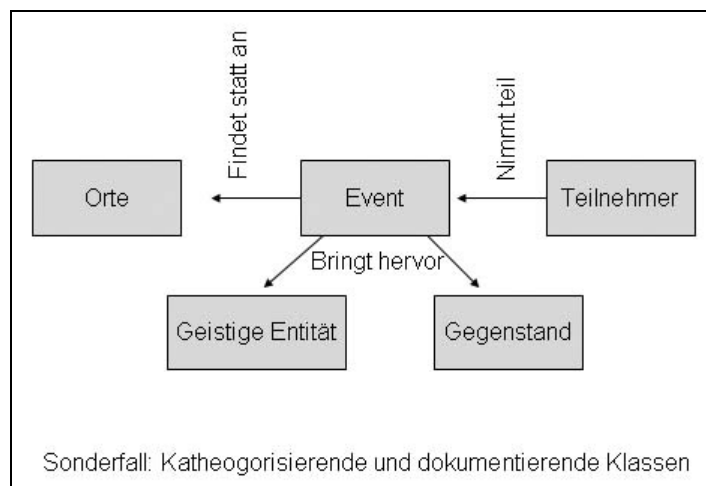


Abb. 2: Struktur des CIDOC CRM

⁷ Ebd.

Eine Sonderrolle nehmen die Beschreibungen ein, welche selbst zu den geistigen Entitäten zählen und somit in das Beschreibungsmodell eingebettet sind. Sie stehen zusätzlich außerhalb des Systems, da sie die später einzufüllenden Daten darstellen. Sie können in beschreibende Daten (Notes) und kategorisierende Daten (Types) eingeteilt werden, zu denen als Sonderfall die Benennungen (Appellations) gehören.⁸

Ein Beispiel für einen solchen Event ist die Niederschrift dieses Aufsatzes. Sie findet in Stuttgart statt, und wird vom Autor als Actor verfertigt. Dabei entsteht das Dokument als physikalischer Gegenstand sowie der Inhalt als geistige Entität. Als Type kann Aufsatz und als Appellation der Titel „Ontologien in den historischen Wissenschaften“ hinzugefügt werden. Auf diese Art und Weise ergibt sich eine Grobstruktur der Beschreibung des zeitlich im Mai 2006 gebundenen Ereignisses, welches daraufhin über Verfeinerungen der Semantik noch näher definiert werden kann.

Hierfür enthält CIDOC CRM 81 Klassen (Entities), von denen jede eine eigene Nummer hat, welche sprachenunabhängig eindeutig definiert ist. Die fünf großen Klassen sind in Unterklassen aufgeteilt, welche in einer Hierarchie stehen. An der Spitze der Hierarchie befindet sich die rein abstrakte „CRM entity“, die alles umfasst und sich immer weiter spezialisieren lässt. Das System von Unterklassen (Subclasses) und Überklassen (Superclasses) lässt eine Vererbung der einzelnen Attribute zu, so dass alle Attribute, die einer höheren Klasse zugeordnet sind, ebenfalls in den darunter liegenden Bereichen gelten.

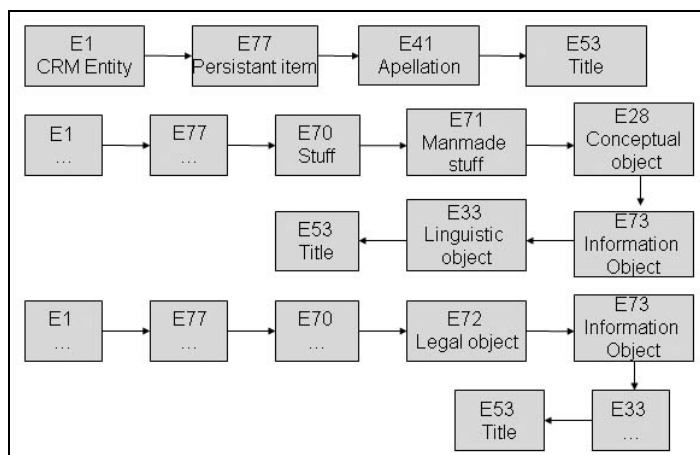


Abb. 3: Vererbung

⁸ Siehe unten.

Zum Beispiel (Abb. 3) ist die Klasse „E53 Title“, die Unterklasse von „E41 Apellation“. Damit wird ausgedrückt, dass ein Titel immer die Benennung eines Gegenstandes darstellt. Hingegen gibt es Benennungen, die kein Titel sind, wie z.B. ein Barcode. „E41 Apellation“ ist wiederum eine Unterklasse von „E77 Persistent Item“, was aussagt, dass eine Benennung eine zeitlich unabhängige, bestehende Entität darstellt und Unterklasse der abstrakten „E1 CRM Entity“ ist, was bedeutet, dass sie der Untersuchungsgegenstand ist.

Durch den Aufbau dieser Hierarchie wurde nun definiert, in welcher Weise der Titel in diesem Beispiel verstanden wurde. Alle Inhalte, die mit „E3 Title“ bezeichnet sind (in diesem Fall die einzelnen Titel die in der Klasse Title zusammengefasst wurden) können nun auch als Benennungen erkannt werden. Ferner müssen alle Kriterien, die auf die Benennung zutreffen, zwangsläufig auch für die Titel gelten. Damit kann man einen Gegenstand in möglichst vielen Facetten dokumentieren, ohne die einzelnen Attribute auf der Detailebene immer wieder aufführen zu müssen.

Properties

Neben den Entities und deren Hierarchie durch Vererbung gibt es im CIDOC CRM noch eine zweite wichtige Gruppe an Begriffen: die Properties.

Durch sie wird die Modellierung der Attribute geschaffen. Das Modell stellt eine Liste von 133 Eigenschaften zur Verfügung, die je zwei Klassen miteinander verbinden.

Beispielsweise verbindet die Relation „P70 documents – is documented in“ die Entität „E32 Authority Document“ mit „E1 CRM Entity“. Diese Verknüpfung sagt aus, dass ein Lexikonartikel einen Untersuchungsgegenstand beschreibt. Durch die bereits dargestellte Vererbung kann daraus modelliert werden, dass eine Person, ein Gegenstand, eine Periode oder auch ein Lexikonartikel selbst wiederum in einem Lexikonartikel beschrieben sein kann.

Der Lexikonartikel ist also selbst Unterklasse einer „E1 CRM Entity“ und steht für sich ebenfalls in einer solchen Hierarchie. Dies führt zu einem Universum miteinander verknüpfter und sich gegenseitig bedingender „E1 CRM Entities“, aus denen nun der Ersteller des einzelnen Modells sich seinen Bereich herausnehmen kann. Ein zweiter Nutzer kann nun ein weiteres Modell zu einem ganz anderen Untersuchungsgegenstand aufbauen. Dennoch sind diese beiden Systeme über die dem CRM immanente Struktur direkt miteinander verbunden, und so können sie von einem Dritten durch das Aufstellen neuer Verknüpfungen kombiniert abgefragt werden. Ein erster Schritt hin zum Semantic Web ist also getan.

Inferenz- und Integritätsregeln des CIDOC CRM

Wie bereits oben erwähnt, genügen jedoch ein gemeinsames Vokabular und die Bedingungen zur Vollständigkeit und Richtigkeit der Begriffe sowie deren Beziehungen in Entities und Relationships nicht, um eine Ontologie zu definieren. Die Inferenz- und Integritätsregeln sind weitere wichtige Kriterien, welche CIDOC CRM erst zu einer Ontologie machen.

Eine erste wichtige Regel ist die der Monotonie. Nach ihr müssen bei der Weiterentwicklung des Systems einmal aufgestellte Strukturen erhalten bleiben, damit im Gegensatz zu Datenbanken oder Programmen die aufgebauten Datenmodelle kein „Verfallsdatum“ haben, sondern jederzeit von der einen Ausgestaltung in Form einer Datenbank in eine andere übertragen werden können.

Des Weiteren muss für die Einzigartigkeit der Entities gesorgt sein. Bei der Modellierung darf keine Entity von einer anderen abgeleitet oder durch eine andere ersetzt werden können. Sollte dies doch der Fall sein, so ist das Modell unscharf gestaltet und die beiden Entitäten müssen zu einer zusammengefasst werden.

Die dritte wichtige Regel ist die des Widerspruchs. Es gibt einige Entities, die grundsätzlich nicht miteinander in Verbindung gesetzt werden können. Diese werden als „disjoint“ bezeichnet. Zwei dieser „Disjoint-Entity-Verbindungen“ sind nun für das System äußerst bedeutend. Dies sind zum einen „E2 temporal Entity“ und „E77 Persistent item“ und zum anderen „E18 Physical stuff“ und „E28 Conceptual stuff“.

Die beiden Gegensätze drücken also aus, dass Fortdauerndes immer von Zeitlichem sowie Materielles immer von Immateriellem getrennt sein muss. Dies bedeutet jedoch nicht, dass Zeitliches nicht auf Fortdauerndes verweisen und somit durch Properties verbunden sein kann und auch nicht, dass Gegenstände nicht gleichzeitig materiell sein und auf Immaterielles verweisen können. Dennoch müssen für das Modell beide Wege getrennt gehalten werden.

Hierbei ist es wichtig, dass die Beziehungen zwischen Entities nicht ausschließlich sind. Dies lässt sich gut an dem bereits dargestellten Beispiel der Vererbung verdeutlichen.⁹

Neben der Bedeutung des Titels als Benennung, kann er auch ein sprachliches Objekt sein (E33 Linguistic object). Als solches ist er ein Informationsträger (E73 Information Object) und somit entweder ein nicht materielles Objekt unseres Geistes (E28 Conceptual Object) oder aus urheberrechtlicher Sicht auch ein rechtliches Objekt (E72 Legal Object) usw.

Die drei Wege betreffen also die gleiche Entity, müssen jedoch streng voneinander getrennt gehalten werden, immer abhängig von der Fragestellung, mit

⁹ Siehe im Folgenden Abb. 2.

welcher man an sie herangeht. Ähnlich verhält es sich mit der Beziehung durch Properties. So sind „E18 Physical Thing“ und „E39 Actor“ über „P49 has former or current keeper (is former or current keeper of)“, „P50 has current keeper (is current keeper of)“, „P51 has former or current owner (is former or current owner of)“ und „P52 has current owner (is current owner of)“ miteinander verbunden. Dies besagt, dass ein Teilnehmer entweder Besitzer oder ehemaliger Besitzer sein kann oder aber auch derjenige, welcher einen Gegenstand in Verwahrung hat. Bedenkt man die Beutekunstdebatte, so kann dieser Unterschied von größter Bedeutung sein.

Ein weiteres wichtiges Prinzip, auf welchem das CRM beruht, ist die Entity „E55 Type“. „E55 Type“ stellt eine Metaklasse dar, die im Datenmodell nicht weiter hinterfragt werden soll. Ein Beispiel für einen Eintrag in solch eine Klasse wären die Längenmaße. Es muss davon ausgegangen werden, dass die Herkunft des Maßes „Meter“ nicht näher hinterfragt werden soll. Dennoch ist es wichtig, die Maßeinheit bei einer Messung anzugeben. Man kann diese Entity auch einsetzen, um eigene Klassen aufzubauen, die man für eine Analyse benötigt, wie zum Beispiel den Typ einer Inschrift. Hierfür könnte es die Werte „Grabstein“, „Weihestein“, „Bauinschrift“ etc. geben. Erst durch eine solche Kategorisierung kann eine Inschrift in einer Datenbank analysiert werden, da man für Computeranalysen immer einheitliche Werte benötigt. Dementsprechend kann an jede Entity ein „E55 Type“ angegliedert werden, um einen Rahmen abzustecken, womit kooperatives Arbeiten erst ermöglicht wird.

Ähnlich wie die Entity „E55 Type“ stellt auch die Property „P3 has note“ ein Prinzip zur Entlastung des Systems dar. Die Entwickler des CRM waren sich von Anfang an klar, dass es unmöglich sein würde, sämtliche zu modellierenden Bereiche abzudecken, trotz der strikten Abgrenzung des Anwendungsbereichs. Aus diesem Grund wurde „P3 has note“ eingeführt. Es führt zu „E62 String“ und lässt einfach zu, dass weitere Anmerkungen eingefügt werden können, die jedoch nicht im Gesamtsystem eingebettet sind und nicht weiter verknüpft werden können. „E62 String“ lässt ferner zu, dass Bilder, Tondokumente oder Videosequenzen zu einem Thema in die Datenbank aufgenommen werden können.

Links und Literaturangaben

Semantic Web:

- Reches, R. / Fikes, R. / Finin T. / Gruber, T. / Patil, R. / Senator, T. / Swartoutt, W.R.: Enabling technology for knowledge sharing. In: AI Magazine, Bd. 13, Nr. 3, 1991.
- Berners-Lee, Tim / Hendler, James / Lassila, Ora: *The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities*. <http://www.sciam.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21> (Mai 2006).
- W3C Semantic Web Activity: Advanced Development <<http://www.w3.org/2000/01/sw/>>.

Ontologien:

- Staab, Steffen / Studer, Rudi (Hg.): *Handbook on Ontologies*, Springer Verlag, Heidelberg, 2004.
- Buffalo Ontology Site. <<http://ontology.buffalo.edu/>>.

CIDOC CRM

- Cidoc Homepage. <<http://cidoc.ics.forth.gr/>>.
- Nix, Markus: Die praktische Einsetzbarkeit des CIDOC CRM in Informationssystemen im Bereich des Kulturerbes. Köln, 2004 (Magisterarbeit). <http://www.hki.uni-koeln.de/studium/MA/MA_nix.pdf>.
- Stein, Regine / Gottschewski, Jürgen / Heuchert, Regine / Ermert, Axel / Hagedorn-Saupe, Monika / Hansen, Hans-Jürgen / Saro, Carlos / Scheffel, Regine / Schulte-Dornberg, Gisela: Das CIDOC Conceptual Reference Model: Eine Hilfe für den Datenaustausch? In: *Mitteilungen und Berichte aus dem Institut für Museumskunde* Nr. 31. Berlin, 2005. <http://www.museumbund.de/cms/fileadmin/fg_doku/publikationen/CIDOC_CRM-Datenaustausch.pdf>.