

ZUR NEUBESTIMMUNG DES DOKUMENTENBEGRIFFS IM REIN DIGITALEN

von Jakob Voß

Ogleich der Wandel vom gedrucktem zum digitalen Wort noch immer nicht abgeschlossen ist und viele Medien weiterhin gedruckt erscheinen, vertrieben und gesammelt werden, ist absehbar, dass in Zukunft alle wesentliche Publikation und Kommunikation digital stattfinden wird. Dieser Wandel, dessen Bedeutung noch am ehesten mit der Einführung des Buchdrucks mit beweglichen Lettern verglichen werden kann, erfordert eine Neubestimmung mehrerer Grundbegriffe der Bibliotheks- und Informationswissenschaft. Eine zentrale Rolle spielen dabei die Begriffe der Information, der Publikation und des Dokumentes.

Publikation im Wandel

Allgemein kann eine Publikation als ein öffentlich zugängliches Werk verstanden werden. Riehm, Bohle und Wingert (2004) definieren publizieren als einen

„Kommunikationsvorgang, der über ein Artefakt, die Publikation vermittelt wird. Nicht jedes Dokument, das erstellt, und nicht jede Information, die verbreitet wird, ist eine Publikation. Eine Publikation ist für die Öffentlichkeit, für ein mehr oder weniger anonymes Publikum bestimmt.“

Bei näherer Betrachtung kann zum einen auf das Publizieren als Vorgang und zum anderen auf die dabei veröffentlichte Publikation als Artefakt eingegangen werden. Dieser Artikel beschränkt sich auf die Frage nach der Publikation als Artefakt und fasst das Publizieren vereinfacht als allgemeinen, über ein digitales Objekt vermittelten, Kommunikationsvorgang auf. Die ebenso wichtige Frage, wann ein Objekt in einer rein digitalen Umgebung als veröffentlicht angesehen werden kann, dass heißt wo die Grenzen zwischen Publizieren und anderen Formen der Kommunikation liegen, ist nicht Gegenstand dieser Untersuchung.

Erforderlich wird die Neubestimmung des Publikationsbegriffs vor allem durch die völlige Loslösung der Publikation von einem konkreten Trägermedium. Sofern nicht als zusätzliche Dimension die Zeit hinzugenommen wird, ist im Digitalen eine Unterscheidung zwischen Original und Kopie bedeutungslos. Ebenso ist die konkrete Infrastruktur zum Abruf und zur Speicherung für die Publikation als Artefakt nebensächlich. Zwar macht es durchaus einen Unterschied, ob zum Beispiel ein Film auf einem hochauflösenden Monitor oder über ein mobiles Endgerät konsumiert wird; von diesen Unterschieden muss jedoch zur Bestimmung des Films als digitalem Objekt abstrahiert werden oder alle Versuche zur Langzeitarchivierung sind von vornherein zum Scheitern verurteilt. Im Gegensatz zu physischen Medien ist bei digitalen Medien die Grenze zwischen Artefakt und Benutzungskontext zwar schwieriger zu ziehen, aber dennoch zwangsläufig vorhanden. Beispielsweise könnten die hochauflösende und die mobile Version des Films als unterschiedliche Objekte aufgefasst werden, während davon abstrahiert wird, unter welchem Betriebssystem der Film abgespielt wird.

Forschungsfragen

Statt von digitalen Objekten wird im Folgenden von Dokumenten gesprochen, wobei beide Begriffe synonym verwendet werden können. In der digitalen Welt ist eine Publikation immer ein Dokument und tritt in Form von Daten auf, die verlustfrei kopiert und einfach modifiziert werden können. Die Daten selber sind dabei während der Übertragung, Speicherung und Darstellung vielfältigen technischen Umwandlungen unterworfen, während das eigentliche Dokument gleich bleibt oder lediglich in identische Repräsentationen umgewandelt wird. Diese Unterscheidung wirft einige Fragen auf. Digitale Dokumente lassen sich zudem viel einfacher als physische Objekte aufteilen, abändern und neu kombinieren – bis hin zur Zerstückelung in die atomaren, einzeln adressierbaren Bestandteile eines Dokumentes wie Sätzen, Wörtern, Relationen und Werten (umgesetzt auf der technischen Ebene,

beispielsweise als XML-Elemente, Datenfelder oder RDF-Tripel). Motiviert durch diesen Hintergrund können folgende Forschungsfragen formuliert werden:

- I. Wann kann eine Sammlung von Daten sinnvoll als Dokument betrachtet werden?
- II. Wann sind zwei Dokumente identisch?
- III. Worin unterscheiden sich zwei nicht-identische Dokumente?

Die Fragen können auf syntaktischer, semantischer und pragmatischer Ebene beantwortet werden. Syntaktisch kann jede beliebige Datensammlung als Dokument zusammengefasst werden und zwei Dokumente unterscheiden sich, sobald auch nur ein Datum ihrer Bestandteile unterschiedlich ist. Aus Sicht der Pragmatik sind zwei Dokumente identisch, wenn sie als gleichwertig betrachtet werden können, d.h. wenn sie den gleichen Zweck erfüllen und ein Dokument eine Sammlung von Daten ist, die überhaupt irgendeinen Zweck erfüllt. Auf semantischer Ebene hängen die Antworten schließlich davon ab, auf welcher Abstraktionsstufe ein Dokument betrachtet wird. Im Folgenden soll ein erweitertes Informationsmodell skizziert werden, dass unter Beibehaltung der Möglichkeit syntaktischer Vergleiche Zwecke und Abstraktionsstufen integriert und so den gesamten semiotischen Rahmen einbezieht.

Information

Information gehört in einer Vielzahl von Fächern zu den Grundbegriffen, was eine allgemein gültige Definition nahezu unmöglich macht. Bereits Shannon (1953) bemerkte „it is hardly to be expected that a single concept of information would satisfactorily account for the numerous possible applications of this general field“, um darauf seinen Informationsbegriff auf eine spezielle Art von Kommunikationsproblemen einzuschränken: „the present note outlines a new approach to information theory which is aimed specifically at the analysis of certain communication problems“. Der kommunikationstheoretische Ansatz von Shannon und Weaver (1949) gehört inzwischen zu den Grundlagen der Informationstheorie. Er beinhaltet eine Definition des Bits als Basiseinheit zur Messung von Information und basiert auf einem einfachen Sender-Empfänger-Modell: Information in Form einer Nachricht wird vom Sender als Signal kodiert, über einen Kanal übertragen und dabei ggf. gestört, vom Empfänger dekodiert und wieder in eine Nachricht umgewandelt (Abbildung 1).

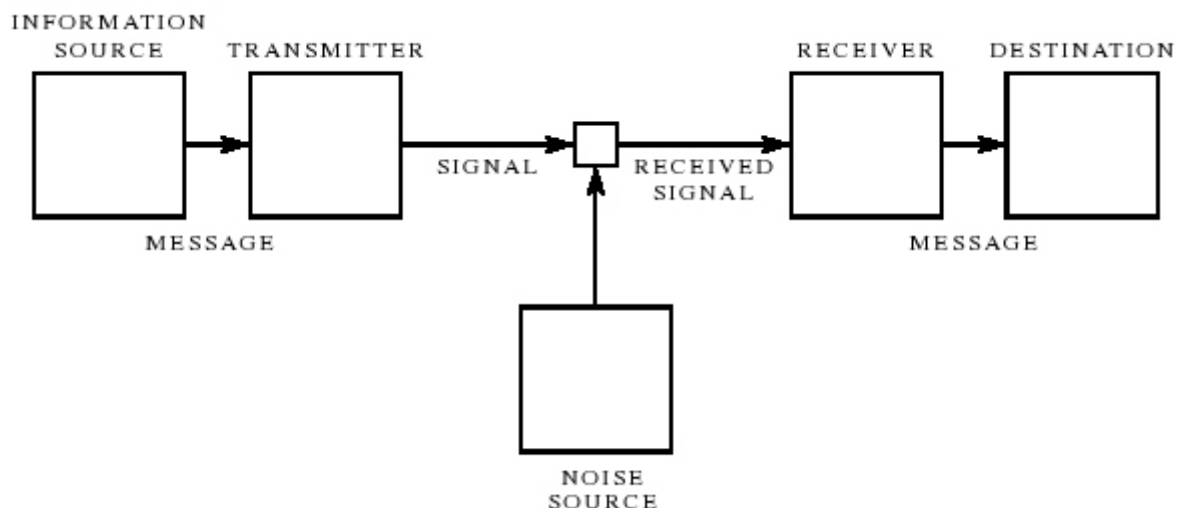


Abbildung 3: Modell eines Kommunikationssystems nach Shannon und Weaver (1949)

Wie Shannon (1949) explizit anmerkt, beschränkt sich dieser Ansatz auf die syntaktische Ebene von Information:

„frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem“.

Welche Bedeutung eine Information (Semantik) hat und ob sie zweckdienlich eingesetzt werden kann (Pragmatik), wird in der Informationstheorie ausgeklammert. Um diese Lücke zu schließen, sind von der Informationsphilosophie verschiedene Modelle vorgeschlagen worden. Wie Floridi (2005) darlegt, lässt sich der inzwischen übliche Informationsbegriff, die *Standard Definition of Information* (SDI), auf die Gleichung Information = Daten + Bedeutung bringen (die in der deutschsprachigen Informationswissenschaft bekannte Definition Rainer Kuhlens von „Information als Wissen in Aktion“ hat im internationalen Diskurs wenig Rezeption erfahren, kann allerdings als Variante der SDI aufgefasst werden). Floridi vertritt dabei den philosophischen Standpunkt, dass zusätzlich ein positiver Wahrheitsgehalt der Information angenommen werden muss, um von „semantischer Information“ sprechen zu können; dieser erkenntnistheoretische Aspekt von Information soll zunächst ausgeklammert bleiben.

Daten und Dokumente

Die Vorstellung von Informationen als bedeutungstragende Daten (SDI) lässt sich mit dem Shannon'schen Kommunikationsmodell vereinbaren, indem Signale Daten entsprechen. Die Nachrichten, welche Informationen zusammenfassen, können als identische Dokumente aufgefasst werden. Während Daten direkt messbar sind und letztendlich als eine eindeutige Folge von Bits ausgedrückt werden können, kann der Inhalt eines Dokumentes auf verschiedene Weise aus Daten zusammengesetzt sein. Fragen nach der Identität von Dokumenten (Forschungsfragen II und III) können deshalb nur mit Bezug auf die (semantischen) Informationen beantwortet werden, welche in den jeweiligen Dokumenten kodiert sind. Mit zunehmender Digitalisierung sollte weniger die konkrete physische Publikationsform, wie zum Beispiel das Buch, sondern das publizierte Dokument als Materialisierung von Information, Gegenstand der Bibliothekswissenschaft sein. Damit wird die Bibliotheks- und Informationswissenschaft wie Floridi (2002) beschreibt zur angewandten Informationsphilosophie. Aktuelle Forschungsbeispiele für eine solche Auseinandersetzung mit Dokumenten bieten unter anderem die *Functional Requirements for Bibliographic Records* (FRBR), das *OAI Object Exchange and Reuse* Modell (OAI-ORE), sowie das Problem der Migration von Dokumenten im Rahmen der Langzeitarchivierung.

Formate und Standards

Welche Daten in der Praxis jeweils als ein Dokument zusammengefasst werden (Forschungsfrage I), hängt von konkreten Vereinbarungen ab, die in Form von Standard und Formaten ausgedrückt sind. Diese Vereinbarungen können von informellen Traditionen über detaillierte Erschließungsrichtlinien bis hin zu rein formalen Formatvorgaben reichen. In jedem Fall impliziert die Vereinbarung eine gemeinsame Vorstellung (d.h. ein Modell), was im jeweiligen Kontext als ein Dokument anzusehen ist und welche Informationen ein gültiges Dokument enthalten kann. Ein nüchterner Blick auf bibliothekarische Datenformate und Standards und deren Anwendung in der Praxis zeigt allerdings, dass beide Begriff eher dehnbar verwendet werden: viele Daten genügen nicht einmal den syntaktischen Regeln ihrer Formate und die Daten verschiedener Bibliotheken und verschiedener Bibliothekssysteme weisen trotz gemeinsamer Regelwerke oft deutliche Unterschiede auf. Die Auseinandersetzung mit den Daten geschieht fast ausschließlich auf Ebene der Eingabe und Anzeige, während die eigentliche Datenverarbeitung in geschlossene Systeme verbannt wird. Für den bibliothekarischen Umgang mit Daten symptomatisch ist, dass das letzte deutschsprachige Standardwerk von Eversberg (1999) bereits zehn Jahre alt ist.

Die Ursachen des oberflächlichen Umgangs mit Datenformaten sind nicht nur in mangelhafter informatischer Ausbildung, sondern auch darin zu suchen, dass bibliothekarische Regelwerke von der semantischen Ebene, Datenformate (d.h. formale Sprachen) dagegen von der syntaktischen Ebene ausgehen und beide häufig nicht richtig zueinander finden. Da die technische Umsetzung immer an

syntaktische Datenverarbeitung gebunden ist, reicht es nicht aus, auf der Bedeutungsebene stehen zu bleiben. Stattdessen müssen konkrete Formate und Regeln gemeinsam analysiert, adaptiert und geschaffen werden. Die Grundlagen hierfür werden häufig von außerhalb des Bibliotheks- und Informationswesens durch andere Standards vorgegeben (z. B. die Web-Standards des W3C). Ein Beispiel für die mangelnde praktische Umsetzung der semantischen Ebene liefern die FRBR: nachdem sie bereits 1998 publiziert wurden, gab es zwar theoretische Diskussionen und Vorschläge in der Fachwelt; konkrete Umsetzungen in Software und Formaten blieben jedoch weitgehend aus (vgl. Gradmann, 2005 für einen Vorschlag eines RDF-Formates der FRBR).

Da Dokumente und Informationen letztendlich immer nach Maßgaben bestimmter Formate und Standards auftreten, reicht es nicht aus, die Struktur von Informationen semantisch zu beschreiben, sondern es muss die syntaktische Umsetzung in Form von Formaten mit gedacht und umgesetzt werden. Die Informatik stellt hierzu das Handwerkszeug und die Theorie der formalen Sprachen und Grammatiken bereit. Zur Definition von Formaten gibt es eine Vielzahl von Schemasprachen wie zum Beispiel XML Schema und DTD für XML sowie RDFS und OWL für RDF. Die Bezeichnung einzelner Formate als „semantisch“ oder als „Ontologie“ ist dabei irreführend, weil eine so genannte Ontologie – sei sie in OWL, Prädikatenlogik oder einer beliebigen anderen Ontologiesprache verfasst – immer auf der syntaktischen Ebene verbleibt, sofern sie nicht gleichzeitig mit einem Modell verbunden ist. Modelle bilden also das Bindeglied zwischen syntaktischer und semantischer Ebene.

Modelle

Viele Formatbeschreibungen konzentrieren sich auf syntaktische Regeln, so dass der Zweck (Pragmatik) und die Bedeutung (Semantik) der im Format kodierten Informationen in den Hintergrund geraten. Dabei sollte jedem Format implizit oder besser explizit ein Modell zugrunde liegen. So sind beispielsweise MARCXML und RDF/XML zwei verschiedene Formate, die beide auf der Extensible Markup Language (XML) und dem zugrunde liegenden *XML Information Set* Modell basieren. Die in MARCXML bzw. RDF/XML formulierten Informationen können jedoch auch in anderen Formaten, zum Beispiel MARC 21 bzw. Notation3, kodiert werden. Wesentlich ist also nicht das konkrete Format, sondern das dahinter stehende Daten- bzw. Informationsmodell.

Typische Verfahren und Sprachen zur Formulierung von Modellen sind das *Entity Relationship* Modell (ERM), *Object Role Modeling* (ORM) sowie die *Unified Modeling Language* (UML). Diese Modellierungssprachen abstrahieren von konkreten syntaktischen Regeln und beschreiben stattdessen, welche Entitäten existieren können beziehungsweise müssen und wie diese Entitäten durch Relationen miteinander verknüpft sind. Datenmodelle können wiederum durch Metamodellierung beispielsweise als Hypergraph beschrieben werden (vgl. Boyd und McBrien, 2005); wesentlich ist jedoch, dass eine Verbindung zwischen Entitäten und Relationen des Modells mit Bedeutungen der realen bzw. zu modellierenden Welt ausgedrückt werden. Während Formate Mengen von Daten beschreiben, beschreiben Datenmodelle Mengen von Informationen (vertikale Linien in Abbildung 2).

Die übliche Vorgehensweise bei der Datenmodellierung besteht darin, dass ausgehend von einer zu modellierenden Diskurswelt, dem „Universe of Discourse“, ein Datenmodell formuliert wird, das anschließend in Formaten ausgedrückt werden kann. Das Datenmodell wird auch als semantisches oder konzeptionelles Schema und das Format als logisches Schema bezeichnet. Aus semiotischer Sicht kann die Diskurswelt mit der semantischen, und das logische Schema mit der syntaktischen Ebene identifiziert werden, während das Modell eine vermittelnde Stellung einnimmt. Die bis hier vorgestellten Begriffe lassen sich abschließend in einem gemeinsamen Informationsmodell zusammenbringen:

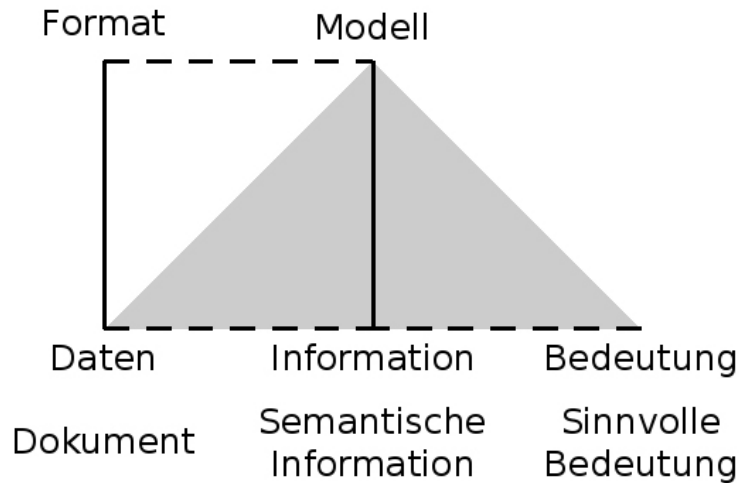


Abbildung 4: Erweitertes Informationsmodell

Der Begriff von Information als bedeutungstragende Daten (SDI) steht dabei im Mittelpunkt. Er ist in das semiotische Dreieck aus Daten (Signifikanten oder Zeichen), Modellen (Signifikaten oder Begriffen) und Bedeutungen (bezeichneten Objekten) eingebettet und stellt eine Verbindung zwischen Zeichen und Objekten her. Um Bedeutung zu besitzen, müssen Informationen dabei einem Modell entsprechen. Entsprechend müssen die Daten, in denen Informationen kodiert sind, einem Format entsprechen, welches das Modell in konkrete syntaktische Regeln umsetzt. Prinzipiell kann jede Menge von informationstragenden Daten als Dokument aufgefasst werden. Aus Sicht der Pragmatik müssen die Informationen, die durch das Dokument ausgedrückt sind, jedoch eine sinnvolle Bedeutung besitzen; erst dann kann auch von semantischer Information gesprochen werden. Die erste der eingangs gestellten Forschungsfragen kann also folgendermaßen beantwortet werden:

- I. Eine Sammlung von Daten kann genau dann sinnvoll als Dokument betrachtet werden, wenn sie in einem Format vorliegt, dem ein sinnvolles Modell zugrunde liegt.

Es sei darauf hingewiesen, dass es sich bei den hier skizzierten Zusammenhängen zwischen Daten, Informationen, Formaten, Modellen und Bedeutungen weniger um ein vollständiges Ergebnis, sondern eher um einen ersten Entwurf im Rahmen eines größeren Forschungsvorhabens handelt. Eine genauere Ausarbeitung muss zum einen die einzelnen Beziehungen zwischen den Entitäten des erweiterten Informationsmodells genauer analysieren (z. B. die zwischen Formaten und Modellen), und zum anderen zeigen, wie sich das Modell in der Praxis gewinnbringend auf tatsächliche Daten, Informationen, Formate, Modelle und Bedeutungen anwenden lässt.

Literatur

Bernhard Eversberg (1999). Was sind und was sollen Bibliothekarische Datenformate. 3. Aufl. <http://www.allegro-c.de/formate/formate.htm>.

Michael Boyd und Peter McBrien (2005): Comparing and Transforming Between Data Models Via an Intermediate Hypergraph Data Model. In: Journal of Data Semantics, Band 4, S. 69-109.

Luciano Floridi (2009). Trends in the Philosophy of Information. In: Handbook of Philosophy of Information. Amsterdam [u.A.]: Elsevier. S. 113-132.

Luciano Floridi (2005). Is Information Meaningful Data? In: Philosophy and Phenomenological Research, Band 70, Nummer 2, S. 351-370.

Luciano Floridi (2002). On defining library and information science as applied philosophy of information. In: Epistemology, Band 16, Nummer 1, S. 37-49.

Stefan Gradmann (2005). rdfs:frbr: Towards an Implementation Model for Library Catalogs Using Semantic Web Technology. In: Cataloging & Classification Quarterly, *Band* 39, Nummer 3/4.

Ulrich Riehm, Knud Bohle und Bernd Wingert (2004): Elektronisches Publizieren. In: Kuhlen, Seeger, Strauch (Hrsg.): Grundlagen der praktischen Information und Dokumentation. 5. Aufl. München: Saur. S. 549-560.

Claude E. Shannon, Warren Weaver (1949). The Mathematical Theory of Communication. Urbana: University of Illinois Press.

Claude E. Shannon (1953) The Lattice Theory of Information. In: Transactions of the IRE Professional Group on Information Theory. Band 1, Nummer 1, S. 105-107.

IFLA Study Group on the Functional Requirements for Bibliographic Records (1998). Functional requirements for bibliographic records: Final report. München: Saur.
<http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records>