

Bericht der Arbeitsgruppe Metadaten

Wolf-Fritz Riekert¹

Metadaten sind Daten über Daten. Auf den Anwendungsfall Umwelt bezogen, bedeutet das: Während sich die Umwelt mit Hilfe von Umweltdaten beschreiben läßt, lassen sich Umweltdaten selbst, auf einer abstrakteren Ebene, mit Hilfe von Metadaten beschreiben. Metadaten sind in der Informatik, und speziell in der Umweltinformatik seit längerem ein wichtiger Gegenstand der Forschung (Radermacher 1991, Jaeschke et al. 1992). Da Metadaten insbesondere auch eine Schlüsselfunktion für die Integration von Umweltdaten besitzen, ist es nicht verwunderlich, daß sie von Anfang an einen thematischen Schwerpunkt des GI-Arbeitskreises „Integration von Umweltdaten“ darstellten. Verschiedene Autoren unternahmen den Versuch, den Begriff „Metadaten“ abzugrenzen und zu konkretisieren. So gibt G. Wiederhold (1995) in einem Glossar zum Thema „Intelligente Integration von Information (I³)“ eine Definition der Begriffe *Metadaten*, *Metainformation* und *Metawissen* (siehe Tafel 1). In einem Umweltinformatik-Lehrbuch unterscheiden R. Denzer und R. Güttler (1995) zwischen *syntaktischer*, *semantischer*, *struktureller* und *navigatorischer* Metainformation. Eine weitere Möglichkeit der Differenzierung ist die Unterscheidung zwischen *Zusatzdaten* (auch *Codaten* genannt) und eigentlichen Metadaten (Riekert 1995).

<p>data = recorded facts, typically maintainable by clerks</p> <p>information = data useful to a customer</p> <p>knowledge = meta-data, relationships among terms, processing paradigms, ..., needed to transform data to information</p> <p>meta-data = descriptive information about data in a resource, includes domain assignment, ownership, access restrictions, database model</p> <p>meta-knowledge = descriptive information about knowledge in a resource, includes ontology, domain coverage, ownership, access restrictions, representation</p> <p>meta-information = descriptive information about services, as owner, capabilities, cost</p>
--

Tafel 1: Definitionen aus Wiederhold (1995)

Etwa die Hälfte der Teilnehmer des 3. Workshops „Integration von Umweltdaten“ interessierte sich für eine intensivere Auseinandersetzung mit dieser Thematik und

¹ Forschungsinstitut für anwendungsorientierte Wissensverarbeitung (FAW) an der Universität Ulm, Postfach 2060, D-89010 Ulm
Tel.: (0731) 501-500, Fax: (0731) 501-999, mailto: riekert@faw.uni-ulm.de

entschied sich für die Mitarbeit in der Arbeitsgruppe Metadaten. Im einzelnen beteiligten sich: G. Behling (GSF Oberschleißheim), R. Denzer (HTW Saarbrücken), G. Dostert (Universität Saarbrücken), W. Fabricius (Bundesinstitut für gesundheitlichen Verbraucherschutz, Berlin), K. Großmann (Zentralstelle für Agrardokumentation und -information, Bonn), R. Güttler (HTW Saarbrücken), D. Hack (Innenministerium Saarbrücken), R. Kramer (FZI Karlsruhe), W.-F. Riekert (FAW Ulm), J. Risau (OFFIS Oldenburg), R. Schorr (HTW Saarbrücken), A. Schuch (Arkusa Ingersheim), W. Sinowski (GSF, Oberschleißheim) und M. Tischendorf (Arkusa Ingersheim).

Die Arbeitsgruppe näherte sich dem Thema anhand von konkreten Fragestellungen. Die Teilnehmer der Arbeitsgruppe stellten zunächst sich selbst und ihr Arbeitsgebiet vor und gaben dann konkrete Beispiele aus ihrem Arbeitsgebiet, in denen Metadaten eine Rolle spielen. Das Ziel war es, zunächst eine Vielzahl unterschiedlicher Anwendungsfälle anzuschauen, ohne sich dabei gleich ein formales Korsett anzulegen. So hat die Gruppe zunächst keine Definition des Begriffs Metadaten versucht und auch keine Unterscheidung getroffen zwischen den verschiedenen, in der Literatur beschriebenen Kategorien von Metadaten, Metainformationen, Metawissen, Zusatzdaten usw., wie sie oben aufgezählt wurden. Auf informelle Weise und ausgehend von einem intuitiven Verständnis wurde die Rolle von Metainformationen in verschiedenen Systemen diskutiert, an denen einzelne Teilnehmer arbeiten. Unter anderem wurde die Datenbank des Forschungsverbundes Agrarökosysteme München (FAM), das Umweltforschungsinformationssystem UFIS, der Einsatz von Informationstechnik in einer Beratungsstelle für Umwelt und gesundheitliche Fragen sowie der Aufbau eines Informationsservers für die Zentralstelle für Agrardokumentation und Information (ZADI) unter dem Aspekt der Metadaten diskutiert.

Um die Inhalte dieser Diskussionen zu strukturieren, war eine Unterteilung sehr hilfreich, die Ralf Denzer einbrachte. Ralf Denzer unterschied drei wichtige Anwendungsgebiete der Metainformation.

- Das erste Anwendungsgebiet betraf die *Navigation* und *Recherche*, also die Nutzung von Metainformation, um an den Datenbestand zu gelangen, der einen interessiert.
- Das zweite Anwendungsgebiet bezog sich auf das Thema *Dokumentation* und *Präsentation*. So kann Metainformation dazu dienen, Inhalt und Funktionalität eines Systems an sich zu dokumentieren; mit dieser Information ist es insbesondere möglich, automatisch Präsentationen der Inhalte und der Funktionalität eines solchen Systems zu generieren, z. B. in Form von Bildschirmmasken.

- Das dritte Anwendungsgebiet ist die *Wiederverwendbarkeit* und *Bestandssicherung* von Daten. Hier liegt die Überlegung zugrunde, daß Daten, die in einem Archiv abgelegt werden, später nur dann wieder nutzbar sind, wenn alle Meta-informationen vorhanden sind, um sie zu interpretieren.

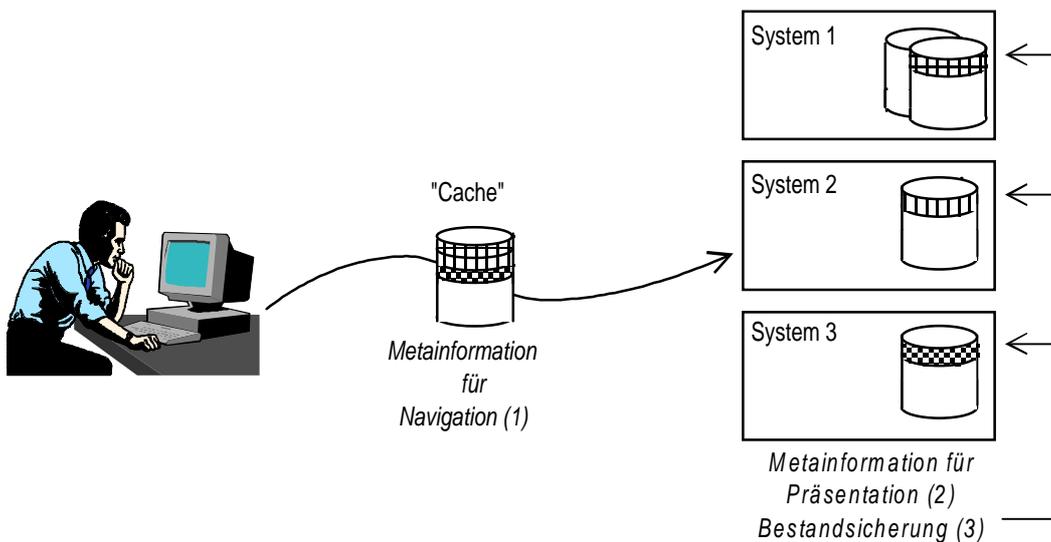


Abbildung 1 : Drei Anwendungsfälle für Metainformation

Die drei Anwendungsfälle sind dargestellt in Abbildung 1. Zu sehen ist ein Mensch – es könnte aber auch ein Computer sein –, der auf bestimmte Datenbestände zugreifen möchte. Diese Datenbestände sind versteckt in bestimmten Informationssystemen, die der Informationssuchende nicht kennt, und das Problem ist jetzt, diese aufzufinden. Benötigt wird hierfür eine zentrale Instanz (dargestellt in der Mitte der Abbildung), die die Metainformation bereitstellt, die erforderlich ist, dorthin zu navigieren. An Ort und Stelle bei den Informationssystemen wird dann Metainformation benötigt für die Präsentation und die Bestandssicherung der Daten. Die Arbeitsgruppe hat hierzu dann Anforderungen formuliert. Zum einen war sie der Meinung, daß die datenhaltende Stelle gemeinsam mit den Daten auch die zugehörigen Metainformationen pflegen sollte. Diese Forderung gilt prinzipiell auch für die Metainformation, die an zentraler Stelle zur Navigation zur Unterstützung vorgehalten wird. Hier wurde die Vorstellung vertreten, daß es sich bei diesen Metadaten um eine Art *Cache* handelt, der Auszüge aus den Metadatenbeständen der Einzelsysteme als Kopien vorhält, so daß für die Zwecke der Navigation und Recherche nicht alle Einzelsysteme konsultiert werden müssen.

Überhaupt nahm das Thema *Navigationsunterstützung* einen großen Raum in den Diskussionen der Arbeitsgruppe ein. Systeme, die das Auffinden von Umwelt-

datenbeständen unterstützen, wie z.B. der Umweltdatenkatalog (UDK), wurden als sehr wichtig empfunden; dennoch wurden große Defizite beim UDK in seiner derzeitigen Form als Mittel zur Navigationsunterstützung festgestellt. Zum einen wurde kritisiert, daß die Pflege der Metadaten im UDK zentral erfolgt und nicht dezentral durch die datenhaltenden Stellen, wie dies oben gefordert wurde. Zum anderen wurde gefordert, daß es möglich sein muß, auf die Originaldaten und die Originaldienste, die in einem Metainformationssystem beschrieben sind, direkt durchzugreifen, so daß die Recherche nicht einfach bei einer Adresse einer Behörde endet, die für die Daten zuständig ist. Man möchte eigentlich sofort auf die Originaldaten durchgreifen können und dabei auch verifizieren können, ob das Gefundene wirklich mit dem übereinstimmt, was man gesucht hat.

Die Arbeitsgruppe hat dann versucht, sich einer Definition des Begriffs *Metadaten* ein Stück weit zu nähern. Nach ersten Versuchen, bestimmte Beispiele von Metadaten aufzuzählen, wurde die Kritik laut, eigentlich gäbe es keine umfassende Liste von Beispielen. Anstelle einer aufzählenden Definition wurde deshalb dann eine intentionale Definition versucht. In der Formulierung von Ralf Kramer liest sich das so: „*Metadaten sind die Daten, die – in Abhängigkeit von der jeweiligen Fragestellung – die Navigation in den Daten gestatten und die korrekte Interpretation der Daten für die Fragestellung gestatten.*“ Ein wesentliches Merkmal von Metadaten ist also, daß sie als Mittel zur Interpretation von Daten dienen können.

Schließlich wurde doch noch der Versuch unternommen, Beispiele aufzuzählen für die verschiedenen Erscheinungsformen, in denen Metainformation anfällt. Das Ergebnis ist in Abbildung 2 gezeigt. In jeder – in Informationssystemen vorkommenden – Aggregationsstufe von Daten und Funktionalitäten fallen Metainformationen an. Diese Metainformationen umfassen *Beschreibungen* von *Computernetzen*, von *Systemen*, die sich in diesen Netzen befinden, von *Datenbeständen*, die diese Systeme enthalten und von den *Datenstrukturen*, aus denen diese Datenbestände aufgebaut sind. In relationalen Systemen sind dies bspw. *Tabellen*, die aus *Spalten* bestehen, und auch hierfür werden zusätzliche Beschreibungen benötigt. Das setzt sich fort bis zum einzelnen *Meßwert* oder *Zahlenwert*, zu dem noch Zusatzdaten vorhanden sein können. Darüber hinaus werden auch Metadaten benötigt, um *Methoden*, *Dienste* oder *Modelle* zu beschreiben, die ebenfalls Komponenten eines Informationssystems darstellen; dieser Fall wurde jedoch von der Arbeitsgruppe nicht eingehend untersucht.

Die hier angeführten Beispiele für Metainformationen sind gewiß nicht vollständig. Nichtsdestoweniger erschienen sie als Kandidaten gut geeignet, wenn es darum geht, in einer konkreten Anwendung einen sinnvollen Bestand an Metainformationen zu identifizieren. Ein Gedanke, der in der Arbeitsgruppe geäußert wurde, war,

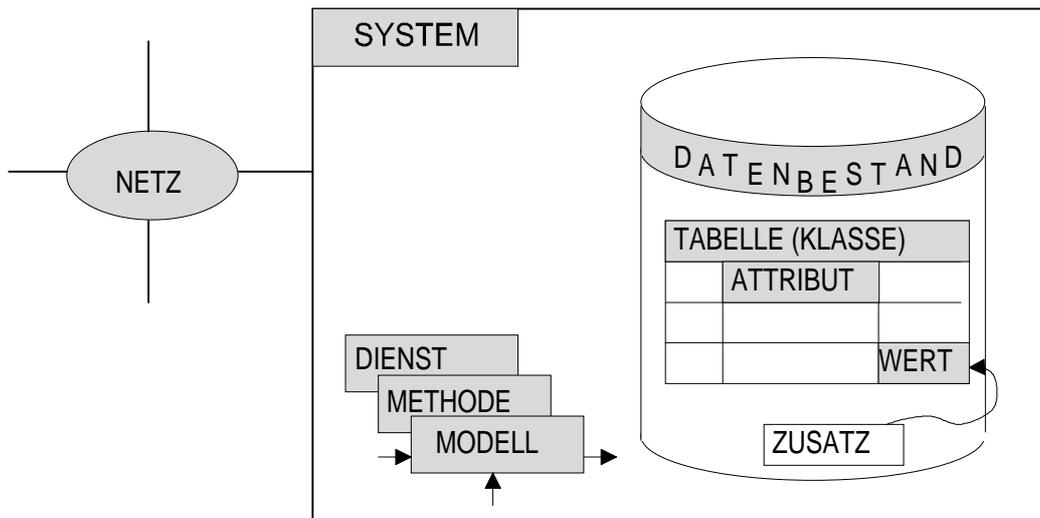


Abbildung 2 : Systemkomponenten, zu denen Metainformationen existieren können.

die Liste dieser Kandidaten zu einer „Checkliste“ zu erweitern. Im Extremfall benötigt man Beschreibungen zu allen genannten Komponenten eines Systems, und zwar hinsichtlich besonderer Merkmale oder besser gesagt *Metamerkmale*. Dazu zählen Kriterien wie *Fachbezug*, *Ortsbezug*, *Zeitbezug*, *datenhaltende Stelle*, *Genauigkeit* und *Maßeinheiten*. Hinzu kommen *allgemeine Zusatzbemerkungen*, die fast jeder Aggregationsstufe bis hinab zum einzelnen Meßwert assoziiert werden können, wie bspw. der Sachverhalt: „*Dieser Meßwert wurde bei Regenwetter gemessen.*“ Darüber hinaus kann es noch Metadaten geben, die Querbezüge zwischen den obengenannten Systemkomponenten beschreiben.

Eine große Zahl von Fragen konnte nicht behandelt werden und mußte offen bleiben. Dennoch hat die Gruppe versucht, diese offenen Fragen festzuhalten. Hierzu zählte die Frage „*Wie kann ich Metainformationen einsetzen für Fragen des Business Reengineering in der Umweltverwaltung?*“, die D. Hack eingebracht hat. Ein anderes Thema, das kaum diskutiert werden konnte, war die Frage der Langzeitar Archivierung von Daten. Metainformationen über Modelle, Methoden, Dienste usw. wurden, wie bereits angeführt, ebenfalls nur peripher behandelt. Eine tiefere Beschäftigung mit diesem Thema wurde zwar als notwendig angesehen, war aber im Rahmen der Arbeitsgruppe nicht leistbar.

Eine weitere offene Frage bleiben *Standards* für Metainformationen. Da Metadaten insbesondere auch die Brücke zwischen unterschiedlichen Systemen bilden sollen, ist es natürlich ausgesprochen wünschenswert, daß wenigstens auf der Metaebene systemübergreifende Standards existieren. Metadaten-Standards sind auch erforderlich für die Konsistenzerhaltung; dies ist insbesondere dann wichtig, wenn an unterschiedlichen Stellen dezentral Metainformationen beigesteuert werden sollen

für ein gemeinsames Metadatenchema. Erforderlich ist zum einen, daß die Metadaten den Datenbestand wirklich korrekt beschreiben, zum anderen aber auch, daß sie im Vergleich zu Beschreibungen anderer Datenbestände nach einheitlichen Kriterien abgefaßt sind, so daß eine Interoperabilität der verschiedenen Metainformationen gegeben ist.

Angesichts der Bedeutung der Metainformationen wurde die eindringliche Forderung aufgestellt, zu allen Umweltdaten generell auch die zugehörigen Metadaten bereitzustellen. Dies wurde als so wichtig angesehen, daß insbesondere gefordert wurde, in allen Umweltprojekten zuerst die Metakonzepte zu klären, bevor begonnen wird, Untersuchungen vorzunehmen und konkrete Umweltdaten zu erheben. Wenn man das aber will, dann braucht man dazu auch einen organisatorischen Formalismus, wie man zu einem Metadatenkonzept kommt. Es würde ein Vorgehensmodell gefordert, das zu einer Konzeption für die Metadaten führt, und es wurde vorgeschlagen, diese Frage zum Gegenstand einer der nächsten Sitzungen des Arbeitskreises zu machen.

Ein interessantes Thema, gerade auch für Informatiker, ist die Entwicklung von Automatismen zur Gewinnung von Metainformation. In dem Vortrag des Berichterstatters wurde gezeigt, wie man aus in Textform vorliegenden Umweltinformationen mit Hilfe eines Thesaurus Metainformation gewinnen kann. Die Frage steht jetzt an, wie man aus existierenden Datenbanken Metainformationen gewinnen kann. Inwieweit ist es bspw. möglich, durch Analyse eines Data Dictionary oder gar von konkreten Daten eines Informationssystems nützliche Metadaten zu gewinnen, etwa im Sinne eines „Data Mining“?

Die Arbeitsgruppe war sich einig über die hohe Bedeutung, die Metadaten in der Umweltinformatik besitzen. Umweltdaten ohne Metadaten haben ein kurzes Verfallsdatum und sind auf die Dauer von geringem Nutzen. Die Erkenntnis zu verbreiten, daß die Erhebung von Umweltdaten stets von einer Erhebung der zugehörigen Metadaten begleitet sein muß, wurde einhellig als eine Kernmission für die Umweltinformatik gesehen.

Literatur:

Denzer, R. / Güttler, R. (1995): Integration von Umweltdaten, in: Page, B. / Hilty, L.M. (Hrsg.): Umweltinformatik – Informatikmethoden für Umweltschutz und Umweltforschung; 2. Auflage. R. Oldenbourg Verlag, München – Wien, S. 165–189.

- Jaeschke, A. / Keitel, A. / Mayer-Föll, R. / Radermacher, F.J. / Seggelke, J. (1992): Metawissen als Teil von Umweltinformationssystemen, in: Günther, O. / Kuhn, H.; Mayer-Föll, R. / Radermacher, F.J. (Hrsg.): Konzeption und Einsatz von Umweltinformationssystemen, Informatik-Fachberichte 301, Springer-Verlag, Berlin – Heidelberg – New York, S. 115–130.
- Radermacher, F.J. (1991): The Importance of Metaknowledge for Environmental Information Systems, in: Günther, O. / Schek, H.-J. (Hrsg.): Advances in Spatial Databases, Lecture Notes in Computer Science 525, Springer-Verlag, Berlin – Heidelberg – New York. S. 35–44.
- Riekert W.-F. (1995): Cooperative Management of Data and Services for Environmental Applications, in: Huber-Wäschle, F. / Schauer, H. / Widmayer, P.(Hrsg.): GIS 95 – Herausforderungen eines globalen Informationsverbundes für die Informatik. Informatik aktuell. Springer-Verlag, Berlin – Heidelberg – New York, S. 618–625.
- Wiederhold, G. (1995): I³ Vocabulary, in: Hull, R. / King R. (Hrsg.) (1995): Reference Architecture for the Intelligent Integration of Information. Prepared by the Program on Intelligent Integration of Information (I³). Advanced Research Projects Agency. 1995. Appendix B. URL: http://isse.gmu.edu/I3_Arch/index.html.