

## **CARMEN: Content Analysis, Retrieval and Metadata: Effective Net-working**

Ein Halbzeitbericht

**Inka Tappenbeck, Carola Wessel**

Das Projekt CARMEN<sup>1</sup> startete als Sonderfördermaßnahme im Rahmen von *Global Info*<sup>2</sup> im Oktober 1999 mit einer geplanten Laufzeit von 29 Monaten. Der Schwerpunkt des Projekts liegt in der Weiterentwicklung von Konzepten und Verfahren der Dokumenterschließung, die den Zugriff auf heterogene, dezentral verteilte Informationsbestände und deren Verwaltung nach gemeinsamen Prinzipien ermöglichen sollen. Dabei geht CARMEN gezielt einen anderen Weg als die meisten bisherigen Ansätze in diesem Bereich, die versuchen, Homogenität und Konsistenz in einer dezentralen Informationslandschaft technikorientiert herzustellen, indem Verfahren entwickelt werden, durch die physikalisch auf verschiedene Dokumentenräume gleichzeitig zugegriffen werden kann. Eine rein technische Parallelisierung von Zugriffsmöglichkeiten reicht jedoch nicht aus, denn das Hauptproblem der inhaltlichen, strukturellen und konzeptionellen Differenz der einzelnen Datenbestände wird damit nicht gelöst. Um diese Differenzen zu kompensieren, werden Problemlösungen und Weiterentwicklungen innerhalb des Projekts CARMEN in drei Bereichen erarbeitet:

- Metadaten (Dokumentbeschreibung, Retrieval, Verwaltung, Archivierung)
- Methoden des Umgangs mit der verbleibenden Heterogenität der Datenbestände
- Retrieval für strukturierte Dokumente mit Metadaten und heterogenen Datentypen

Diese drei Aufgabenbereiche hängen eng zusammen. Durch die Entwicklungen im Bereich der Metadaten soll einerseits die verlorengegangene Konsistenz partiell wiederhergestellt und auf eine den neuen Medien gerechte Basis gestellt werden. Andererseits sollen durch Verfahren zur Heterogenitätsbehandlung Dokumente mit unterschiedlicher Datenrelevanz und Inhalterschließung aufeinander bezogen und retrievalseitig durch ein Rechercheverfahren ergänzt werden, das den unterschiedlichen Datentypen gerecht wird.

---

1 <<http://www.mathematik.uni-osnabrueck.de/projects/carmen/>>

2 <<http://www.global-info.org>>

Innerhalb des Gesamtprojekts CARMEN werden diese Aspekte arbeitsteilig behandelt. Acht Arbeitspakete (APs)<sup>3</sup> befassen sich in Abstimmung miteinander mit je verschiedenen Schwerpunkten. Um die Koordination der Arbeiten der verschiedenen APs untereinander zu unterstützen, trafen sich die ca. 40 Projektbearbeiter am 1. und 2. Februar 2001 zum „CARMEN middle OfTheRoad Workshop“<sup>4</sup> in Bonn. Anlässlich dieses Workshops wurden die inhaltlichen und technischen Ergebnisse, die in der ersten Hälfte der Projektlaufzeit von den einzelnen APs erzielt worden sind, in insgesamt 17 Präsentationen<sup>5</sup> vorgestellt.

In CARMEN AP 1<sup>6</sup> wird ein Dokumenten-Upload für ein verteiltes Informationssystem entwickelt. Um dem Nutzer das Finden und Lesen eines Dokumentes zu erleichtern und dessen Authentizität zu gewährleisten, wird ein Instrument zum Upload von Dokumenten auf WWW-Server, zur Dokumenten-Beschreibung mit Metadaten (RDF) und zum Signieren der Dokumente benötigt. *Thomas Severiens*<sup>7</sup> (Universität Oldenburg) stellte mit CURD (Carmen Uploader with RDF and Digsig) ein solches Instrument vor. Möglichkeiten des Einsatzes digitaler Signaturen zur Gewährleistung der Authentizität von Dokumenten werden von *Michael Kaplan* (Technische Universität München) entwickelt, der eine modifizierte Version des OpenSSL-Systems erläuterte. Neben den dort verwendeten privaten und öffentlichen Schlüsseln („private key“ und „public key“) dient auch das Hash-Verfahren der Sicherung von Daten. Hierbei wird eine individuelle Quersumme über Daten erstellt, vergleichbar einem Fingerabdruck. Bei jeder Veränderung des Dokuments ändert sich auch der Hash-Wert und zeigt dadurch die Veränderung an.

Ziel des Projektteils CARMEN AP 2/5<sup>8</sup> ist es, Konzepte, Definitionen und Beispiele für Metadaten zur Archivierung und Rechteverwaltung sowie einen Workflow für deren dezentrale Erstellung zu entwickeln. Damit ist AP 2/5 das einzige Arbeitspaket, in dem neues Metadatenvokabular entwickelt wird. Den Projektverlauf und die bisher erzielten Ergebnisse präsentierten *Alexander Huber*<sup>9</sup> und *Inka Tappenbeck*<sup>10</sup> (beide SUB Göttingen) in zwei Vorträgen. Darin

---

3 Von den anfangs zwölf konzipierten Arbeitspaketen wurden acht realisiert, davon AP 2/5 als Doppelpaket. <<http://www.mathematik.uni-osnabrueck.de/projects/carmen/APabstracts.shtml>>

4 <<http://www.mathematik.uni-osnabrueck.de/projects/carmen/ws2/index.shtml>>

5 <<http://www.mathematik.uni-osnabrueck.de/projects/carmen/ws2/program1.shtml>>

6 <<http://www.physik.uni-oldenburg.de/carmen/ap1/>>

7 <<http://www.mathematik.uni-osnabrueck.de/projects/carmen/ws2/talks/ap1/talk1/>>

8 <<http://HARVEST.sub.uni-goettingen.de/carmen/Projektinhalt/projektinhalt.html>>

9 <<http://www.mathematik.uni-osnabrueck.de/projects/carmen/ws2/talks/ap25/talk1/>>

erläuterten sie, in welcher Weise in CARMEN AP 2/5 auf der Grundlage des „Reference Model for an Open Archival Information System“ (OAIS)<sup>11</sup> Metadaten für die Bereiche Rechteverwaltung und Archivierung entwickelt werden. Ferner stellten sie den Metadatenprototyp vor, dessen Implementierung in XML/RDF in der zweiten Hälfte der Projektlaufzeit realisiert werden wird. Das in AP 2/5 entwickelte Metadatenvokabular wurde in einer Sitzung der AG Metadaten ausführlich diskutiert.

CARMEN AP 4<sup>12</sup> beschäftigt sich mit „Persistent Identifiern and Metadata Management in Science“. Da digitale Objekte nur dann zitierfähig sind, wenn man sie dauerhaft nachweisen kann, stellt die mangelnde Stabilität von URLs und anderen WWW-Adressen ein großes Problem für die wissenschaftliche Kommunikation dar. In diesem AP sollen daher Werkzeuge zur Nutzung stabiler WWW-Adressen erstellt und installiert werden. *Bernd Diekmann*<sup>13</sup> (BIS Oldenburg) nannte die Anforderungen an ein Verwaltungssystem für Persistent Identifier: Es sollte Eindeutigkeit, Referenzierbarkeit, Authentizität, Verfügbarkeit, Administrierbarkeit und Integrierbarkeit gewährleisten. Dies wird in AP 4 durch ein URN/URL-Resolvingverfahren realisiert. Erstanwendungen dieses Verfahrens am FIZ Karlsruhe und an der SUB Göttingen wurden von *Michael Derr*<sup>14</sup> (FIZ Karlsruhe) am Beispiel mathematischer Zeitschriften vorgestellt. Auch Die Deutsche Bibliothek gehört zu den Erstanwendern; *Kathrin Schröder*<sup>15</sup> (DDB) erläuterte die Vergabe von Persistent Identifier durch diese Institution.

Das AP 6<sup>16</sup> befasst sich mit der Verbindung verschiedener von den Anbietern von Informationen verwendeten Metadaten schemata in einem spezifischen, qualitativ hochwertigen Index. *Roland Schwänzl*<sup>17</sup> und *Hartmut Polzer*<sup>18</sup> (beide Universität Osnabrück) erläuterten die in diesem AP zu lösenden Aufgaben, die sich von der Erstellung der Metadaten in Verlagen und Bibliotheken über die Konversion von Metadaten, die Zusammenführung heterogener Dokumentenmengen, die Annotation von Metadaten und deren Relationen untereinander bis zum Einsatz der Metadaten in Suchmaschinen im bibliothekarischen

---

10 <<http://www.mathematik.uni-osnabrueck.de/projects/carmen/ws2/talks/ap25/talk2/>>

11 <<http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-1.pdf>>

12 <[http://www.bis.uni-oldenburg.de/carmen\\_ap4/index.html](http://www.bis.uni-oldenburg.de/carmen_ap4/index.html)>

13 <<http://www.mathematik.uni-osnabrueck.de/projects/carmen/ws2/talks/ap4/talk0/>>

14 <<http://www.mathematik.uni-osnabrueck.de/projects/carmen/ws2/talks/ap4/talk1/>>

15 <<http://www.mathematik.uni-osnabrueck.de/projects/carmen/ws2/talks/ap4/talk2/>>

16 <<http://www.mathematik.uni-osnabrueck.de/projects/carmen/AP6/>>

17 <<http://www.mathematik.uni-osnabrueck.de/projects/carmen/ws2/talks/ap6/talk1/>>

18 <<http://www.mathematik.uni-osnabrueck.de/projects/carmen/ws2/talks/ap6/talk2/>>

und fachwissenschaftlichen Bereich erstrecken. In den nächsten Monaten sollen sich die Aktivitäten in diesem AP insbesondere auf die Entwicklung neuer Konvertierungstechniken, den Einbau von Broker- und Gatherer-Komponenten und die Einbeziehung von RDF-Quellen mit dem Ziel der bibliothekarischen Weiternutzung von Verlagsdaten konzentrieren.

Thema von AP 7<sup>19</sup> ist „A Document Referencing and Linking System“. Strukturierte Dokumente und reichhaltige Metadaten können nur dann sinnvoll genutzt werden, wenn hierfür geeignete Retrieval- und Navigationsfunktionen zur Verfügung stehen. In diesem AP wird daher ein integriertes Retrieval- und Hypertextsystem realisiert, das Metadaten und Volltexte verwaltet. Grundsätzlich wird als Datenformat XML genutzt sowie für die Metadaten die W3C Recommendation RDF. Zum Suchen und Browsen wurde die Software CAP7 erstellt, die gegenwärtig zu CARA (CARMen RDF Application programming interface) weiterentwickelt wird. CARA basiert auf dem Graphenmodell von RDF und wurde von *Stefan Kokkelink*<sup>20</sup> (Universität Osnabrück) vorgestellt. Eine andere in diesem AP entwickelte Software ist HyREX (Hypertext Retrieval Engine for XML). *Norbert Gövert*<sup>21</sup> (Universität Dortmund) erläuterte in diesem Zusammenhang auch die Anfragesprache XIRQL (XML IR Query Language), die struktur- und inhaltsorientierte Anfragen kombiniert.

In CARMEN AP 9<sup>22</sup> geht es um die Entwicklung eines Prototypen für ein verteiltes, fachübergreifendes Informationssystem. Ziel des Arbeitspaketes ist es, am Beispiel der Informationssysteme MathNet und PhysNet die bestehenden Begrenzungen der fachspezifischen Dienste zu überwinden und eine verteilte Suche in interdisziplinären Datenpools zu ermöglichen. Ein wesentlicher Schritt auf dem Weg zu diesem Ziel ist die Zusammenführung der verschiedenen Klassifikationsschemata. In ihren Vorträgen erläuterten *Thomas Severiens*<sup>23</sup> (Universität Oldenburg) und *Judith Plümer*<sup>24</sup> (Universität Osnabrück) die Unterschiede der in MathNet und PhysNet eingesetzten Klassifikationsschemata MSC und PACS und die Möglichkeiten ihrer Zusammenführung. Ferner stellten sie ein Autorenwerkzeug zur Erstellung von Metadaten vor, Mathematic Metadata Markup (MMM), das die Verbreitung qualitativ hochwertiger und in automatisierter Form konvertierbarer Metadaten erlaubt. Ein wesentlicher Teil der Arbeiten in der zweiten Projekthälfte wird im Bereich der Konvertierung erfolgen.

---

19 <<http://ls6-www.cs.uni-dortmund.de/ir/projects/carmen/wp7.html.de>>

20 <<http://www.mathematik.uni-osnabrueck.de/projects/carmen/ws2/talks/ap7/talk1/>>

21 <<http://www.mathematik.uni-osnabrueck.de/projects/carmen/ws2/talks/ap7/talk2/>>

22 <<http://www.physik.uni-oldenburg.de/carmen/ap9/>>

23 <<http://www.mathematik.uni-osnabrueck.de/projects/carmen/ws2/talks/ap9/talk2/>>

24 <<http://www.mathematik.uni-osnabrueck.de/projects/carmen/ws2/talks/ap9/talk1/>>

Gegenstand der Arbeiten in CARMEN AP 11<sup>25</sup> ist es, Verfahren zur Heterogenitätsbehandlung bei textueller Information verschiedener Datentypen und Inhaltserschließungsverfahren zu entwickeln. Die Heterogenität reicht dabei von unstrukturierten Text-Dokumenten über Dokumente in Markup-Formaten bis zu XML-Dokumenten mit RDF-Metadaten. Ziel des AP 11 ist es, exemplarisch für Dokumente der Mathematik und Sozialwissenschaften Transfer-Komponenten zur Kompensation dieser Heterogenität zu erarbeiten. In diesem Zusammenhang sprach *Stefan Kokkelink*<sup>26</sup> (Universität Osnabrück) über die Extraktion von Metadaten aus verschiedenen Datenformaten und stellte das heuristische Verfahren vor, durch das aus Postskript- (Mathematik) und HTML-Dokumenten (Sozialwissenschaften) in automatisierter Form Metadaten erzeugt werden. Bisher ist es gelungen, eine Heuristik für die Extraktion von Abstracts, Keywords und MSC Klassifikation (Mathematik) bzw. von Titel, Keywords und Abstracts (Sozialwissenschaften) zu entwickeln. Ziele für 2001 sind die Extraktion weiterer Metadaten sowie die Integration der Extraktions-Software in CARMEN AP 7. *Robert Strötgen*<sup>27</sup> (IZ Sozialwissenschaften) erläuterte die Architektur und Arbeitsweise der Transfermodule im Bereich der statistisch-quantitativen und der intellektuellen Transfers über Crosskonkordanzen und Thesauribeziehungen. Ende 2001 soll eine lauffähige Version vorgestellt und in ersten praktischen Tests eingesetzt werden.

CARMEN AP 12<sup>28</sup> beschäftigt sich mit der Konkordanz von Klassifikationen und Thesauri. Die unterschiedliche Verwendung dieser Systeme in Bibliotheken und Fachinformationssystemen erschwert die fach- und datenbankübergreifende Suche, da der Nutzer jeweils mit anderen Suchbegriffen und eigener Suchlogik arbeiten muss. Ziel dieses APs ist es daher, eine integrierte Suche nach sachlichen Gesichtspunkten in verteilten Datenbeständen mit unterschiedlichen inhaltlichen Schwerpunkten zu ermöglichen. Dies soll mit Hilfe von Crosskonkordanzen geschehen.

Bei der Erstellung von Crosskonkordanzen für Klassifikationen wurde entschieden, nicht alle auf eine Klassifikation abzubilden, sondern die Konkordanzen jeweils zwischen zwei Klassifikationen zu erstellen. Wie *Albert Schröder*<sup>29</sup> (Universitätsbibliothek Regensburg) berichtete, erfolgt die Verknüpfung

---

25 <<http://www.bonn.iz-soz.de/research/information/carmen/ap11/>>

26 <<http://www.mathematik.uni-osnabrueck.de/projects/carmen/ws2/talks/ap11/talk1/>>

27 <<http://www.mathematik.uni-osnabrueck.de/projects/carmen/ws2/talks/ap11/talk2/>>

28 <<http://www.bibliothek.uni-regensburg.de/projects/carmen12/index.html.de>>

29 <<http://www.mathematik.uni-osnabrueck.de/projects/carmen/ws2/talks/ap12/talk1/>>

über die Notation. Inzwischen sind die Klassifikationen MSC, PACS (in zwei Versionen), RVK, BK, DDC und die des IZ Sozialwissenschaften dafür vorbereitet worden. Zur verteilten Erstellung von Crosskonkordanzen wurde das Werkzeug Carmen X entwickelt.

Die Erstellung von Crosskonkordanzen für Thesauri erläuterte *Hannelore Schott*<sup>30</sup> (IZ Sozialwissenschaften). Hier werden die SWD, der Thesaurus Sozialwissenschaften und das Schlagwortmaterial des Deutschen Instituts für Internationale Pädagogische Forschung berücksichtigt. Bisher wurde die Software SIS-TMS (Thesaurus Management System for Distributed Digital Collections des ICS-FORTH) verwendet.

In beiden AP-Teilbereichen wurden Entsprechungen benannt, die die Äquivalenz, Enger-weiter-Beziehungen (in beide Richtungen) oder die Ähnlichkeit von Begriffen beschreiben; bei der Wahl der Termini muss noch eine Abstimmung erfolgen. Außerdem sollen die Möglichkeiten von Crosskonkordanzen zwischen Thesauri und Klassifikationen erforscht werden.

Im Anschluss an die Vorträge und in den Sitzungen der Arbeitsgruppen „Metadaten“ und „Heterogeneity/Retrieval“ wurden die Ergebnisse ebenso wie die Desiderate und weiteren Arbeitsplanungen konstruktiv diskutiert. Dabei konnten die einzelnen Bereiche des Projekts CARMEN neu aufeinander abgestimmt und die Perspektiven für die kommenden Monate deutlich gemacht werden. Es ist geplant, die in der zweiten Hälfte der Projektlaufzeit erzielten Ergebnisse in einem Abschlussworkshop zu präsentieren. Obwohl dieser in Analogie zum Halbzeittreffen dann möglicherweise den Titel „CARMEN end OfTheRoad“ tragen wird, ist schon jetzt absehbar, dass dieses Projekt keinesfalls in eine Sackgasse führt, sondern Ergebnisse hervorbringt, die in Bibliotheken und anderen Informationseinrichtungen vielfältig nachnutzbar sein werden.



---

30 <<http://www.mathematik.uni-osnabrueck.de/projects/carmen/ws2/talks/ap12/talk2/>>