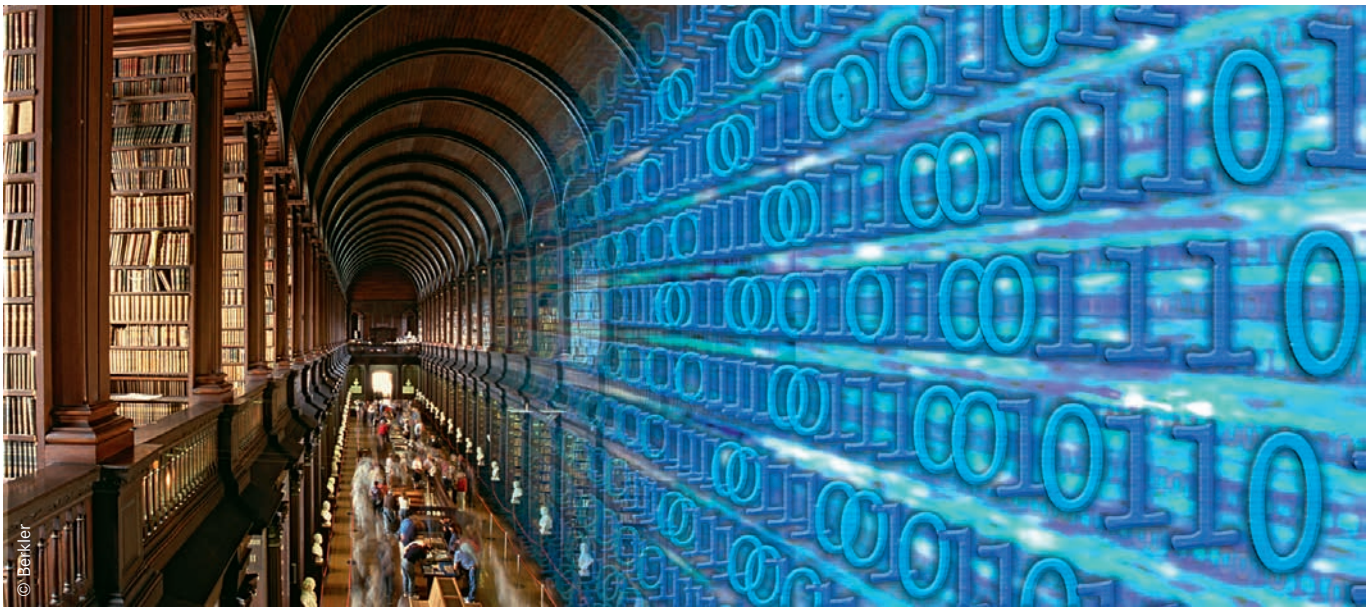


Eine Maschine, die Kunst und Kultur in Zahlen verarbeitet

Vera Münch

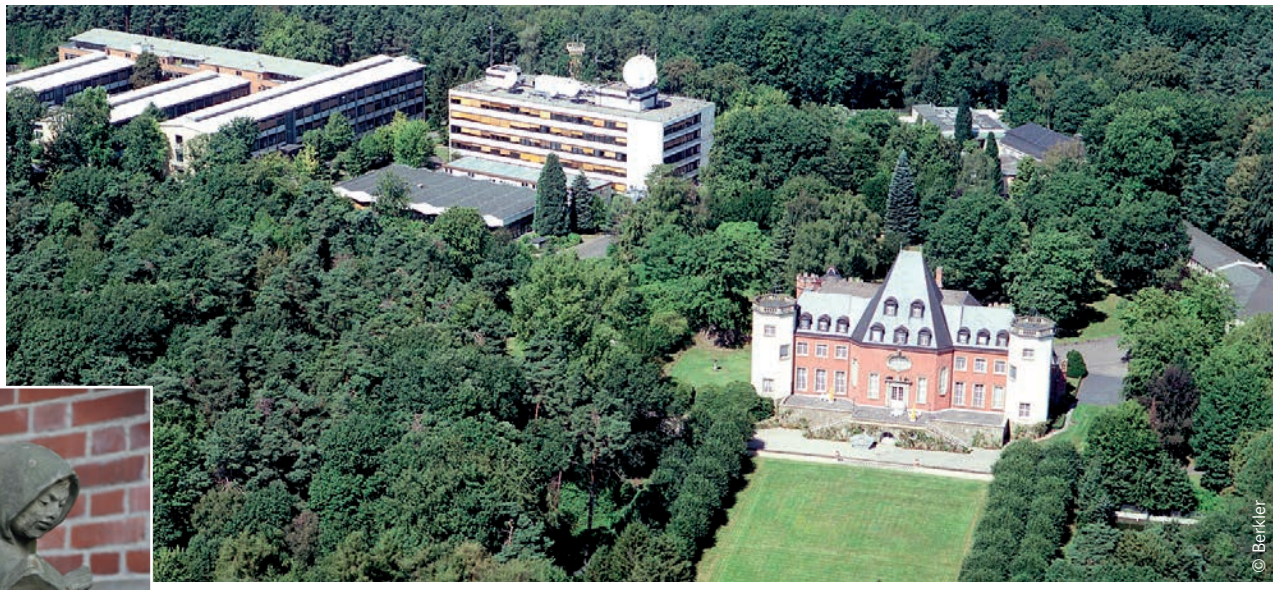
Das Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS hat mit der Softwareinfrastruktur IAIS-CORTEX ein automatisches Datenlogistikzentrum entwickelt, das als Systemkern der Deutschen Digitalen Bibliothek seit Ende 2011 bei FIZ Karlsruhe im Pilotbetrieb läuft.

Die Deutsche Digitale Bibliothek soll zukünftig die Kulturschätze, die in Deutschlands Kultur- und Wissenschaftseinrichtungen aufbewahrt, ausgestellt oder ausgeliehen werden, kostenfrei über das Internet für alle Bürgerinnen und Bürger zugänglich machen. Dieses ambitionierte Ziel haben sich die Protagonisten des zentralen deutschen Portals für Kultur und Wissenschaft auf die Fahnen geschrieben. Der Aufbau der ersten Ausbaustufe wird vom Beauftragten der Bundesregierung für Kultur und Medien (BKM) mit Mitteln aus dem „IT-Investitionsprogramm" des Deutschen Bundestages finanziert.

Lange bevor die Konzeptionsphase für die Deutsche Digitale Bibliothek Mitte 2010 abgeschlossen war, begannen landauf landab in Fachkreisen heftige Diskussionen darüber, ob das Jahrhundertwerk eine reale Chance hat, jemals zu einem Teil des gesellschaftlichen Alltags zu werden. Auf Podien auf der Buchmesse 2010 und 2011 (Fachbuchjournal berichtete in 6/2011, S.4-10, http://www.fachbuchjournal.de/journal/taxonomy/term/4) auf den Bibliothekartagen, Fachkonferenzen und in unzähligen Gesprächen dazwischen wurden Pro und Kontra eines Zentralportals mit allen anhaftenden Grundsatzfragen wie nationale Digitalisierungsstrategie, Urheberrecht, Zugriffsrechte, wer finanziert die Digitalisierung der Kulturgüter? usw. beleuchtet. Alle diese Fragen sind noch nicht abschließend gelöst, weshalb die Skepsis nach wie vor groß ist, obwohl „alle Parteien hinter der Sache stehen", wie Wendelin Bieser (BKM) auf der Buchmesse 2011 betonte. Eine bessere politische Ausgangslage kann man sich eigentlich nicht wünschen. Nun müssen den Worten nur noch die Taten folgen.

Der technische Kern des Deutschen Kultur- und Wissenschaftsportals ist fertig

Während draußen im Land die Diskussionen liefen und weiter laufen, haben das mit der technischen Gesamtkonzeption der Deutschen Digitalen Bibliothek sowie der Koordination der Arbeiten zu ihrer Realisierung beauftragte Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS und seine Entwicklungspartner, das Leibniz-Institut für Informationsinfrastruktur FIZ Karlsruhe auf Seiten des tech-



nischen Betriebes sowie Bibliotheken, Archive, Mediatheken, Museen, wissenschaftliche Einrichtungen und Einrichtungen der Denkmalpflege als Partner für die Inhalte (Content und Metadaten) das Kernsystem für den Betrieb der Deutschen Digitalen Bibliothek gebaut. Es heißt IAIS-CORTEX und ist ein Logistikzentrum für die automatische Organisation und Bereitstellung von Datenbeständen. Nach dem Softwarekonstruktionsprinzip einer Service-orientierten Architektur (SOA) modular ausgelegt und mit offenen Schnittstellen versehen, die auf Standards basieren, lässt sich das System flexibel an den jeweiligen Einsatzzweck anpassen. Das Kernsystem umfasst die Services Ingest, Search und Access; also Module für die Datenaufnahme, die Suche sowie den Zugang zu den Informationen über das Webportal. Module für die Zugriffskontrolle und die Langzeitarchivierung digitaler Originale sind als Option vorgesehen; letzteres ist vor allem als Angebot für kleinere Einrichtungen gedacht, die auch die Archivierung der Digitalisate auslagern wollen.

Konzipiert für den großen Anwendungsfall Deutsche Digitale Bibliothek ist die Plattform dafür ausgelegt, die digitalisierten Bestände aus allen potentiell als Contentpartner in Frage kommenden Einrichtungen so aufzubereiten, dass die Kulturobjekte für alle Bürgerinnen und Bürger einfach aber gezielt auffindbar sind und bei den Sucherergebnissen jederzeit nachvollziehbar ist, woher der vorgeschlagene Content und auch die weiteren dazu verfügbaren Information stammen.

Dauerhafte Quellenreferenz zu jedem Informationsobjekt

IAIS-CORTEX kann diese Anforderungen nach Aussage des technischen Leiters der Entwicklungen am Fraunhofer IAIS, Dr. Kai Stalmann, erfüllen. Das Geheimnis hinter diesen Fähigkeiten erläutert er wie folgt: „IAIS-CORTEX macht aus Metadaten logisch verknüpfte, semantische Wissensnetze. Man kann so auf vielen verschiedenen Wegen auf die Inhalte zugreifen; entweder ganz gezielt, dafür stellen wir in der Suchmaschine alle bekannten Suchfunktionen und zusätzlich starke Interessensfilter zur Eingrenzung der Suchräume bereit. Oder man stöbert in den Beständen und entdeckt interessante Inhalte explorativ. Die Referenz zur Originalquelle bleibt immer erhalten, unabhängig davon, über welchen Weg man zu den Inhalten gelangt.“ Interessant ist in diesem Zusammenhang, dass die Entwickler sowohl binären Content, also Digitalisate und Originaldaten, als auch Metadatenobjekte und Referenzobjekte als eigenständige Informationsobjekte betrachten.

Pilotbetrieb bei FIZ Karlsruhe angelaufen

Die Entwicklung von IAIS-CORTEX dauerte anderthalb Jahre. In Spitzenzeiten arbeiteten bis zu 60 Wissenschaftlerinnen und Wissenschaftler des Fraunhofer IAIS daran und rund 30 Einrichtungen stellten ersten Content bereit. Für die vorangestellte Analyse der Anforderungen brauchte das Fraunhofer-Team sechs Monate. Untersucht wurde, welche Kriterien aus politischer, rechtlicher und funktionaler/technischer Sicht ein webbasiertes Informationssystem wie die Deutsche Digitale Bibliothek erfüllen muss. Die Zusammenfassung der Erkenntnisse ist 106 Seiten stark.

Mit der Auslieferung der Software an den Auftraggeber BKM Ende 2011 war der Projektauftrag für Fraunhofer IAIS erfüllt. Nun läuft das Kernsystem bei FIZ Karlsruhe im Pilotbetrieb mit zunächst 30 bis 100 Content- und Anwendungspartnern. Die Freischal-

tung zur Nutzung durch jedermann im Laufe des Jahres 2012 wird vorbereitet.

Produktionsstraßen für digitale Informationsbereitstellung

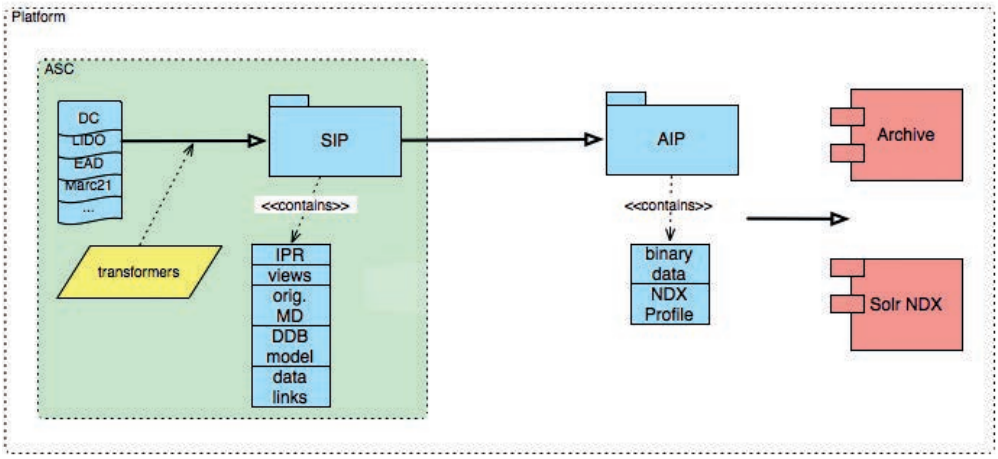
Datenmanagement in dieser Größenordnung muss nach Ansicht der Fraunhofer-Wissenschaftlerinnen und Wissenschaftler mit industriellen Maßstäben gemessen werden. Das heißt, die Forscherinnen und Forscher streben an, möglichst viele Teilabschnitte der Produktionsstraße für digitale Informationsbereitstellung vollständig zu automatisieren. Am Fraunhofer IAIS beschäftigen sich mehrere Forschungsgruppen mit dieser Aufgabe. Sie haben bereits sichtbare Erfolge erzielt, unter anderem hat ein Team um Dr. Stefan Paal und Dr. Stefan Eickeler die multimediale Erschließung von Digitalisaten automatisiert. Ihre Diensteplattform spielt mit dem System des Teams Deutsche Digitale Bibliothek nahtlos zusammen; kann ihm also zur Metadatengewinnung vorangestellt werden.

Die Softwaremaschine IAIS-CORTEX erledigt als zentrale Infrastrukturkomponente für die Datenlogistik alles, was nach dem Digitalisieren und Erstellen der Metadaten durch Bibliothekare, Archivare und Kuratoren – mit oder ohne Unterstützung der Diensteplattform – zu tun ist, um die digitalisierten Objekte über ein Webportal im Internet bereitzustellen. Nach einer vorbereitenden Analyse und Erfassung der Metadatenformate, die vom Content-Partner zur Beschreibung seiner Informationsobjekte eingesetzt werden, läuft der gesamte nachfolgende Prozess voll automatisiert.

Verlustfreie Transformation von Formaten

Das Ingest-Modul von IAIS-CORTEX liest die von den Content-Partnern zu deren Digitalisaten, Archivalien und/oder Objekten gelieferten Metadaten automatisch in die Plattform ein. Auf Wunsch können auch die digitalen Originale, die zu den Metadaten gehören (binärer Content in Form von Digitalisaten, Derivaten, Volltexten, Audiodateien, Bildern, Filmen) eingelesen und zur Archivierung in ein Repository mit Zugangskontrolle abgelegt werden.

Metadaten, die so unterschiedliche digitalisierte Informationsbestände wie Bücher, Archivalien, Filme, Audio, Bilder usw. beschreiben, sind in beinahe unzähligen Formaten kodiert und die Beschreibungen variieren stark – in der Beschreibungstiefe wie im Datenvolumen. Die Frage, wie man diese Vielfalt an Datenformaten und Volumina einer automatischen Weiterverarbeitung zuführen kann, war eine der größten Herausforderungen im Projekt. Zu den bekanntesten Formaten, die im Umfeld der Deutschen Digitalen Bibliothek vorkommen, zählen beispielsweise Marc, EAD, museumdat (MUDA), METS, MODS, Lido und DC, um nur einige zu nennen. Beschreibungsvolumina, die in der Entwicklungsphase angetroffen wurden, reichten von weniger als 1 KB bis zu 100 MB für ein Objekt.



CORTEX-Datenfluss: Der Augmented SIP Creator (ASC) transformiert die Ursprungsdatenformate verlustfrei nach CIDOC CRM (mapping). Grundlage der Datenverarbeitung im ASC sind Transformer Skripte, die in einer XSLT Library verwaltet werden. Der Ingestservice der Plattform erzeugt in der Folge aus den SIPs Archival Information Packages, mit denen das Archiv (Cloud), der Suchmaschinenindex (Solr) und der Node Store (Solr) befüllt werden. Beim Ingest erkennt das System vorhandene, zusammengehörige oder verwandte Informationen anhand von Heuristiken und stellt Verbindungen zwischen neuen und vorhandenen Objekten her.

Das Fraunhofer-Team hat zur Lösung dieser Frage den sogenannten Augmented SIP Creator (ASC) entwickelt. Diese in das Ingest-Modul integrierte Software transformiert die verschiedenen Metadatenformate der Content-Partner ohne Informationsverlust in die IAIS-CORTEX-Plattform (Mapping). SIP steht für Submission Information Package und ist ein Fachterminus aus dem OAIS Referenzmodell (siehe Abbildung „Cortex-Datenfluss“). Nach diesem Schritt werden die Daten maschinell so aufbereitet, dass bei der späteren Benutzung des Portals für Interessenten explorativ Entdecken von Wissen ebenso möglich wird wie der gezielte Zugriff, wenn man weiß, was man sucht; beispielsweise ein Exemplar des Theuerdank aus dem frühen 16. Jahrhundert oder eine chinesische Vase aus der Zeit zwischen 1880 und 1915. Für die Suche stehen alle heute bekannten technischen Möglichkeiten zur Verfügung: einfache und fortgeschrittene Suche ebenso wie eine hoch entwickelte, browserbasierte facettierte Suche.

IAIS-CORTEX veredelt die Daten bei der Verarbeitung

Bei der Datenverarbeitung im Ingestprozess versieht IAIS-CORTEX jedes Informationsobjekt mit einer eindeutigen Kennung (Persistenter Identifikator / Persistent Identifier / PID), unter der es im Repository der Deutschen Digitalen Bibliothek (oder jedem anderen Repository) verwaltet wird. Gegebenenfalls gibt es weitere Kennungen, die auf Herkunft bzw. Standort des digitalen Originals zurückverweisen. Interessant ist an dieser Stelle wie gesagt, dass sowohl binärer Content als auch Metadatenobjekte und Referenzobjekte PID-gekennzeichnete Informationsobjekte sein können.

Durch semantische Anreicherung werden die Metadaten bei der Aufbereitung in IAIS-CORTEX veredelt. Das heißt, man gewinnt durch das Herstellen von Relationen zum Umfeld Informationen über die Bedeutung des Inhaltes, der im Objekt abgespeichert ist. Dieser Versuch, die menschliche Fähigkeit, aus einer Situation logische Rückschlüsse zu ziehen, auf Informationssysteme zu übertragen, ist die Grundlage dafür, um später die komfortablen Suchfunktionen bereitstellen zu können.

Wissensnetze: Aus Triples werden Triples erzeugt

Aus den mit Bedeutung versehenen Informationsobjekten kann die Softwaremaschine nun logische Wissensnetze knüpfen. Wie das geht, ist für Informatiker schnell erklärt: Aus Triples werden Triples erzeugt, allerdings werden die Triples nicht in einem Triplestore verwaltet, sondern vorberechnet und im wesentlich schnelleren Suchmaschinenindex Solr indexiert. Für Nicht-Informatiker dauert die Erklärung etwas länger. Jeder Gegenstand, jedes Ereignis, jedes Datum ist mit weiterem Wissen verbunden, das im menschlichen Gehirn aktiviert wird, wenn die Situation auftritt. Sieht ein Mensch eine Vase, weiß er sofort, dass sie den Zweck hat, Blumen aufzunehmen, erkennt, dass sie ein chinesisches Design hat und aus Porzellan ist. Er sieht, ob sie leer oder mit Blumen gefüllt, alt oder neu ist. Solche Bezüge können Maschinen heute noch nicht herstellen. Semantische Datenaufbereitung versucht, es ihnen beizubringen, indem sie die Objekte mit zusätzlichen Informationen zu ihrer Bedeutung verbindet. Verständlich machen lässt sich das Prinzip dieser semantischen Aufbereitung von Information an einem Beispiel, das so zwar im Web noch nicht umgesetzt ist, die Arbeitsweise aber anschaulich erklärt. Semantische Anreicherung verbindet den Begriff „Hamburg“ mit den weiteren Begriffen Stadt und Deutschland und Norden und Regen und schön und weltoffen und Regenschirm. Solche Relationen werden in sogenannten Triples abgebildet und gespeichert. In der Fachsprache heißen diese Zusatzinformationen Entitäten (Stadt, Deutschland, Norden, Regen, Regenschirm) und Attribute (schön, weltoffen).

Fraunhofer hat in IAIS-CORTEX nicht nur die semantische Erschließung und Anreicherung von Metadaten durch Verknüpfung mit Entitäten und Attributen eingebaut, sondern erzeugt über eine Verknüpfung der Persistenten Identifikatoren aus den Triples auch noch ein Wissensnetz. Dieses Wissensnetz stellt die Informationsobjekte im System als Wissensressourcen dar (abgebildet in einem Graph). Um die Indexierung nicht zu komplex werden zu lassen, haben die Wissenschaftlerinnen und Wissenschaftler ein Verfahren entwickelt, mit dem sie die Eigenschaften der Triples auf ein kompaktes Vokabular reduzieren, das als ISO Standard vorliegt und in zahlreichen Projekten verwendet wurde, das CIDOC CRM.

Der Grundstein ist gelegt

Auch wenn sich IAIS-CORTEX im Betrieb unter hoher Last erst noch dauerhaft bewähren muss, hat Fraunhofer mit dieser Softwareinfrastruktur einen technisch beeindruckenden Grundstein für die Deutsche Digitale Bibliothek gelegt. Bleibt zu hoffen, dass neben der derzeit gesicherten Finanzierung für den Betrieb auch die Mittel für die Digitalisierung der Inhalte durch die Content-Partner bereitgestellt werden. Denn es wird zuallererst von der Attraktivität der Inhalte für die Bürgerinnen und Bürger abhängen, ob die Deutsche Digitale Bibliothek ein Erfolg wird oder nicht. Gelder werden zudem für die kontinuierliche Weiterentwicklung und den weiteren technischen Ausbau gebraucht. Vorgesehen hat das Fraunhofer IAIS-Team zum Beispiel Web 2.0- und Community-Funktionen im Konzept schon. Sie können nur noch nicht realisiert werden, weil die Mittel fehlen. I

Auf der CeBIT 2012 (Hannover, 6. - 10. März) wird IAIS-Cortex in Halle 9 am Fraunhofer Gemeinschaftsstand E 08 und in Halle 7 im Public Sector Parc am Stand des Bundesministeriums für Inneres (BMI) vorgestellt.

Vera Münch
ist freie Journalistin und
PR-Beraterin/PR+Texte
Leinkampstraße 3
31141 Hildesheim
vera-muench@t-online.de