Integration von nicht-textuellen Objekten in das GetInfo Portal der Technischen Informationsbibliothek

Jan Brase, Ina Blümel

jan.brase@tib.uni-hannover.de,ina.bluemel@tib.uni-hannover.de

Abstract: Nicht-textuelle Objekte wie Forschungsdaten aber auch 3D-Modelle sind mittlerweile eine immer wichtigere Quelle des Wissenserwerbes. Die Technische Informationsbibliothek (TIB) als Deutsche Zentrale Fachbibliothek für Technik sowie Architektur, Chemie, Informatik, Mathematik und Physik hat in verschiedenen Projekten begonnen über ihr Portal GetInfo auch solche Objekte suchbar und verfügbar zu machen. Das vorliegende Paper stellt die aktuellsten Beispiele dieser Integration von nicht-textuellen Objekten in ein Bibliotheksportal vor.

1 Einleitung

Wissenschaftliche Information wird immer komplexer und basiert immer häufiger auf großen Datensammlungen. Die übliche Verbreitungsform für Wissen ist immer noch der Wissenschaftliche Artikel, der aber nur den letzten Schritt in einem Prozess darstellt, an dessen Anfang Forschungsdaten gestanden haben. In ihrem 2007 Bericht Cyberinfrastructure Vision for 21st Century Discovery schrieb die National Science Foundation aus den USA:

Science and engineering research and education have become increasingly data-intensive as a result of the proliferation of digital technologies, instrumentation, and pervasive networks through which data are collected, generated, shared and analyzed. Worldwide, scientists and engineers are producing, accessing, analyzing, integrating and storing terabytes of digital data daily through experimentation, observation and simulation. Moreover, the dynamic integration of data generated through observation and simulation is enabling the development of new scientific methods that adapt intelligently to evolving conditions to reveal new understanding. ([8]).

Im dem kürzlich veröffentlichten Buch *The Fourth Paradigm: Data-Intensive Scientific Discovery* [5] von Microsoft Research wird ebenfalls darauf hingewiesen, wie sich die Wissenschaft generell in den letzten Jahrhunderten von einer Empirische zu einer Daten-intensiven Disziplin gewandelt hat. Aus der Sicht von Bibliotheken werden dort ebenfalls neue Aspekte identifiziert:

Scientific communication, including peer review, is also undergoing fundamental changes. Public digital libraries are taking over the role of holding publications from conventional librariesbecause of the expense, the need for timeliness, and the need to keep experimental

data and documents about the data together.

Wissenschaftliche Information tritt heutzutage nicht mehr nur in der Form eines Artikels oder eines Buches in Erscheinung. Daher müssen auch Bibliotheken ihre Kataloge für andere Inhaltsformen öffnen, wenn sie ihren Aufgaben der Informationsversorgung gerecht werden wollen. Dieses bedeutet aber auch gerade im Umgang mit teilweise sehr großen Datensätzen, dass Bibliotheken nicht mehr alle Inhalte selber vorhalten müssen. Entscheidend ist vielmehr den Katalog der Zukunft als ein Portal zu verstehen. Externe Inhalte wie große Datensätze aber auch Filme oder beliebige Lernobjekte werden katalogisiert, um dem Nutzer passende Ergebnisse zu seinen Suchen anzubieten. Der Zugang zu den Inhalten erfolgt dann aber über eine stabile Verlinkung zu den externen, vertrauenswürdigen Archiven und Datenbanken, die auf die Speicherung dieser Inhalte spezialisiert sind. (siehe Abbildung 1).

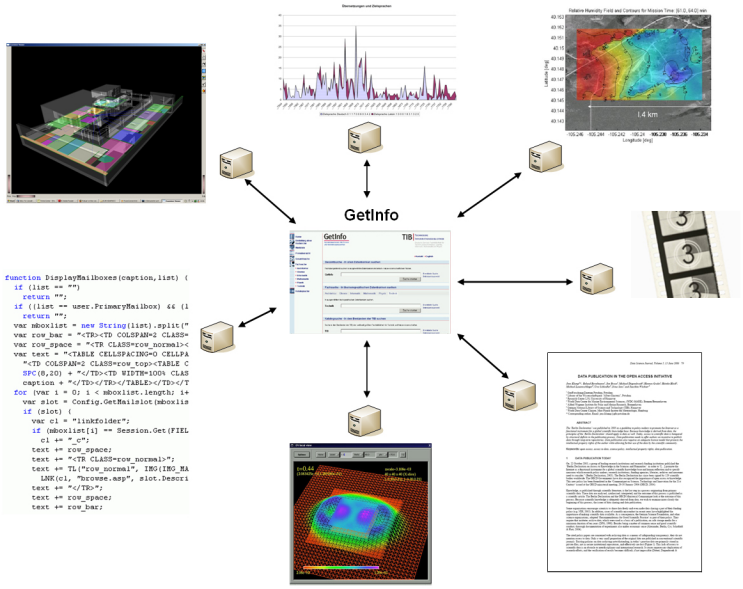

Abbildung 1: Das Bibliotheksportal der Zukunft, beispielhaft für GetInfo der TIB, basierend auf einen Neztwerk von vertrauenswürigen Inhaltsanbietern

2 Zugang zu Forschungsdaten durch GetInfo: Beispiele

Die Technische Informationsbibliothek (TIB) ist die Deutsche Zentrale Fachbibliothek für Technik sowie Architektur, Chemie, Informatik, Mathematik und Physik. GetInfo ist das Portal für technisch- naturwissenschaftliche Fach- und Forschungsinformationen, es bündelt den Zugang zu führenden Fachdatenbanken, Verlagsangeboten und Bibliothekskatalogen mit integrierter Volltextlieferung. GetInfo bietet damit einen weltweit einzigar-

tigen Bestand an technisch-naturwissenschaftlicher Fachinformation. Derzeit ist GetInfo das einzige große Bibliotheksportal in Europa, das auch den Zugang zu Forschungsdaten bietet. Anbei drei Beispiele:

- Bibliotheksportale sind die klassische Quelle für Information [7]. Auf der Suche nach bestimmten Themen mag den Nutzer nicht nur Interesse an allen Publikationen zu dem Thema haben, sondern auch an den Datensätzen, die ein bestimmter Wissenschaftler gesammelt hat. Durch die Vergabe von persistenten Identifiern werden diese Daten direkt über den Katalog zugänglich. Derzeit sind bereits über 2.000 Datensätze, die Grundlage einer Wissenschaftlichen Publikation sind, direkt als eigenständige Objekte zugänglich über GetInfo und den Gemeinsamen Bibliotheksverbund (GBV) [3]. Abbildung 2 zeigt Daten zu einem Bohrkern als Ergebnis einer Suche in GetInfo. Durch Auflösen des persistenten Identifier, gelangt man auf eine Vorschauseite des Datenzentrums, die weitere Metadaten zu dem Datensatz enthält, sowie Download-Links zu einzelnen Teilen des Datensatzes oder den gesamten Daten (Abbildung 3). Dieser Workflow entspricht der traditionellen Verwendung von DOI Namen bei Wissenschaftlichen Zeitschriften.
- Ein weiteres Beispiel nicht-textueller sind geologische Karten. Auch solche sind über GetInfo verfügbar. Die Auflösung des DOI-Namen führt hier ebenfalls auf eine Vorschauseite beim Datenzentrum (Abbildung 4), von der dann die eigentlich Karte zugänglich ist (Abbildung 5)
- Innerhalb des Bereichs der nicht-textuellen Information ist in den letzten Jahren ein starker Anstieg von mulitmedialen Inhalten zu beobachten. Beispielsweise wird die Verwendung von 3D-Modellen im Bereich der Ingenieurswissenschaften immer wichtiger. Mittlerweile sind auch 3d-Modelle über GetInfo suchbar (Abbildungen 6 und 7). Die Darstellung und Suche erfolgt über die Datenbank PROBADO3D (siehe nächster Abschnitt) als eingebundene externe Quelle bei GetInfo (siehe [2])

3 Inhaltsabhängieg Indexierung

Nicht-textuelle Materialien enthalten in der Regel keine beschreibenden Metadaten, wenn Sie nicht extra katalogisiert werden. Heutzutage werden nicht-textuelle Objekte, wenn überhaupt, basierend auf textuellen Metadaten in Bibliotheken indiziert und präsentiert. Eine von der TIB durchgeführte Umfrage unter Architekten ergab, dass Architekten überwiegend mit textuellen Schlagwörtern nach 3D-Modellen suchen und nicht mit visuellen Interfaces. Daher ist es wichtig solche textuellen informationen aus den Modellen zu extrahieren. Im DFG geförderten Projekt PROBADO wurden an der TIB verschieden Verfahren untersucht und implementiert ([1]).

Derzeit verwendete inhaltsbasierte Indexierung-Werkzeuge für 3D-Modelle basieren entweder auf der geometrischen Form ([6]) der Objekte, wodurch eine schnelle Suche durch große Datenbestände möglich ist; oder auf Verbindungsgraphen für die einzelnen Räume. Letzteres Verfahren ist hervorragend für den Einsatz im Architektonischen Kontext geeig-

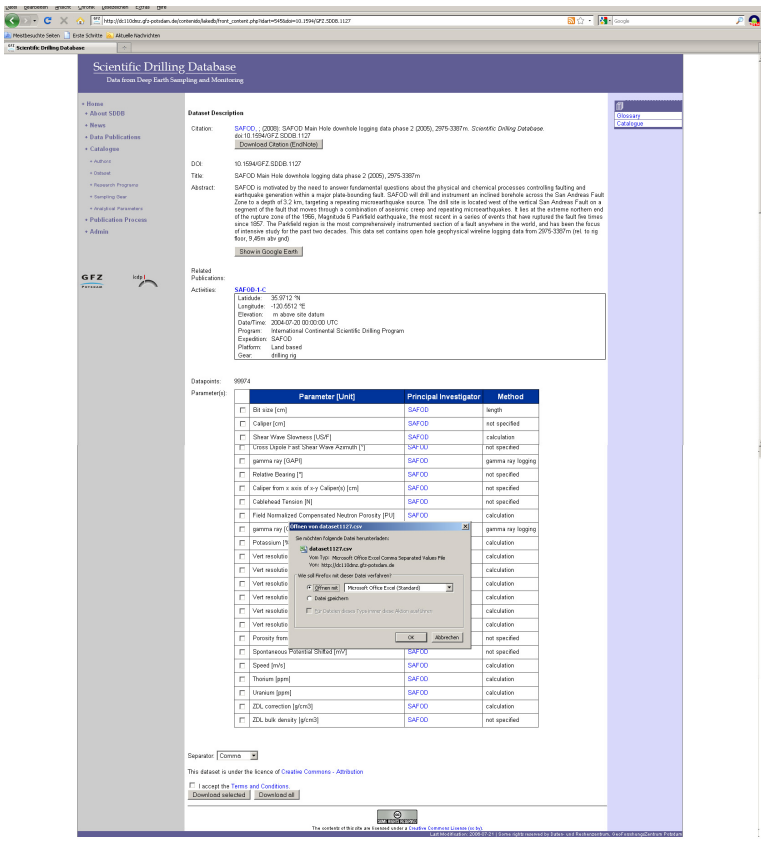

Abbildung 2: Kataloganzeige eines Datensatzes in GetInfo

net und eignet sich sehr gut für die Charakterisierung von Gebäuden ([9]). Räume werden hierbei durch Knoten repräsentiert; Verbindungen wie Türen, Fenster, Treppen usw. durch Kanten (siehe Abbildung 8). Die Suchanfrage kann durch Attributierung des Graphen präzisiert werden, etwa durch Quadratmeter oder Höhe von einzelnen Räumen.

4 Visuelle Suche

Design und Usability einer Suchoberfläche sind wichtige Aspekte für die Akzeptanz von Multimedialen Objekte als Wissenschaftliche Informationsobjekte. Das PROBADO Nutzer-Interface ermöglicht für 3D-Modelle (siehe Abbildung 9):

- Suche in den Metadaten (Titel, Autor, Beschreibung usw.)
- Verwendung von Filtern (Kategorie, Quelle usw.)

Abbildung 3: Präsentation des Datensatzes beim Datenzentrum mit entsprechendem Download-link

- Hochladen von eigenen Modellen und 3D-Skizzieren für Query-by-example Suchen
- Skizzieren und Suche mit Raumverbindungsgraphen in Gebäudemodellen

Für die Möglichkeit über nach den in GetInfo verfügbaren Forschungsdaten visuell zu suchen hat 2010 das Projekt *Visueller Zugang zu Forschungsdaten* gestartet. (http://www.tib-hannover.de/de/die-tib/projekte/visueller-zugang-zu-forschungsdaten/) Ziel ist die Entwicklung und prototypische Umsetzung von innovativen Ansätzen für den interaktiven, graphischen Zugang zu Forschungsdaten, um diese optimal im Information Retrieval Prozess darstell- und suchbar zu machen. Im Projekt werden hierzu Verfahren zur Datenanalyse sowie für visuelle Suchsysteme untersucht und weiterentwickelt, sowie deren prototypische Umsetzung in das Fachportal GetInfo evaluiert. Ziel ist neben der metadatenbasierten Suche auch die unmittelbare Suche in den Forschungsdaten. Die Abbildung 10 visualisiert die graphisch-interaktive Suche nach einem vom Nutzer vorgegebene Kurvenverlaufsmuster innerhalb eines großen Bestandes von Verlaufsmustern. Eine visuelle Clusteranalyse des Gesamtbestandes bildet die Ausgangsbasis. Die Suchanfrage wird über die interaktive Skizzierung des gesuchten Kurvenverlaufs gestellt (siehe links oben). Ein hell-dunkles Colormapping über dem Gesamtbestand visualisiert den Grad der Übereinstimmung des vorgegebenen Verlaufsmusters mit den Mustern innerhalb des gesamten Datenbestandes.

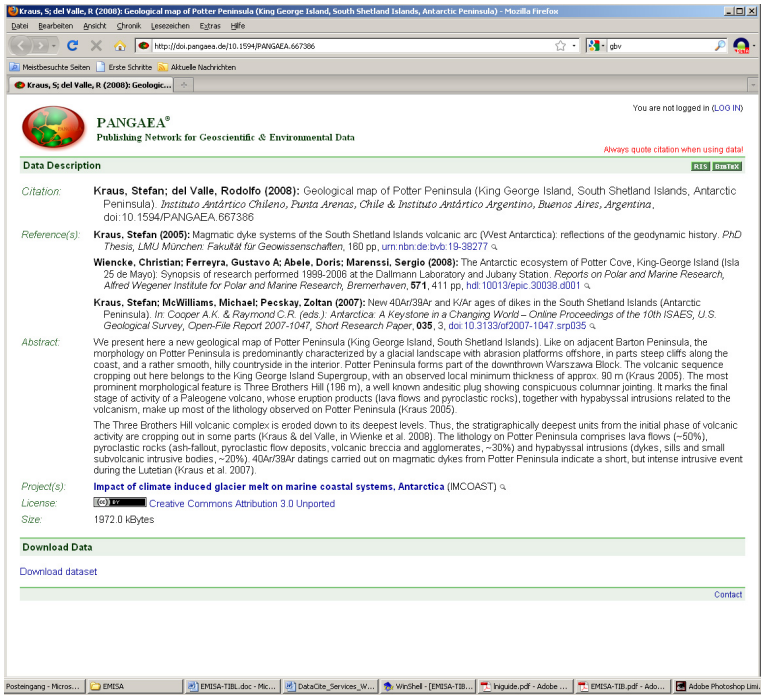

Abbildung 4: Vorschauseite einer geologischen Karte beim Datenzentrum PANGAEA

5 Persistente Identifizierung

Integration von Daten in Publikationen ist ein wichtiger Bestandteil wissenschaftlicher Zusammenarbeit. Es ermöglicht die Verifizierung wissenschaftlicher Ergebnisse und aktiven Wissensaustausch von Forschern. Im wissenschaftlichen Bereich besteht zwar grundsätzlich Bereitschaft, Daten für eine interdisziplinäre Nutzung zur Verfügung zu stellen, aber es ist zur Zeit unüblich, dass die erforderliche Mehrarbeit für Aufbereitung, Kontextdokumentation und Qualitätssicherung im Wissenschaftsbetrieb anerkannt wird. Die klassische Form der Verbreitung wissenschaftlicher Ergebnisse ist ihre Veröffentlichung in Fachzeitschriften, normalerweise ohne Veröffentlichung der zugrunde liegenden Daten. Diese klassische Publikation wird im ”Citation Indexërfasst. Dieser Index wird zur Leistungsbewertung von Wissenschaftlern herangezogen. Datenveröffentlichungen werden darin bisher nicht berücksichtigt. In dem an der Technischen Informationsbibliothek (TIB) Hannover durchgeführten Projekt Publikation und Zitierfähigkeit wissenschaftlicher Forschungsdaten wurde eine Infrastruktur zur Registrierung von DOI-Namen und URNs für Wissenschaftliche Datensätze geschaffen und erfolgreich getestet. Mit dem System wurden an der TIB bereits über 650.000 Datensätze aus dem Bereich der Geowissenschaften mit persistenten Identifiern versehen.

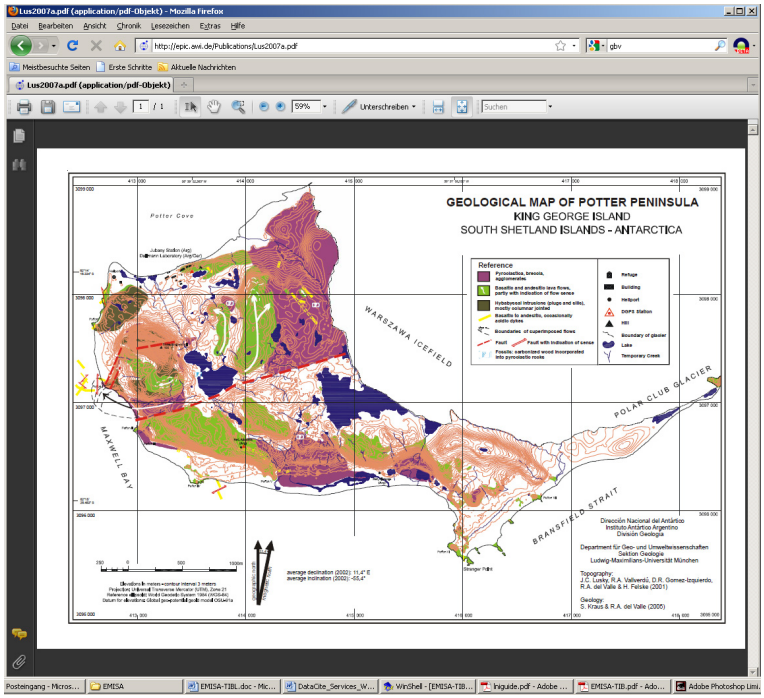

Abbildung 5: Anzeige der eigentlichen Karte

Die Verwendung von DOI-Namen als Identifier ermöglicht eine elegante Verlinkung zwischen einem wissenschaftlichen Artikel und den im Artikel analysierten Forschungsdaten. Artikel und Datensatz sind durch ihre jeweiligen DOI Namen in gleicher Weise eigenständig zitierbar. Diese Form der Zitierung und Verlinkung bietet sich insbesondere bei Forschungsdaten an, die in direkter Beziehung zu Wissenschaftlichen Artikeln stehen, sogenannte suppplementary data. So wird beispielsweise der Datensatz:

Kuhlmann, H et al. (2009):
Age models, iron intensity, magnetic susceptibility records and dry bulk density of sediment cores from around the Canary Islands.
PANGAEA, Bremen
doi:10.1594/PANGAEA.727522

in folgendem Artikel verwendet:

Kuhlmann, Holger; Freudenthal, Tim; Helmke, Peer; Meggers, Helge (2004):
Reconstruction of paleoceanography off NW Africa during the last 40,000 years: influence of local and regional factors on sediment accumulation.
Marine Geology, 207(1-4), 209-224,
doi:10.1016/j.margeo.2004.03.017

Der im Dezember 2009 gegründete Verein DataCite [4] hat sich zum Ziel gesetzt, Wissen-

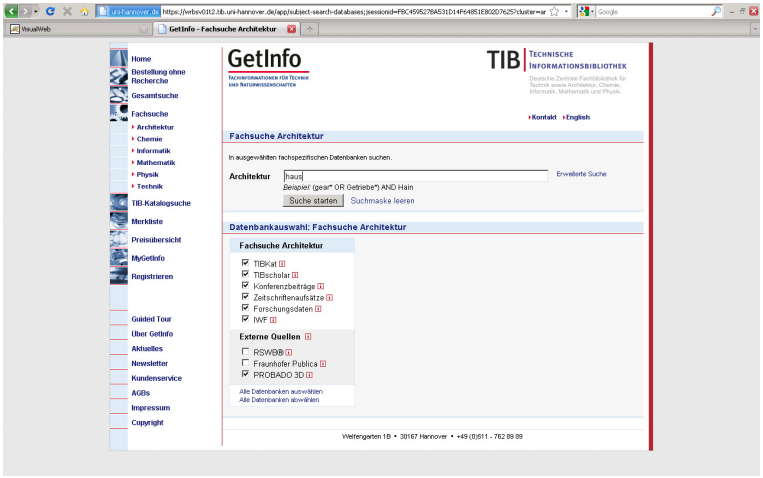

Abbildung 6: Auswahl der Datenbank PROBADO3D als externe Quelle bei GetInfo

schaftlern den Zugang zu Forschungsdaten über das Internet zu erleichtern, die Akzeptanz von Forschungsdaten als eigenständige, zitierfähige wissenschaftliche Objekte zu steigern und somit die Einhaltung der Regeln guter wissenschaftlicher Praxis zu gewährleisten. Bis heute haben sich 12 Partner aus 9 Ländern unter dem Dach von DataCite zusammengefunden: die British Library, das französische LInstitut de lInformation Scientifique et Technique (INIST), das Technical Information Center of Denmark, die TU Delft Bibliothek aus den Niederlanden, das Canada Institute for Scientific and Technical Information (CISTI), die California Digital Library (USA), der Australian National Data Service (ANDS) die Purdue University (USA) und die Eidgenössische Technische Hochschule in Zürich. Deutsche Mitglieder sind neben der Technischen Informationsbibliothek (TIB), die Goportis Partner Deutsche Zentralbibliothek für Medizin (ZB MED) und voraussichtlich ab dem 01.01.2011 die Deutsche Zentralbibliothek für Wirtschaftswissenschaften, sowie das Leibniz-Institut für Sozialwissenschaften GESIS.

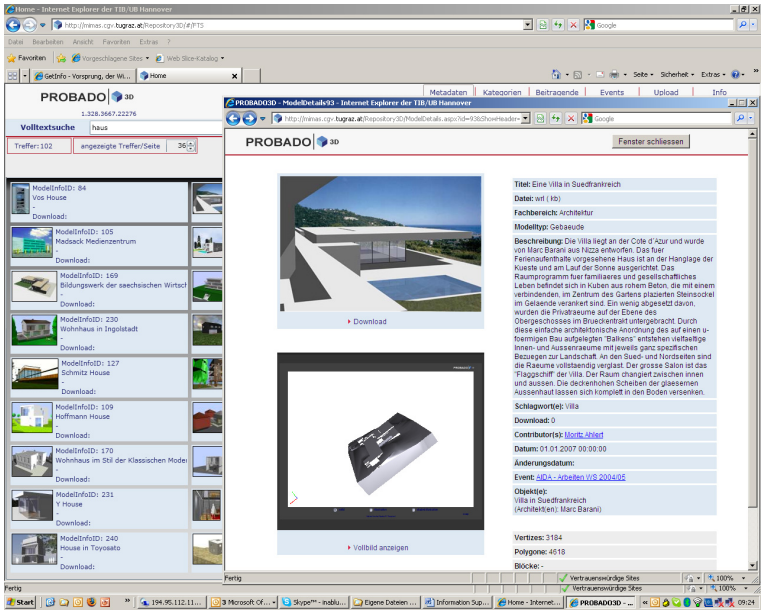

Abbildung 7: Anzeige eines 3D Modells als Suchergebnis

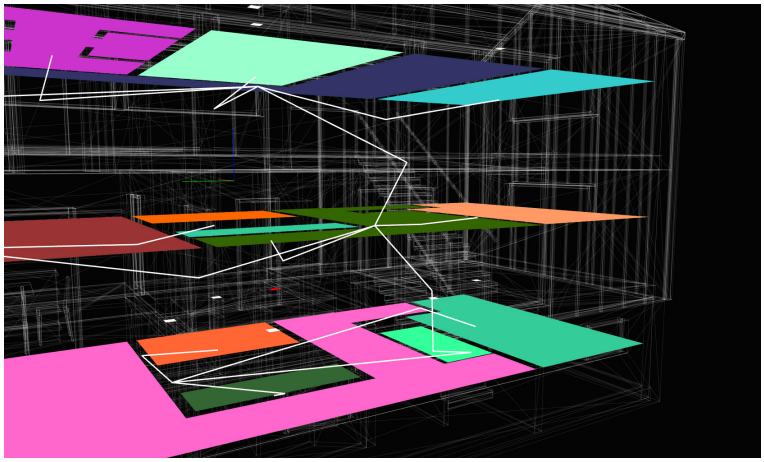

Abbildung 8: Automatisch extrahierter Raumverbindungsgraph eines Gebäudes

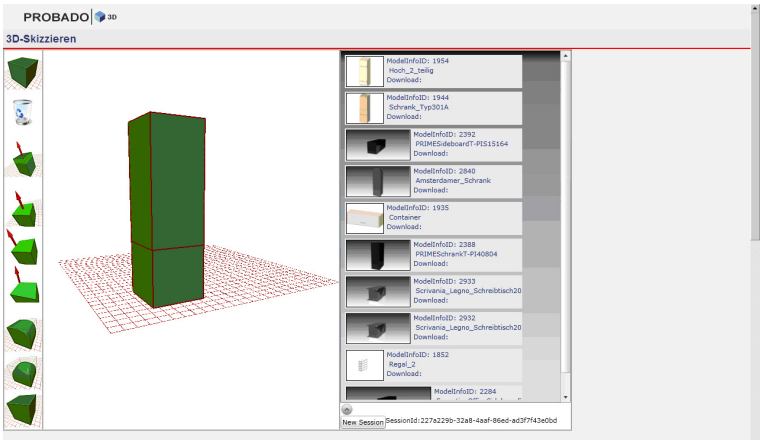

Abbildung 9: Online Zeichenoberfläche für Query-by-example Suchen mit Egebnissen

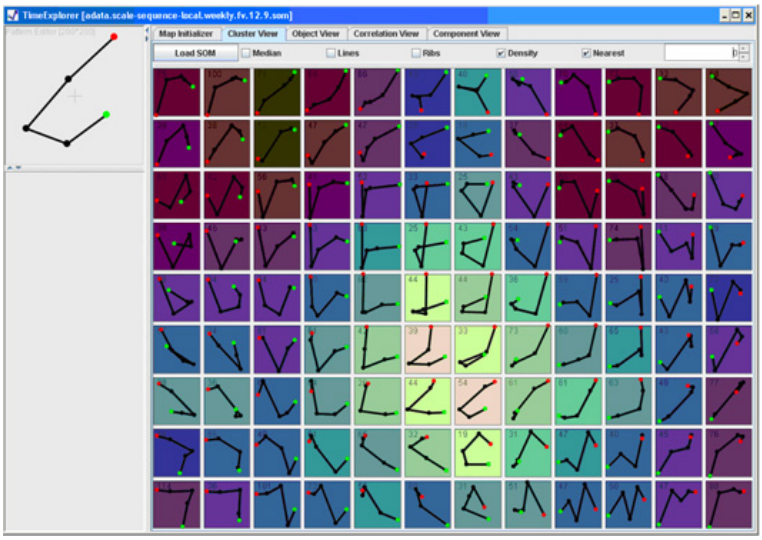

Abbildung 10: Prototyp einer graphisch-interaktiven Suchoberfläche über Forschungsdaten

Literatur

[1] BLÜMEL, I. DIET, J. KROTTMAIER, H. (2008), *Integrating Multimedia Repositories into the PROBADO Framework*. Third International Conference on Digital Information Management (ICDIM), 2008.

[2] BLÜMEL, I. SENS (2009), *Das PROBADO-Projekt: Integration von nichttextuellen Dokumenten am Beispiel von 3DObjekten in das Dienstleistungsangebot von Bibliotheken*. S. 79 ZfBB, Heft 2, 2009, Klostermann, Frankfurt am Main.

[3] BRASE, J. (2004), *Using Digital Library Techniques - Registration of Scientific Primary Data*. Lecture Notes in Computer Science, No 3232, pp 488-494.

[4] BRASE, J. (2010), *DataCite - A global registration agency for research data*, Working Paper 149/2010 German Council for Social and Economic Data (RatSWD)

[5] HEY, T. TANSLEY, S, TOLLE, K. EDS. (2009), *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, ISBN 978-0-9825442-0-4

[6] JOHNSON, A. (1997), *Spin-Images: A Representation for 3-D Surface Matching*, PhD thesis, Robotics Institute, Carnegie Mellon University,

[7] INGER, S. GARDNER, T. (2008), *How Readers Navigate to Scholarly Content*, available at http://www.sic.ox14.com/howreadersnavigatetoscholarlycontent.pdf accessed 1.9.10

[8] NSF. (2007), *Cyberinfrastructure Vision for 21st Century Discovery*, Arlington/VA: National Science Foundation (NSF), Cyberinfrastructure Council (CIC).

[9] WESSEL, R. BLÜMEL, I. AND KLEIN, R. (2008), *The Room Connectivity Graph: Shape Retrieval in the Architectural Domain*, The 16th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, Feb. Available at http://cg.cs.uni-bonn.de/en/publications/paper-details/wessel-2008-the-room/ Accessed 1.9.10