# Aufbau einer verteilten digitalen Bibliothek für nichttextuelle Dokumente - Ansatz und Erfahrungen des PROBADO Projekts

*Autoren:*

René Berndt (1), Ina Blümel (2), Michael Clausen (3), David Damm (3), Jürgen Diet (4), Dieter Fellner (1,5,6), Christian Fremerey (3), Reinhard Klein (3), Maximilian Scherer (5), Tobias Schreck(5), Irina Sens (2), Verena Thomas (3), Raoul Wessel (3)

*Institutionen und Adressen:*

1 – Technische Universität Graz, Inffeldgasse 16c, A-8010 Graz, Österreich.
2 – Technisch Informationsbibliothek, Hannover, Welfengarten 1B, D-30167 Hannover
3 – Universität Bonn, Römerstraße 164, D-53117 Bonn
4 – Bayerische Staatsbibliothek München, Ludwigstraße 16, D-80539 München
5 – Technische Universität Darmstadt, Fraunhoferstraße 5, D-64283 Darmstadt
6 – Fraunhofer Institut für Computer Graphik, Darmstadt, Fraunhoferstraße 5, D-64283 Darmstadt

***Zusammenfassung***

Das PROBADO Projekt ist ein von der DFG gefördertes Leistungszentrum für Forschungsinformation mit dem Ziel des prototypischen Aufbaus und Betriebs einer verteilten digitalen Bibliothek für heterogene, nicht-textuelle Dokumente. Betrachtet werden in diesem Projekt alle Schritte der bibliothekarischen Verwertungskette vom Bestandsaufbau über semi-automatische Inhaltserschließung, bis hin zu visuell-interaktiver Suche und Präsentation sowie Betriebsaspekten. In diesem Beitrag werden der gewählte Ansatz beschrieben und die bislang im Projekt gemachten praktischen und konzeptionellen Erfahrungen systematisiert und eingeordnet.

***Abstract***

The PROBADO project is a collaborative research effort aiming at prototypical development and operation of a distributed digital library system for heterogeneous, nontextual documents. The project considers all steps of the processing chain in libraries ranging from repository compilation, semi-automatic, content-based indexing, to visual-interactive search and presentation, and operational aspects. In this paper, we describe the chosen approach, specifically detailing on experiences and lessons learned up to now in the course of this project.

# 1. Einleitung

Digitale Bibliothekssysteme sollen einen effizienten und effektiven Zugang zu Dokumenten ermöglichen, die in digitaler Form vorliegen, um möglichst vielen Nutzererwartungen gerecht zu werden. Durch das Vorliegen digitaler Inhalte ergeben sich prinzipiell viele neue und hochinnovative Einsatzmöglichkeiten, wie visuelle Suche und Ergebnispräsentation oder die Vernetzung von Inhalten. Während sich in der Praxis digitale Bibliothekssysteme für *textuelle* Dokumentarten mehr und mehr etablieren und Standardisierungsentwicklungen beobachtet werden können, trifft dies auf den Bereich der *nichttextuellen* (allg.: Multimedia-) Dokumente bei weitem nicht zu. Aufgrund der Vielzahl der vorliegenden Dokumentarten (z.B. Videodokumente, Audiodokumente, 2- und 3-dimensionale Bilddaten, Primärdaten, etc.) und Fachgebiete (z.B. wissenschaftlicher Film, klassische Musik, 3D-Architekturdaten) ist dieser Bereich aktuell stark von Forschungs- und Prototypentwicklung gekennzeichnet.
Das von der DFG geförderte Leistungszentrum PROBADO[1] entwickelt in Zusammenarbeit zwischen Universitäts- und Bibliothekspartnern digitalen Bibliothekssupport für ausgewählte nichttextuelle Dokumente und Fachgebiete. Das Ziel ist es, einerseits Referenzlösungen für die ausgewählten Bereiche zu entwickeln und auch prototypisch zu betreiben. Andererseits sollen systemseitige und konzeptionelle Ergebnisse dazu beitragen, den Einsatz von nichttextuellen digitalen Bibliothekssystemen in der deutschsprachigen und internationalen Bibliothekslandschaft zu etablieren. PROBADO ist ein Folgeprojekt des DFG Schwerpunktprogramms „Verteilte Verarbeitung und Vermittlung digitaler Dokumente“ (1997-2004). Während $V^3D^2$ auf Grundlagenforschungsergebnisse ausgerichtet war, ist es das Ziel von PROBADO, die zuvor eingeführten Methodiken weiterzuentwickeln und in die Praxis zu transferieren. Besondere Beachtung finden hierbei aktuellen Anforderungen und Erwartungen aus Nutzersicht, sowie die Anforderungen des bibliothekarischen Arbeitsablaufes und Betriebes.
Wir berichten in diesem Papier über den in PROBADO gewählten Ansatz und über eine Bandbreite an Erfahrungen, die während den ersten beiden Projektphasen gemacht werden konnten. Im Besonderen systematisieren und diskutieren wir die gemachten Erfahrungen und Herausforderungen und stellen unsere Lösungsvorschläge ausführlich dar. Dieses Papier leistet damit einen konzeptionellen Beitrag und eine praktische Perspektive auf ein großes Digital Library Projekt, dessen Ergebnisse und Erfahrungen für die Weiterentwicklung digitaler Bibliotheksdienste relevant sind. Wir geben zunächst in Abschnitt 2 einen Überblick über wichtige verwandte Projekte und Arbeiten. In Abschnitt 3 schließt sich eine Darstellung der PROBADO Technologien an, bevor Abschnitt 4 die gemachten Erfahrungen systematisiert und diskutiert. Abschnitt 5 fasst diesen Beitrag zusammen und stellt die nächste Schritte dar.

[1] PROBADO ist ein Akronym für „Prototypischer Betrieb allgemeiner Dokumente“.

## 2. Verwandte Systeme und Projekte

Eine Reihe von praktisch relevanten Softwaresystemen existiert, die für den Aufbau digitaler Bibliothekssysteme in Frage kommen, wie z.B. Fedora (Lagoze et al, 2006), Greenstone (Witten et al, 2000), oder DLib (Castelli et al, 2002). Diese Systeme sind primär für textuelle bzw. strukturierte Dokumente vorgesehen und bieten keine native Unterstützung von nichttextuellen bzw. multimedialen Dokumenten. Entsprechende Funktionalität muss z.B. durch externe inhaltsbasierte Indexierer oder Visualisierungsmodule bereit gestellt werden. In PROBADO ist das Ziel, nativen Support für inhaltsbasierte automatische Erschließung und Zugriff auf ausgewählten, nichttextuellen Dokumenten zur Verfügung zu stellen. Das DelosDLMS System (Agosti et al, 2007) stellte einen Demonstrator von bis dato sehr umfangreicher Funktionalität für Indexierung, Retrieval und Visualisierung heterogener Dokumentarten dar, ist aktuell unseres Wissens nach aber nicht im praktischen Einsatz. Das Variations3 Projekt (Dunn et al, 2006) entwickelt und betreibt ein Digitales Bibliothekssystem für Musikdokumente. Im VICTORY Projekt (Daras, 2008) wird Unterstützung für inhaltsbasierte Suche in verteilten 3D Beständen entwickelt, basierend auf einer Peer-to-Peer Architektur. Das PROBADO Projekt folgt in seinem Ansatz diesen Systemen, insoweit in PROBADO die semiautomatische Erstellung von inhaltsbasierten Indices für die nichttextuellen Dokumente integriert ist. PROBADO ist jedoch nicht auf die ausgewählten Dokumentarten und Fachgebiete beschränkt, sondern kann durch seine Architektur im Prinzip beliebige Dokumentarten unterstützen (vgl. Abschnitt 3.1).
Auch in der kommerziellen Domäne nehmen inhaltsbasierte und visuelle Ansätze zur Suche zu. Die Internet Suchmaschine Google (http://www.google.com) unterstützt z.B. schon seit längerer Zeit die Suche in 2D Bilddaten, sowie über das sog. Google Warehouse (http://sketchup.google.com/3dwarehouse/) auch 3D Modelldaten. Während die Suche bis vor kurzem auf textuelle Metadaten, welche aus den umgebenden Webseiten extrahiert wurden, basierte, werden in Zukunft auch zunehmend inhaltsbasierte Ansätze zum Einsatz kommen. Ein Beispiel ist das Goggles Projekt von Google (http://www.google.com/mobile/goggles/). Während ‚klassische' Multimedia Dokumente wie Bilder, Videos und Audio schon seit längerem Gegenstand auch von Digitalen Bibliotheken sind, treten mittlerweile auch Forschungsprimärdaten in den Fokus des Interesses. Diese können ebenfalls als nichttextuelle Dokumente aufgefasst werden, die mit bibliothekarischen Mitteln behandelt werden können, angefangen von der Aufnahme, Bereitstellung, Suche, Präsentation und Langzeitarchivierung. Im Projekt KoLaWiss (Society for Scientic Data Processing Goettingen, 2009) wurde eine breite Reihe von organisatorischen, technischen und betriebswirtschaftlichen Anforderungen an eine kollaborative Primärdateninfrastruktur identifiziert. Eine Reihe von Systemen hält bereits jetzt Primärdaten für den öffentlichen Zugriff bereit, z.B. die Systeme PANGAEA (http://www.pangaea.de/), PsychData (http://psychdata.zpid.de/), oder Dryad (http://www.datadryad.org/). Während in diesen Bereichen vermutlich zunächst Fragen der Informationsinfrastruktur im Vordergrund stehen dürften, erwarten wir, dass auch hier langfristig inhaltsbasierte Suche- und Visualisierungslösungen benötigt werden. Ein aktuelles Projekt, was letzteren Fragen nachgeht, wird in (Bernard et al, 2010) beschrieben.

# 3 Das PROBADO Konzept

Das PROBADO System ist als verteiltes multimediales Digital Library System ausgelegt. Es unterstützt sowohl metadatenbasierte als auch inhaltsbasierte Suche in 3D und Musik-Dokumenten. In diesem Abschnitt stellen wir die wesentlichen Teiltechniken vor und beleuchten unseren Ansatz zum Technologietransfer.

## *3.1 Systemarchitektur*

Die PROBADO Systemarchitektur ist als grundsätzlich verteiltes Multimedia Datenbanksystem auf eine einfache Erweiterbarkeit ausgerichtet. Dokument- und domänenspezifische Funktionalität wird dabei in Spezialschichten gekapselt, während generische Datenbankdienste wie z.B. ein konsolidierter Metadatenindex und die Weiterleitung von Suchanfragen an in Frage kommende Datenbanken von einer Kernschicht bedient werden. Das Modell zerfällt damit in drei Schichten: Backend, Middleware und Frontend (Krottmaier et al, 2007).
In PROBADO formulieren Benutzer inhaltsbasierte Suchanfragen mittels dokumentyp-spezifischen Benutzerschnittstellen auf der Frontendschicht. Diese Anfragen werden mittels des PROBADO Webservice Protokolls an die Middleware (sog. Kernschicht) weitergeleitet, wo sie ausgewertet und entweder direkt beantwortet oder an entsprechende Backends weitergeleitet werden. Anfragen auf dem konsolidierten Metadatenindex können direkt auf der Kernschicht beantwortet werden, während Anfragen auf dem (dokumenttyp-spezifischen) erweiterten Metadatenset an die entsprechenden Backends weitergeleitet werden. Ein Synchronisierungsmechanismus hält den Metadatenindex in der Kernschicht sowie eine Tabelle mit zu jedem Zeitpunkt angeschlossenen und erreichbaren Backendrepositories aktuell. Aufgabe der Repositories ist es, die inhaltsbasierten Indices aktuell zu halten, inhaltsbasierte Anfragen auszuwerten, und Dokumente auszuliefern. Abschnitte 3.2 und 3.3 detaillieren Front- und Backendfunktionalität in zwei aktuell in PROBADO berücksichtigten Dokumentarten. Abb. 1 veranschaulicht die Architektur in einem Diagramm.

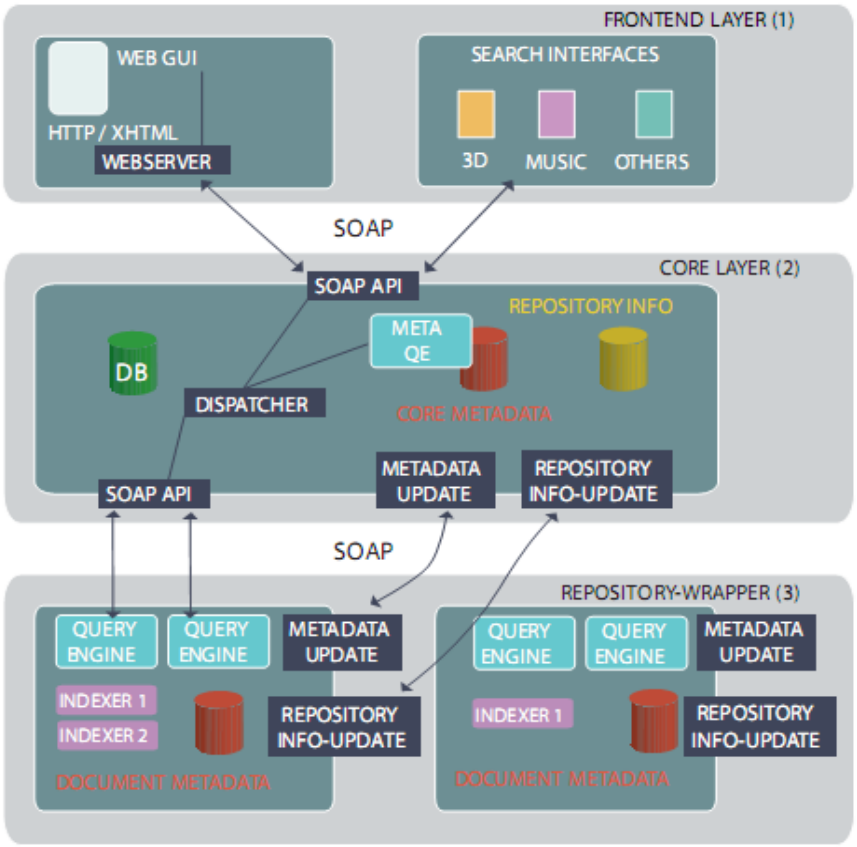


Abb. 1: Die PROBADO Drei-Schichten Architektur.

*3.2 PROBADO 3D Front- und Backendfunktionalität*

Eine von zwei ausgewählten Domänen ist die Architekturdomäne, innerhalb der exemplarisch der Bereich der 3-dimensionalen Architekturmodellobjekte betrachtet wird. Das Ziel hier ist es, den architektonischen Entwurfsprozess mittels eines umfangreichen Bestandes an Beispielmodellen, welche auf verschiedene Art durchsucht werden können, zu unterstützen. Typischer Inhalt von entsprechenden Beständen sind etwa 3D-Modelle von Einrichtungsgegenständen (z.B. Möbel) und Elementen von Außenanlagen (z.B. Bäume, Bänke, etc.). Vor allem interessant sind jedoch Modelle von Gebäudeformen, sowohl in Hüllendarstellung, als auch mit Innenstrukturen. Einer von uns durchgeführten Benutzerstudie (Blümel und Sens, 2009) zufolge, sind Architekten sowohl an Suche mit Metadaten als auch inhaltsbasiert interessiert. Die metadatenbasierte Suche setzt praktischer Weise auf den vorhandenen Metadaten auf, die ggf. vom Bibliothekar im Zuge der manuellen Erschließung vergeben werden. Insbesondere sind wir durch verschiedene inhaltsbasierte Analysealgorithmen, welche wir entwickeln konnten, beim PROBADO 3D *Backend* in der Lage inhaltsbasierte Deskriptoren zu berechnen und im Index abzulegen. Beispiele hierfür sind etwa globale 3D Formdeskriptoren (Tangelder und Veltkamp, 2008), aber auch Angaben über Raumstrukturen wie Anzahl der Geschosse oder Zimmer, welche vollautomatisch aus vorliegenden 3D Gebäudemodellen extrahiert werden können. Ein besonderer Deskriptor, der

in Zusammenarbeit mit Domänenexperten neu entwickelt werden konnte, bezieht sich auf die Erkennung von sog. Raumstrukturgraphen (Wessel et al, 2008). Hierbei werden Art und Topologie der Raumkonfigurationen welche in einem Gebäude vorliegen, erkannt und abgelegt. Der Nutzer ist dann in der Lage, über die Spezifikation von zu suchenden Strukturen, ihn interessierende Gebäudemodelle zu finden.

Das *Frontend* in diesem Bereich beinhaltet im wesentlichen Funktionalität für die Spezifikation der Suchanfragen, die Präsentation der Suchergebnisse, und das Navigieren in den Dokumenten. Zur Spezifikation von Suchanfragen stehen visuelle Editoren zur Verfügung, welche die Angabe einer globalen Formskizze oder einer Raumstruktur ermöglichen. Die Ergebnispräsentation zeigt herkömmliche Listenansichten, aber auch projektionsbasierte Visualisierungen von Voransichten. Die Bestände können zudem in Baum-orientierten Darstellungen durchstöbert werden, wobei die zugrundeliegenden Objektklassifikationen entweder manuell vorgegeben, oder automatisch berechnet werden können. Vorgesehen ist auch eine Auslieferung von Suchergebnissen, also 3D Modellen, via Download. Während das System intern unterschiedliche 3D Dateiformate beherrscht, haben wir uns für den Export für das PDF3D Format entschieden, da es, ähnlich wie bereits das PDF Format für Text- und Bilddokumente geworden ist, ein robustes und verbreitetes 3D Format mit hoher Nutzerakzeptanz ist bzw. dieses erwartet wird. Abb. 2 illustriert einige Ansichten auf das PROBADO3D Frontend.

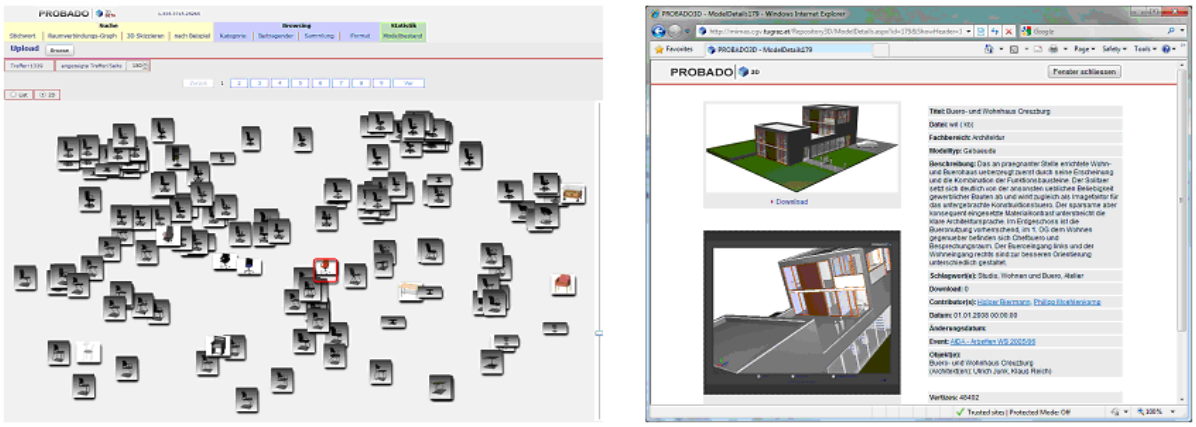


Abb. 2: Suchergebnis-Visualisierung (links) und Detailansicht auf ausgewählten Suchtreffer (rechts).

### *3.3 PROBADO Musik Front- und Backendfunktionalität*

Das PROBADO Musik Repository unterstützt inhaltsbasierte Indexierung und Suchen in klassischer Musik. Das Dokumentkonzept ist multimodal aufgebaut und beinhaltet eine Gesamtbetrachtung bestehend aus digitalem Audio und Scans der den Stücken zugrundeliegenden Partituren, sowie Coverscans der Tonträger. An der Bayerischen Staatsbibliothek wurde hierzu durch Digitalisierung eine digitale Kollektion von Audioaufnahmen und Partituren von westlicher klassischer Musik mit aktuell rund 96.000 Notenseiten und zugehörigen Audiodateien erstellt.

Das PROBADO Musik *Backend* beinhaltet die Speicherung der Digitaldaten und Methoden zur inhaltsbasierten Analyse (Transformation der Partituren und Audioströme in symbolische Repräsentationen) und zum Retrieval (Synchronisierung von Audioformen mit Partiturpositionen, sowie Funktionen zur Berechnung der Ähnlichkeit von Audioausschnitten). Hierzu wurden State-of-the-Art Verfahren aus dem Music Information Retrieval (Damm et al, 2009; Kurth und Müller, 2008; Suyoto et al, 2008) adaptiert bzw. weiterentwickelt. Ein Workflowmodell wurde aufgesetzt, welches effizient manuelle und automatische Datenverarbeitung kombiniert (sog. ContentCreator, vgl. Thomas et al, 2009).
Das *Frontend* offeriert visuelle Benutzerschnittstellen zur Anfrageformulierung und Ergebnispräsentation. Suchanfragen können wie gewohnt auf textuellen Metadaten gestellt werden und darüber hinaus inhaltsbasiert sein. Zu letzterem ist vorgesehen, auch Teile einer Partitur zu markieren und dann die ähnlichsten Stücke im Bestand aufzufinden. Zur Wiedergabe wurde ein Player entwickelt, welcher zur Audiowiedergabe die entsprechenden Partituren einblendet und die exakte Position darin markiert (vgl. Abb. 3 für ein Beispiel).

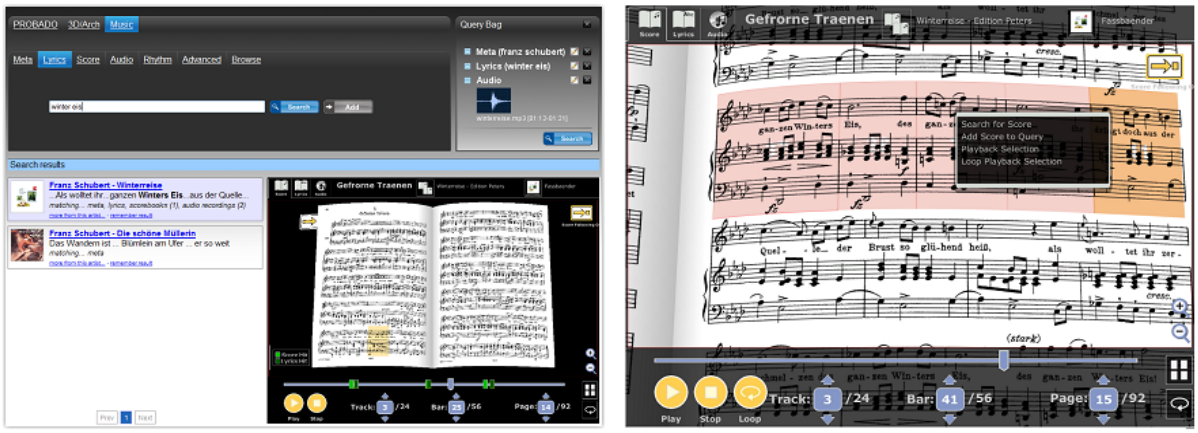


Abb. 3: PROBADO Musik Retrieval Frontend mit Score Audio Player (links). Die Partituransicht (rechts) erlaubt das Mitlesen der Audiowiedergabe in den Noten.

### *3.4 Technologietransfer*

Ein integraler Bestandteil des PROBADO Projektes ist die Bereitstellung und der Betrieb von prototypischen Systemen bei den beteiligten Bibliothekspartnern. Hierzu ist ein sorgfältiger und abgestimmter Plan zum Transfer von Technologie und Know-How zwischen Universitäts-Entwicklern und der zuständigen Fach- und IT-Abteilungen bei den Bibliotheken nötig. Die Hauptherausforderung darin besteht, verglichen mit herkömmlichen Software-Entwicklungsprojekten, in einer hohen Komplexität der Systeme und einem sehr dynamischen Entwicklungsprozess. Unser Entwicklungs- und Transfermodell besteht aus zwei Phasen. In der ersten Phase definieren und entwickeln Forschungs- und Bibliothekspartner gemeinsam die anzustrebenden Retrieval- und Präsentationstechniken für die jeweiligen Domänen. Die Bibliothekspartner nehmen dabei die fachliche Sichtweise ein, während die Entwicklungspartner die Kontrolle über die sich entwickelnden Prototypen haben. Die Entwicklung der Funktionalität in dieser Phase ist dynamisch und von

wiederholenden Zyklen von Idee-Implementierung-Test-Idee gekennzeichnet. In Phase zwei findet der eigentliche Transfer von Technologie und Know-How zu den Betreibern statt. Wichtige Teilschritte in dieser Phase beinhalten (a) die Auswahl und Konsolidierung von Kernfunktionalität aus einem größeren Bestand an (vor)prototypisch entwickelter Funktionalität, (b) Anpassung der Schnittstellen an Betreiberanforderungen, (c) Dokumentation der Technologien und Unterweisung von Mitarbeitern und (d) Testbetrieb mit Usability Feedback.

In jeder dieser beiden Phasen sind Herausforderungen zu adressieren. In Phase eins ist zu einem bestimmten Zeitpunkt ein Schnitt zu machen zwischen jüngsten Funktionen, die möglicherweise hochinnovativ, aber noch nicht ausgereift sind, und einem Kern an robuster Funktionalität. Dieser Kern kann aus Wartungsgründen nicht den Vollumfang der in der universitären Entwicklung betrachteten Funktionen aufweisen. In Phase zwei ist insbesondere die Abstimmung der Schnittstellen relevant, da die Entwicklungs- und Betriebsumgebungen typischerweise gewisse Abweichungen enthalten hinsichtlich Entwicklungssprache, Datenbank- und Betriebssystemsoftware, unterstützter Webtechnologien etc. Ein erfolgreicher Betrieb erfordert eine Identifikation und ein Grundverständnis der Betreiber bzw. deren Administratoren mit den zugrundeliegenden Technologien. Hier sind Kommunikation und Dokumentation ein erheblicher Erfolgsfaktor.

In PROBADO werden in den beiden Teilbereichen unterschiedliche Modelle verwendet. Im Bereich Musik wird der In-House Betrieb beim Bibliothekspartner angestrebt, während im Bereich 3D Architekturdaten angestrebt wird, den Betrieb an einen externen Betreiber auszugliedern. Der Transfer ist für Herbst 2010 vorgesehen, und wir rechnen mit umfangreichen praktischen Ergebnissen und einer Evaluierung, wie gut das gewählte Modell sich bewährt hat.

## 4. Projekterfahrungen

Wir systematisieren und besprechen hier wichtige Erfahrungen, die wir während der ersten vier Jahren des Projektverlaufs machen konnten.

### *4.1 Systemarchitektur*

Architekturseitig sind bei der Realisierung eines verteilten digitalen und heterogenen Bibliotheksystemes verschiedene Herausforderungen zu bewältigen. Wir befassen uns hier mit der Metadatenabstraktion, Relevance Feedback und mit dem Inter-Domain Retrieval.

**Metadatenabstraktion**. Zur Integration von heterogenen Dokumentbeständen ist ein konsolidierter Metadatenindex erforderlich. Hierzu ist aus der Vielzahl an spezialisierten Domänen-orientierten Metadatenschemata ein kleinster gemeinsamer Nenner zu bilden. Hierzu sind pro Domäne Kompromisse zwischen hochdeskriptiven aber spezialisierten, und weniger deskriptiven aber generischeren Feldern zu treffen. Generische Felder wie Titel, Autor oder Datum sind domänenübergreifend einsetzbar, haben jedoch oftmals domänenspezifische Interpretationen bzw. Restriktionen bezüglich des Kontexts. Andererseits sind spezialisierte Felder oftmals nur auf bestimmten Domänen sinnvoll und würden bei gemeinsamer Betrachtung für viele Dokumente Nullwerte annehmen müssen. In PROBADO

konnte nach intensiven Diskussionen zwischen den Domänen Architekturmodelle und Klassische Musik eine Entscheidung für eine kompakte Dublin-Core-Teilmenge getroffen werden, welche aller Voraussicht nach die Erweiterbarkeit von PROBADO für weitere Domänen sicherstellt. Diese Teilmenge ist jedoch nur für den konsolidierten Metadatenindex maßgeblich. In den Backend Schichten existieren reichhaltigere Metadatenschemata in Anlehnung an domänenübliche Schemata, wie z.B. FRBR (Diet und Kurth, 2007) im Musik-Teilbereich, und können durch Spezialanfragen erreicht werden. Anfragen auf den Dublin Core Metadaten werden in der Kernschicht ausgewertet und liefern prinzipiell bereichsübergreifende Ergebnismengen zurück.

**Relevance Feedback**. Diese Technologie (Novotni et al, 2005) ist seit mehreren Jahrzehnten in der Forschung bekannt. Sie sieht vor, dass der Benutzer Ergebnislisten einer initialen Suchanfrage mit seiner Präferenz bezüglich der Relevanz für die gestellte Suchanfrage bewertet. Aufbauend hierauf optimiert das System dann mittels Verfahren des Machine Learnings die Suchfunktion, um (so die Erwartung) weitere relevante Antworten zurückliefern zu können. Hierzu sind jedoch weitreichende Kenntnisse bezüglich der verwendeten Suchfunktion und vorhandenen Datenbestände erforderlich. In einem verteilten und heterogenen Umfeld ist dies schwierig, da die einzelnen Repositories annahmegemäß keine Kenntnisse von anderen Datenbeständen oder Interna der Suchalgorithmen aufweisen müssen. Insofern kann Relevance Feedback zunächst nur Repository-lokal zum Einsatz kommen. Eine Alternative wäre die Einführung einer weiteren Relevance Feedback Abstraktionsschicht für jede zu unterstützende Dokumentendomäne. Dies wäre mit einem erheblich gesteigerten Implementierungsaufwand für ein Relevance Feedback Protokoll verbunden, von dem wir aus praktischen Gesichtspunkten zunächst in PROBADO absehen.

**Inter-Domänen Retrieval**. PROBADO ermöglicht dem Benutzer in allen an das Kernsystem angeschlossenen Repositorien, über Domänengrenzen hinweg, zu suchen. Dies ist mit textueller bzw. metadatenbasierter Suche leicht möglich. Bezüglich inhaltsbasierter Suche ist dies zunächst nicht möglich, wenn Dokumentarten vorliegen, in denen sinnvollerweise nicht mit den gleichen Suchkonzepten gearbeitet werden kann. Zum Beispiel ist es nicht möglich, mit einer geometrischen Formskizze in einem Audiostück zu suchen, da die Syntax der inhaltsbasierten Suchen jeweils nicht kompatibel ist. Ein Ausweg besteht eventuell darin, über automatisch erzeugte, textuelle Metadaten zu suchen. Etwa könnte eine Suche nach dem Term „Barock“ einerseits Barockmusik, andererseits 3D-Modelle von Gebäuden aus dieser Epoche zurückliefern.

### *4.2 Repository Erstellung*

In beiden Bereichen, Musik und 3D, mussten zu Projektbeginn geeignete Datenbestände erzeugt werden, sowohl zum Testen der Suchfunktionalität, zu Demonstrationszwecken und zur Sammlung von Erfahrung mit Digitalisierungs- und Datenintegrationsarbeiten. Hierbei wurden im Projekt zwei unterschiedliche Wege beschritten:

**Institutioneller Ansatz**. Hierbei werden die Dokumente durch einen Digitalisierungsprozess von Originalquellen erzeugt. In der BSB wird hierzu, als Teil einer größeren Massendigitalisierungsinitiative, ein umfangreicher Bestand and CDs in Audiodateien migriert, sowie Partituren gescannt und maschinell gelesen. Zu letzterem wurde ein Optical-Music-Recognition Ansatz gewählt, welcher die Partiturdaten in digitale symbolische Repräsentation übersetzt. Zum Retrieval auf unterschiedlichen Granularitätsebenen werden zu den Stücken umfangreiche Metadaten erfasst, welche eine Zuordnung z.B. auf Satz-, Werk-

und Komponistenebene erlaubt. Ein spezialisiertes Metadatenmodell basierend auf dem FRBR-Datenmodell wurde hierzu erstellt.

**Provider-Orientierte Ansatz.** Im Bereich 3D wurde ein umfangreicher Testdatenbestand aus ganz unterschiedlichen Quellen zusammengetragen. Wir kontaktierten Architekturfakultäten, Internetportalbetreiber und freie und kommerzielle Sammlungen. Wir erhielten eine Vielzahl an Modellbeiträgen in unterschiedlichen Formaten, Qualitätsstufen und bezüglich unterschiedlicher Inhalte. Lediglich 10% der erhaltenen Modelle enthielt Metadaten. Ein spezialisiertes Metadatenmodell wurde entwickelt welches Dublin Core Entitäten sowie Gruppen aus dem FRBR-Datenmodell enthält.

Wenn wir die beiden Ansätze vergleichen, stellen wir fest, dass der institutionelle Ansatz ein hochstrukturierter Prozess ist, der volle Kontrolle über Mengen, Inhalte und Qualität der digitalen Materialien erlaubt. Er nutzt bestehende Digitalisierungsworkflows, und das erstellte Dokumentenrepository ist hochgradig homogen. Der Provider-orientierte Ansatz ist gekennzeichnet durch mehr Heterogenität der Bestände, da diese aus ganz unterschiedlichen Quellen stammen. Die Heterogenität erstreckt sich auf typische wichtige Aspekte im Zusammenhang mit 3D Modelldaten, einschließlich deren Auflösung, Dateiformat, Qualität, Metadaten sowie Inhalte. Jeder dieser Prozesse kann für unterschiedliche Zielsetzungen jeweils ideal geeignet sein. Der institutionalisierte Ansatz wird in unserem Fall begünstigt durch den vergleichsweise etablierten Status der zugehörigen Digitalisierungsschritte (bzgl. 2D Bild- und Audiomaterial), der zugehörigen Austauschformate, und des Vorhandenseins eines Bestandes an zu digitalisierenden Materialien. Im Bereich 3D besteht aktuell vergleichsweise weniger Standardisierung sowohl bei Verfahren der 3D Digitalisierung als auch bei 3D Austauschformaten. Aufgrund dieser Tatsache und im Zusammenhang damit, dass zu Projektbeginn noch kein größerer Fundus an digitalen 3D Architekturmodellen vorlag, wurde hier der Provider-orientierte Ansatz gewählt.

### *4.3 Betriebsaspekte*

Wie in Abschnitt 3.4 beschrieben, ist PROBADO ein kooperatives Bibliotheks- und Universitätsprojekt. Die Entwicklung ist durch die Universitätspartner zu leisten, während der Betrieb nach Projektende durch die Bibliothekspartner sicherzustellen ist. Bereits in dieser Projektphase wurden zwei Betriebsmodelle entwickelt, die in der letzten Projektphase ab Herbst 2010 umgesetzt werden. In den beiden Bereichen Musik und 3D werden verschiedene Modelle angestrebt.
Im Bereich Musik wird ein zentraler Inhouse-Betrieb an der BSB angestrebt. Die entwickelten Systeme werden dort betrieben, ebenso wird der erschlossene Digitalbestand zentral dort verwaltet. Im Bereich 3D wird ein dezentraler und Host-basierter Betrieb angestrebt. Das entwickelte Vollsystem wird an einen Dienstleistungsprovider übergeben, der den Betrieb durchführt. Das System wird an die Portallösung GetInfo an der TIB angebunden. Das Modell sieht vor, die Bestände auch dezentral bei den Bestandsbesitzern zu halten, und durch PROBADO freigegebene Voransichten der Modelle zu zeigen, während für weitere Anfragen (u.a. hochauflösenden oder texturierte Modelle) der Anwender auf die Portalseite der jeweiligen Bestandsbesitzer durchgeleitet werden. Die Gesamt-Qualitätskontrolle der eingepflegten Bestände aus Fachsicht obliegt weithin der TIB.
PROBADO leistet durch diese beiden Modelle ein weiteres Experiment, in dem unterschiedliche Betriebsmodelle praktisch untersucht werden, und soll hierdurch Vorbild für

andere bibliothekarische Institutionen sein, die am Betrieb entsprechender Digitaler Bibliothekssysteme interessiert sind.

## 5. Zusammenfassung und Ausblick

Wir haben in diesem Papier den Ansatz und unsere bisherigen Erfahrungen im Projekt PROBADO vorgestellt. Ziel des Projektes ist es, prototypische Digitale Bibliothekssysteme für ausgewählte Multimedia Dokumente in Kooperation zwischen bibliothekarischer Praxis und universitärer Forschung zu schaffen. Es wurde ein Spektrum an inhaltsbasierten Such- und Navigationsmöglichkeiten geschaffen, von welchen eine Auswahl zusammen mit den aufgebauten Beständen in den prototypischen Betrieb gehen werden. Hierzu wurden Optionen für den Technologietransfer und den Betrieb entwickelt. Durch die Umsetzung dieser Schritte werden zusätzliche, wegweisende Erfahrungen mit praktischer Relevanz für den Bereich der Digitalen Bibliotheken für Multimedia Dokumente erwartet.

## Danksagung

Das PROBADO Projekt startete in 2006, und seine Projektlaufzeit ist bis 2011 ausgelegt. Partner sind die Universität Bonn, die Technische Universität Darmstadt, die Technische Universität Graz, die Technische Informationsbibliothek Hannover und die Bayerische Staatsbibliothek München. PROBADO wird von der Deutschen Forschungsgemeinschaft DFG im Programm für Wissenschaftliche Literaturversorgung und Informationssysteme gefördert. Für weitere Informationen besuchen Sie bitte die Projektwebseite unter www.probado.de. Dieses Papier ist eine erweiterte Version von (Berndt et al, 2010).

## Referenzen

Agosti, M., Berretti, S., Brettlecker, G., Bimbo, A.D., Ferro, N., Fuhr, N., Keim, D.A., Klas, C.P., Lidy, T., Milano, D., Norrie, M.C., Ranaldi, P., Rauber, A., Schek, H.J., Schreck, T., Schuldt, H., Signer, B., Springmann, M.: Delosdlms – the integrated delos digital library management system. (In: DELOS Conference, Post- Proceedings, Springer LNCS 2007).

Bernard, J., Brase, J., Fellner, D., Koepler, O., Kohlhammer, J., Ruppert, T, Schreck, T., Sens, I.: A Visual Digital Library Approach for Time-Oriented Research Data. European Conference on Digital Libraries 2010.

Berndt, R., Blümel, I., Clausen, M., Damm, D., Diet, J., Fellner, D., Fremerey, C., Klein, R., Krahl, F., Scherer, M., Schreck, T., Sens, I., Thomas, V., and Wessel, R.: The PROBADO Project - Approach and Lessons Learned in Building a Digital Library System for Heterogeneous Non-textual Documents. European Conference on Digital Libraries 2010.

Blümel, I., Sens, I.: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.

Castelli, D., Pagano, P.: Opendlib: A digital library service system. In: ECDL (2002).

Damm, D., Kurth, F., Fremerey, C., Clausen, M.: A concept for using combined multimodal queries in digital music libraries. In: 13th ECDL. (2009).

Daras, P., Tzovaras, D., Dobravec, S., Trnkoczy, J., Sanna, A., Paravati, G., Traphoener, R., Franz, J., Kastrinogiannis, T., Malavazos, C., Ploskas, N., Gumz, M., Geramani, K., Wintterle, G.J.: Victory: a 3d search engine over p2p and wireless p2p networks. (In: 4th International Conference on Wireless Internet, 2008).

Diet, J., Kurth, F.: The PROBADO music repository at the bavarian state library. In: 8th International Conference on Music Information Retrieval. (2007).

Dunn, J.W., Byrd, D., Notess, M., Scherle, R.: Variations2: Retrieving and using music in an academic setting. Communications of the ACM 49 (2006).

Krottmaier, H., Kurth, F., Steenweg, T., Appelrath, H.J., Fellner, D.: PROBADO - a generic repository integration framework. In: Pr. of the 11th European conference on Research and Advanced Technology for Digital Libraries. (2007).

Kurth, F., Müller, M.: Efficient index-based audio matching. IEEE Transactions on Audio, Speech, and Language Processing 16 (2008) 382-395.

Lagoze, C., Payette, S., Shin, E., Wilper, C.: Fedora: an architecture for complex objects and their relationships. Int. J. Digit. Libr. 6 (2006) 124-138.

Novotni, M., Park, G.J., Wessel, R., Klein, R.: Evaluation of kernel based methods for relevance feedback in 3d shape retrieval. In: The Fourth InternationalWorkshop on Content-Based Multimedia Indexing (CBMI'05). (2005).

Society for Scientic Data Processing Goettingen: Cooperative long-term preservation for research centers (in German). Project Report (2009).

Suyoto, I., Uitdenbogerd, A., Scholer, F.: Searching musical audio using symbolic queries. IEEE Transactions on Audio, Speech, and Language Processing 16 (2008).

Tangelder, J.W., Veltkamp, R.C.: A survey of content based 3d shape retrieval methods. Multimedia Tools and Applications 39 (2008) 441-471.

Thomas, V., Fremerey, C., Damm, D., Clausen, M.: SLAVE: a Score-Lyrics-Audio-Video-Explorer. In: Proceedings of the 10th ISMIR. (2009).

Wessel, R., Blümel, I., Klein, R.: The room connectivity graph: Shape retrieval in the architectural domain. In: WSCG. (2008).

Witten, I.H., Mcnab, R.J., Boddie, S.J., Bainbridge, D.: Greenstone: A comprehensive open-source digital library software system. In: Proceedings of the Fifth ACM International Conference on Digital Libraries. (2000).