IR-Unterstützung für die Digital Humanities: Problemstellungen und erste Lösungsideen

Tobias Gradl und Andreas Henrich
Otto-Friedrich-Universität Bamberg, Lehrstuhl für Medieninformatik
96045 Bamberg
{tobias.gradl|andreas.henrich@uni-bamberg.de}

Abstract

Die vom BMBF geförderte Initiative DARIAH-DE[1] hat eine Stärkung digitaler Methoden in den Geisteswissenschaften zum Ziel. DARIAH-DE verknüpft vorhandene digitale Ressourcen, Dienste und Erkenntnisse über einzelne Disziplinen und Forschungsfragen hinweg und errichtet und erforscht eine dezentrale technische Infrastruktur, auf der geisteswissenschaftliche Methoden implementiert und genutzt werden können. Ein wichtiger Bestandteil davon sind Suchdienste, die von spezialisierten Suchmaschinen für einzelne Domänen bis zu generischen, übergreifenden Suchdiensten reichen. Das vorliegende Papier gibt zu einem frühen Zeitpunkt einen Überblick über Zielsetzungen, Umfeld und Problemstellungen.

1 Motivation

Der Anteil digital repräsentierter Forschungsdaten nimmt in den Kultur- und Geisteswissenschaften stetig zu. Fortschritte in den Bereichen der Digitalisierungstechnik, Bildverarbeitung und digitaler Archive führen dazu, dass unterschiedlichste Fachdisziplinen eine signifikante Menge von Informationen produzieren und bereitstellen. Die Ausprägungen von Forschungsdaten in den Kultur- und Geisteswissenschaften umspannen dabei die volle Medienbandbreite - von Texten bis hin zu multimedialen Daten. Derzeit als zentralisierte Speicher fungierende Archive können wie folgt klassifiziert werden: *Horizontale Archive* fokussieren bestimmte Medientypen, beinhalten jedoch potenziell fachübergreifende Forschungsdaten, während *vertikale Archive* unterschiedliche Medientypen einer abgegrenzten Forschungsdomäne beinhalten. Die gemeinsame Eigenschaft digitaler Sammlungen besteht jedoch darin, dass durch Einschränkung der unterstützten Medientypen oder des Anwendungsbereichs ein möglichst hoher Grad der Unterstützung konkreter Zielgruppen und deren Arbeitsweisen erwirkt werden soll.

Während eine derartige Einschränkung dazu führt, dass die Komplexität der Archive in einem beherrschbaren Rahmen gehalten werden kann, geht von einer übergreifenden Integration ein hohes Potenzial für eine fachübergreifende Kollaboration und die Erkennung neuer Zusammenhänge in Forschungsdaten unterschiedlicher Disziplinen aus. An das Information Retrieval (IR), welches bereits einen hohen Grad der Unterstützung für digitale Archive bereitstellt, sind für eine derartige Integration komplexe Anforderungen zu stellen, um eine organisations- und disziplinübergreifende Suche zu ermöglichen.

Nach einer Einordnung des Begriffs der Digital Humanities und der Forschungsinitiative DARIAH, die eine derartige Integration versucht, möchte diese Arbeit einige besondere Problemstellungen für die Anwendung von Methoden des IR erläutern und erste Ideen für eine Konzeption im Rahmen von DARIAH zur Diskussion stellen.

Digital Humanities

Der Begriff der Digital Humanities oder auch e-Humanities trägt dem Umstand Rechnung, dass eine Adaption der Angewandten Informatik an die Fragestellungen der Kultur- und Geisteswissenschaften zwar seit Jahrzehnten erfolgt (vgl. z. B. [Hockey, 2004]), eine fundamentale, eigenständige Umsetzung in der Forschung und Lehre jedoch bisher weitgehend ausblieb. Neuartige Studiengänge und Kursmodule sowie Forschungsinitiativen, wie das in dieser Arbeit vorgestellte ESFRI[2]-Projekt „Digital Research Infrastuctures for the Arts and Humanities (DARIAH)“ verdeutlichen jedoch, dass das Bewusstsein für die besonderen Arbeitsweisen und Anforderungen an die Digital Humanities wächst.

DARIAH

Die Forschungsinitiative DARIAH wird in Deutschland durch das BMBF gefördert und beschreibt ein umfangreiches Projekt an der Schnittstelle zwischen den Strategien zur Etablierung der Digital Humanities und der Bereitstellung von Forschungsinfrastrukturen. Aus der technologischen Perspektive verfolgt DARIAH das Teilziel, eine technologische Infrastruktur zu schaffen, die die Grenzen zwischen isolierten digitalen Archiven auflöst und eine Basis für interorganisationale und interdisziplinäre Forschungsarbeiten bildet.

Ein wesentlicher Aspekt dieser technologischen Basis, die sich seit Mai 2011 in ihrer Konzeptionsphase befindet, besteht in der Förderung der Interoperabilität durch die Umsetzung einer Plattform, die durch generische und fachspezifische Dienste erweitert werden kann. Die Methoden des IR können dabei sowohl als Mechanismen zur Generierung dieser Interoperabilität, als auch als adaptive, in die Arbeitsweisen der Wissenschaftler integrierte Dienste zur fachübergreifenden Suche verstanden werden.

Standards und Umfeld

Die Reife informationstechnischer Methoden im Umfeld digitaler Bibliotheken und Sammlungen hat zur Spezifikation zahlreicher generischer und fachspezifischer Standards für die Strukturierung von Metadaten und den Zu-

[1] http://de.dariah.eu/

[2] http://ec.europa.eu/research/infrastructures/

griff auf entsprechende Archive geführt. Obwohl aufgrund des angestrebten, generischen Charakters von DARIAH keine Priorisierung der zu unterstützenden Standards möglich ist, können dennoch Protokolle identifiziert werden, deren Umsetzung aufgrund ihrer weitreichenden Nutzung möglichst kurzfristig erfolgen sollte. Im Hinblick auf die Anwendbarkeit bestehender Modelle des IR sind dabei insbesondere die verschiedenen Zugriffsstrategien hervorzuheben, deren Implementierung jeweils von konkreten Föderationsendpunkten abhängig ist: In Bibliotheken und Sammlungen häufig unterstützt wird das *Protocol for Metadata Harvesting* der Open Archives Initiative (OAI-PMH, [Lagoze et al, 2008a]), bei dem verarbeitende Aggregatoren Metadaten zentral integrieren und Suchmechanismen auf lokal verwalteten Indizes arbeiten. Mit dem *Object Reuse and Exchange* (OAI-ORE, [Lagoze et al, 2008b]) Protokoll verfolgt die OAI das Ziel, Assoziationen zwischen Objekten bei deren Austausch zu erhalten. OAI-ORE wird im Gegensatz zu PMH durch Formen des Distributed Queryings angesprochen. Im Kontext von DARIAH werden insbesondere auch das im Bibliothekswesen gängige Z39.50[3] bzw. dessen semantische Nachfolger Search & Retrieve Web Service (SRW) und Search & Retrieve URL Service (SRU) eine besondere Bedeutung einnehmen.[4]

Neben Protokollen für den Zugriff auf die standardisierten Schnittstellen digitaler Archive sind für eine Föderation von Forschungsdaten im Kontext von DARIAH insbesondere die in den Kultur- und Geisteswissenschaften verwendeten Metadaten-Schemata von herausragender Bedeutung. Als Beispiel eines allgemeinen Metadaten-Standards kann der, auch in Verbindung mit OAI-PMH (vgl. ‚oai_dc', [Lagoze et al, 2008a]) weit verbreitete Dublin Core [DCMI, 2010a] Standard angeführt werden, welcher die Annotationstiefe grundsätzlich auf 15 unterstützte Elemente reduziert. Fachspezifische Schemata reichen hingegen von angepassten Dublin Core Profilen bis hin zu feingranularen Protokollen spezifischer Fachbereiche. Entsprechende Beispiele sind hierbei EpiDoc als TEI XML-basierter Standard für epigraphische Dokumente oder auch der MEI-Standard der Music Encoding Initiative für musikwissenschaftliche Disziplinen.

Aufgrund erfolgreicher Vor- und Parallelprojekte, wie beispielsweise TextGrid[5] oder EHRI[6] kann sich die Konzeption von DARIAH auf Erfahrungen stützen, die im Umgang mit den Kultur- und Geisteswissenschaften, wie auch mit der Komplexität des Kontexts digitaler Archive gesammelt wurden.

2 Informationsbedürfnisse

Die Aufgaben des IR können im Kontext digitaler Archive grundsätzlich anhand der drei Phasen typischer Suchprozesse klassifiziert werden (vgl. [Adams and Blandford, 2005]). *Initiierung*: Eine Interaktion oder ein Ereignis führt zur Definition eines Informationsbedürfnisses. *Gewinnung*: Die gezielte Informationsgewinnung wird durch ein System oder eine Person erleichtert. *Interpretation*: Gewonnene Daten werden durch Anreicherung kontextbasierter Informationen interpretiert.

Die Ergebnisse von Adams und Blandford zeigen, dass die Unterstützung von Benutzern digitaler Bibliotheken durch das IR vorwiegend in der Phase der Informationsgewinnung erfolgt. Die Initiierung und besonders die Interpretation hingegen werden durch eine persönliche Interaktion ohne Systembeteiligung erreicht. Während die Untersuchungen von Adams und Blandford im medizinischen und akademischen Umfeld erarbeitet wurden und die Nutzung isolierter digitaler Archive betrachten, muss im Kontext von DARIAH eine weiterführende Unterstützung durch das IR erreicht werden. Dies gilt insbesondere für die Phase der Interpretation, da der konkrete Anwenderkontext bei interdisziplinären Suchen nicht implizit vorhanden ist und mögliche Anwendungsfälle demnach generisch unterstützt werden müssen.
Die Ergebnisse einer ersten Analyse der Fragestellungen/Informationsbedürfnisse, die im Rahmen von DARIAH betrachtet werden sollten fasst Tabelle 1 zusammen.

Dimension	**Ausprägungen**		
Ziel der Suche	Sammlung	Ressource	Experte
Datenebene	Metadaten	Inhalt	
Medientyp	Text	Bild	…
Medienformat	XML	SGML	JPEG
Objekttyp	Brief	Portrait	Epigraph
Suchraum	Global	fachliche Teilmenge	benutzer-definierte Teilmenge
Multilingualität	nein	Metadaten	Inhalt
Abfragetyp	exakt	vage	
Abfrageziel	bekanntes Ziel	explorativ	
Datenqualität	Rohdaten	annotiert	aufbereitet
Rechte	frei	keine	Metadaten

Tabelle 1: Grober Morphologischer Kasten zu Informationsbedürfnissen

Aufgrund der föderierten Struktur der Informationslandschaft von DARIAH kann das zu ermittelnde Ziel eines Nutzers sowohl in einer digitalen Sammlung als auch in den beinhalteten Ressourcen bestehen. Die Wortfolge „Jüdische Grabsteine" kann somit je nach Anwenderkontext im Rahmen einer generischen Suche sowohl auf die Gewinnung von Arbeiten z. B. zu der Symbolik jüdischer Grabsteine, als auch auf die Identifizierung epigraphischer Datenbanken jüdischer Grabsteine in Deutschland abzielen.

Auch die Dimension der Datenebene muss in Abhängigkeit von der Anwendersicht interpretiert werden. Während eine Suche auf der Basis von Metadaten zu einer hohen Präzision der Suchergebnisse führt, beschränkt diese den Umfang der Suchergebnisse auf den Kontext, in dem die Metadaten generiert wurden, während die Suche über Inhalten zu einem hohen Recall bei interdisziplinären Anwendungsfällen führen kann. So könnten die jüdischen Grabsteine, die im Rahmen der Judaistik durch eine aufwendige, epigraphische Annotation beschrieben werden, auch für die Analyse einer geographisch bedingten Verwitterung im Rahmen der Denkmalkunde verwendet werden. Die Dimensionen Medientyp, Medienformat und Objekttyp würden in diesem disziplinübergreifenden Anwendungsfall nicht die qualitativ hochwertigen, epigraphischen Metadaten fokussieren, sondern den Inhalt, wie Fotografien in typischen Bildformaten oder Freitexte.

[3] http://www.loc.gov/z3950/agency/
[4] http://www.loc.gov/standards/sru/
[5] http://www.textgrid.de/
[6] http://www.ehri-project.eu/

Analog muss auch der Suchraum als Nutzerpräferenz von Recall oder Precision interpretiert werden. So wird eine globale Suche über alle föderierten Sammlungen zu einem hohen Recall führen, der gerade die Forschungsdaten in einer Ergebnismenge miteinschließt, die bei einer explorativen und fachübergreifenden Suche von besonderer Bedeutung sein können. Eine Einschränkung des Suchraums auf eine Teilmenge fachlich verwandter oder durch den Benutzer spezifizierter Sammlungen erhöht dagegen die Präzision der Suche, wird aber in der Regel einen geringeren Recall bedingen.

Im Hinblick auf eine organisations- und disziplinübergreifende Föderation digitaler Archive führen die beiden Dimensionen der Datenqualität und Rechte zu besonderen Anforderungen an das IR, da sich beide nur in einer Kombination von Merkmalen auf Sammlungs- und Ressourcenebene beantworten lassen. Die Untersuchung der Dimension „Datenqualität" muss hierbei auf unterschiedlichen Ebenen betrachtet werden.

(1) Es können Klassifikationsebenen der Nutzbarkeit von Forschungsdaten für das IR unterschieden werden, die auch bei isolierten Sammlungen auftreten. *Nichtannotierte Forschungsdaten*: Eine Suche über die Forschungsdaten kann inhaltsbasiert erfolgen. *Interpretierbare Metadaten* liegen ohne Verweis auf die Daten (z. B. nicht ausreichende Rechte) vor und können je nach Umfang und Vollständigkeit die Präzision der Suche erhöhen. *Metadaten und Forschungsdaten* können kombinierten Suchmechanismen bereitgestellt werden.

(2) Neben der horizontalen Bewertung der Qualität vorliegender Daten muss je nach Anwenderperspektive auch eine vertikale Dimension analysiert und in den Anwendungen des IR berücksichtigt werden. Wird die Qualität von Metadaten an dieser Stelle definiert als die Granularität der Beschreibungsdaten, so gilt diese stets nur für die Perspektive, aus der eine entsprechende Annotation erfolgt ist. An das obige Beispiel der Judaistik anknüpfend ist eine hohe Qualität der Metadaten zu einem Grabstein definiert durch die Vollständigkeit der epigraphischen Beschreibung. Bei Betrachtung aus der Perspektive der Denkmalkunde weisen diese Metadaten jedoch nur einen geringen Informationswert auf.

3 Problemstellungen

Eine wesentliche Fragestellung für die Konzeption der Föderationsarchitektur und die Anwendung der Methoden des IR ergibt sich aus dem interorganisationalen und interdisziplinären Charakter der Anwendungsdomäne von DARIAH und der hieraus resultierenden Heterogenität.

Kontext

Wie auch Palmer, Zavalina und Mustafoff [2007] im Rahmen ihrer Analysen zur Föderation digitaler Archive von Bibliotheken und Museen feststellten, existiert bereits eine Vielzahl von Metadaten-Schemata, Infrastrukturen und Best Practices, die in isolierter Betrachtung einzelner Domänen als ausgereift bewertet werden können.

Eine wichtige Voraussetzung hierfür ist jedoch die Begrenzung auf einen ausgewählten Kontext, der den Anwendern bei Nutzung der Sammlung bewusst sein muss und somit implizites Wissen voraussetzt. Erst durch die Einordnung von Suchergebnissen in diesen Kontext kann eine solide Interpretation gewonnener Daten erfolgen.

Während implizites Wissen in einem fachlich homogenen Umfeld vorausgesetzt werden kann, muss dieses im Rahmen einer interdisziplinären Föderation expliziert werden, um den Kontext und damit den Informationsgehalt von Meta- und Forschungsdaten zu bewahren.

Dieser Kontext ist zum einen erforderlich, um eine kontextsensitive Analyse im Rahmen des IR zu ermöglichen, zum anderen muss bei einer Sammlung von Forschungsdaten aus unterschiedlichen Quellen auch stets ein Erhalt von Autoren- bzw. Urheberbeziehungen möglich sein. Unterschiedliche Arbeiten zur Assoziation von Metadaten auf Objekt- und Sammlungsebene (z. B. [Palmer et al, 2007], [Foulonneau et al, 2005]) verdeutlichen die Notwendigkeit der Verknüpfung von Metadaten auf diesen Ebenen.

Metadaten und Medien

Durch die Heterogenität des Kontexts und die Vielzahl der zu unterstützenden Medientypen ergibt sich für die Anwendungslandschaft der Digital Humanities eine breite Menge generischer und spezialisierter Metadaten-Schemata, die im Rahmen des IR zu verarbeiten sind. Während hierbei der Dublin Core Standard zunächst als geeignetes Medium der Interoperabilität für eine fach- und medienübergreifende Föderation erscheint, zeigen bereits erste Analysen, dass dessen Beschreibungstiefe für viele Anwendungsfälle nicht ausreicht. In den DCMI Metadata Terms [DCMI, 2010b] wird aus diesem Grund empfohlen, die wenigen Elemente des Dublin Core Standards durch eine fachspezifische Untergliederung zu konkretisieren, um auf wertvolle, feingranulare Beschreibungen nicht verzichten zu müssen. Hierdurch wird jedoch auch der Aspekt der Interoperabilität des Standards abgeschwächt. Fachspezifische Schemata, wie epiDoc oder MEI sowie organisationseigene Standards weisen zwar eine inhärent hohe Granularität auf, erschweren dabei aber die Interoperabilität und führen zu der Notwendigkeit definierter Transformationsregeln zwischen Schemata. Eine Aufgabe des IR muss demnach im Kontext von DARIAH in der Unterstützung und Ausnutzung dieser so genannten Crosswalks liegen und die Fragestellung lösen, wann eine Transformation zwischen den Schemata optimal durchgeführt bzw. genutzt werden kann. Zentral scheint an dieser Stelle das Problem, eine geeignete Strategie der Verwaltung, Versionierung und Zuordnung von Schemata und Crosswalks zu entwickeln. Während die Auswahl definierter Zielformate die Anzahl der Transformationsregeln vermindert, erhöhen direkte n:m-Crosswalks die Präzision der Mappings, führen jedoch zu einem Verwaltungsaufwand, der bei der Vielzahl möglicher Schemata, Profile und Anpassungen zu einem erhöhten Aufwand für die Erweiterbarkeit und Wartung der Infrastruktur führt.

Zugriff und Rahmenbedingungen

Eine weitere Ebene der Heterogenität, welche durch die Föderationsarchitektur behandelt werden muss, besteht in der Unterstützung unterschiedlicher Zugriffsstrategien und –protokolle anzubindender Archive. Während die Implementierung des Harvesting Protokolls OAI-PMH laut den Untersuchungen von Brogan [2006] einem bemerkenswerten Wachstum unterliegt, muss auch anerkannt werden, dass OAI-PMH und dessen Fokussierung auf den einfachen Dublin Core Metadaten-Standard Grenzen aufweist, welche sich insbesondere bei Versuchen der Repräsentation komplexer Objekte und ihrer Assoziationen zeigen. Die Nutzung umfangreicherer Metadaten-Standards erlaubt zwar eine Erweiterung dieser Beschreibungsgrenzen. Dennoch darf sich die Konzeption der

Föderationsinfrastruktur nicht auf die Erwartung stützen, dass OAI-PMH und Dublin Core als Minimalstandards stets durch digitale Archive angeboten werden können. Gesammelte Erfahrungen eines Projekts der NSDL[7] zur Integration digitaler Archive durch die Sammlung und Aggregation von Metadaten [Lagoze et al, 2006] zeigen, dass auch die Umsetzung der einfachen Standards OAI-PMH und Dublin Core durch die angebundenen Datenquellen nicht vorausgesetzt werden kann. Die Forscher der Cornell University führten diesen Umstand auf fehlende personelle Ressourcen mit Wissen über die fachliche Domäne, die Metadaten selbst und über die technische Umsetzung zurück. Als logische Konsequenz muss für DARIAH vorausgesetzt werden, dass neben OAI-PMH auch eine Anbindung von Quellen berücksichtigt werden muss, die keine derart standardisierte Schnittstelle aufweisen. Sammlungen können demnach im Wesentlichen klassifiziert werden durch die Art der bereitgestellten Schnittstellen.

- *Harvesting*: Metadaten können durch einen zentralen Index für das IR bereitgestellt werden.
- *Querying*: Sammlungen erlauben nur Zugriffe in Form von Distributed Queries.
- *Keine*: Sammlungen können ausschließlich durch Crawling-Mechanismen verarbeitet werden.

Die Bereitstellung von Schnittstellen durch die verteilten Archive ist dabei auch abhängig von der Bereitschaft oder den rechtlichen Bedingungen der bereitgestellten Forschungsdaten und Metadaten. Für Lagoze et al. konnte nur die Verpflichtung der angebundenen Quellen des NSF zu der Bereitstellung von Metadaten führen, die im Rahmen von DARIAH aufgrund der Autonomie anzubindender Organisationen und Archive nicht möglich sein kann.

4 Suchdienste

In DARIAH-DE wurden die Suchdienste zunächst primär unter dem Aspekt einer generischen, übergreifenden (Volltext-)Suche gesehen. Eine genauere Betrachtung bringt aber sehr schnell die verschiedensten Anforderungen zutage:

Verteiltes IR: DARIAH verfolgt im Hinblick auf die Kollektionen primär den Gedanken einer Föderation mit zentralen Registierungspunkten. Viele Systeme (wie z. B. Fedora Commons[8]) verfügen dabei auch über eigene Suchdienste. Für entsprechende Informationsbedürfnisse sind daher zunächst auch Techniken der Quellenauswahl relevant.

Metasuche: Dem Sinn einer Föderation entspricht auch der Gedanke der Metasuche. Probleme bereiten dabei aber Kollektionen ohne hinreichende eigene Suchfunktionalität und insgesamt Schwächen der individuell bereitgestellten Suchdienste.

Domänenspezifische Suchdienste: Um die Leistungsfähigkeit des Konzepts zu demonstrieren erscheint es sinnvoll auch leistungsfähige spezifische Suchdienste anzubieten, die den spezifischen Dokumenten und Anforderungen einzelner Disziplinen Rechnung tragen (ein Beispiel findet sich in [Ernst-Gerlach and Fuhr, 2010]).

Generische und übergreifende Suchdienste: Das primäre Augenmerk von DARIAH liegt in der Interoperabilität und in der übergreifenden Perspektive. Ein Suchdienst kann unerwartete Zusammenhänge zwischen Kollektionen verschiedener Disziplinen aufdecken.

5 Fazit

Insgesamt bieten die Suchdienste im Kontext von DARIAH-DE vielfältige Aufgabenstellungen und Herausforderungen. Im Rahmen des Projektes können dabei nur ausgewählte Aspekte wie die Nutzung und Unterstützung von Crosswalks in der Suche adressiert werden.

[7] National Science Digital Library, http://nsdl.org/

[8] http://fedora-commons.org/

References

[Adams and Blandford, 2005] Anne Adams and Ann Blandford. Digital Libraries' Support for the User's 'Information Journey'. In *Proceedings of the 5th ACM/IEEE-CS joint conf. on Digital Libraries*, p. 160-169, Denver, Colorado, 2005.

[Hockey, 2004] Susan Hockey. The History of Humani-ties Computing. In Susan Schreibman, Ray Siemens and John Unsworth (eds) *A Companion to Digital Humanities*. Blackwell Publishing, 2004

[DCMI, 2010a] *Dublin Core Metadata Element Set, Version 1.1*. Dublin Core Metadata Initiative, 2010. http://dublincore.org/documents/2010/10/11/dces/

[DCMI, 2010b] *DCMI Metadata Terms*. Dublin Core Metadata Initiative, 2010. http://dublincore.org/documents/2010/10/11/dcmi-terms/

[Ernst-Gerlach and Fuhr, 2010] Andrea Ernst-Gerlach and Norbert Fuhr. Semiautomatische Konstruktion von Trainingsdaten für historische Dokumente. In *Lehren – Wissen – Adaptivität (LWA 2010).* Kassel, 2010.

[Lagoze et al, 2008a] Carl Lagoze, Herbert Van de Sompel, Michael Nelson, and Simeon Warner. *The Open Archives Initiative Protocol for Metadata Harvesting - v.2.0*. Open Archives Initiative, 2008. http://www.openarchives.org/OAI/openarchivesprotocol.html

[Lagoze et al, 2008b] Carl Lagoze, Herbert Van de Sompel, Pete Johnston, Michael Nelson, Robert Sanderson, Simeon Warner. *ORE User Guide – Primer*. Open Archives Initiative, 2008.

[Palmer et al, 2007] Carole L. Palmer, Oksana L. Zavalina, and Megan Mustafoff. Trends in metadata practices: a longitudinal study of collection federation. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries* (JCDL '07), p. 386-395. New York, NY, 2007. ACM.

[Foulonneau et al, 2005] Muriel Foulonneau, Timothy W. Cole, Thomas G. Habing, and Sarah L. Shreeves. 2005. Using collection descriptions to enhance an aggregation of harvested item-level metadata. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries (JCDL '05)*, p. 32-41. New York, NY, 2005. ACM.

[Lagoze et al, 2006] Carl Lagoze, Dean Krafft, Tim Cornwell, Naomi Dushay, Dean Eckstrom, and John Saylor. 2006. Metadata aggregation and "automated digital libraries": a retrospective on the NSDL experience. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries (JCDL '06)*, p. 230-239. New York, NY, 2005. ACM.

[Brogan, 2006] Martha Brogan. Contexts and Contributions: *Building the Distributed Library*. Washington, DC, 2006, Digital Library Federation.