

LZI+DBLP

Zwischenbericht Mai 2011

Oliver Hoffmann Marc Herbstritt Christian Lindig
Schloss Dagstuhl – Leibniz-Zentrum für Informatik
www.dagstuhl.de

Zusammenfassung

Die Informatik benötigt eine belastbare Datenbasis zum Nachweis und zur Evaluierung wissenschaftlicher Literatur. Eine von der Leibniz-Gemeinschaft geförderte Zusammenarbeit von Schloss Dagstuhl – Leibniz-Zentrum für Informatik und der Literaturdatenbank DBLP entwickelt dafür die technischen, inhaltlichen und organisatorischen Strukturen. Vor dem offiziellen Projektbeginn im Juni 2011 ermöglichte eine Spende der Klaus Tschira Stiftung erste Vorarbeiten. Dieser Zwischenbericht erläutert den Stand nach dem ersten Jahr der informellen Zusammenarbeit. Der Schwerpunkt liegt auf für Nutzer sichtbaren Veränderungen bei Inhalt, Umfang und Aktualität: so konnte die Aufnahme neuer Literatur durch Automatisierung um 79% gesteigert werden.

1 Einleitung

Die spezielle Publikationskultur in der Informatik, die traditionell den Schwerpunkt auf Konferenzpublikationen legt, erfordert eine belastbare Datenbasis für wissenschaftliche Publikationen, der eine nachhaltige Infrastruktur zugrunde liegt. Schloss Dagstuhl – Leibniz-Zentrum für Informatik (LZI) und die Literaturdatenbank DBLP an der Universität Trier arbeiten seit Anfang 2010 zusammen mit dem gemeinsamen Ziel, die Literaturdatenbank DBLP gemäß ihrer nationalen und internationalen Bedeutung als zentralen Nachweis für wissenschaftliche Literatur in der Informatik und der Informatik nahestehenden Gebieten inhaltlich, organisatorisch und technisch zu stärken und auszubauen.

DBLP besteht seit 1993 und wurde seitdem von seinem Gründer Dr. Michael Ley betrieben, wobei ausschließlich wissenschaftliche Hilfskräfte zur Unterstützung zur Verfügung standen. Durch einen gemeinsamen Projektantrag des LZI mit der DBLP konnte nun für eine Profilierungsphase von zwei Jahren die personelle Situation durch die Verstärkung von zwei Vollzeit-Mitarbeitern deutlich verbessert werden. Eine zusätzliche Spende ermöglichte, dass vorab ein Projektmitarbeiter bereits ein halbes Jahr vor dem offiziellen Projektbeginn mit seiner Arbeit beginnen konnte.

Dieser Zwischenbericht dient daher (1) als Bilanz für die erzielten Ergebnisse nach dem ersten Jahr der Zusammenarbeit und (2) als Referenz für das eigentliche Projekt.

2 Organisatorische Struktur

Der Wissenschaftliche Beirat, das Industrielle Kuratorium und das wissenschaftliche Direktorium von Schloss Dagstuhl unterstützen nachdrücklich eine Kooperation zwischen Schloss Dagstuhl und DBLP. Ebenso positiv war die Resonanz durch den Fakultätentag Informatik¹ und den Beirat der Universitätsprofessoren der Gesellschaft für Informatik (GIBU)², denen die Zusammenarbeit 2010 vorgestellt wurde.

Schloss Dagstuhl hat deshalb im März 2010 beim Senatsausschuss Wettbewerb der Leibniz-Gemeinschaft³ einen Antrag auf Förderung gestellt, der im Dezember 2010 bewilligt wurde und im Wesentlichen die Finanzierung von zwei wissenschaftlichen Stellen für die Dauer von zwei Jahren vorsieht. Mit zunächst einer durch Dipl.-Inform. Oliver Hoffmann besetzten Stelle beginnt das Projekt offiziell im Juni 2011 und wird ab August 2011 durch Dr. Marcel Ackermann verstärkt. Beide arbeiten an der Universität Trier mit Dr. Michael Ley zusammen.

Vorab wurde das Vorhaben durch eine Spende in Höhe von 25.000 € von der Klaus Tschira Stiftung gefördert, die durch Prof. Dr.-Ing. Dr. h.c. Andreas Reuter als Mitglied des Industriellen Kuratoriums von Schloss Dagstuhl vermittelt wurde. Die Spende ging im November 2010 ein und diente der Finanzierung der Stelle von Herrn Oliver Hoffmann ab diesem Zeitpunkt. Dadurch konnten Vorarbeiten für das Projekt bereits vor dem offiziellen Projektbeginn begonnen werden, deren Ergebnis wir nachfolgend dokumentieren.

Zusätzlich zum technischen Ausbau der DBLP ist es ein Ziel des Projekts, eine wissenschaftliche Aufsicht zu etablieren, welche Qualitätskriterien für die Aufnahme von Publikationen in die DBLP entwickeln soll, die mehr Transparenz und Systematik zum Ziel haben. Vor dem Hintergrund steuert Schloss Dagstuhl sein Wissen und seine Kontakte für den systematischen Aufbau einer wissenschaftlichen Aufsicht bei.

3 Projektfortschritt November 2010 – April 2011

Die Literaturdatenbank DBLP verzeichnet wissenschaftliche Literatur in der Informatik auf der Ebene einzelner Beiträge. DBLP konzentriert sich bei der Wahl der Datenquellen auf Verlage und Bibliotheken und verfolgt das

¹<http://www.ft-informatik.de/>

²<http://www.gibu.gi-ev.de/>

³<http://www.wgl.de/?nid=wet>

Ziel, ganze Reihen und Serien und damit Themengebiete wissenschaftlicher Literatur vollständig zu erfassen. Anfang des Jahres 2010 enthielt DBLP etwa 1,3 Millionen Datensätze.

Über Jahre hinweg wurden diese Daten mit hohem manuellen Anteil erfasst. Dies führte zu einer anerkannt hohen Qualität des Datenbestandes, gefährdete aber seine Aktualität. In einem ersten Schritt wurde daher im Rahmen der Diplomarbeit von Herrn Oliver Hoffmann eine Suchmaschine entwickelt, die Rohdaten zu Literatur von Web-Seiten wissenschaftlicher Verlage und Bibliotheken sammelt [1]. Dies entspricht dem Vorgehen einer Suchmaschine wie Google, die durch *Crawler* Web-Seiten automatisch besucht und indexiert.

Zunächst wurden hierzu sämtliche in DBLP bereits indexierten Zeitschriften (journals) und ihre Verleger systematisch als Datei erfasst. Die Crawler-Software kann die Datei regelmäßig abarbeiten, die Web-Seiten der einzelnen Zeitschriften besuchen und überprüfen, ob seit dem letzten Besuch neue Daten veröffentlicht wurden. Ist dies der Fall, so werden die entsprechenden Daten extrahiert, normalisiert und in ein internes Format gebracht, das zur Integration in DBLP erforderlich ist. Diesen Schritt automatisiert eine *Wrapper*-Software, die ebenfalls von Herrn Hoffmann im Rahmen seiner Diplomarbeit entwickelt wurde.

3.1 Steigerung der Produktivität

Die Wrapper-Software extrahiert Rohdaten aus den Web-Seiten von Verlegern und Bibliotheken und muss an deren individuellen (und sich oftmals ändernden) Aufbau angepasst werden. Eine regelbasierte Architektur der Wrapper-Software ermöglicht die schnelle Anpassung an die individuellen Gegebenheiten. Im Oktober 2010 verfügte DBLP über insgesamt 34 solcher individualisierter Wrapper. Im Rahmen der ersten Etappe wurde diese Software erweitert und verbessert: Die Anzahl an Wrappern wurde nahezu verdoppelt und auf insgesamt 61 (Stand: Mai 2011) erhöht, einige bereits bestehende Wrapper mussten zudem an sich ändernde Gegebenheiten angepasst werden.

Im Vergleich zum Vorjahr wurde die Produktivität bei der Erfassung neuer Datensätze gesteigert. Dies resultiert einerseits aus der Automatisierung einzelner Arbeitsabläufe und der gesteigerten Anzahl an Wrappern. Andererseits war es dank erhöhter „Manpower“ auch möglich, die Aufnahme der nach wie vor manuell zu erfassenden Daten zu beschleunigen. Dies zeigt sich auch an anderer Stelle: Oftmals erhält DBLP E-Mails von Verlegern, die ihre Daten bereits in einem mehr oder weniger strukturierten Format einreichen. Diese Daten müssen vor einer Aufnahme in den Bestand von DBLP stets manuell aufbereitet werden, was in der Vergangenheit aus Zeitgründen jedoch nicht immer möglich war. Oftmals mussten sie sehr lange auf eine Aufnahme warten oder konnten gar nicht erfasst werden.

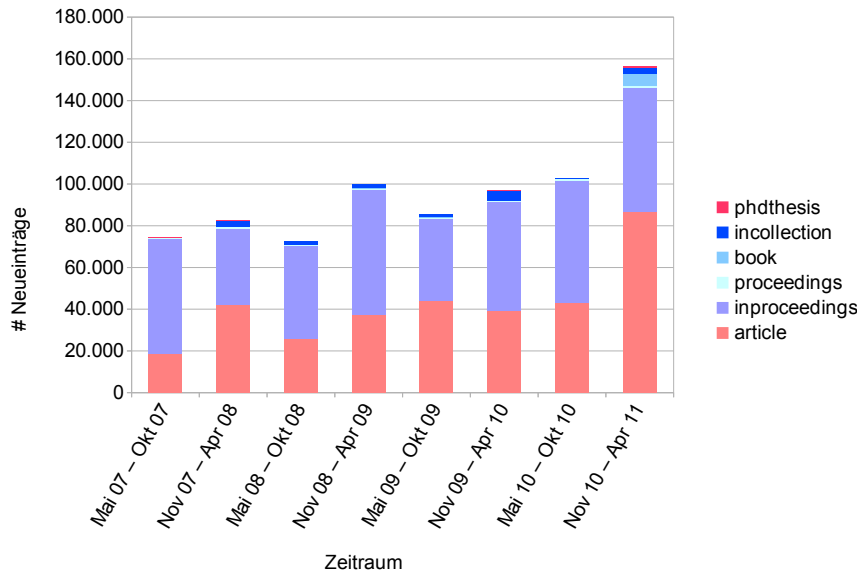


Abbildung 1: Anzahl der neu erfassten Records in Perioden von 6 Monaten. Die Vorarbeiten für LZI+DBLP starteten im November 2010.

Abbildung 1 verdeutlicht den Zuwachs an Produktivität. Man erkennt, dass im Projektzeitraum⁴ die Anzahl der neu erfassten Einträge die entsprechenden Werte vergangener Perioden signifikant übersteigt. Vor Projektbeginn wurden in 6 Monaten durchschnittlich 87.780 neue Einträge in DBLP aufgenommen. In den 6 Monaten von November 2010 bis April 2011 wurden 156.390 neue Einträge aufgenommen – dies entspricht einer Steigerung der Neuaufnahmen von 79%. Dabei wurde die Aufnahme von Zeitschriftenartikeln (Typ *article*) überproportional sogar um 142% gesteigert.

3.2 Aktualität

Die Literaturdatenbank DBLP soll neue Literatur möglichst schnell und vollständig erfassen, um die Nachfrage insbesondere bei der Recherche von aktuellen Publikationen befriedigen zu können. Ein Maß für die Aktualität einer Veröffentlichung ist das Alter bei ihrer Erfassung. Es ist der Zeitraum zwischen dem Erscheinen der Veröffentlichung und ihrem Eintrag in DBLP. Um die Entwicklung der Aktualität zu messen, haben wir das Alter von neu erfassten Veröffentlichungen verglichen. Wir berücksichtigten dazu die

⁴Als *Projektzeitraum* wird im folgenden die Zeit von November 2010 bis April 2011 bezeichnet, in der Herr Hoffmann aufgrund der Spende vorab für das Projekt tätig werden konnte. Das auf zwei Jahre geförderte Projekt der Leibniz-Gemeinschaft hingegen beginnt offiziell im Juni 2011.

Tabelle 1: Klassifizierung des Alters von neuen Einträgen in DBLP im Projekt- und im Vergleichszeitraum basierend auf dem Erscheinungsjahr.

Aufnahme in DBLP				
Nov 2010 bis Apr 2011		Nov 2009 bis Apr 2010		Alter
Jahr*	Anzahl	Jahr*	Anzahl	
≥ 2010	94894	≥ 2009	72280	aktuell
2009	18450	2008	9572	1 Jahr
2008	6913	2007	3319	2 Jahre
≤ 2007	36121	≤ 2006	11487	älter

* Erscheinungsjahr der Publikation

im Projektzeitraum November 2010 bis April 2011 in DBLP erfasste Literatur und die im Vorjahreszeitraum (November 2009 bis April 2010) erfasste Literatur.

Das Alter eines Literatureintrags wurde entsprechend Tabelle 1 klassifiziert. Die Altersverteilung der im Projektzeitraum erfassten Einträge im Vergleich zum Vorjahr zeigt Abbildung 2. In beiden Zeiträumen wurde aktuelle wie ältere Literatur in vergleichbaren Anteilen in DBLP aufgenommen.

Es ist zu erwarten, dass das Verhältnis von aktueller zu älterer Literatur stabil bleibt. Eine Steigerung der Aktualität wäre derzeit nur durch den Verzicht auf die Aufnahme älterer Literatur möglich. DBLP nimmt aber weiterhin relevante Zeitschriften und Reihen neu auf und damit auch darin enthaltene zurückliegenden Ausgaben.

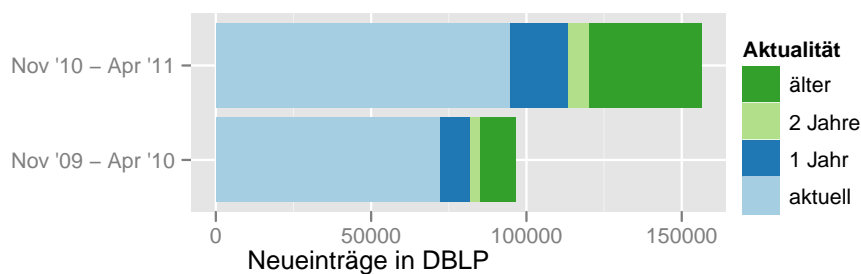


Abbildung 2: Aktualität von Neueinträgen im Projekt- und Vergleichszeitraum gemessen am Alter einer Publikation bei der Aufnahme in DBLP.

3.3 Integration der Bibliothek von Schloss Dagstuhl

Ein weiteres Ziel der ersten Projektetappe war die Integration des Buchbestands der Bibliothek von Schloss Dagstuhl im Bereich von Monographien und Sammelbänden. Hierzu wurde ein Gesamtbestand von 22.833 Büchern aus dem Bibliothekskatalog exportiert und vorverarbeitet. Die Aufnahme in DBLP musste dennoch manuellen Kontrollen unterzogen werden, um Redundanzen im Datenbestand von DBLP zu vermeiden und Fehler im Datenbestand der Bibliothek aufzudecken. Obwohl es sich hierbei um einen äußerst gut gepflegten Datensatz handelt, konnten zahlreiche Korrekturen und Ergänzungen durchgeführt werden. Neben der Erweiterung von DBLP wurde somit auch der Datenbestand des Bibliothekskatalogs von Schloss Dagstuhl aufgewertet.

Die Grafik in Abbildung 3 zeigt eine Aufteilung des genannten Bestands von 22.833 Monographien und Sammelbänden, aufgeteilt in die folgenden Kategorien:

Bücher – übernommen Diese Bücher wurden manuell überprüft und in den Datenbestand von DBLP aufgenommen.

Bücher – ausgeschlossen Ein kleiner Teil der überprüften Bücher wurde bewusst nicht aufgenommen. Hierbei handelt es sich einerseits um Bücher, die bereits zu früheren Zeitpunkten in DBLP erfasst wurden, andererseits um solche, die außerhalb des Themengebiets der Informatik liegen.

Bücher – bereits erfasst Die Dagstuhl-Bibliothek verzeichnet einige große Buchserien – allen voran die LNCS-Bände von Springer –, die ohnehin in DBLP erfasst werden. Diese sind für den Export unerheblich und werden daher ignoriert.

Sammelbände Sammelbände können nicht direkt in DBLP aufgenommen werden, da hier der Fokus auf der Erfassung der einzelnen Artikel liegt, während in der Bibliothek lediglich die Bücher selbst verzeichnet sind. Diese Bände können daher zwar als Anregung dienen, ihre Daten müssen jedoch von anderer Stelle besorgt werden.

Bücher – ausstehend Der restliche Bestand wurde noch nicht bearbeitet und keiner Sortierung unterzogen. Die Bücher dieser Kategorie müssen noch untersucht und einer der vier anderen Kategorien zugeordnet werden. Dies ist für die nächste Projektetappe geplant.

Die Integration der Daten aus der Bibliothek von Schloss Dagstuhl konnte im Projektzeitraum bereits zu großen Teilen abgeschlossen werden: Lediglich 28% des Bestandes wurde noch nicht untersucht (ausstehend), bei weiteren 13% muss eine manuelle Erfassung auf Artikelebene erfolgen.

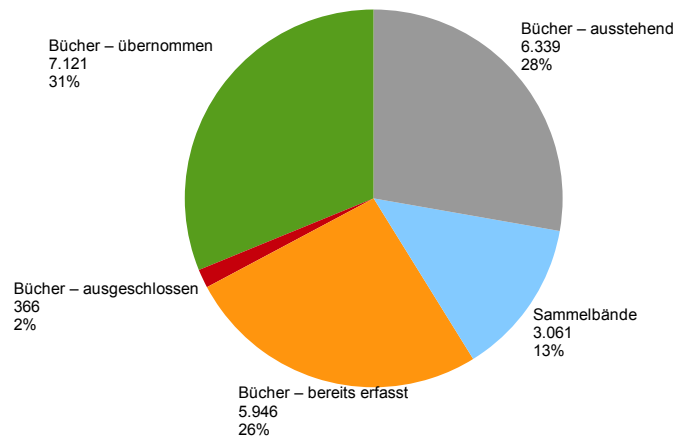


Abbildung 3: Integration der Daten aus der Bibliothek Schloss Dagstuhl. Etwa 30% des Bestandes wurde bereits übernommen und weitere 28% sind für eine Übernahme vorgesehen. Der restliche Bestand ist für DBLP nicht relevant oder wird aus anderen Quellen erschlossen.

Durch die Sichtbarkeit der Integration der Bibliothek auch nach Außen (die Primärschlüssel aller Datensätze beinhalten ihre Herkunft „daglib“) soll Autoren und Verlegern in Zukunft ein hoher Anreiz geboten werden, der Bibliothek Bücherspenden zukommen zu lassen, um deren Metadaten somit schnell in DBLP wiederzufinden. Einige Bücher und eher „exotische“ Zeitschriften aus dem Bestand von DBLP wurden der Bibliothek gespendet.

3.4 Verzeichnis aller Zeitschriften

Alle in DBLP erfassten Zeitschriften können auch über ein Stichwortregister⁵ gefunden werden. Dieses wurde bisher manuell gepflegt: eine neu aufgenommene Zeitschrift musste unter den jeweils relevanten Stichworten eingetragen werden. Besonders die manuelle Suche der korrekten Position innerhalb der alphabetischen Liste war sowohl mühsam als auch fehleranfällig; und wurde der Eintrag vergessen, so war die Zeitschrift über diesen Index nicht auffindbar.

Im Zuge der Systematisierung der Zeitschriften wurde nun die Möglichkeit geschaffen, dieses Register vollautomatisch generieren zu lassen. Stichworte, unter denen die Zeitschrift geführt werden soll, können in einer zentralen Datei bequem markiert werden. Die Sortierung der Einträge und Erstellung der Web-Seiten übernimmt anschließend die Software.

⁵<http://dblp.uni-trier.de/db/journals/>

Zudem wurde ein weiterer Index „by Publisher“ erstellt⁶, über den die erfassten Zeitschriften großer Verlage oder Fachgesellschaften eingesehen werden können. Dies steigert die Transparenz und kann bei der Verhandlung mit den entsprechenden Organisationen (vgl. Abschnitt 3.6) äußerst hilfreich sein, um zu überprüfen, wie viele und welche Journale bereits aufgenommen wurden.

3.5 Aufnahme neuer Zeitschriften

Neben der Aktualisierung und Ergänzung bereits bekannter Zeitschriften wurde ein weiterer Schwerpunkt auf die Aufnahme neuer Zeitschriften in DBLP gesetzt. Bei der Entscheidung, welche Zeitschriften hier bevorzugt behandelt werden sollten, wurden die Listen zweier internationaler Journal-Rankings hinzugezogen:

- Ranking des Australian Council of Professors and Heads of Information Systems (ACPHIS)⁷
- Ausgewählte Informatik-Kategorien des SCImago Journal & Country Rank (SJR) 2009⁸

Diese Listen wurden systematisch durchgesehen. Fehlende Zeitschriften hoher Ränge (A* und A bzw. $SJR \geq 4$) wurden ergänzt, sofern dies mit vertretbarem Aufwand möglich war, d.h. bereits ein Wrapper für den publizierenden Verlag existierte oder ein solcher schnell erstellt werden konnte.

Vorwiegend beim SJR wurden einige interdisziplinäre Zeitschriften gefunden, die das Gebiet der Informatik nur tangieren. Daher wurde zuvor ein Abgleich der Autorennamen durchgeführt: wurden zahlreiche Autoren einer als Stichprobe ausgewählten Ausgabe auch in DBLP gefunden, so wurde die Zeitschrift aufgenommen. War die Abdeckung jedoch nur gering, so wurde die Zeitschrift vorerst nicht aufgenommen.

Abbildung 4 zeigt den Zuwachs an Zeitschriften, aufgeteilt nach Verlagen. Der höchste Zuwachs kann hier bei IGI Global festgestellt werden; Dies hat jedoch andere Gründe, die in Abschnitt 3.6 erläutert werden. Man erkennt außerdem die mengenmäßige Dominanz der Zeitschriften, die von kleinen Verlegern oder teilweise nur auf speziellen Web-Seiten publiziert werden (in der Grafik als „einzelne“ bezeichnet). Diese Anzahl wurde durch die Aufnahme der hoch bewerteten Zeitschriften um 24,12% erhöht. Jene Daten sind oftmals auf manuell erstellten Web-Seiten zu finden, für die das Schreiben eines Wrappers nur dann Sinn ergibt, wenn die Zeitschrift eine besonders hohe Qualität und/oder Quantität der Daten beinhaltet. Hieraus resultiert auch der in Abschnitt 3.1 beschriebene starke Zuwachs an Wrappern.

⁶<http://dblp.uni-trier.de/db/journals/publ/>

⁷<http://www.acphis.org.au>

⁸<http://www.scimagojr.com/journalrank.php?area=1700&year=2009>

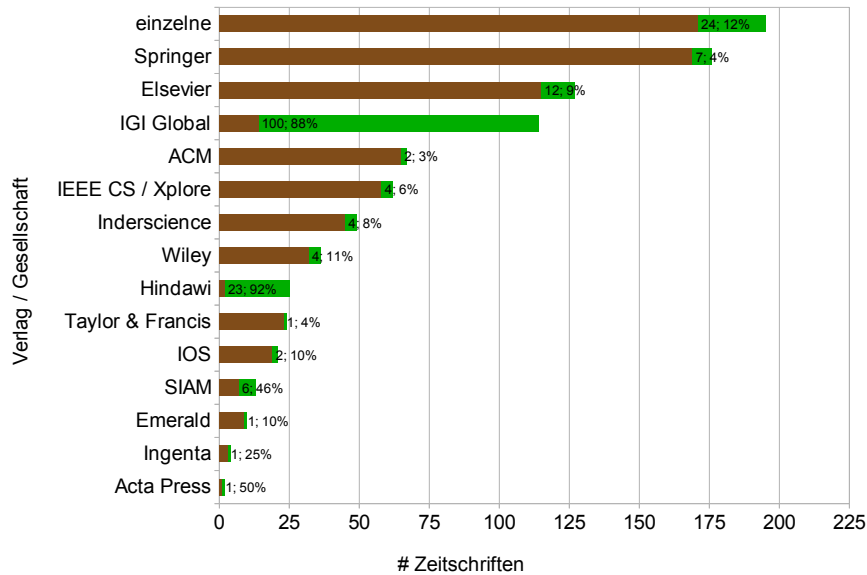


Abbildung 4: Zuwachs an Zeitschriften im Projektzeitraum. Die Bestände von IGI Global und Hindawi wurden stark erhöht, zudem erfolgte eine Aufnahme zahlreicher Zeitschriften kleiner Verlage.

Ein weiterer signifikanter Zuwachs an Zeitschriften kann beim Verleger Hindawi festgestellt werden. Hier finden sich vorwiegend hochwertige Beiträge, die innerhalb der „EURASIP Journal on ...“-Serien publiziert werden. Der gesamte online verfügbare Datenbestand dieser Serien wurde in DBLP integriert.

3.6 Kontakte mit Verlagen als Datenlieferanten

Neben der Datenproduktion sollte auch die Sichtbarkeit der Arbeit nach Außen verbessert werden. Daher wurde begonnen, gezielt Kontakte zu Verlagen herzustellen oder auszubauen. Erwähnenswert sind hier vor allem

IOS Durch eine Kooperation mit IOS wurden zahlreiche Zeitschriften rückergänzt. Da die meisten für die Informatik wichtigen Zeitschriften von IOS bereits in DBLP integriert waren, spiegelt sich diese Kooperation jedoch nicht in Abbildung 4 wider.

IGI Global Fast das komplette Verlagsprogramm wurde in DBLP aufgenommen. Viele der Zeitschriften wurden allerdings in den Jahren 2009 bis 2011 neu gegründet und verfügen nur über wenige Artikel; es bleibt

abzuwarten, wie viele der Zeitschriften sich langfristig durchsetzen werden.

Springer Science+Business Media Zum Ende des Projektzeitraums wurde auch ein Treffen mit Alfred Hofmann, Leiter der Informatik-Abteilung bei Springer Science+Business Media, auf Schloss Dagstuhl organisiert. Dieses Treffen fand am 10. Mai 2011 statt. Da dieser Termin außerhalb des Projektzeitraums liegt, werden die Ergebnisse an anderer Stelle dokumentiert.

Wie bereits in Abschnitt 3.1 beschrieben, erhält DBLP oftmals auch E-Mails von Verlegern, Editoren oder Autoren, mit der Bitte, Daten aufzunehmen oder zu verändern. Einige Verleger liefern auch bereits Metadaten neuer Publikationen in einem strukturierten Eingabeformat. In der Vergangenheit wurden solche E-Mails oftmals erst spät oder gar nicht bearbeitet. Fehler wurden zwar stets zügig korrigiert, eine Benachrichtigung des Senders blieb jedoch aus Zeitgründen meist aus.

Durch die erhöhte „Manpower“ konnte während des Projektes damit begonnen werden, derartige Anfragen konsequent zu beantworten. Die Sender erhielten stets eine Rückmeldung, dass ihre Daten aufgenommen wurden, oder aus welchen Gründen eine Aufnahme nicht stattfinden konnte. Sender, die einen Fehler meldeten, erhielten eine kurze Benachrichtigung über die erfolgte Korrektur. Dieses Vorgehen soll auch in Zukunft beibehalten und ausgebaut werden. Zum einen wird dadurch die Transparenz der Aufnahme von Daten erhöht und damit die Seriosität von DBLP nach außen hin untermauert. Zum anderen erhalten die Sender positive Signale, die sie motivieren sollen, auch zukünftig bei der Erweiterung und Verbesserung des Datenbestandes aktive Hilfe zu leisten.

4 Fazit

Die Zusammenarbeit von Schloss Dagstuhl mit der Literaturdatenbank DBLP an der Universität Trier ist auf einem guten Weg. Bereits vor dem Start der Förderung durch den Senatsausschuss Wettbewerb der Leibniz-Gemeinschaft konnten durch eine Spende der Klaus Tschira Stiftung wesentliche Vorarbeiten geleistet werden:

1. Die Zusammenarbeit findet weltweit breite Unterstützung innerhalb der Informatik.
2. Die technische Infrastruktur wurde verbessert, so dass bereits jetzt die Produktivität um 79% verbessert wurde.
3. Durch den Abgleich der Kataloge von DBLP und der Bibliothek von Schloss Dagstuhl wurden verstärkt Lehrbücher in DBLP aufgenommen.

4. Für die Projektarbeit wurden beide vorgesehenen wissenschaftlichen Stellen besetzt.
5. Die Zahl der in DBLP nachgewiesenen Zeitschriften wurde um 192 Zeitschriften erhöht.
6. Eine DBLP-Initiativgruppe wurde gebildet, die den Grundstein für die wissenschaftliche Aufsicht legen wird. Diese Gruppe hatte ihr erstes Treffen am 29. Mai 2011.

Die bereits jetzt erzielten Ergebnisse zeichnen ein zuversichtliches Bild für die nun beginnende offizielle Projektarbeit.

Literatur

- [1] Oliver Hoffmann. Regelbasierte Extraktion und asymmetrische Fusion bibliographischer Informationen. Diplomarbeit, Universität Trier, 2009.