LZI+DBLP: Konsolidierung der bibliometrischen Datenbasis in der Informatik

Abschlussbericht Juni 2013

Marcel R. Ackermann* Marc Herbstritt* Oliver Hoffmann*
Michael Ley† Christian Lindig* Michael Wagner*

Zusammenfassung

Die Informatik benötigt eine belastbare Datenbasis zum Nachweis und zur Evaluierung wissenschaftlicher Literatur. In dem von der Leibniz-Gemeinschaft geförderten Projekt „LZI+DBLP“ entwickeln Schloss Dagstuhl – Leibniz-Zentrum für Informatik und die Literaturdatenbank *dblp* gemeinsam die dafür notwendigen technischen, inhaltlichen und organisatorischen Strukturen. Das Projekt wird zudem durch die Klaus Tschira Stiftung gefördert.

Dieser Bericht dokumentiert die Ergebnisse des Projektes zum Zeitpunkt des offiziellen Projektendes im Mai 2013. Er berichtet von den Fortschritten bei der Etablierung neuer Richtlinien und Strukturen zur Sicherung der Datenqualität von *dblp* und der kontinuierlichen Weiterentwicklung der Datenbasis.

1 Einleitung

Die Evaluierung wissenschaftlicher Literatur erfordert eine belastbare Datenbasis für wissenschaftliche Publikationen, der eine nachhaltige Infrastruktur zugrunde liegt. Die einzigartige Publikationskultur in der Informatik, die einen starken Schwerpunkt auf Konferenzpublikationen legt, stellt dabei besondere Anforderungen, denen allgemeine Literaturdatenbanken in der Regel nicht gerecht werden. *Schloss Dagstuhl – Leibniz-Zentrum für Informatik (LZI)* und die *Universität Trier* arbeiten seit November 2010 zusammen mit dem gemeinsamen Ziel, die Literaturdatenbank *dblp* gemäß ihrer nationalen und internationalen Bedeutung als zentralen Nachweis für wissenschaftliche Veröffentlichungen in der Informatik und der Informatik nahestehenden Gebieten inhaltlich, organisatorisch und technisch zu stärken und auszubauen.

*Schloss Dagstuhl – Leibniz-Zentrum für Informatik, http://www.dagstuhl.de/
†*dblp* computer science bibliography, Universität Trier, http://www.dblp.org/

Zu diesem Zweck wurde das über den Pakt für Forschung und Innovation[1] von der *Leibniz-Gemeinschaft*[2] geförderte Projekt „LZI+DBLP: Konsolidierung der bibliometrischen Datenbasis in der Informatik“ im Juni 2011 offiziell gestartet. Das Projekt wurde zudem bereits ab November 2012 durch eine Spende der *Klaus Tschira Stiftung*[3] signifikant gefördert. Dieser Bericht dokumentiert die erzielten Ergebnisse von der Projektvorbereitungsphase im November 2010 bis zum offiziellen Projektende im Mai 2013.

2 Hintergrund und organisatorische Struktur

Die Datenbank *dblp* besteht seit 1993 und wird seitdem von ihrem Gründer Dr. Michael Ley gepflegt und erweitert. Zunächst als Ein-Personen-Projekt gestartet und später durch gelegentliche Spendenmittel, wissenschaftliche Mitarbeiter und studentische Hilfskräfte unterstützt, entwickelte sich *dblp* über die Jahre zu der internationalen Referenzdatenbank bibliographischer Metadaten in der Informatik. Die Personalstärke von *dblp* blieb dabei jedoch stets deutlich unterhalb der gemäß Aufgabe und Bedeutung angemessenen Notwendigkeit.

Um den Nutzen von *dblp* für die nationale und internationale Informatik-Forschung langfristig zu konsolidieren und zu stärken, wurde Anfang 2010 eine Zusammenarbeit zwischen *dblp* und Schloss Dagstuhl – Leibniz-Zentrum für Informatik initiiert. Der wissenschaftliche Beirat, das industrielle Kuratorium und das wissenschaftliche Direktorium von Schloss Dagstuhl sprachen sich dabei nachdrücklich für eine Kooperation zwischen Schloss Dagstuhl und *dblp* aus. Ebenso positiv war die Resonanz durch den *Fakultätentag Informatik*[4] und den *Beirat der Universitätsprofessoren der Gesellschaft für Informatik (GIBU)*[5], denen die Zusammenarbeit 2010 vorgestellt wurde.

Schloss Dagstuhl hat deshalb im März 2010 beim Senatsausschuss Wettbewerb[6] der Leibniz-Gemeinschaft 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 vorsah.

Dank der Finanzierung war es möglich, ein Team unter der Koordination von Dr. Michael Ley an der Universität Trier zu formieren. Mit zunächst einer durch Dipl.-Inform. Oliver Hoffmann besetzten Stelle begann das Pro-

[1] http://www.pakt-fuer-forschung.de/
[2] http://www.leibniz-gemeinschaft.de/
[3] http://www.klaus-tschira-stiftung.de/
[4] http://www.ft-informatik.de/
[5] http://www.gibu.gi-ev.de/
[6] http://www.leibniz-gemeinschaft.de/ueber-uns/leibniz-wettbewerb/

jekt offiziell im Juni 2011 und wurde ab August 2011 durch Dr. Marcel R. Ackermann verstärkt. Die Datenerfassung und Qualitätskontrolle wird zudem seit Juli 2011 halbtags von Frau Stefanie von Keutz unterstützt.

Bereits vorab wurde das Vorhaben durch eine Spende in Höhe von 25.000 € von der Klaus Tschira Stiftung gefördert. Die Spende ging im November 2010 ein und diente der Finanzierung der Stelle von Herrn Oliver Hoffmann während der Projektvorbereitungsphase von November 2010 bis Mai 2011. Dadurch konnten Vorarbeiten für das Projekt bereits vor dem offiziellen Projektbeginn initiiert werden.

Dank einer weiteren Spende der Klaus Tschira Stiftung in Höhe von insgesamt 120.000 € konnte im Januar 2012 mit Dr. Michael Wagner eine weitere Vollzeitkraft zunächst für zwei Jahre für das Projekt gewonnen werden.

Seit März 2012 wird das *dblp*-Team bei der Administration der *dblp*-Server von Herrn Christopher Perrin (studentische Hilfskraft) unterstützt.

3 Projektergebnisse

3.1 Wissenschaftliche Aufsicht

Eine wesentliches Ziel des Projektes war die Formierung eines wissenschaftlichen Beirates unter dem Dach von Schloss Dagstuhl, welcher *dblp* in strategischen Entscheidungen hinsichtlich der strategischen Entwicklung berät und Rahmenrichtlinien vorgibt. Das *dblp* Advisory Board hat sich im November 2011 in Saarbrücken konstituiert. Dabei ist es gelungen, für das Board namhafte Persönlichkeiten aus verschiedenen Disziplinen der Informatik zu gewinnen. Dem *dblp* Advisory Board gehören an:

- Prof. Dr. Andreas Butz (LMU München, *Media Informatics and Human-Machine-Interaction*),
- Prof. Dr. Dietmar Saupe (Universität Konstanz, *Multimedia Signal Processing*),
- Prof. Dr. Hannah Bast (Universität Freiburg, *Algorithms and Data Structures*),
- Prof. Dr. Hans-Peter Lenhof (Universität des Saarlandes, *Bioinformatics*),
- Prof. Dr.-Ing. Jürgen Teich (Universität Erlangen-Nürnberg, *Hardware-Software-Codesign*),
- Prof. Dr. Mila Majster-Cederbaum (Universität Mannheim, *Complex Systems*),

- Prof. Oliver Günther, Ph.D. (Universität Potsdam, *Information Systems*),
- Prof. Dr. Dr. h.c. Otto Spaniol (RWTH Aachen, *Communication and Distributed Systems*),
- Prof. Dr. Dr. h.c. Reinhard Wilhelm (Universität des Saarlandes, *Programming Languages and Compiler Construction*),
- Prof. Dr.-Ing. Rüdiger Dillmann (Karlsruher Institut für Technologie, *Humanoids and Intelligence Systems*),
- Prof. Dr. Rüdiger Reischuk (Universität zu Lübeck, *Theoretical Computer Science*).

Die Mitglieder des *dblp* Advisory Boards sind zunächst auf 4 Jahre berufen. Sprecher des Boards ist Hannah Bast. Zu den Aufgaben des *dblp* Advisory Boards gehören insbesondere:

- Festlegung von Richtlinien, Qualitäts- und Prioritätsvorgaben zur inhaltlichen Ausrichtung von *dblp*,
- wissenschaftliche Aufsicht über die inhaltliche Qualität des Datenbestandes,
- ideelle Förderung von *dblp* durch das Einbringen von Fachkompetenz und
- Repräsentation von *dblp* gegenüber der wissenschaftlichen Gemeinschaft.

Nach seiner Konstituierung im November 2011 hat das *dblp* Advisory Board 2012 seine Arbeit voll aufgenommen. Dabei standen insbesondere zunächst zwei Themen im Mittelpunkt: Die Vorgabe von Rahmenrichtlinien für den Aufbau einer bibliometrischen Infrastruktur (siehe Abschnitt 3.8) sowie die Erarbeitung eines Kataloges an Mindeststandards für die Indizierung von Publikationsreihen in *dblp* (siehe Abschnitt 3.4). Zudem hat der *dblp* Beirat das *dblp*-Team bei der Erweiterung von *dblp* beraten und durch das Einbringen von Fachwissen unterstützt. So konnte auf Anregung des Advisory Board die Abdeckung gerade in multidisziplinären Teilgebieten wie Wirtschaftsinformatik, Bioinformatik und weiteren Gebieten verbessert und ausgeweitet werden.

3.2 Weiterentwicklung der Daten-Wrapper

Die Literaturdatenbank *dblp* verzeichnet wissenschaftliche Literatur in der Informatik auf der Ebene einzelner Beiträge. Sie konzentriert sich bei der Wahl der Datenquellen auf Verlage und Bibliotheken und verfolgt das Ziel, ganze Reihen und Serien und damit Themengebiete wissenschaftlicher Literatur möglichst vollständig zu erfassen.

Über Jahre hinweg wurden diese Daten stets mit hohem manuellen Anteil erfasst. Dieses Vorgehen führte zu einer anerkannt hohen Qualität des Da-

tenbestandes, stellt aber einen hohen Arbeitsaufwand dar und kann der sich stetig ausweitenden Publikationslandschaft nicht adäquat Rechnung tragen. Bereits in der Projektvorbereitungsphase wurde daher eine spezialisierte Software, so genannte *Wrapper*, entwickelt, die bibliographische Rohdaten 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 zur Indexierung vorbereitet.

Diese Software wurde im Rahmen des Projektes stetig erweitert und ausgebaut. Derzeit (Stand Ende Mai 2013) kommen Wrapper bei bereits 111 verschiedenen digitalen Bibliotheken und Verlagssystemen zum Einsatz und stellen heute einen integralen Teil des Arbeitsablaufes bei der Neuaufnahme von Daten dar. Die abschließende Qualitätskontrolle und Fehlerbeseitigung hingegen wird zwar von Hilfssoftware unterstützt; die Endabnahme erfolgt jedoch auch weiterhin bewusst von Hand. Eine Übersicht über alle zum formalen Projektende verfügbaren Wrapper kann Tabelle 1 entnommen werden.

In einem gemeinsamen DFG-geförderten Projekt[7] von *dblp* und *GESIS – Leibniz-Institut für Sozialwissenschaften*[8] wird zudem seit Dezember 2012 die Schaffung von Werkzeugen und Verfahren für das wissenschaftliche Informationsmanagement erforscht. Das Projekt baut dabei fundamental auf den im SAW-Projekt geschaffenen Wrappern auf. Ein Ziel ist es, die bisher rein regelbasierte Wrapper-Software durch generische, adaptive oder lernende Verfahren zu erweitern, um diese noch breiter einsetzen zu können. Von solchen „smart harvesting“ Verfahren wird *dblp* in Zukunft in besonderem Maße profitieren.

3.3 Produktivität und Aktualität der Datenakquise

Bei Beginn der Zusammenarbeit von Schloss Dagstuhl und *dblp*, Anfang November 2010, indexierte *dblp* etwa 1,49 Millionen Publikationen. Das Aufnahmevolumen in den Jahren davor Betrug dabei zwischen 100.000 und 150.000 Publikationen pro Jahr. Demgegenüber ist *dblp* zum offiziellen Projektende im Mai 2013 um etwa 53% gewachsen und indexiert nun über 2.3 Millionen Publikationen. Dabei konnte bereits im ersten Projektjahr ein Aufnahmevolumen von nun über 300.000 neuen Publikationen pro Jahr etabliert und verstetigt werden. Das anvisierte Projektziel von etwa 200.000

[7] *„Smart Harvesting: Verbesserung des Open Access-Zugangs; Erhöhung der Qualität der Metadaten“*, DFG LIS-Förderprogramm „Werkzeuge und Verfahren des wissenschaftlichen Informationsmanagements“, Fördernummer WA 1267/2-1

[8] http://www.gesis.org/

- Academy Publisher
- ACL Anthology
- ACTA Press
- Advanced Institute of Convergence IT
- American Institute of Mathematical Sciences
- American Mathematical Society
- Ars Combinatoria
- Association for Computational Linguistics and Chinese Language Processing
- Association for Computing Machinery
- Association for Information Systems
- Association for the Advancement of Artificial Intelligence
- Association of College & Research Libraries
- Atypon
- BCS, The Chartered Institute for IT
- BioMed Central
- BMJ Group
- Cambridge University Press
- Central Europe - Workshop Proceedings
- Conferences in Research and Practice in Information Technology
- CRC Press
- Dagstuhl Publishing
- De Gruyter
- Digital Information Research Foundation
- D-Lib Magazine
- DLINE Journals Portal
- doiSerbia
- dpunkt.verlag
- Edinburgh University Press
- Ed/ITLib
- Electronic Colloquium on Computational Complexity
- Electronic Proceedings in Theoretical Computer Science
- Elsevier
- Emerald
- Emis (LNI der GI)
- Eurasip Journals
- Eurographics
- European Conference on Information Systems
- European Symposium on Artificial Neural Networks
- European Union Digital Library
- Experimental Wrapper for Conferences
- Hermes Science Publications
- Hindawi Publishing Corporation
- IBM Technical Journals
- IEEE Xplore
- IEEE Communications Society
- IEEE Computer Society
- IGI Global
- Inderscience
- Information Research
- Ingenta
- Institute for Computer Sciences, Social Informatics and Telecommunications Engineering
- Institute for Operations Research and the Management Sciences
- Institute of Electronics, Information and Communication Engineers
- International Association for Cryptologic Research
- International Association of Computer Science in Sport
- International Journal of Interactive Multimedia and Artificial Intelligence
- IOS Press
- ISCA-Speech
- Journal of Artificial Intelligence Research
- Journal of Communications
- Journal of Information Processing Systems
- Journal of Management Information Systems
- Journal of Research and Practice in Information Technology
- Journal of Telecommunications and High Technology Law
- Journal of the Artivicial Societies and Social Simulation
- Journal of the Association for Information Systems
- Journal of Universal Computer Science
- Journal STORage
- KOREA Science
- Liebert Open Access
- Linkoeping Electronic Conference Proceedings
- Machine Translation Archive
- Massachusetts Institute of Technology
- Mensch und Computer
- MetaPress
- MIS Quarterly
- Molecular Diversity Preservation International
- NIPS
- NOW Publishers
- Old City Publishing
- Oldenbourg Verlagsgruppe
- Open Journal System
- Oxford University Press
- Palgrave Macmillan
- PLOS
- Project Euclid
- Project Muse
- Rinton Press
- Robotics: Science and Systems Conference
- SAGE Publications
- Scholarpedia
- Slovak Academy of Sciences
- Society for Industrial and Applied Mathematics
- Softwaretechnik-Trends
- Sourceforge
- Springer science+business media
- Taylor & Francis
- The Electronic Journal of Combinatorics
- The Journal of Object Technology
- The Prague Bulletin of Mathematical Linguistics
- TIIS Wrapper
- TinyTocs Journal
- TREC Conference
- UbiCC.org
- Universite de Provence
- University of Pittsburgh
- VDE-Verlag
- Versita
- Webology
- Wiley
- World Scientific

Tabelle 1: Wrapper-Software. Die Tabelle gibt eine Übersicht über die Verlage und digitalen Bibliotheken, für welche zum Projektende Daten-Wrapper zur verfügung stehen (Stand: Mai 2013).

Neuaufnahmen pro Jahr wurde demnach deutlich übertroffen. Die Produktivitätsentwicklung von *dblp* ist in Abbildung 1 dargestellt.

Die Literaturdatenbank *dblp* soll neue Literatur möglichst schnell und vollständig erfassen, um die große Nachfrage insbesondere bei der Recherche von aktuellen Publikationen befriedigen zu können. Ein Maß für die Aktualität einer Veröffentlichung ist das Alter bei ihrer Erfassung, d.h., 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. Tabelle 2 fasst diese Daten zusammen. Dabei lassen sich im Wesentlichen zwei Beobachtungen machen. Zum einen ist die Anzahl an Publikationen, die binnen eines Jahres in *dblp* aufgenommen werden, signifikant gestiegen, während gleichzeitig die Anzahl der im zweiten oder dritten Jahr aufgenommenen Publikationen gesunken ist. So wurden zwischen Mai und Oktober 2012 etwa 55% mehr Publikationen binnen eines Jahres indexiert als noch zwei Jahre zuvor. Dies legt eine deutliche Steigerung der Aktualität der in *dblp* verfügbaren Daten nahe. Des Weiteren ist zu erkennen, dass die Aufnahme von Publikationen nach mehr als 3 Jahren ebenfalls sehr stark angestiegen ist. Dies spiegelt die erfolgreichen Bemühungen des *dblp*-Teams wieder, gezielt fehlende, ältere Ausgaben von wichtigen Publikationsreihen zu vervollständigen (vgl. Abschnitt 3.5).

3.4 Neuaufnahme von Publikationsreihen

Trotz der hohen Reputation von *dblp* erwies sich in einer Studie von 2010 [2] sowohl die thematische Breite in den Randgebieten der Informatik als auch die Abdeckung einzelner Themengebiete der Kern-Informatik noch als unzureichend. In der Vergangenheit führten vor allem fehlende Ressourcen dazu, dass viele Publikationsreihen nicht berücksichtigt werden konnten. Als Auswahlkriterien zur Aufnahme neuer Reihen wurden lange Zeit vor allem die eigene Erfahrung sowie persönliche Kontakte herangezogen. Diesen Prozess galt es, auf ein solideres Fundament zu stellen.

Da *dblp* bisher noch keine eigene Infrastruktur zur Bewertung von Publikationsreihen etabliert hat, wurde hierzu verstärkt externe Sachkompetenz herangezogen. Zum einen konnte mit dem neu konstituierten *dblp* Advisory Board erstmals eine Runde von Experten aus verschiedenen Disziplinen der Informatik direkt ihre Expertise in den Auswahlprozess einbringen. Mitglieder des Boards beraten das *dblp*-Team bei Entscheidungsfragen und weisen zudem aktiv auf Fehlstände im Datenbestand hin. Das *dblp* Advisory Board hat dabei einen Katalog an Mindeststandards definiert, die i.d.R. für die Aufnahme in *dblp* vorausgesetzt werden sollen:

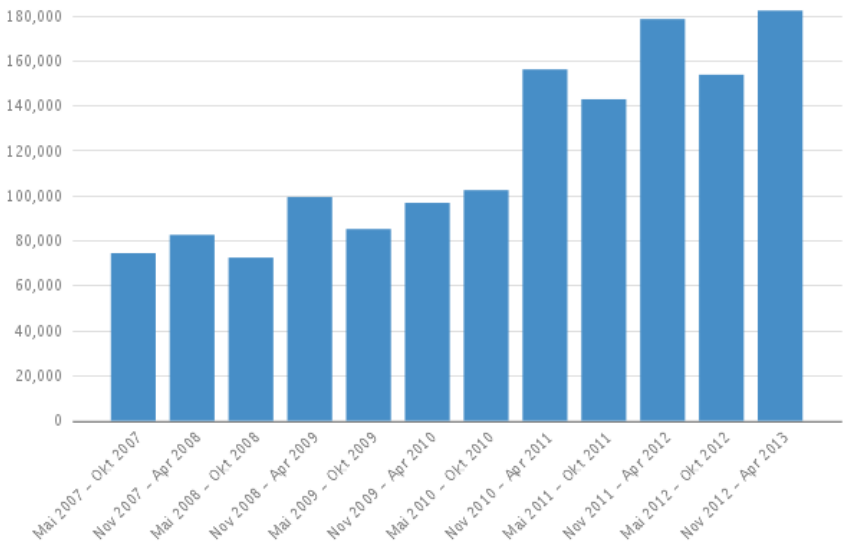

Abbildung 1: Produktivitätsentwicklung *dblp*. Die Grafik gibt eine Übersicht über die Anzahl neu aufgenommener Publikationseinträge in den Sechsmonatsintervallen vor und seit Projektbeginn (November 2010). Der nur scheinbare Produktivitätsrückgang in den Sommermonaten ist durch saisonale Schwankungen im Publikationskalender und Urlaubszeiten begründet.

Anzahl neuer Publikationen in *dblp*														
Nov. 2009 bis Apr. 2010		Mai 2010 bis Okt. 2010		Nov. 2010 bis Apr. 2011		Mai 2011 bis Okt. 2011		Nov. 2011 bis Apr. 2012		Mai 2012 bis Okt. 2012		Nov. 2012 bis Apr. 2013		
Anzahl	Jahr	Anzahl	Jahr	Anzahl	Jahr	Anzahl	Jahr	Anzahl	Jahr	Anzahl	Jahr	Anzahl	Jahr	Alter
0	2011	544	2011	1	2012	704	2012	0	2013	1097	2013	0	2014	zukünftig
16753	2010	59079	2010	26154	2011	87869	2011	38078	2012	93146	2012	37014	2013	aktuell
55830	2009	25333	2009	68647	2010	21557	2010	68107	2011	12143	2011	66766	2012	-1 Jahr
9564	2008	6211	2008	18441	2009	8733	2009	15181	2010	6363	2010	11188	2011	-2 Jahre
3269	2007	2979	2007	6911	2008	5064	2008	9512	2009	5460	2009	5872	2010	-3 Jahre
11483	älter	8422	älter	36114	älter	19064	älter	47765	älter	35690	älter	61576	älter	älter
96899	total	102568	total	156268	total	142991	total	178643	total	153894	total	182416	total	

Tabelle 2: Aktualität von Neuaufnahmen *dblp*. Die Tabelle gibt eine Übersicht über die neu aufgenommener Publikationseinträge, aufgeschlüsselt nach ihrem jeweiligen offiziellen Publikationsjahr. Es ist zu beachten, dass auf Grund der *online first* Strategie einiger Verleger in *dblp* auch Publikationen gelistet sind, deren Publikationsdatum erst in einem kommenden Jahr liegt.

- Aspekte der Publikationsreihe
  - Die Reihe soll vorwiegend Themen der Informatik behandeln.
  - Die Reihe soll eine klar definierte thematische Ausrichtung innerhalb der Informatik besitzen.
  - Die Reihe soll etabliert sein und regelmäßig erscheinen.
  - Die Reihe soll durch einen etablierten Herausgeber oder eine etablierte Fachgesellschaft unterstützt werden.
- Aspekte der Editoren und Autoren
  - Das Editorial Board bzw. die Steuerungs- und Programmkomitees sollen mit innerhalb des thematischen Schwerpunktes herausragenden Wissenschaftlern besetzt sein.
  - Unter den Autoren sollen sich herausragende Wissenschaftler des thematischen Schwerpunktgebietes befinden.
  - Der Autorenkreis der Reihe soll international sein.
- Aspekte der Beiträge
  - Beiträge der Reihe sollen wissenschaftliche Originalarbeiten sein, die nicht bereits zuvor in anderen Reihen publiziert wurden.
  - Beiträge der Reihe müssen in einem seriösen Peer-Reviewing-Verfahren überprüft werden.
  - Beiträge der Reihe sollen eine adäquate typographische Aufbereitung und Struktur besitzen.
  - Beiträge der Reihe sollen den gängigen Konventionen im Bezug auf Quellenangaben, Literaturverweise und der Darstellung des Standes der Forschung folgen.
  - Beiträge der Reihe sollen sich an eine internationale Leserschaft richten.
- Aspekte der Daten-Zugänglichkeit
  - Metadaten und Zusammenfassungen aller Beiträge sollen frei zugänglich sein.
  - Volltexte aller Beiträge sollen frei oder als bezahlte Dienstleistung zugänglich sein.
  - Elektronische Versionen aller Beiträge der Reihe sollen in einer digitalen Bibliothek im Web derart aufbereitet sein, so dass diese auch noch in Jahrzehnten verfügbar sind.
  - Alle Beiträge sollen einen eindeutigen und dauerhaften elektronischen Identifikator (z.B. DOI) besitzen.
  - Digitalen Bibliotheken sollten die Metadaten der Ausgaben und Beiträge vollständig und fehlerfrei anbieten und das elektronische Auslesen und Sammeln von Metadaten erlauben.

Darüber hinaus konnten in vielen Einzelgesprächen mit Teilnehmern von Dagstuhl-Seminaren Einschätzungen und Empfehlungen von anerkannten Fachexperten gewonnen werden. In diesem Austausch offenbarte sich ein

weiterer, großer Vorteil der Kooperation von *dblp* und Schloss Dagstuhl. Zudem wurden als weiteres, regelmäßiges Werkzeug eine Befragung unter allen Seminarteilnehmern mittels eines Fragebogens etabliert. Die Auswertung der Fragebögen liefert wertvolle Hinweise die helfen, Fehlstände im *dblp*-Datenbestand auszumerzen. Der Fragebogen hat zudem als Erprobung einer allgemeinen Befragung unter Informatik-Experten fungiert, wie sie seit Juni 2013 in größerem Umfang stattfindet (siehe Abschnitt 3.8).

Neben der Integration von Expertenmeinungen in den Auswahlprozess haben sich Entscheidungen zudem an anerkannten Informatik-Rankings orientiert. Hier sind vor allem das CORE-Ranking[9] der *Computing Research and Education Association of Australasia* sowie das CAPES-Ranking[10] des brasilianischen Forschungsministeriums hervorzuheben. Beide Rankings bieten — wenngleich auch regional geprägt — eine wertvolle Hilfestellung bei der Suche nach wichtigen Publikationsreihen. Um eine angemessene Datengrundlage aus Sicht der deutschen Informatik zu erhalten, wurde zudem in Zusammenarbeit mit der Gesellschaft für Informatik e.V. und dem Fakultätentag Informatik ein eigenes Rating aus Sicht der deutschen Informatik-Experten auf den Weg gebracht (siehe Abschnitt 3.8).

Somit konnten seit Projektbeginn über 900 neue Reihen identifiziert und aufgenommen werden. Eine detaillierte Aufschlüsselung der neu aufgenommenen Reihen kann Tabelle 3 entnommen werden.

3.5 Vollständigkeit bestehender Reihen

Neben der Aufnahme neuer Publikationsreihen liegt ein weiterer Schwerpunkt auf der Vervollständigung bereits indexierter Zeitschriften und Konferenzserien. Hier wurden im Verlaufe des Projekts verstärkt Anstrengungen unternommen, insbesondere die wissenschaftlich bedeutendsten Reihen lückenlos in *dblp* zu erfassen.

Als Auswahlkriterien wurden dabei ebenfalls die etablierten CORE- und CAPES-Rankings sowie die Eingaben des *dblp*-Beirats und der Besucher der Dastuhl-Seminare herangezogen. Ziel der Bemühungen war dabei insbesondere, die Reihen der Spitzenkategorien vollständig, sowie hochklassige Reihen zumindest bis in das Jahr 2000 vollständig zu ergänzen. Hier konnten inzwischen große Lücken in der Abdeckung geschlossen werden. Die Vervollständigung aller prestigeträchtigen Reihen ist dabei ein Prozess, der noch immer andauert.

[9] `http://core.edu.au/`

[10] `http://qualis.capes.gov.br/webqualis/`

Anzahl Publikationen/Reihen in *dblp*						
	Journale			Konferenzen/Workshops		
	Reihen	Ausgaben	Artikel	Reihen	Bände	Papiere
Okt. 2010	852	14031	577743	2877	16184	877320
Apr. 2011	1044	15874	664129	2950	16939	937194
Okt. 2011	1091	16591	722161	3094	18138	1019094
Apr. 2012	1179	18139	806739	3204	19311	1100778
Okt. 2012	1213	19345	879053	3279	20464	1174361
Apr. 2013	1275	21671	989474	3363	21675	1284735
Zuwachs	+50%	+54%	+71%	+16%	+34%	+46%

Tabelle 3: Journale und Konferenz-/Workshopreihen in *dblp*. Die Tabelle beschreibt die Trendentwicklung bei der Anzahl von Publikationsreihen, Ausgaben und Artikeln in *dblp*. Es ist zu beachten, dass die Aufteilung in Reihen und Bände insbesondere bei Konferenzen und Workshops nicht immer eindeutig ist. Die vorliegende Zählung betrachtet daher die Aufteilung in Übersichtsseiten (Reihen), Inhaltsverzeichnisse (Ausgaben/Bände) und bibliographische Einträge (Artikel/Papiere) wie sie auf der *dblp*-Webseite abgebildet ist, und spiegelt damit die allgemeine Tendenz wieder.

3.6 Neuaufnahme von Monographien

Ein weiteres Ziel des Projektes ist die Verbreiterung der Datenbasis im Bereich von Monographien und Sammelbänden. Hierbei wurde bereit der in Projektvorbereitungsphase ab November 2010 eine Zusammenarbeit mit der Informatik-Fachbibliothek von Schloss Dagstuhl initiiert. Nachdem im Zeitraum vor offiziellem Projektbeginn zunächst die Integration des Ist-Bestandes der Bibliothek im Vordergrund stand, wurde seitdem ein regelmäßiger Datenaustausch zwischen Bibliothek und *dblp* installiert, bei dem die monatlichen Bibliothek-Neuzugänge weitgehend in *dblp* integriert werden. Diese Zusammenarbeit wurde im Laufe des Projektes gepflegt und intensiviert. Insgesamt wurden so bis zum Mai 2013 über 8,100 neue Bände in den Datenbestand übernommen. Die konkrete Entwicklung der Indexierung von Büchern und Sammelbänden kann Tabelle 4 entnommen werden.

Ähnlich zum Fall der Journale und Konferenzen stellt die Aufnahme von Fachbüchern und Sammelbänden die Frage nach den Kriterien für die Auswahl geeigneter Bände. Eine Aufnahme ganzer Bibliotheks- oder Verlagskataloge scheint nicht sinnvoll, da auch hier Bedeutung und wissenschaftliche Qualität sehr inhomogen sind. Zudem erfordert die Bearbeitung bei der Neuaufnahme im Vergleich zu Journalen und Konferenzen einen überproportional hohen Arbeitsaufwand des *dblp*-Teams bei relativ geringer Anzahl aufzunehmender Datensätze. Nichts desto trotz hat sich das *dblp* Advisory

Anzahl Monographien in *dblp*		
	Bücher und Sammelbände	Dissertationen
Okt. 2010	1675	91
Apr. 2011	7484	808
Okt. 2011	8851	829
Apr. 2012	9263	6898
Okt. 2012	9502	6909
Apr. 2013	9841	6927
Zuwachs	+588%	+7612%

Tabelle 4: Monographien in *dblp*. Die Tabelle fasst die Anzahl in *dblp* indexierter Monographien im Sechsmonatsabstand zusammen. Der Sprung von Oktober 2010 zu April 2011 geht auf die Integration des Ist-Bestandes der Informatik-Fachbibliothek von Schloss Dagstuhl zurück. Der Anstieg bei den Dissertationen zum April 2012 wurde durch das gezielte Sammeln von Metadaten über die Webseiten deutscher Hochschulen und der Deutschen Nationalbibliothek erreicht.

Board für eine Ausweitung der Aufnahmebemühungen ausgesprochen. Die genaue Ausgestaltung der Arbeitsprozesse und Qualitätskriterien sind jedoch zum formalen Ende des Projektes ein offenes Problem, dem sich *dblp* noch stellen muss.

Darüber hinaus bildet die Neuaufnahme von Dissertationen einen weiteren Schwerpunkt. Dank der Freigabe der Katalogdaten der *Deutschen Nationalbibliothek*[11] als OpenData unter der Creative Commons Zero 1.0 Universal[12] Lizenz konnte *dblp* im Laufe des Projektes um über 6,800 Dissertationen von zumeist deutschen Hochschulen erweitert werden. Es bleibt ein Ziel von *dblp*, die Aufnahme von Dissertationen zu intensivieren und zudem auf den internationalen Bereich auszudehnen. Jedoch stellt zu Projektende insbesondere die Akquise von internationalen Dissertation noch immer ein offenes Problem dar.

3.7 Kontakte mit Verlagen als Datenlieferanten

Der Kontakt zu Verlagen und Fachgesellschaften ist notwendig, um über Neuerscheinungen informiert zu bleiben, sowie um möglichst zeitnah an bibliographische Metadaten zu gelangen. Im Idealfall erweisen sich Verlage

[11] http://www.dnb.de/

[12] http://creativecommons.org/publicdomain/zero/1.0/

dabei als kontinuierliche Datenlieferanten. Zu den neu gewonnenen Partnern von *dblp* gehören zum Zeitpunkt des formalen Projektendes folgende Partner:

- Bereits im Mai 2011 konnte mit *Springer Science+Business Media*[13] einer der für die Informatik wichtigsten Verleger als Partner und Datenlieferant gewonnen werden. Seitdem beliefert Springer *dblp* regelmäßig mit den Metadaten zu allen Neuerscheinungen. Die Daten werden tagesaktuell in den *dblp* Datenbestand eingepflegt.
- Seit Oktober 2011 ist es zudem gelungen, IEEE als regelmäßigen Datenlieferant für *dblp* zu gewinnen. *dblp* erhält wöchentlich alle Neuaufnahmen der digitalen Bibliothek *IEEE Xplore*[14] und hat zudem auch vollen Zugriff auf deren Altbestand.
- Im Dezember 2012 konnte ferner mit *IOS Press*[15] eine Vereinbarung über regelmäßige Datenlieferungen getroffen werden. Die Metadaten der Informatik-Sparte des Verlagsprogrammes werden seit dem über eine FTP-Schnittstelle bereit gestellt.
- Darüber hinaus befinden sich zum Zeitpunkt des formalen Projektendes Kooperationen mit der *Association for Computing Machinery (ACM)*[16] und der *USENIX Association*[17] in Verhandlung. Beide Partner haben sich bereits positiv über die Einrichtung regelmäßiger Datenlieferungen geäußert.
- Ferner werden auch die OpenAccess-Publikationen von Schloss Dagstuhl LZI[18] von *dblp* indexiert. Über eine eigens bereitgestellte Schnittstelle werden bereits seit 2011 die Metadaten an *dblp* übermittelt.

Die etablierten Partnerschaften erwiesen sich für *dblp* als äußerst vorteilhaft und trugen erheblich zur Steigerung von Produktivität und Aktualität der Neuaufnahmen bei (vgl. Abschnitte 3.3). Es ist auch in Zukunft beabsichtigt, Kontakte mit weiteren Großverlagen zu intensivieren. Aber auch für die Autoren der Verlage dürfte sich die verbesserte Indexierung als Vorteil erweisen: Nicht selten werden inzwischen Publikation in *dblp* indexiert, noch bevor sie in den digitalen Bibliotheken der Verlage gelistet werden.

3.8 Aufbau einer bibliometrischen Infrastruktur

Die hohe Datenqualität der Literaturdatenbank *dblp* ist das Ergebnis eines arbeitsintensiven Datenpflegeprozesses. Für ein ressourcenbeschränktes

[13] http://www.springerlink.com/
[14] http://ieeexplore.ieee.org/
[15] http://www.iospress.nl/
[16] http://dl.acm.org/
[17] https://www.usenix.org/
[18] http://drops.dagstuhl.de/

Projekt wie *dblp* ist es daher unvermeidlich, sich bei der Auswahl zu indexierender Reihen auf eine repräsentative Teilmenge der Zehntausende an existierenden Publikationsreihen zu beschränken. *dblp* verfolgt dabei den Ansatz, zentrale und wissenschaftlich hochwertige Reihen bei der Aufnahme zu bevorzugen. Ein wesentliches Ziel des Projektes war daher auch die Bildung einer bibliometrischen Infrastruktur, die mittelfristig eine möglichst objektive Auswahl zu indexierender Publikationsreihen ermöglicht.

Im ersten Projektjahr wurden in enger Absprache mit dem *dblp* Advisory Board sowie dem Fakultätentag Informatik die Anforderungen an eine solche Infrastruktur spezifiziert:

- Einheit der Evaluation sind Publikationsreihen, nicht Autoren oder Institutionen.
- Grundlage der Evaluation soll die Beurteilung durch die Forschenden in der Informatik sein, und kein starres, auf statistischen Daten basierendes Formelgerüst.
- Die Beurteilung soll auf der Basis von qualifizierten Mindeststandards erfolgen.
- Das Resultat der Evaluation soll kein Ranking oder numerisches Rating sein, sondern eine Klassifikation in eine geringe Anzahl grober Qualitätskategorien (z.B.: „Klasse A“ bis „Klasse D“).
- Die Beurteilung der Reihen soll relativ zum jeweilig einschlägigen Teilfeld der Informatik erfolgen.
- Die Ergebnisse und Grundlagen der Evaluation sollen frei zugänglich und für jedermann nachvollziehbar sein.
- Die Evaluation soll in periodischen Abständen wiederholt werden.

Das *dblp* Advisory Board sprach sich zu diesem Zweck für die Durchführung einer großflächigen Umfrage unter Informatik-Forschenden aus. Die erhobenen Daten sollen die Grundlagen für eine Klassifikation nach wissenschaftlichem Stellenwert und thematischer Ausrichtung bilden.

Als erster Schritt zur Erprobung einer solchen Erhebung wurde seit Anfang 2012 eine Serie von Umfragen unter den Seminarteilnehmern von Schloss Dagstuhl initiiert. Mittels der Umfragen konnten wertvolle Erfahrungen für eine Befragung im größeren Rahmen gesammelt werden. Dabei liefern die Dagstuhl-Umfragen schon jetzt wertvolle Hinweise auf Fehlstände im *dblp*-Datenbestand.

Mitte 2012 wurde zudem in Zusammenarbeit mit der Gesellschaft für Informatik e.V. (GI) und dem Fakultätentag Informatik mit der Planung der ersten deutschlandweiten Erhebung zum Stellenwert von nationalen und internationalen wissenschaftlichen Publikationsreihen in der Informatik begonnen. Ziel der geplanten Erhebung ist eine qualifizierte Übersicht über die

in der deutschen Informatik genutzten Publikationskanäle in Form von internationalen und nationalen wissenschaftlichen Konferenzen und Zeitschriften.

Im Juni 2013 wurde schließlich die Online-Befragung aller Mitglieder der Gesellschaft für Informatik gestartet. GI-Mitglieder wurden dazu mit einem anonymisierten Zugangscode persönlichen auf dem Postweg eingeladen. Inhaltlich ergründet die Befragung das persönlichen Publikationsverhalten der Befragten (Welche Publikationsreihen werden verwendet?) sowie deren Einschätzung der bekannten Reihen bezüglich verschiedener Qualitätsdimensionen (Einzigartigkeit, Originalität der Ergebnisse, wissenschaftliche Methodik, Qualität der Präsentation, Einfluss auf die eigene Forschungsarbeit, Bedeutung als Kommunikationsorgan) und ihrer thematischen Kategorisierung. Die Befragung soll mindestens bis Ende Juli andauern.

Zur Qualitätssicherung der Umfrage und zur Auswertung der Ergebnisse wurde ein Editorial Board, bestehend aus Vertretern der Kooperationspartner, etabliert:

- Prof. Dr. Hannah Bast (Sprecherin *dblp*-Beirat)
- Prof. Dr.-Ing. Peter Liggesmeyer (Vizepräsident Gesellschaft für Informatik)
- Prof. Dr. Rüdiger Reischuk (stellvertrender Vorsitzender Fakultätentag Informatik)
- Prof. Dr. Dr. h.c. Reinhard Wilhelm (Wissenschaftlicher Direktor Schloss Dagstuhl)
- Dr. Michael Wagner (Projekt „LZI+DBLP“, Koordination Umfrage)

Das Editorial Board wird zudem von Prof. Dr. Uwe Brinkschulte (Gesellschaft für Informatik, FB Technische Informatik), Prof. Dr. Oliver Deussen (Gesellschaft für Informatik, FB Graphische Datenverarbeitung), Prof. Dr. Gregor Engels (Gesellschaft für Informatik, FB SWT), Prof. Dr. Erhard Rahm (Gesellschaft für Informatik, FB DBIS) und Prof. Dr.-Ing. Dr. h.c. Manfred Nagl (Fakultätentag Informatik) und Alexander Rabe (Leiter Hauptstadtbüro GI) beraten und unterstützt.

Mit den Ergebnissen und deren Veröffentlichung ist Ende 2013 zu rechnen.

3.9 Server-Infrastruktur des Webdienstes

Unter der URL `dblp.dagstuhl.de` wurde Mitte 2012 ein neuer, leistungsfähiger *dblp*-Server in Schloss Dagstuhl in Betrieb genommen. Bei der Administration arbeitet das *dblp*-Team mit der IT-Abteilung von Schloss Dagstuhl unter der Leitung von Dipl.-Inform. Thomas Schillo zusammen.

Der neue Server fungiert derzeit noch als Mirror der beiden bestehenden *dblp*-Server in Trier, wird aber in naher Zukunft unter der Domain `dblp.org` zum zentralen *dblp*-Server ausgebaut und in den gängigen Suchmaschinen etabliert werden. Dazu wurde im Juni 2013 auch formal der Ankauf der Domain `dblp.org`, die derzeit noch treuhänderisch von Prof. Hannah Bast verwaltet wird, in die Wege geleitet.

3.10 Anbindung an das semantische Web als OpenData

OpenSource, OpenAccess und OpenData haben sich in den vergangenen Jahren als wegweisende Konzepte der Wissensvermittlung im Web etabliert. Das semantische Web mit seinem Bedarf an Linked Open Data (LOD) stellt dabei eine der großen Zukunftschancen für eine freie Wissensgesellschaft dar. Obwohl *dblp* bereits seiner Gründung 1993 seine Daten frei im Web zur Verfügung stellt, gab es bisher keine technische Anbindung der *dblp* Stammdaten an die LOD-Cloud. Die existierenden *dblp*-Komponenten des semantischen Webs basieren viel mehr auf veralteten, statischen Exporten des *dblp*-Datensatzes, was zwangsläufig zu Inkonsistenzen gegenüber dem aktuellen, „lebendigen“ *dblp*-Datensatz zur Folge hat.

Entsprechend der Empfehlungen[19] der *Open Knowledge Foundation*[20] und der *Deutschen Initiative für Netzwerkinformationen*[21] zu offenen bibliographischen Daten wurde der komplette *dblp*-Datensatz daher seit November 2011 unter der Open Data Commons Attribution License v1.0[22] auch formal als OpenData veröffentlicht. Die ohnehin seit Jahren etablierte Nutzung des *dblp*-Datensatz steht somit der Allgemeinheit in Zukunft auch rechtssicher zur Verfügung.

Des weiteren wurde mit der Etablierung eines RDF/XML-basierten Datensicht die Anbindung an das semantische Web mit der Bereitstellung von Linked Open Data in die Wege geleitet. Zudem wurde die HTML-Sicht gemäß RDFa 1.1 annotiert. Mit Hilfe dieser neuen Infrastruktur können die *dblp*-Daten nun mit den Werkzeugen des semantischen Webs (z.B. SPARQL-Browsern) erschlossen werden. Die RDF-Schnittstelle von *dblp* befindet sich zu Projektende zwar noch in der internen Erprobungsphase, wird aber im dritten Quartal 2013 in den öffentlichen Betrieb übergehen.

[19] `http://openbiblio.net/principles/`
[20] `http://okfn.org/`
[21] `http://www.dini.de/ag/standards/`
[22] `http://opendatacommons.org/licenses/by/summary/`

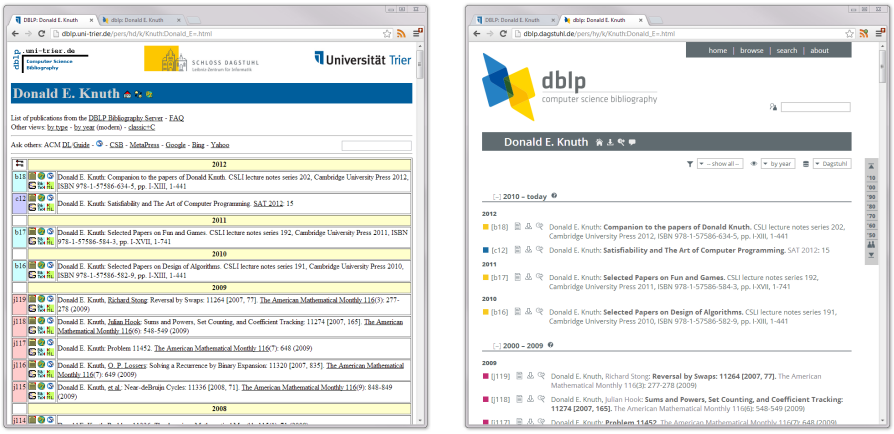

Abbildung 2: Das alte und das neue Layout des *dblp*-Webservice.

3.11 Nutzerorientierte Neugestaltung der Webseiten

Die *dblp*-Webseiten wurden 1993 als einfach Sammlung statischer HTML-Seiten geschaffen. Seitdem hat sich die Erscheinung des Webfrontends von *dblp* in Prinzip nur wenig gewandelt. Das rustikale Tabellenlayout wirkt heute veraltet und fällt im optischen Vergleich und in Sachen Benutzerführung gegenüber modernen Websystemen wie *Google Scholar*[23] oder *Microsoft Academic Search*[24] zurück.

Aus diesem Grund wurde das Webfrontend überarbeitet und im Sinne eines leichtgewichtigen, modernen Designs unter Beibehaltung der inhaltlichen Aussagekraft erneuert. Neben einer optischen Generalüberholung stand dabei insbesondere auch die nutzungsgerechte Aufbereitung der bibliographischen Daten im Vordergrund. So wird einer häufig geäußerten Kritik Rechnung getragen und neben der bisherigen chronologischen Sortierung der Autorenseiten auch eine kategorisierte Sortierung nach Publikationstyp angeboten. Die neuen Leistungsmerkmale umfassen die folgenden Punkte:

- Volle HTML5- und CSS2-Konformität.
- Neues *dblp*-Logo und ein klares und aufgeräumtes Layout. Das klassische Layout ist als Variante weiter verfügbar.
- Integration der bisher extern betriebenen CompleteSearch-Suchmaschine, inklusive erweiterter Such- und Sortieroptionen.
- Neue Kategorien-Sicht auf Autorenseiten als Standard; die herkömmliche chronologische Sicht ist weiterhin verfügbar.
- Einfache Seitennavigation durch Sektionsmenüs, sowie ein- und aufklappbare Abschnitte.

[23] http://scholar.google.de/
[24] http://academic.research.microsoft.com/

- Neue Datenexport-Optionen, wie etwa einen RSS-Feed und eine neue XML-Sicht für Autorenseiten.
- Anbindung an das semantische Web als Linked Open Data mittels RDF/XML-API für Autoren- und Publikations-Entitäten.
- Neue Statistik-Seiten geben tagesaktuell Auskunft über die Entwicklung von *dblp*.
- Die F.A.Q.-Seiten wurden optisch und inhaltlich überarbeitet. Kontextuell relevante Fragen sind nun auch direkt aus jeder Webseitensicht heraus erreichbar.

Des Weiteren ist bei der Restrukturierung auf einen modularen Aufbau geachtet worden, der in Zukunft die Erweiterung um zusätzliche Komponenten und Verlinkungen, sowie die Anpassung an unterschiedlichste Ausgabegeräte (z.B. Drucker oder Mobilgeräte) erleichtern wird.

Seit Anfang 2013 hat das neue Layout die alten Seiten abgelöst und befindet sich nun im öffentlichen Testbetrieb.

3.12 Verstetigung der Zusammenarbeit von LZI und *dblp*

Die Zusammenarbeit von Schloss Dagstuhl LZI und *dblp* fand von Anfang an in der internationalen Wissenschaftsgemeinschaft großen Anklang. In Gesprächen mit Seminarteilnehmern auf Schloss Dagstuhl wurde uns immer wieder positiver Zuspruch zugetragen. Sowohl Schloss Dagstuhl LZI als auch *dblp* gelten in der internationalen Informatikforschung als zwei wichtige Institutionen mit tadellosem Ruf, die sich zudem hervorragend ergänzen.

Wie die Projektergebnisse berichten, konnte dank der personellen Verstärkung des *dblp*-Teams die Literaturdatenbank sowohl inhaltlich weiterentwickelt als auch organisatorisch gestärkt werden. Dabei konnten vor allem bei der Weiterentwicklung der organisatorischen Struktur vielerlei Synergien genutzt werden. Zum Beispiel konnte beim Aufbau des *dblp* Advisory Boards von der hervorragenden Einbindung des LZI in die deutsche Informatikgemeinschaft profitiert werden. Die Gespräche und Befragung der Seminarteilnehmer von Schloss Dagstuhl haben zudem wertvolle Hinweise für die inhaltliche Ausgestaltung von *dblp* geben können. Umgekehrt ist *dblp* ein Werkzeug, das Organisatoren, Teilnehmer und den wissenschaftlichen Stab des LZI bei der Vorbereitung und Durchführung der Dagstuhl-Seminare unterstützt.

Zur Verstetigung der Zusammenarbeit über das formale Projektende hinaus wurden deshalb unter dem Dach von Schloss Dagstuhl LZI drei feste Mitarbeiterstellen für die Arbeit an *dblp* geschaffen. Diese Entwicklung wurde ebenfalls allenthalben positiv aufgenommen.

4 Fazit

Die Zusammenarbeit von Schloss Dagstuhl mit der Literaturdatenbank *dblp* hat bereits nach nur zweieinhalb Jahren substantielle Erfolge vorzuweisen. Die durch Personalstärke und technische Weiterentwicklung ermöglichte Produktivitätssteigerung hat schon jetzt die ursprünglichen Erwartungen übertroffen. Dank der Verstetigung der Zusammenarbeit von LZI und *dblp* besteht auch die Absicht, dieses Niveau in Zukunft zu halten und darauf aufzubauen.

Gleichzeitig wurde aber nicht nur die Quantität sondern auch die Vollständigkeit des Datenbestandes verbessert. Die Etablierung des *dblp* Advisory Board bedeutet einen wesentlichen Schritt vorwärts, um die wissenschaftliche Aufsicht über den *dblp* Datenbestand auf solide Beine zu stellen. Die für die Datenaufnahme nötige Priorisierung von Publikationsreihen hat durch die Einbeziehung von Experten und externer Rankings eine qualitative Verbesserung erfahren. Die Hinzunahme von externem Sachverstand hat zudem zu einer verbesserten Abdeckung gerade in den multidisziplinären Randgebieten der Informatik geführt. Außerdem wurden bestehende Reihen der Kerninformatik bezüglich ihrer Abdeckung in *dblp* geprüft und Fehlstände konsequent behoben.

Der Aufbau einer eigenen bibliometrischen Infrastruktur wurde in die Wege geleitet. Die Befragung der Seminarteilnehmer von Schloss Dagstuhl LZI liefert bereits regelmäßig wertvolle Hinweise auf übersehene oder unterschätzte Reihen. Mit der großflächigen Befragung unter den Mitgliedern der Gesellschaft für Informatik (GI) e.V. wird zudem in der zweiten Hälfte von 2013 eine wertvolle Datenbasis für die zukünftige Neuaufnahme und Priorisierung von Publikationsreihen zur Verfügung stehen.

Es ist zudem hervorzuheben, dass die Förderung durch die Klaus Tschira Stiftung ihren signifikanten Anteil an dem Erfolg des Projekts hat. Durch die Förderung der Projektvorbereitungsphase von November 2010 bis Mai 2011 konnten bereits erhebliche Teilziele im Bereich der technischen Infrastruktur vor dem offiziellen Projektbeginn realisiert werden. Zudem hat erst die Förderung der dritten Vollzeitstelle die Fortschritte insbesondere beim Aufbau der eigenen bibliometrischen Infrastruktur in dem erreichten Maße möglich gemacht.

5 Ausblick

Die Entwicklung von *dblp* ist mit dem Ende des Projektes „LZI+DBLP“ selbstverständlich noch nicht abgeschlossen. Neben der tagtäglichen Neu-

aufnahme von Publikationen und dem steten Streben nach einer möglichst vollständigen Abdeckung bleiben noch eine Vielzahl von konzeptionellen Herausforderungen bestehen.

So gilt die Gewinnung von weiteren Verlagen als Partner bei der Datenakquise unverändert als ein Ziel von *dblp*. Die etablierten Kooperationen haben sich als sehr förderlich für die Effizienz und Aktualität der Datenakquise erwiesen. Auf den im Rahmen des Projektes gemachten ersten Erfahrungen soll weiter aufgebaut werden.

Ein offenes Problem bleibt zudem die Behandlung von Monographien in *dblp*. Das Selektions- und Priorisierungsproblem von Monographien steht dem von Journalen und Konferenzreihen in nichts nach, jedoch sind die jeweiligen Arbeitsabläufe bei Monographien notwendigerweise um einiges individueller und deren Ertrag um Größenordnungen geringer als bei den periodisch erscheinenden Reihen. Die Bedeutung der Indexierung von Monographien ist jedoch unbestritten. Hier gilt es in Zukunft eine Infrastruktur zu finden, die eine effiziente Selektion und Akquise analog zu den für Journale und Konferenzen etablierten Arbeitsabläufen ermöglicht. Zudem sollen (sowohl nationale als auch internationale) Dissertationen in Zukunft eine wichtigere Rolle bei *dblp* spielen. Erste Erfahrungen wurden mit Hilfe von Daten der Deutschen Nationalbibliothek gemacht; auf diesen Erfahrungen soll aufgebaut werden.

Mit der deutschlandweiten Befragung der GI-Mitglieder wird in diesem Jahr ein großer, erster Schritt zur Bildung einer eigenen bibliometrischen Infrastruktur geleistet worden sein. Diese Entwicklung gilt es zu verstetigen und auszuweiten. Es ist bereits vorgesehen, die Erhebung im selben oder ähnlichen Expertenkreis mit etwas zeitlichem Abstand periodisch zu wiederholen. Die Ergebnisse der Evaluation sollen so kontinuierlich aktualisiert und weiter entwickelt werden. Aufbauend auf den gemachten Erfahrungen soll aber auch insbesondere eine Einbeziehung von zunächst europäischen (und später auch internationalen) Experten das Ziel sein.

Eine große Herausforderung der Literaturdatenbank ist und bleibt die Sicherung der Datenqualität. Ein besonderes Alleinstellungsmerkmal von *dblp* liegt in der hochqualitativen Identifikation von Autoren-Entitäten an Hand von in der Regel mehrdeutigen Metadaten. Die Routinen und Arbeitsabläufe, die diesem Merkmal zu Grunde liegen, haben jedoch ihre Grenzen und Eigenarten. So stellt insbesondere der asiatische Namensraum *dblp* noch vor große Herausforderungen. Eine Disambiguierung asiatischer Namen ist derzeit meist nur durch sehr arbeitsintensive manuelle Recherche möglich.

Es wird ein zentrales Ziel von *dblp* sein, mittelfristig leistungsfähigere und dem Stand der Forschung entsprechende Methoden zu installieren, die die Qualität der Autorenidentifikation noch weiter steigern und eine effizien-

te Disambiguierung unterstützen. Methoden des maschinellen Lernens und der Computerlinguistik haben in den vergangenen Jahren auf diesem Gebiet bereits einige Fortschritte verzeichnen können. Diese theoretischen Verfahren sind aber in der Regel nicht auf die Bedürfnisse und Besonderheiten eines Einsatzes in einer sich kontinuierlich verändernden und entwickelnden Datenbank ausgerichtet. Um diese Lücke zwischen Theorie und Praxis zu schließen, soll der Kontakt mit einschlägigen Experten verstärkt und das Disambiguierungsproblem mit forschungsnahen Methoden praktisch angegangen werden. Ferner gilt es auch, die Zusammenarbeit mit anderen bibliographischen Informationsdiensten und digitalen Bibliotheken zu intensivieren, um diese große, gemeinsame Herausforderung auch gemeinsam zu meistern.

Zudem gibt es neue Trends in der wissenschaftlichen Kommunikation, für welche die eher traditionelle Struktur von *dblp* derzeit noch nicht aufgestellt ist. So haben verschiedene Reihen in der Informatik damit begonnen, hybride Strukturen zwischen Journalen und Konferenzen zu bilden, die von der *dblp*-eigenen Typisierung nicht eindeutig erfasst werden. „Mega-Journale“ nach dem Modell von *PLOS One*[25] publizieren statt Ausgaben einen kontinuierlichen Strom von Einzelartikeln und entziehen sich zudem gängigen Fächerkategorien. Online-Repositories wie `arXiv.org` verstehen Publikationen als „lebendige“ Dokumente, die sich über die Zeit von Version zu Version wandeln können. Die *DataCite*[26] Initiative bemüht sich, Forschungsdaten verfügbar zu machen und mit zitierfähigen Metadaten zu versehen. Eine ähnliche Fragestellung zeigt sich auch bei dem bibliographischen Nachweis von Programm-Quellcode. In all diesen Bereichen gilt es, zunächst beobachtend die aktuellen Entwicklungen zu begleiten und sich für die Zukunft richtig zu positionieren.

Literatur

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

[2] Florian Reitz and Oliver Hoffmann. An analysis of the evolving coverage of computer science sub-fields in the dblp digital library. In *Research and Advanced Technology for Digital Libraries, 14th European Conference, ECDL 2010, Glasgow, UK, September 6-10, 2010. Proceedings*, pages 216–227, 2010.

[25] `http://www.plosone.org/`
[26] `http://www.datacite.org/`