Vorstellung des JOIN2 Statistikmoduls mit seinen Differenzen und Problemen zum kommenden Kerndatensatz Forschung

Robert Thiele[1,3], Katrin Große[2,3]

[1]Deutsches Elektronen-Synchrotron DESY
[2]GSI Helmholtzzentrum für Schwerionenforschung Darmstadt

[3]für JOIN²

JOIN² – Just anOther INvenio INstance

JOIN² ist eine kollaborativ entwickelte Repositorien-Infrastruktur basierend auf der INVENIO-Software vom CERN. Projekt-Partner sind die Bibliotheken des Deutschen Elektronensynchrotrons DESY (Hamburg/Zeuthen), des Deutschen Krebsforschungszentrums DKFZ (Heidelberg), des Forschungszentrums Jülich (Jülich), des GSI Helmholtzzentrums für Schwerionenforschung (Darmstadt), des Heinz Maier-Leibnitz-Zentrum MLZ (Garching) und der Rheinisch Westfälischen Technischen Hochschule RWTH (Aachen).

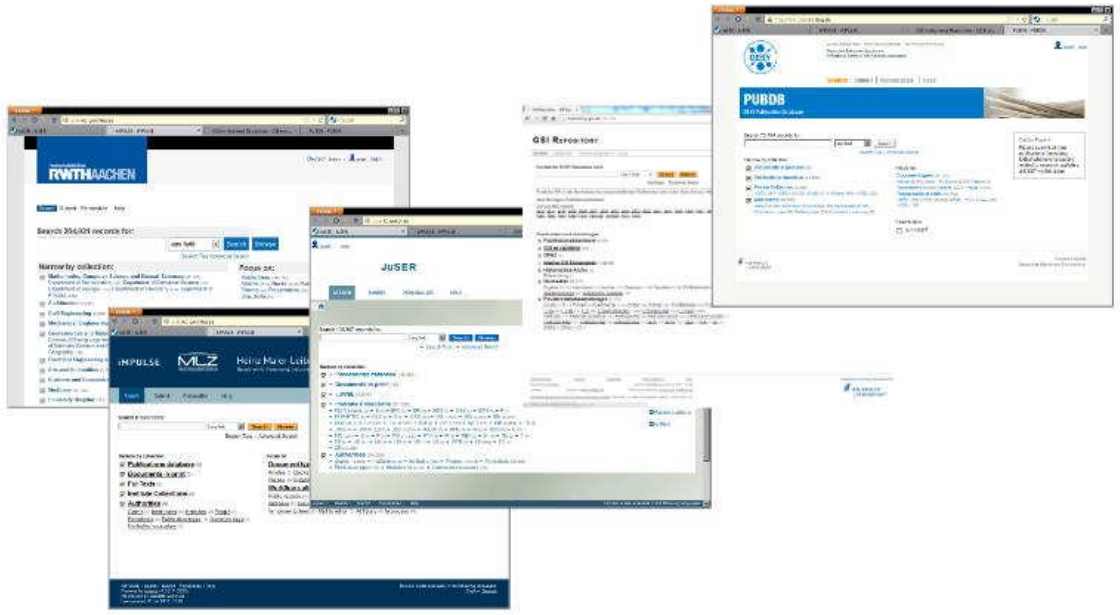

Abbildung 1: Startseite der verschiedenen JOIN² Instanzen

Derzeit laufen beim DESY, bei der GSI, im Forschungszentrum Jülich, dem MLZ, sowie an der RWTH die Systeme live mit derzeit knapp 170.000 gemeinsamen Normdatensätzen. Das DKFZ migriert gerade die Publikationsdaten vom alten System.

Das Deutsche Zentrum für Neurodegenerative Erkrankungen DZNE in Bonn evaluiert derzeit den Eintritt in unser Projekt. Abbildung 1 zeigt typische Startseiten von unseren Repositorien in den verschiedenen JOIN²-Instanzen. Durch die Normdatensätze können zum Beispiel EU-Grants von OpenAIRE <www.openaire.eu> eingesammelt werden. Im Rahmen von HORIZON2020, werden Publikationen, welche durch EU-Gelder gefördert werden, mit dem entsprechenden EU-Grant angereichert und dann automatisch zu OpenAIRE gepusht. Damit wird unseren WissenschaftlerInnen bei der Meldepflicht der Publikationen die Arbeit signifikant erleichtert.

Weitere gemeinsame Normdatensätze sind u.a. Experimente, über 65.000 Zeitschriften-Normdaten, sowie die komplette Programmorientierte Förderung (POF) der Helmholtz-Gemeinschaft (HGF) in allen drei Perioden.

HGFStatistic – Das Statistiktool für die HGF

Die Grundlage belastbarer Statistiken sind gute Datensätze. In JOIN² werden Normdatensätze für Zeitschriften, wie auch für Grants zentral verwaltet. Das macht es den Helmholtz-Zentren möglich, ein gemeinsames Tool zur Erfassung der in der Helmholtz-Gemeinschaft benötigten Kennzahlen, wie für den Zentrenfortschrittsbericht zu benutzen. Hierbei werden nach den Richtlinien der POF die erforderlichen Publikationszahlen nach ISI und Scopus entnommen. Im Rahmen der POFIII wurde die Frage, welche Publikationen (Journals) als referiert gelten dürfen, mit der Auflistung in der Thomson Reuter Master Journal List (TRMJL) sowie bei SCOPUS beantwortet. Um diese Information in die Datensätze zu bekommen, werden einmal im Jahr die Zeitschriften-Normdaten aktualisiert und mit den zugehörigen Statistikschlüsseln (TRMJL, SCOPUS, etc.) versehen, sodass beim Eintragen einer neuen Publikation und der Auswahl des Journals die entsprechenden Schlüssel abgespeichert werden können.

```
915__ $$0StatID:(DE-HGF)0100$$2StatID$$aJCR$$bNUCL INSTRUM METH A : 2014
915__ $$0StatID:(DE-HGF)0200$$2StatID$$aDBCoverage$$bSCOPUS
915__ $$0StatID:(DE-HGF)0300$$2StatID$$aDBCoverage$$bMedline
915__ $$0StatID:(DE-HGF)0310$$2StatID$$aDBCoverage$$bNCBI Molecular Biology Database
915__ $$0StatID:(DE-HGF)0199$$2StatID$$aDBCoverage$$bThomson Reuters Master Journal List
915__ $$0StatID:(DE-HGF)0110$$2StatID$$aWoS$$bScience Citation Index
915__ $$0StatID:(DE-HGF)0150$$2StatID$$aDBCoverage$$bWeb of Science Core Collection
915__ $$0StatID:(DE-HGF)0111$$2StatID$$aWoS$$bScience Citation Index Expanded
```

Abbildung 2: Beispiel der JOIN² Statistikschlüssel

Ein Beispiel dieser Schlüssel ist in Abbildung 2 zu sehen. Bei JOIN² werden alle Daten im MARC21-Standard abgespeichert. Die Qualität der Daten konnte zusätzlich durch den bei JOIN² angebotenen Digital-Object-Identifier-Import, kurz DOI-Import, erhöht werden, so dass in den Daten deutlich weniger Fehler zu finden sind, als noch vor ein paar Jahren. Zusätzlich wird die Eingabe eines Datensatzes einfacher und geht damit deutlich schneller, da viele notwendige Felder, wie Autor, Titel, Journal, etc. schon befüllt wurden.

In Abbildung 3 ist ein Teil unserer Masken direkt nach einem DOI-Import dargestellt, die Pflichtfelder sind jeweils rot mit Stern gekennzeichnet. Wie man sieht, werden nahezu alle Pflichtfelder befüllt. Die Autoren müssen nach dem Import noch bestätigt werden, damit IDs zugeordnet werden können, nach denen man später auch suchen kann.

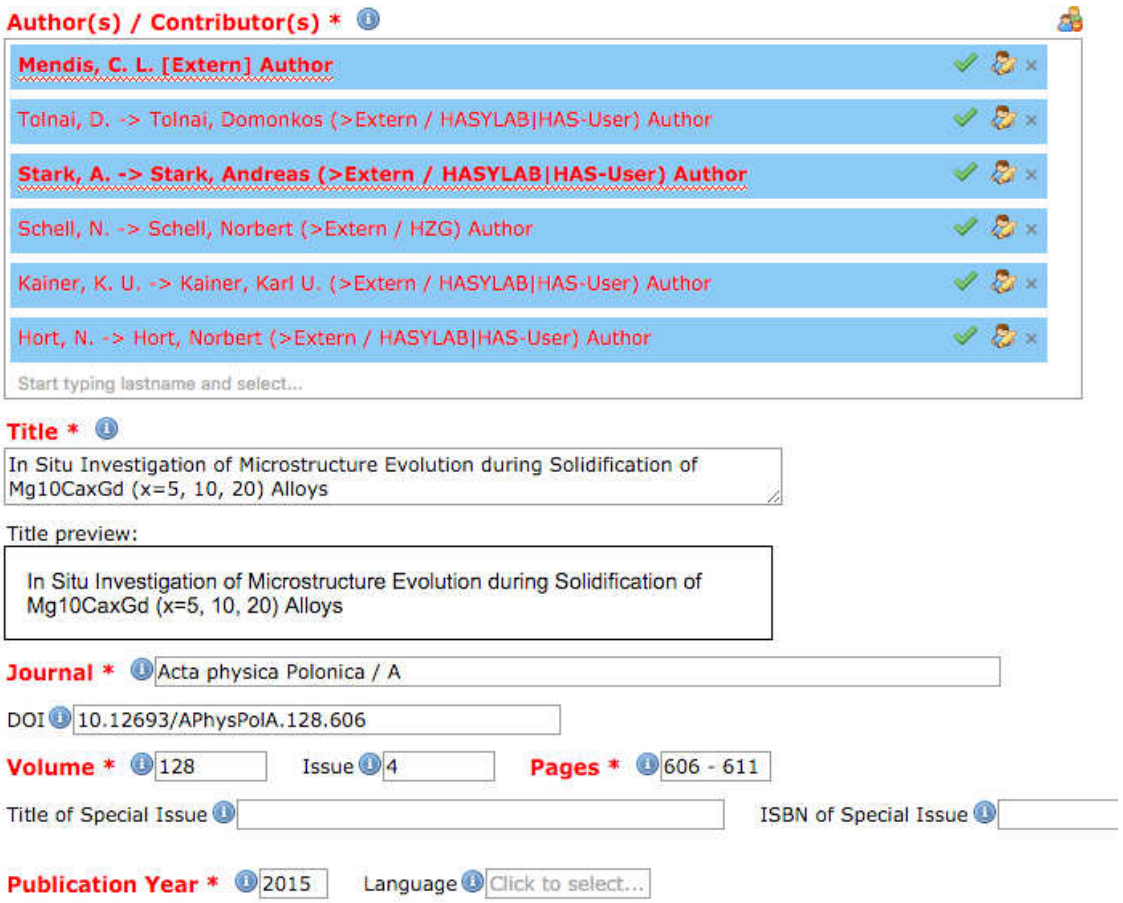

Abbildung 3: Typische Eingabemasken bei JOIN²

Nachdem eine Publikation durch den JOIN²-Workflow (USER-EDITOR-STAFF) gegangen ist und von der Bibliothek abschließend bestätigt wurde, ist er freigegeben, hat den Stempel VDB (Veröffentlichungsdatenbank) und erscheint auf unserer dynamischen Statistikseite.

Abbildung 4 zeigt einen Ausschnitt aus dem Gesamtüberblick der Statistik-Webseite. Die Webseite besteht aus einem Kopf mit allen Zahlen der Einrichtung, sortiert nach Statistikschlüssel, sowie nach den Publikationstypen. Danach kommen die POF-Zahlen, wiederum aufgeschlüsselt nach Statistikschlüssel sowie den Publikationstypen. Als drittes können die gleichen Tabellen auch auf Gruppen- bzw. Institutsebene erhoben werden.

865	records	ISI + Scopus (POF3)	(POF3)	EndNote BibTeX RIS
845	records	WOS listed journal OR entry (POF2)	(WOS_UT)	EndNote BibTeX RIS

Abbildung 4: Auszug aus HGFStatistic

Alle Zahlen sind anklickbar und man bekommt sofort die Suche in der Datenbank aufgelöst. Wenn die Treffermenge leer ist und eine Null in der Tabelle steht, wird nur die Zahl angezeigt, ohne Suche. Zusätzlich können die Suchen gegen die Datenbank in eines der drei Ausgabeformate (Endnote, BibTeX, RIS) gebracht werden, so dass man damit gängige Literaturverwaltungsprogramme füttern kann.

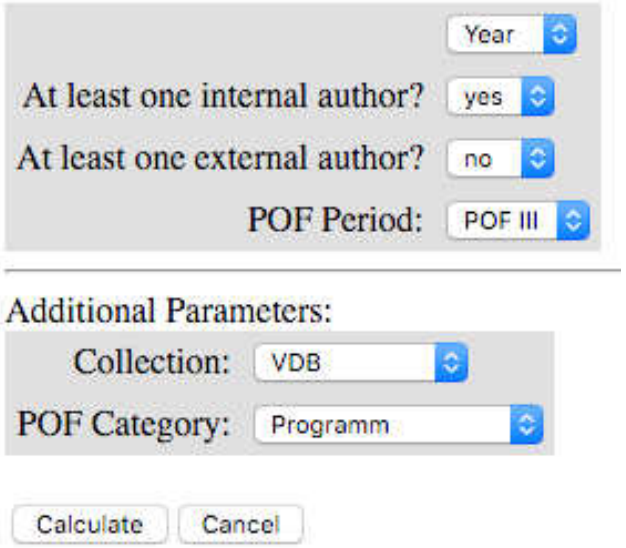

Abbildung 5: Web Frontend zur Auswahl der gewünschten Daten auf der Statistik-Webseite

Mit einem einfachen Web Frontend, wie in Abbildung 5 zu sehen, kann man die Parameter für die Berechnung und Anzeige einstellen. Hier kann man das Jahr, ob alle Autoren von der Anstalt kommen sollen oder nur externe betrachtet werden sollen, sowie die POF-Periode und die Kategorie eingestellt werden. Danach kann die Berechnung gestartet werden. Somit sind auch Fragen nach älteren Jahren, aber in der aktuellen POF möglich und werden beantwortet, solange die Datensätze die dafür benötigten Informationen tragen.

Differenzen und Probleme beim Übergang des bestehenden Evaluierungstools HGFStatistic zum Kerndatensatz Forschung

Nachdem gerade aufgezeigt wurde, welche Kennzahlen wir bereits regelmäßig ermitteln, gehen wir nun im zweiten Teil auf mögliche Realisierungen und Probleme im Hinblick auf den Kerndatensatz Forschung ein. Die „Empfehlungen zur Spezifikation des Kerndatensatz Forschung“ (Wissenschaftsrat, 2016), die auch die Spezifikationen des Kerndatensatzes in der Version 1.0 enthalten, hat der Wissenschaftsrat Anfang dieses Jahres zur Standardisierung der Forschungsberichtserstattung verabschiedet.

Bei unseren Überlegungen beschränken wir uns dabei vor allem auf den Objektbereich Publikationen, der aber auch zahlreiche Verknüpfungen zu allen anderen Objektbereichen aufweist (wie den Objekten Beschäftigte, Nachwuchsförderung, Drittmittel & Finanzen, Patente & Ausgründungen oder den Forschungsinfra-strukturen). Diese Verknüpfungen spiegeln bereits die ersten Schwierigkeiten, denen wir als Bibliotheken gegenüberstehen: Um die funktionierende und nicht gebrochene Verknüpfungen in den Daten vorhalten zu können, müssen wir Zentrums-intern an Standardisierungen und der Etablierung von Identifikatoren arbeiten sowie Mappings und Ontologien aufbauen und pflegen, die wir in der für den Kerndatensatz Forschung gebrauchten Weise noch nicht in hinreichender Form haben. Aus Datenschutzgründen müssen dabei auch Zwischen-Identifikatoren in Betracht gezogen werden, so enthalten die bibliothekarischen Personennorm-datensätze beispielsweise keinesfalls Personalnummern. Die Verknüpfung von internen IDs zu Personenidentifikatoren wie der ORCID, der Scopus-ID, der GND-ID oder auch Fachcommunity-spezifischen Identifikatoren wie der PMID-ID, der JaCOW-ID oder der INSPIRE-ID muss ebenfalls geleistet werden, was nicht ohne manuelle Nacharbeiten geschehen kann.

Auch müssen zur Integrität der Verknüpfungen neue Automatismen aufgebaut werden, um die Verknüpfung der Identifikatoren zu gewährleisten. Neben der Pflege der Verknüpfungen wird für uns als Bibliothek auch die Synchronisierung mit anderen internen Abteilungen deutlich mehr Zeit als bisher in Anspruch nehmen.

Wie Sie bereits gesehen haben, können wir mit unseren bisherigen Evaluierungsabfragen auch Publikationen ermitteln, an denen keine WissenschaftlerInnen unserer Einrichtungen beteiligt waren, die aber an unseren Forschungsinfrastrukturen entstanden sind. Damit kommen wir der Definition, was eine Publikation, Fi14 nach dem Kerndatensatz Forschung ist, bereits nach. Bei Strahlzeitbewilligungen werden die externen ForscherInnen aufgefordert, auf die genutzten Forschungsinfrastrukturen zu verweisen. DOIs für unsere Forschungsinfrastrukturen im Journal of large-scale research facilities (JLSRF) sind in Vorbereitung.

Neben den Publikationen in Listenform werden im Objektbereich von Publikationen die folgenden Ausdifferenzierungen gefordert, die wir nun im Hinblick auf unsere bisherigen JOIN^2-Evaluiserungroutinen darstellen und problematisieren wollen:

- Schöpfer Pu2
- Titel des Werkes Pu5
- Veröffentlichungsjahr Pu84
- Verlag Pu67
- Quelle Pu143 [meint u.a. Zeitschriftentitel, Buchtitel]
- Identifier Pu132
- Format Pu87
- Zugangsrechte Pu19

Bei etwa 90% unserer Veröffentlichungseinträge sind diese Ausdifferenzierungen in der Regel automatisiert über den bereits vorgestellten DOI-Import aus den Verlagsangaben oder anderen ID-Importen (wie PMID oder INSPIRE-ID) aus Datenbanken gefüllt. Bei etwa weiteren 8% der Einträge können wir die von den WissenschaftlerInnen eingetragenen Ausdifferenzierungen im Rahmen unserer Bibliotheksfreigabe recherchieren und ggfs. ändern. Bei den restlichen 2% müssen wir auf die Nutzerangaben vertrauen.

Selbst bei diesen „einfachen“ Angaben, die im Prinzip vollständig abdeckt sind, zeigen sich bei den verschiedenen Ausprägungen erste Probleme in der Praxis. Zum Beispiel beim Schöpfer: Die Eingabe von Herausgebern, von Körperschaften mit Herausgeberfunktionen oder von Gruppen von Herausgebern wird bei Eintragenden vermutlich auf wenig Resonanz stoßen. Und eine Nachrecherche solcher Angaben, insbesondere im Zeitschriftenbereich, werden wir als Bibliothek nicht leisten können. Bei Veröffentlichungen von mehr als tausend AutorInnnen ist es nicht möglich, die Vornamen aller AutorInnen nachzuerfassen; über die Normdatenverknüpfungen erhalten wir nur die Vornamen bei unseren eigenen AutorInnen.

Auch scheinbar einfache Identifikatoren wie ORCID oder das Publikationsjahr haben in der Praxis ihre Tücken. Immer wieder führen wir im Alltag mit den WissenschaftlerInnen Diskussionen, wann etwas als „erschienen“ gilt, weisen online-first aus und pflegen das Publikationsjahr nach. Die Verbreitung von ORCID ist in unseren Zentren gänzlich unterschiedlich und zeigt sich damit in der Praxis als unterschiedlich sinnvoll. Nur das Forschungszentrum Jülich strebt zurzeit von unseren Zentren an, für jeder Wissenschaftler/jede Wissenschaftlerin eine ORCID zu vergeben und unterstützende Prozesse dazu anzubieten. Auch basieren die ORCID-ID-Zuordnungen in der Regel auf den Personenzuordnungen unserer WissenschaftlerInnen. Wenn dort Fehler bei der Zuordnung von Einzelpersonen gemacht werden, was bei mehreren Tausend AutorInnen für eine einzige Veröffentlichung in der Praxis durchaus vorkommt (z.B. bei Veröffentlichungen der ALICE Collaboration <https://repository.gsi.de/record/50756>), sind damit auch die Abbildungen zu den ORCIDs falsch. Bei den Autorenzuordnungen erfassen wir nur die eigenen WissenschaftlerInnen; niemand wird uns die Affiliationen anderer Einrichtungen eintragen wollen. Die Fach-ID, die sich im DFG Research Explorer zu unseren Einrichtungen findet, gibt unsere interdisziplinären Forschungsfeldern nicht wieder, so dass wir vermutlich, wir bisher, die Helmholtz-Systematik der POF als Forschungsfeld Pu141 weiterhin melden werden. Der Import in die Veröffentlichungsdatenbank erfolgt, wie bereits beschrieben, in der Regel über den DOI-Import und auch der Doublettencheck erfolgt bei JOIN² über die DOI, sowie andere über weiterer Identifikatoren, wie zum Beispiel die arXiv-ID. Daher haben wir in fast allen Journal-Datensätzen eine DOI als Identifikator.

Veröffentlichungen ohne DOI werden dabei nicht nur ungern per Hand eingetragen, sie werden auch oftmals von den Eintragenden nicht als „richtige“ Veröffentlichung betrachtet, da sie in der Regel nicht den Evaluierungskriterien von Helmholtz-reviewed entsprechen. Unsere eigenen Instituts-/Abteilungs-ID ist in allen Datensätzen enthalten. Wir erfassen keine fremden Institute und die Institut-ID unserer Einrichtungen beim DFG-Research Explorers spiegelt i.d.R. nur die Institution wider. In unseren Zentren und Instituten gibt es unterschiedliche Kulturen, was beteiligtes Institut meint und welchen Fällen diese eingetragen werden. Wie oben bereits dargestellt wurde, haben wir ein eigenes gemeinsames Normdatenset für Förderprojekte und –Programme. Einen Umstieg nach Fundref <http://www.crossref.org/fundingdata/> ziehen wir bei unseren eigenen Datenanforderungen zurzeit nicht in Betracht, da wir u.a. eine tiefere Erfassung benötigen (z.B. auf Workpackage-Ebene bei Sonderforschungsbereichen). Unter Format versteht der Kerndatensatz Daten, die wir in der Regel bereits automatisiert erhalten können wie Band, Heft oder Seitenbereich. Andere Attribute wie Name der Konferenz werden von unserer WissenschaftlerInnen eingetragen oder auch zwecks Zeitersparnis gerne ausgelassen. Zugangsrechte, sofern sie Open Access, Creative Commons oder die Nationallizenzen widergeben, kommen in der Regel durch den Import mit. Creative Commons können für eigene Verlagsveröffentlichungen von den Bibliotheken nachgetragen werden.

- Sprachcode Pu95

Da unsere Zentren in der Regel in Englisch veröffentlichen, ist dieses Feld für die Eingebenden freiwillig. Die MitarbeiterInnen der Bibliotheken versuchen, fehlende Sprachcodes nachzutragen, bzw. diese automatisiert nachzuerfassen.

- Peer-Reviewed Pu104

Aus dem Bibliothekswesen ist uns keine handhabbare Definition von peer-reviewed mit den Ausprägungen ja-nein bekannt, die auch automatisiert abrufbar ist. Wir können hier, wie bereits vorgestellt, allerdings formelle Ersatz-Definitionen liefern, wie wir sie im Rahmen der Programm-orientierten Förderung über die Journal-Masterlisten von kommerziellen Produkten definieren. Die Definition in der Spezifikation des Kerndatensatzes Forschung ist in der Praxis nicht erklärbar, sowohl für die Bibliothek als auch für die eintragenden WissenschaftlerInnen.

- Qualifikationsschrift Pu146
- Dokumenttyp Pu101
- Publikationstyp Pu6/Pu22-Pu51
- Ressource Pu102
- Förderer Pu86
- Förderkennzeichen Pu90
- Forschungsinfrastrukturen Fi0

Bei den Qualifikationsschriften haben wir ein breiteres Spektrum als die Ausprägungen des Kerndatensatzes (Dissertation und Habilitation). Die Angaben der WissenschafterInnen zur Qualifikationsschrift werden von der Bibliothek um URNs ergänzt und im Detail geprüft. Eine gewissenhafte Prüfung des Publikationstyps und des Dokumenttyps, den der Eintragende bei der Eingabe auswählt, erfolgt ebenfalls durch die Bibliothek. Allerdings können wir bestimmte Ausprägungen wie Bibliografie, Editorial, Arbeitspapier, Quellenedition, Wissenschaftliche Vortragsfolien, „Letter to the Editor“ oder Sonderheft einer Zeitschrift nicht erfassen. Da unsere Eintragungen von WissenschaftlerInnen und Verwaltungspersonal gemacht werden, können wir nicht alle Ausprägungen des Kerndatensatzes Forschung als Eingabemaske anbieten oder abfragen. Wir beschränken uns auf die in unseren Einrichtungen typische Ausprägungen des Publikationstyps, die auch für die Eintragenden verständlich sind. Ausprägungen wie ePaper oder Sammelbandbeitrag erfassen wir gar nicht. Über die Wahl der Eingabemasken können wir bestimmte Ressourcen als Bilder, Daten oder Multimedia (Kerndatensatz: Audio, bewegte Bilder) ausweisen. Schwierig wird es bei Journalveröffentlichungen, die auch Videos oder ähnliches enthalten, diese werden wir automatisiert weiterhin nur als Text kennzeichnen können. Förderinformationen und Förderkennzeichen werden von den WissenschafterInnen eingetragen – oder je nach Institutskultur auch nicht. Informationen über Forschungsinfrastrukturen können wir vermutlich aus anderen Eintragungen wie Beamlines oder Experimenten automatisiert ableiten. Dies wird allerdings von der Definition der Forschungs-infrastrukturen und derer Detailtiefe abhängen.

Die vom Kerndatensatz geforderten Aggregationsmöglichkeiten nach Fach, Organisationseinheit, Publikationstyp, Dokumenttyp, Schöpfer, Peer-reviewed und Veröffentlichungsjahr werden mit den oben ausgeführten Einschränkungen mit JOIN² möglich sein.

Ob JOIN² als Forschungsinformationssystem die XML-Exporte für den Kerndatensatz abliefern wird oder einen Zentrums-eigenen Forschungsinformationssystem die entsprechenden Daten automatisiert zuliefern wird, wird in unseren Zentren vermutlich unterschiedlich realisiert werden.

Zusammenfassend lässt sich sagen, dass wir durch unsere JOIN²-Projektarchitektur und die konsequente Nutzung von Normdatensätzen in der Lage sind, viele Ausdifferenzierungen des Objektes Publikationen in Kerndatensatz Forschung im Prinzip unter den ausgeführten Einschränkungen erfüllen zu können. Nur wenige Ausdifferenzierungen wie peer-reviewed können wir gar nicht liefern. Mit den JOIN²-Instanzen können unsere Zentren daher der Empfehlung des Wissenschaftsrates nachkommen, über die an unseren Zentren entstandenen Veröffentlichungen auskunftsfähig zu sein (Wissenschaftsrat, 2016, S. 41).

Bei der Umsetzung des Kerndatensatzes Forschung dürfen wir unsere WissenschaftlerInnen nicht aus den Augen verlieren, die die Eingaben und vor allem die Verknüpfungen zu Projekten, Abteilungen und Einzelpersonen vornehmen, und deren Aufwand in vertretbarem Rahmen belieben muss. Kennzahlen sind bei unseren JOIN²-Instanzen lediglich ein Nebenprodukt. Aus Sicht unserer WissenschaftlerInnen und unseren Zentren ist die Nachnutzbarkeit für eigene Listen im Web oder für das Schreiben von Veröffentlichungen die zentrale Aufgabe von JOIN². Diese Nutzungsmöglichkeiten für die WissenschaftlerInnen sollten von einer deutlich stärkeren Differenzierung für den Kerndatensatz Forschung nicht eingeschränkt werden.

Referenz

Wissenschaftsrat : *Empfehlungen zur Spezifikation des Kerndatensatz Forschung.* Berlin, 2016. <http://www.wissenschaftsrat.de/download/archiv/5066-16.pdf>