Thesenpapier Nationale Forschungsdateninfrastruktur für die Chemie (NFDI4Chem)

Autoren: Oliver Koepler[a] (contact@nfdi4chem.de), Nicole Jung[b], Angelina Kraft[a], Janna Neumann[a]

Mitwirkende: Sören Auer[a], Felix Bach[c], Thomas Bähr[a], Thomas Engel[d], Carsten Kettner[e], Johanna Kowol-Santen[f], Johannes Liermann[g], Anne Lipp[f], Andrea Porzel[h], Matthias Razum[j], Nils Schlörer[k], Dörte Solle[l], Torsten Winkler[m]

[a] Technische Informationsbibliothek, Hannover
[b] Institut für Organische Chemie, Karlsruher Institut für Technologie (KIT)
[c] Steinbuch Centre for Computing, Karlsruher Institut für Technologie (KIT)
[d] Department Chemie, Ludwig-Maximilians-Universität München
[e] Beilstein-Institut, Frankfurt
[f] Deutsche Forschungsgemeinschaft (DFG)
[g] Institut für Organische Chemie, Johannes Gutenberg-Universität Mainz
[h] Leibniz-Institut für Pflanzenbiochemie, Halle (Saale)
[j] FIZ Karlsruhe, Karlsruhe
[k] Department für Chemie, Universität Köln
[l] Institut für Technische Chemie, Leibniz Universität Hannover
[m] Otto-Diels Institut für Organische Chemie, Christian-Albrechts Universität zu Kiel

DOI: 10.5281/zenodo.1404201

1. Einleitung

Der Rat für Informationsinfrastrukturen (RfII, www.rfii.de) hat 2016 die schrittweise Schaffung einer „Nationalen Forschungsdateninfrastruktur (NFDI)“ vorgeschlagen. In einem Diskussionspapier vom April 2017 empfiehlt der RfII folgendes für die Weiterentwicklung des Forschungsdatenmanagements (FDM) im Rahmen einer NFDI für Deutschland: „*Hier sollen sich wissenschaftliche Communities bzw. Fachgemeinschaften, also die forschenden Nutzer, die bestimmte FDM-Dienste benötigen, in hinreichender Breite formieren und sich für einen Einstieg in die NFDI mit aus ihrer Sicht geeigneten Infrastruktur-Partnern auf längere Zeit zusammentun. Die so entstehenden Verbünde (sog. Konsortien) erhalten Ressourcen, die insbesondere organisatorisch und personell für die erforderlichen FDM-Lösungen eine auch dauerhafte Basis schaffen sollen. Die Konsortien sind die Hauptakteure der Ausgestaltung des Forschungsdatenmanagements – und auch des schrittweisen Aufbaus der NFDI*.“[1]

Zur Umsetzung soll eine Verwaltungsvereinbarung zwischen RfII und GWK, vermutlich im November 2018, geschlossen werden. Diese Vereinbarung soll die näheren Informationen und Schritte zur raschen Umsetzung enthalten. Informell werden folgende Randbedingungen diskutiert:

1. Initial werden Anträge von Fachkonsortien aus allen Fachgesellschaften und Fächern erwartet, wobei in dem jeweiligen Fach zunächst der Fokus auf ein FDM-Thema gelegt werden soll.
2. Ein erfolgreiches Konsortium soll neben einer Breite von beteiligten Institutionen auch eine methodische Breite, sowohl im Hinblick auf die zu Grunde liegenden Use Cases als auch die hierfür genutzten Methoden, eines Forschungsdatenmanagements (FDM) darstellen.
3. In einem zweiten Diskussionspapier vom März 2018 werden weitere Voraussetzungen für die Schaffung der NFDI sowohl auf Ebene der Fachgemeinschaften als auch auf Ebene der Infrastrukturpartner genannt. Demnach soll bereits der Auswahlprozess zweistufig und streng wissenschaftsgeleitet erfolgen, um größtmögliche Akzeptanz der Forscherinnen und Forscher[2] zu gewährleisten.
4. Die NFDI als Gesamtkonstrukt soll über einen Zeitraum von mehreren Jahren stufenweise aufgebaut werden. Insgesamt sei mit einer mittleren zweistelligen Zahl von Konsortien zu rechnen.

[1] Diskussionspapier RfII, “Schritt für Schritt – oder: Was bringt wer mit? Ein Diskussionsimpuls zu Zielstellung und Voraussetzungen für den Einstieg in die Nationale Forschungsdateninfrastruktur (NFDI)”, 2017, www.rfii.de

[2] Aus Gründen der besseren Lesbarkeit wird im Folgenden auf die gleichzeitige Verwendung männlicher und weiblicher Sprachformen verzichtet. Sämtliche Personenbezeichnungen gelten gleichermaßen für beiderlei Geschlecht.

2. Zielsetzung NFDI4Chem

"*Der stufenweise Aufbau einer Nationalen Forschungsdateninfrastruktur in Netzwerkform hat das Ziel, ein verlässliches und nachhaltiges Dienste-Portfolio zu schaffen, welches generische und fachspezifische Bedarfe des Forschungsdatenmanagements in Deutschland abdeckt.*"[3] Für das Fachgebiet Chemie ermöglicht eine solche nationale Forschungsdateninfrastruktur, öffentlich-finanzierte Forschungsdaten effizient zu erheben, standardisiert zu beschreiben, dauerhaft zu speichern und durch Persistent Identifier (PID) eindeutig referenzierbar und auffindbar zu machen. Sie unterstützt gemäß den Vorgaben des RfII die Reproduzierbarkeit und Nachnutzbarkeit der Daten zum Zwecke einer perpetuierten Wissensgenerierung. Mit der Reproduzierbarkeit von Forschungsergebnissen unterstützt eine solche Forschungsdateninfrastruktur den Peer Reviewing Prozess zur Förderung der wissenschaftlichen Selbstkontrolle und erhöht die Datenqualität, insbesondere in wissenschaftlichen Publikationen. Die NFDI4Chem ist ein gemeinschaftlicher Ansatz von Wissenschaftlern aus der Chemie, der Fachgesellschaft Gesellschaft Deutscher Chemiker und deren Fachgruppen, Einrichtungen aus der Forschungsförderung und Infrastruktureinrichtungen (Technische Informationsbibliothek). Eine Gruppe von Vertretern dieser Stakeholder hat sich Ende April 2018 zu einem Auftakttreffen "Fachgespräch NFDI4Chem" in Hannover getroffen. Weitere Stakeholder wie Verlage oder Datenbankanbieter sind im folgenden Diskurs willkommen.

3. Nationales Forschungsdatenmanagement heute

In Laboratorien von Forschungseinrichtungen und Unternehmen werden tagtäglich große Mengen von experimentellen Daten erzeugt, analysiert und Erkenntnisse daraus abgeleitet. Der Umgang mit diesen Forschungsdaten in der Chemie ist sehr heterogen, wie unterschiedliche Erhebungen und Umfragen[4] immer wieder zeigen. Auch eine nicht repräsentative Erhebung zur Handhabung von Forschungsdaten unter den Teilnehmern des Fachgesprächs NFDI bestätigte diese Aussage. Während es in den sogenannten Big Data Wissenschaften viele etablierte und anerkannte Datenzentren und -repositorien gibt, existieren in der Chemie in den wenigsten Forschungseinrichtungen etablierte Strukturen, die sich mit Forschungsdatenmanagement beschäftigen.

3.1. Umgang mit Forschungsdaten, Datenformate

Die Heterogenität der Daten hinsichtlich Inhalt, Umfang und Datenformat ergibt sich unter anderem durch den Einsatz einer Vielzahl analytischer Methoden, deren Messdaten zwar digital aber oft in proprietären Formaten anfallen. Die digitale Vernetzung von Messdaten unterschiedlicher Analysemethoden ist daher oft unzureichend, genauso eine nachhaltige Beschreibung der erhobenen Daten sowie die Verwendung von offenen, standardisierten

[3] Diskussionspapier RfII, "Schritt für Schritt – oder: Was bringt wer mit? Ein Diskussionsimpuls zu Zielstellung und Voraussetzungen für den Einstieg in die Nationale Forschungsdateninfrastruktur (NFDI)", 2017, www.rfii.de

[4] K. Pappenberger, bwFDM-Communities – Wissenschaftliches Datenmanagement an den Universitäten Baden-Württembergs, Bibliothek Forschung und Praxis, 40(1), S. 21-25, 2016, http://nbn-resolving.org/urn:nbn:de:swb:90-832721

Austauschformaten. Der Umgang mit den erzeugten Daten und deren Speicherung ist je nach Institut und Forschungsgruppe individuell geregelt, die Verantwortung tragen in der Regel die Forschungsgruppenleiter. Insbesondere für die nicht datenintensiven (Small Data) Bereiche der experimentellen chemischen Forschung werden Forschungsdaten aktuell auf lokalen Rechnern, USB-Sticks, DVDs oder Institutsservern gespeichert. Die Speicherung in Repositorien ist, bis auf wenige Ausnahmen wie z.B. der *in silico* Chemie und der Kristallographie, nicht üblich. Die lokale Verwaltung und Speicherung der Daten impliziert eine Kuratierung der Daten nach individuellen Gesichtspunkten des einzelnen Forschers oder der Forschungsgruppe, standardisierte und normalisierte Daten sowie beschreibende Metadaten spielen nur eine untergeordnete Rolle.

3.2. Publikationsprozesse

In der Regel erfolgt die Publikation wissenschaftlicher Ergebnisse in Form von Zeitschriftenaufsätzen, in denen sich eine reduzierte Darstellung der Forschungsdaten zur Beweisführung wiederfindet. Ergänzt wird dies häufig durch Supplementary Materials mit detaillierten, experimentellen Daten in Tabellen, Strukturformeln und Abbildungen von Spektren. Jedoch werden diese Supplementary Materials überwiegend als PDF-Dateien veröffentlicht, ihre Inhalte sind nicht direkt referenzierbar und nicht maschinenlesbar. Für Zeitschriftenpublikationen werden die Anforderungen zur Veröffentlichung von Daten in den sogenannten “Research data publication guidelines” beschrieben. Die Bereitstellung dieser erweiterten Supplementary Materials ist teils verpflichtend, teils optional. Bei Daten aus spektroskopischen Methoden wird die Einarbeitung im Manuskript oder als Supplementary Material gefordert, in manchen Vorgaben ist die Mitveröffentlichung auch optional. Als positives Beispiel kann die Handhabung von Kristallstrukturdaten hervorgehoben werden. Fast alle Zeitschriften fordern parallel oder vor Veröffentlichung eines Artikels eine Veröffentlichung dieser Daten in einem Repository wie der Cambridge Structural Database (CSD).[5] Wie für die Chemie die Veröffentlichung von wissenschaftlichen Ergebnissen in einem Zeitschriftenaufsatz mit kombinierter Publikation von Forschungsdaten in einem Repository erreicht werden kann, zeigt das Beispiel von N. Jung et al.[6] In den Supplementary Materials sind für NMR-, IR- und Massenspektren DOI (Digital Object Identifier)-Verlinkungen zu den Datensätzen in einem Open Access Repositorium hinterlegt. Die Datensätze werden inklusive Metadaten wie chemischer Struktur und Details zur Analysemethode gespeichert und visualisiert.

3.3. Verfügbare Infrastrukturen für Forschungsdaten

In Deutschland finden sich nur wenig verfügbare Repositorien bzw. nur wenige Initiativen, die eine Infrastruktur mit Speicher- oder Nachnutzungsmodellen für Forschungsdaten aus der Chemie zur Verfügung stellen. Es können zwei Projekte genannt werden, die jeweils durch DFG-Förderung eine frei nutzbare disziplin-spezifische Forschungsdateninfrastruktur aufbauen konnten: Das elektronische Labor-Notebook (ELN) Chemotion des KIT Karlsruhe

[5] https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/, zuletzt aufgerufen 08.08.2018.

[6] N. Jung, S. Grässle, D. S. Lütjohann, S. Bräse, *Org. Lett.*, **2014**, 16 (4), 1036–1039. DOI: https://doi.org/10.1021/ol403313h

unterstützt den Wissenschaftler bei der Planung und Dokumentation von Experimenten und entstehenden Forschungsdaten[7]. Das ELN ist an ein Open Access Repositorium zur Speicherung und Veröffentlichung von Forschungsdaten angebunden, welches alle analytischen Datensätze in Korrelation zu dem identifizierten Molekül erfasst. Die offene NMR-Datenbank NMRShiftDB[8], gehostet von der Universität Köln, ermöglicht seit 2002 die Publikation von ^{1}H- und ^{13}C-NMR Spektren mit ergänzenden Informationen wie Strukturdaten und Signalzuordnungen. Neu hinzugefügte Spektren durchlaufen eine Plausibilitätsprüfung. Die Datenbank bietet als Referenzdatenbank eine Spektren- und Struktursuche sowie die Vorhersage von Spektren. Im internationalen Kontext sind die bereits erwähnten Röntgenstrukturdatenbanken Cambridge Structural Database (CSD) des Cambridge Crystallographic Data Centre (CCDC) sowie die Crystallography Open Database (COD) seit Jahrzehnten fest etabliert und können als Vorbild für den gesamten Forschungsbereich dienen. Innerhalb der CSD wird ein seit 1965 gewachsenes System zur Kuration der Daten genutzt.[9] Eine wesentliche Voraussetzung dieser Kuration ist die Bereitstellung der Daten als „Crystallographic Information File“ (CIF), welches sich als Standard durchgesetzt hat. Die von Autoren eingestellten CIF-Daten werden schrittweise geprüft und auf eine dauerhafte Speicherung vorbereitet. Teil dieses Prozesses ist die Integration der Workflows der Verlage, eine Bereitstellung der Struktur-Information für Gutachter und die Erfassung deren korrigierter Strukturvorschläge.
Des weiteren gibt es disziplinübergreifende Repositorien wie z.B. RADAR[10] des FIZ Karlsruhe. Dieser Repository-Service verfolgt einen generischen Ansatz für Disziplinen, die noch keine fachspezifischen Lösungen entwickelt haben und einen ganzheitlichen Service für die Archivierung und Publikation ihrer Daten suchen. Ebenfalls generisch ausgerichtet sind spezielle Landesprojekte zur Nutzung von Forschungsdaten-Diensten (Beispiel Baden-Württemberg: Integration eines RDMO Dienstes[11], bwDIM[12], bwDataArchive[13], bwSync&Share[14]).

3.4. Nationale und internationale Initiativen

Die Defizite und das Verbesserungspotential im Umgang mit Forschungsdaten sind vermehrt sichtbar und dringen zunehmend in das Bewusstsein der Wissenschaftler. Bezogen auf deutsche Aktivitäten wurde 2016 das Projekt IDNMR[15] (Initiative Datenqualität in der NMR-Spektroskopie) initiiert, um ein digitales Veröffentlichungsformat für NMR-Daten zu erarbeiten und damit eine automatische Plausibilitätskontrolle sowie die Hinterlegung publizierter Daten in öffentlichen Datenbanken zu ermöglichen. Seit 2013 kombinieren die Chemotion-Projekte des KIT Karlsruhe die Entwicklung von Software für

[7] P. Tremouilhac, A. Nguyen, Y.-C. Huang, S. Kotov, D. S. Lütjohann, F. Hübsch, N. Jung, S. Bräse, J. Chemoinfo., **2017**, 9 (54), DOI https://doi.org/10.1186/s13321-017-0240-0
[8] C. Steinbeck, S. Krause, S. Kuhn, *J. Chem. Inf. Comput. Sci.*, **2003**, 43 (6), 1733-1739. DOI: https://doi.org/10.1021/ci0341363
[9] I.J. Bruno, C. R. Groom., *J. Comput. Aided Mol. Des.*, **2014**, 28 (10): 1015–22. DOI: https://doi.org/10.1007/s10822-014-9780-9
[10] https://www.radar-service.eu/de, zuletzt aufgerufen 07.08.2018.
[11] https://rdmo.aip.de/, zuletzt aufgerufen 13.08.2018.
[12] https://www.scc.kit.edu/forschung/10898.php, zuletzt aufgerufen 13.08.2018.
[13] https://www.rda.kit.edu/, zuletzt aufgerufen 13.08.2018.
[14] https://bwsyncandshare.kit.edu/login, zuletzt aufgerufen 13.08.2018.
[15] IDNMR-Projekt Webseite www.idnmr.uni-koeln.de, zuletzt aufgerufen 08.08.2018.

chemische Forschungsprozesse mit neuen Methoden der Datenspeicherung und Referenzierung. Das bereits genannte elektronische Labor-Notebook (chemotion-ELN) des KIT Karlsruhe wird kontinuierlich weiterentwickelt.
Die Bemühungen für eine systematische Verbesserung im Umgang im Forschungsdaten mit Hilfe von Digitalisierungsstrategien vereint seit vielen Jahren auch auf internationaler Ebene engagierte Wissenschaftler der Disziplin Chemie. Beispiele sind das IUPAC Committee on Publications and Cheminformatics Data Standards (CPCDS)[16], die Chemistry Research Data Interest Group (CRDIG)[17] der Research Data Alliance (RDA)[18] und das GO FAIR Chemistry Implementation Network (ChIN).[19] Am 16. Februar 2018 konstituierte sich die Data Interest Group/Chemistry (DIGChem),[20] in der u.a. Vertreter der oben genannten Institutionen Datenstandards und Anforderungen für Daten-Repositorien in der Chemie diskutieren.[21]

4. Anforderungen an die NFDI4Chem

Beim Auftakttreffen des Fachgesprächs NFDI4Chem wurden der aktuelle Umgang mit Forschungsdaten, notwendige Veränderungen sowie Chancen innerhalb einer nationalen Forschungsdateninfrastruktur für die Chemie diskutiert. Es besteht Konsens, dass für ein erfolgreiches Forschungsdatenmanagement und die breite wissenschaftliche Akzeptanz einer NFDI4Chem eine Reihe von Anforderungen umgesetzt werden müssen. Diese Anforderungen, die hiermit für eine breite Fachcommunity zum Austausch zum zur offenen Diskussion gestellt werden, sind:

- Gesicherte, langfristige finanzielle Unterstützung des FDM an Forschungseinrichtungen (Universitäten, Hochschulen und außerhochschulischen Forschungseinrichtungen)
- Diskussion und Gestaltung von Policies zum FDM mit Forschungsförderungs-Organisationen und Fachgesellschaften
- Hohe Datensicherheit sowie differenzierte Betrachtung von Rollen und Rechten für den Zugriff auf Forschungsdaten (Embargo-Regelungen, Datenaustausch nur auf Projektpartner-Ebene)
- Policies und Guidelines zur Annotation von Forschungsdaten
- Niedrigschwellige Nutzung sowohl der Forschungsdateninfrastruktur als auch peripherer Tools und Plattformen
- Digitalisierung von Workflows in der Chemie: Von der Datenerhebung, Analyse, Interpretation bis zur Ergebnispräsentation (Vernetzung über elektronische Laborjournale)
- Schaffung bzw. Beteiligung an einer offenen, konstruktiven und internationalen Diskussion über wichtige Punkte wie der Datenvernetzung in der Chemie,

[16] https://iupac.org/who-we-are/committees/committee-details/?body_code=024, zuletzt aufgerufen 08.08.2018.
[17] https://www.rd-alliance.org/groups/chemistry-research-data-interest-group.html , zuletzt aufgerufen 08.08.2018.
[18] https://www.rd-alliance.org/ , zuletzt aufgerufen 08.08.2018.
[19] https://www.go-fair.org/implementation-networks/overview/chemistry/, zuletzt aufgerufen 08.08.2018.
[20] https://iupac.org/digchem-a-vision-for-chemical-data-standards/, zuletzt aufgerufen 08.08.2018.
[21] DIGChem—a vision for chemical data standards, Chemistry International, **2018**, 40 (3), 31–32. DOI: https://doi.org/10.1515/ci-2018-0315

Digitalisierung von Workflows, Datenformate, Datenerfassung und Workflow-Änderungen
- Gemeinsame Diskussion und Gestaltung neuer Publikationsprozesse unter Einbindung von Forschungsdaten mit Verlagen und weiteren Stakeholdern

Die dafür notwendige Diskussion sollte möglichst vielen Stakeholdern offen sein, um wichtige Themen auf breiter Basis diskutieren zu können. Der Austausch geht über nationale Grenzen hinaus, Impulse aus den Arbeitsgruppen von FORCE11[22], RDA, IUPAC und DigChem sollen daher Eingang finden und gleichermaßen Erkenntnisse aus der NFDI4Chem international geteilt werden. Dabei sollte allerdings nicht das Ziel der NFDI4Chem außer Augen verloren gehen, schon frühzeitig eine funktionsfähige Basisinfrastruktur zu erschaffen, um für Akzeptanz einer NFDI4Chem zu werben. Eine solche Basisinfrastruktur soll durch Modularität und Erweiterbarkeit die kontinuierliche Diskussion und Manifestation des FDMs in der Chemie begleiten und mit ihr wachsen.

5. Forschungsdateninfrastruktur: Digitaler Wandel in der chemischen Forschung

> These: Für ein erfolgreiches Forschungsdatenmanagement ist eine digitale Unterstützung aller Prozesse, beginnend bei der Erhebung von Daten, über deren Aufbereitung und Analyse bis zur Publikation notwendig. Die vollständige Verzahnung und intuitive Benutzbarkeit von Repositories und digitalen Tools ist essentiell. Mehraufwände für den Wissenschaftler müssen minimal gehalten werden.

Die Bereitstellung von Repositorien für Forschungsdaten bzw. deren Vernetzung gehört zu den zentralen Instrumenten, um die in Abschnitt 4 formulierten Anforderungen umsetzen und ein erfolgreiches FDM etablieren zu können.
Für die zahlreichen Teildisziplinen der Chemie sind auf nationaler Ebene nur wenige Repositorien bekannt, die als Knoten für ein zukünftiges Netzwerk einer Forschungsdateninfrastruktur dienen können. Dies sollte allerdings nicht als Hemmnis oder als Nachteil gesehen werden, vielmehr beinhaltet dies für eine NFDI die Chance eine interoperable Infrastruktur zu initiieren. Dabei soll der Bedarf u.a. an einem übergreifenden Rollen- und Rechtemanagement für die Bereitstellung und den Austausch von Forschungsdaten so frühzeitig wie möglich berücksichtigt werden. Beispiele für existierende nationale Repositorien sind die bereits erwähnten NMRShiftDB2 und Chemotion. Diese fungieren für die Teildisziplin Organische Chemie als Repositorien für Daten aus der Spektroskopie bzw. für chemische Entitäten und zugehörige Forschungsdatensätze.
Aus dem Bereich der Biochemie ist die STRENDA DB für enzymologische Daten des Beilstein Instituts zu nennen.[23] In der Datenbank abgelegte Daten werden gegen die STRENDA ("**St**andards for **R**eporting **En**zymology **Da**ta") Richtlinien validiert und bei Erfolg

[22] https://www.force11.org/ , zuletzt aufgerufen 08.08.2018.
[23] http://www.beilstein-institut.de/projekte/strenda, zuletzt aufgerufen 08.08.2018.

mit einer STRENDA Registry Number (SRN) versehen, mit der in einer Zeitschriftenpublikation auf den Datensatz referenziert werden kann.

Auch nicht chemie-spezifische Projekte können zum erfolgreichen Aufbau der NFDI beitragen. Generische Repositorien wie das bereits erwähnte RADAR (FIZ Karlsruhe), die einen ganzheitlichen Service für die Archivierung und Publikation von Daten anbieten, stehen bereits zur Verfügung. Diese verfügen nicht über die nötige fachspezifische Funktionalität, könnten im Sinne einer Synergienutzung jedoch als alternativer oder zusätzlicher Backend-Speicher in ein Konzept integriert werden. Weiterhin gibt es national als auch international eine Reihe von interdisziplinären Repositorien, die Forschungsdaten aus der Chemie beinhalten, jedoch auf ganz bestimmte Themengebiete spezialisiert sind. Das Repositorienverzeichnis re3data.org gibt eine detaillierte Auskunft hierüber. Beispiele für solche u.a. in Deutschland betriebenen Repositorien bilden PANGAEA[24], GEOROC[25], die European MassBank[26], die HALO database[27], die Spectral Database for Organic Compounds (SDBS)[28], und weitere. Durch Vernetzung könnten disziplin-verwandte Repositorien mit dem NFDI4Chem Verbund einen deutlichen Mehrwert durch interdisziplinären Informationsaustausch schaffen. Aufgabe der NFDI4Chem wird es sein, generische und disziplin-verwandte Repositorien auf ihre Kompatibilität und auf ihren Nutzen für die nationale Infrastruktur hin zu prüfen.
Nicht alle Repositorien sind offen für das Hochladen von eigenen Datensätzen, sondern stellen wie SDBS oder MassBank Referenzdatenbanken dar, die von einem Forschungsinstitut oder Forschungsverbund gepflegt werden. Verweise auf Forschungsdaten finden sich auch in proprietären Fachdatenbanken wie SciFinder oder Reaxys. Diese Referenzen verweisen in der Regel auf Zeitschriftenpublikationen und deren Supplementary Materials. In der offenen Reaktionsdatenbank Chemspider Synthetic Pages[29] werden bei experimentellen Vorschriften zu den Substanzen Spektrendaten als ergänzende Informationen zur Verfügung gestellt.

Im Verlauf des Fachgesprächs NFDI4Chem wurde deutlich, dass die Verfügbarkeit von Repositorien nur eine Komponente einer erfolgreichen NFDI sein wird. Vielmehr muss ganzheitlich die Digitalisierung aller Prozesse in der chemischen Forschung betrachtet werden:

- Durchgehende Digitalisierung der Datenerhebung und Dokumentation chemischer Experimente
- Digitale Zusammenführung erhobener Forschungsdaten
- Abgeleitete Korrelationen und Analysenergebnisse
- Annotation von Forschungsdaten mit fachspezifischen und semantischen Metadaten
- Überführung in offene Datenformate
- Publikation in Repositorien und Registrierung mit Persistenten Identifiern (PID)

[24] https://pangaea.de/, zuletzt aufgerufen 08.08.2018.
[25] http://georoc.mpch-mainz.gwdg.de/georoc/Entry.html, zuletzt aufgerufen 08.08.2018.
[26] https://massbank.eu/MassBank/, zuletzt aufgerufen 08.08.2018.
[27] https://halo-db.pa.op.dlr.de/, zuletzt aufgerufen 08.08.2018.
[28] http://sdbs.db.aist.go.jp/sdbs/cgi-bin/cre_index.cgi, zuletzt aufgerufen 08.08.2018.
[29] http://cssp.chemspider.com/, zuletzt aufgerufen 08.08.2018.

Ein einzelnes Repositorium kann nicht alle für den beschriebenen Prozeß notwendigen Funktionen zur Verfügung stellen. Daher basiert eine NFDI4Chem auf einem Netzwerk aus zentralen oder dezentralen Dokumentations- und Informationssystemen (z.B. elektronische Labor-Notebooks) sowie Repositorien, Langzeitarchivierungssystemen und Publikationssystemen. Eine konsequente Digitalisierungsstrategie soll für alle Forscher darüber hinaus durch die Entwicklung von Modellen zur Integration der Laborgeräte realisiert werden. Die NFDI unterstützt die Forschungsprozesse daher auf verschiedensten Ebenen. Die technische Infrastruktur wird zusätzlich gestützt durch Policies für Prozesse, Daten-Standards sowie differenzierte Rollen- und Rechtemanagement-Systeme für Repositorien.

5.1. Digitalisierung der Datenerhebung und Dokumentation, vernetzte Laborinfrastruktur und ELNs

Am Beispiel der synthetischen Forschung lassen sich sehr gut die Herausforderungen für eine Forschungsdateninfrastruktur und ihre Bausteine beschreiben. So werden an unterschiedlichsten Analysegeräten Daten in einer Vielzahl von Datenformaten erfasst. Zusätzlich werden experimentelle Vorschriften und Beobachtungen manuell erhoben, zusammengeführt und ausgewertet. Die Dokumentation erfolgt auch heute noch überwiegend analog in handschriftlichen Laborjournalen.
Ein modernes FDM soll nach Wunsch der forschenden Wissenschaftler zu möglichst geringen Mehraufwänden führen. Lösungsansätze hierfür müssen bereits frühzeitig bei der Datenerhebung und Dokumentation greifen und rücken Tools und Workflows zur Digitalisierung aller relevanten Schritte in das Zentrum: Laborinformationssysteme (LIMS) und Elektronische Labor-Notebooks (ELNs). Eine Vernetzung von Laborgeräten, Dokumentationssystemen, Repositorien, Langzeitarchivierungssystemen und Publikationssystemen stellen den Schlüssel für eine Digitalisierung der chemischen Forschung dar. Innerhalb eines modernen Laborinformationssystems werden an vernetzten Analysegeräten gemessene oder beobachtete Forschungsdaten in geeignete Austauschformate überführt und mit beschreibenden Metadaten angereichert. Die Daten werden durchgehend elektronisch in die digitale Dokumentation eines elektronischen Labor-Notebooks importiert. Das ELN dokumentiert Planung, Durchführung und Ergebnisse von Experimenten und akkumuliert kontinuierlich Daten und Metadaten zu Experimenten und chemischen Entitäten. Module zur Erstellung von Berichten und Dokumentbausteinen für Publikationen ermöglichen die Bereitstellung und Nachnutzung von Forschungsdaten sowie der daraus abgeleiteten Auswertungen. Über Schnittstellen können Forschungsdaten direkt in Repositorien exportiert werden.

5.2. Datenformate

Die Verwendung von proprietären Datenformaten von Forschungsdaten erschwert die Weiterverarbeitung in offenen Infrastrukturen und die Nachnutzbarkeit publizierter Daten durch andere. Wann immer möglich, sollten daher in der NFDI4Chem offene Datenformate

zur Speicherung von Forschungsdaten verwendet werden. Beispiele sind das JCAMP-DX[30] oder AnIML Datenformat bzw. Datenstandard für spektroskopische Daten. JCAMP-DX steht in zahlreichen Derivaten als freies, maschinenlesbares Austauschformat für unterschiedliche Spektrendaten-Typen zur Verfügung, jedoch befürworten viele Spektroskopiker eine Überarbeitung dieses älteren Formats (Einführung 1988).
Einen Schritt weiter geht die NMReDATA Initiative[31] mit einem offenen Datenstandard zur Verknüpfung von NMR-Parametern wie chemische Verschiebung, Kopplungskonstante, 2D Korrelation mit Strukturdaten. Informationen werden in einem erweiterten Molfile-Format abgespeichert. Für die Speicherung in Repositorien werden NMReDATA Datensatz und gemessene Spektrendaten in einem NMR-Datensatz zusammengeführt. Das NMReData Format kann sowohl von Menschen als auch Maschinen gelesen werden. Es unterstützt die Sichtung und Bewertung von Spektrendaten, da experimentelle Daten und Zuordnungsdaten für die Strukturbestimmung gemeinsam gespeichert werden. Mit diesem Datenformat werden erhobene Daten und daraus abgeleitete Erkenntnisse gemeinsam digital erfasst und gespeichert. Insgesamt ist die NFDI4Chem Diskussion über geeignete Datenformate und Datenstandards im internationalen Kontext zu führen.

5.3. Annotation von Forschungsdaten

Der Informationsgehalt eines publizierten, frei verfügbaren Forschungsdatensatzes ist ohne beschreibende Metadaten mit Angaben zu verwendeten Geräten, Versuchsparametern oder chemischer Strukturinformation stark eingeschränkt. Die Dokumentation und Beschreibung erhobener Forschungsdaten wurde im Fachgespräch NFDI als Herausforderung für den wissenschaftlichen Alltag identifiziert und zieht sich als roter Faden durch die Diskussion. Bereits bei der Datenerhebung mittels Analysegeräten sollten beschreibende Metadaten mit erfasst, in die digitale Dokumentation eines elektronischen Labor-Notebooks importiert werden und dort jederzeit erweiterbar sein. Als Metadaten können auch chemische Strukturinformationen wie beispielsweise der offene IUPAC International Chemical Identifier (InChI) verstanden werden. Die Verwendung der Identifier ermöglicht eine Vereinfachung z.B. der für die Chemie üblichen Rechercheverfahren in Repositorien und kann für eine entitätenbasierten Vernetzung von Repositorien hilfreich sein. Sie ermöglichen auch die Vernetzung mit etablierten Repositorien chemischer Entitäten wie PubChem oder Chemspider.

5.4. Bereitstellung und Publikation von Forschungsdaten, Persistent Identifier

In Zukunft können, durch die NFDI unterstützt, ELNs oder Tools die Bereitstellung von Forschungsdaten sowie der daraus abgeleiteten Auswertungen z.B. durch Erstellung von Berichten und Dokumentbausteinen für Publikationen initiieren. Validierungsfunktionen

[30] R. S. McDonald and Paul A. Wilks, *Appl. Spec.*, **1988**, 1, 151-162. DOI: https://doi.org/10.1366/0003702884428734
[31] Pupier M, Nuzillard J-M, Wist J, et al., *Magn Reson Chem*. **2018,** 1–13. DOI: https://doi.org/10.1002/mrc.4737

unterstützen den Wissenschaftler bei der Aufbereitung von Peaklisten für den experimentellen Teil von Publikationen oder Supplementary Materials. Über Schnittstellen können Forschungsdaten in Repositorien exportiert und dabei mit einem Persistent Identifier versehen werden. Persistent Identifier wie ein DOI ermöglichen die eindeutige Referenzierbarkeit und Auffindbarkeit eines Forschungsdatensatzes, sowie die Verknüpfung mit einer zugehörigen Zeitschriftenpublikation.

Die aufgeführten Aspekte und Lösungsmöglichkeiten zeigen, wie ein vernetztes wissenschaftliches Forschungsdatenmanagement innerhalb der Disziplin Chemie erreicht werden kann. Es wird deutlich, dass bisherige Workflows innerhalb eines Labors modernisiert und digitalisiert werden müssen. Die NFDI soll in der Lage sein, diese Punkte zu adressieren und eine Infrastruktur mit notwendigen Tools zur Verfügung zu stellen. Die NFDI muss darüber hinaus auch die Transformation der Workflows organisatorisch und strukturell begleiten. Die Zusammenarbeit mit der Fach-Community wird nachfolgend beschrieben.

6. Fach-Community

Der Erfolg einer NFDI4Chem liegt in einer gemeinsamen Anstrengung von wissenschaftlicher Fachgemeinschaft, Förderorganisationen und unterstützenden Infrastruktureinrichtungen. Alle können und sollen sich beim Aufbau einer Forschungsdateninfrastruktur gegenseitig befruchten und Impulse liefern. Nachfolgend werden aus Sicht des Fachgesprächs NFDI erfolgskritische Aspekte mit Fokus auf die Fachgemeinschaft diskutiert.

6.1. Bewußtsein in der Fach-Community schaffen

> These: Mechanismen, Werkzeuge und Vorteile eines nachhaltigen Forschungsdatenmanagements sind Wissenschaftlern an Forschungseinrichtungen nur unzureichend bekannt. Die Mehraufwände für das FDM werden befürchtet und abgelehnt, weil Mehrwerte nicht bekannt sind.

Übergeordnetes Ziel der NFDI4Chem muss es sein, ein Bewusstsein und Akzeptanz für die Relevanz des Themas FDM in der Chemie zu schaffen und Mehrwerte trotz Mehraufwänden aufzuzeigen.
Trotz der breiten wissenschaftspolitischen Diskussionen zu Policies, Standards und Umsetzung eines nachhaltigen FDM ist der Umgang mit Forschungsdaten in der wissenschaftlichen Praxis der Chemie praktisch unverändert. Werkzeuge und Möglichkeiten des FDM wie die Veröffentlichung in Daten-Repositorien mit Vergabe von Persistent Identifiern (z. B. DOI) sowie die Verknüpfung von Zeitschriftenpublikationen mit Datenpublikationen sind einer Vielzahl von Wissenschaftlern in der Chemie schlicht nicht bekannt und spielen für ihre wissenschaftliche Reputation nur eine untergeordnete Rolle. Auch dort wo die Möglichkeiten von FDM bekannt sind, wird eine Integration in die Forschungsarbeit aufgrund der befürchteten Mehraufwände abgelehnt. Diesen

Herausforderungen hat sich eine NFDI4Chem zu stellen, Vorteile müssen herausgestellt und für Mehraufwände geworben werden.
Der Aufbau einer NFDI bietet die Möglichkeit einer kritischen Auseinandersetzung im Umgang mit Forschungsdaten im Forschungsprozess in der Chemie. Über die Kommunikationskanäle der Gesellschaft Deutscher Chemiker, ihren Fachgruppen und Regionalstrukturen kann in der Breite ein Diskurs zum FDM angestoßen werden. Begleitet durch eine Umfrage zum Umgang mit Forschungsdaten an Forschungseinrichtungen könnten Bedarfe identifiziert und Use Cases für Teildisziplinen der Chemie formuliert werden.
Ziel muss es sein, die Mehrwerte einer NFDI herauszustellen, Wissenschaftler zur aktiven Teilnahme zu animieren und FDM in den Forschungsprozess fest zu verankern. Die Bedeutung der NFDI als langfristige, übergreifende Lösung gegenüber verbreiteten Infrastruktur-Projektfinanzierung sollte deutlich kommuniziert werden.
Auf Ansätze zur Animierung der Wissenschaftler wird unter Anreize für die Fach-Community im Detail eingegangen. Über Datenmanager und Data Scientists können Datenkompetenzen in Forschungseinrichtungen und beim Wissenschaftler selbst aufgebaut werden. Deren Funktion und Rolle sollte zwischen Wissenschaftlern und Infrastruktureinrichtungen wie Bibliotheken, die dieses Themenfeld schon länger besetzen, vereinbart werden.

6.2. Community Training

> These: Es gibt bisher zu wenig Wissenschaftler in der Chemie mit Forschungsdatenmanagement-Kompetenzen. Diejenigen, die sich mit FDM auseinandersetzen, erhalten dafür kaum Anerkennung aus den eigenen Reihen. Sie brauchen bei der Einführung eines FDM externe Beratung und Hilfestellungen.

Es dürfte unstrittig sein, dass Werkzeuge und Methoden eines nachhaltigen FDM den meisten Wissenschaftlern an Forschungseinrichtungen nur unzureichend bekannt sind. Eine langfristige Verbesserung muss die Aufnahme des Themas FDM in das Curriculum sein, um die Datenkompetenz der Wissenschaftler zu erhöhen. Der Umgang mit digitalen Werkzeugen, elektronischen Labor-Notebooks und die nachhaltige Aufbereitung von Forschungsdaten müssen bereits in der wissenschaftlichen Ausbildung vermittelt werden.
Mittelfristiger Lösungsansatz ist die Etablierung von Data Scientists und Datenmanagern als Ansprechpartner für die Wissenschaftler. Die Rolle des Datenmanagers sollte das Beste aus zwei Welten vereinen. Dieser sollte auf der einen Seite eine hohe Kompetenz im FDM aus Infrastruktursicht mitbringen, auf der anderen Seite Wissen aus der Fachcommunity und Kenntnisse über dortige Prozesse und Bedarfe haben. Er versteht sich als Dienstleister für die Wissenschaft. Ein Data Scientist könnte eher direkt an Forschungseinrichtungen lokalisiert sein und unmittelbar aus der Fach-Community rekrutiert werden. Beide Rollen dienen als Ansprechpartner für konkrete, praktische Fragen zur Umsetzungen des Datenmanagements.
Die NFDI4Chem kann die notwendigen Diskussionen zu Aspekten des FDM anstoßen, Argumente liefern und Veränderungsprozesse begleiten. Sie kann für mehr Anerkennung

der Aufwände für ein FDM werben und dies durch Anreiz- und Belohnungskonzepte langfristig verankern. Die NFDI kann im Kontext Curriculum, Datenmanager und Data Scientists bei der Ausarbeitung von Schulungskonzepten und der anschließenden Umsetzung Forschungseinrichtungen unterstützen und mithelfen eine Grundlage für zukünftige Datenkompetenz zu schaffen. Ein Wissenstransfer kann in erster Instanz aus Infrastruktureinrichtungen wie Bibliotheken oder aus den Gremien zum FDM der Universitäten erfolgen.

6.3. Anreize für die Fach-Community

> These: Eine NFDI4Chem wird nur in Kombination mit Anreizen und sanftem Druck von Seiten der Forschungsförderung langfristig Erfolg haben. Wissenschaftler werden nicht freiwillig die Mehraufwände eines Forschungsdatenmanagements aufbringen.

Die Etablierung einer NFDI geht mit einer Veränderung von etablierten Prozessen einher. Für eine solche Veränderung muss geworben werden, es müssen Anreize aber auch Anforderungen gestellt werden. Kurzfristiger Anreize kann zusätzliche Förderung von Datenmanagement-Arbeitspaketen in Projekten sein, wie sie beispielsweise in den Teilprojekten Informationsinfrastruktur eines Sonderforschungsbereiches der DFG zu finden ist. Auch ausgelobte Preise für erfolgreiches FDM können einen Anreiz darstellen.
Mit dem Einsatz von LIMS und ELNs und erweiterbaren Tools kann der Wissenschaftler bei der Aufbereitung von Forschungsdaten und abgeleiteten Ergebnissen bei der Manuskripterstellung unterstützt werden. Die Zusammenstellung von experimentellen Parametern, Peaklisten, Datenabgleich und Formatierung stellen eine wertvolle Hilfe und einen Zeitgewinn in Aussicht.
Langfristige Anreize können durch das Zusammenwirken mit Verlagen gesetzt werden So ist es vorstellbar, dass eingereichte Publikationen, mit in Daten-Repositorien hinterlegten Datensätzen, schneller bearbeitet werden als solche ohne verfügbare Daten. Für Gutachter stellt die Kombination aus verfügbaren Forschungsdatensätzen und unterstützenden Tools ferner eine Erleichterung im Bewertungsprozess einer wissenschaftlichen Arbeit dar. In Kombination mit Guidelines und Policies für die Veröffentlichung von Forschungsdaten kann ein intelligentes Repositorium eine Validierung der Daten hinsichtlich Vollständigkeit und Plausibilität vornehmen. Existierende Beispiele sind hier STRENDA DB in Kombination mit den STRENDA Guidelines, die Plausibilitätsprüfung der Signalzuordnungen in NMRShiftDB2 und die NMReDATA-Initiative. Mit Verfügbarkeit einer kritischen Masse an Daten in einer NFDI4Chem ergeben sich langfristig Mehrwerte durch die Nutzung verfügbarer Forschungsdaten für Simulationen, Vorhersagemodelle oder Syntheseplanungen. Die Nachnutzungsszenarien sollten dabei so ausbalanciert werden, dass es sowohl Anreize für die Bereitstellung von Daten als auch die Datennutzung gibt.

Veröffentlichte Forschungsdatensätze werden langfristig die Sichtbarkeit der eigenen Forschung erhöhen, diese könnte durch einen Score zum Umfang von Datenpublikationen ausgedrückt werden. Schon heute werden analytische Daten in den Referenzdatenbanken

SciFinder und Reaxys nachgewiesen und mit den Zeitschriftenpublikationen bzw. Supplementary Materials verknüpft. Dies kann leicht um eine DOI-Verlinkung eines Forschungsdatensatzes in einem Repository erweitert werden. Für einen Übergang zu einer verpflichtenden Nutzung einer nachhaltigen Forschungsdateninfrastruktur sollte letztendlich ein umfassendes FDM und die Nutzung einer NFDI eine wichtige Komponente bei Entscheidungen der Forschungsförderung sein, einhergehend mit der Berücksichtigung für die Reputation der wissenschaftlichen Arbeit. Dies gilt auch für die Aufnahme einer Empfehlung besser noch Verpflichtung zu einer parallelen Publikation von Forschungsdaten und die Verknüpfung von Manuskript- und Repositoriendaten über persistente Identifier wi durch die Richtlinien der Verlage.

7. Internationale Vernetzung: Standards und Policies

These: Der Aufgabenbereich der NFDI muss die Zusammenarbeit mit international agierenden Gremien einschließen, um Prozesse zu reflektieren, die nur im internationalen Kontext erarbeitet werden können. Diskussionen zu Daten-Standards, Guidelines und Policies für die Datenpublikation oder Veränderungen des Publikationsprozess sind international zu führen und national zu implementieren.

Die NFDI kann nationale Fragen der Infrastruktur adressieren, ist aber z.B. in den Bereichen Standards und international etablierte Dokumentationsprozesse nicht unabhängig handlungsfähig. Dies gilt insbesondere, da die Publikation von Forschungsergebnissen internationalen Konventionen entsprechen muss. Diskussionen zu Daten-Standards, Guidelines und Policies für die Datenpublikation oder Veränderungen des Publikationsprozesses sind international zu führen und national zu implementieren.
Das Ziel ist es daher, innerhalb von NFDI4Chem ein/mehrere Gremium/Gremien zur internationalen Zusammenarbeit auf geeigneten Fach-Ebenen (z.B. bzgl. Datenformate, Standards, Policies) zu etablieren. Einige Vertreter des Fachgesprächs sind bereits in internationale Gremien eingebunden. Vertreter der NFDI4Chem sollen in Zukunft nationale Entwicklungen mit den Anforderungen und Initiativen des Committee on Publications and Chemoinformatics Data Standards, (CPCDS) und der CRDIG Chemistry Research Data Interest Group der RDA (DIGChem Project) koordinieren.

8. Zusammenfassung

Die Defizite der Chemie im Forschungsdatenmanagement und das Fehlen von etablierten Repositorien sind für die NFD4Chem als große Chance zu sehen. Erfahrungen anderer Disziplinen können früh berücksichtigt werden, viele Herausforderungen und Brennpunkte sind bekannt. Infrastruktureinrichtungen, Fachgesellschaften, Forschungsförderer und Wissenschaftler können für die Chemie gemeinsam den digitalen Wandel im Umgang mit Forschungsdaten von der Entstehung bis zur Publikation gestalten. Unter der Voraussetzung einer gesicherten, langfristigen finanziellen Unterstützung des FDM kann eine nationale Infrastruktur von Repositorien aufgebaut werden und es können gemeinsam

Richtlinien zu deren Nutzung, zur Datensicherheit und zum Rollen- und Rechtemanagement entwickelt werden. Eine breit aufgestellte Diskussion begleitet die Digitalisierung von Workflows in der Chemie; von der Datenerhebung, Analyse, Interpretation bis zur Ergebnispräsentation. Diese Digitalisierung im Labor wird unterstützt durch die niedrigschwellige Nutzung sowohl der NFDI als auch Tools wie z.B. ELNs. Flankierende Vorgaben und Anreizsysteme in der Forschungsförderung und eine Wertschätzung des FDM für die wissenschaftliche Reputation verankern das Datenmanagement in der Fachcommunity. In Repositorien offen verfügbare Forschungsdaten erhöhen so Reproduzierbarkeit und Nachprüfbarkeit von Forschungsergebnissen, sie vermeiden redundante Forschungsprojekte und bieten die Chance auf eine Nachnutzbarkeit von Daten zum Zwecke einer perpetuierten Wissensgenerierung. Alle beteiligten Stakeholder sind eingeladen sich an der Diskussion und der Gestaltung von Lösungsansätzen zu beteiligen.