Konzeptstudie
Vernetzte Primärdaten-Infrastruktur für den Wissenschaftler-Arbeitsplatz in der Chemie

2010

TIB | Technische Informationsbibliothek

FIZ CHEMIE BERLIN
Fachinformationszentrum Chemie GmbH

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Inhaltsverzeichnis

Hintergrund

Konzeptstudie "Vernetzte Primärdaten-Infrastruktur für den Wissenschaftler-Arbeitsplatz in der Chemie"

In dem Projekt „Vernetzte Primärdaten-Infrastruktur für den Wissenschaftler-Arbeitsplatz in der Chemie“, das von der DFG vom 01.06.2009 bis zum 31.05.2010 mit Personal- und Sachmitteln gefördert wurde, ist von der Technischen Informationsbibliothek (TIB), dem Fachinformationszentrum Chemie (FIZ Chemie) und dem AK Fels des Department Chemie der Universität Paderborn (AK Fels) die vorgelegte Konzeptstudie erstellt worden. Zielsetzung des Projektes war es, die Grundlagen zur Schaffung einer vernetzten Primärdateninfrastruktur zu erarbeiten, damit Primärdaten aus der Chemie

- in einem zentralen Datenspeicher dauerhaft und qualitätsgesichert gespeichert,
- durch DOI-Vergabe (DOI: digital object identifier) zitierfähig und verlinkbar,
- zugänglich und
- gezielt suchbar gemacht

werden können. Bisher gibt es im Gegensatz zu anderen Disziplinen wie der Kristallographie oder den Erde und Umweltwissenschaften kaum Möglichkeiten primäre Forschungsdaten in der Chemie persistent in einem Repositorum zu speichern und einer Nachnutzung zuzuführen. Da die Chemie keine homogene Disziplin ist, sondern aus vielen Teildisziplinen besteht, die z. T. stark abweichende Anforderungen an die Forschungsdaten und damit verbundenen Metadaten haben, beschränkt sich die Studie vorerst auf die Forschungsdaten der synthetisch arbeitenden Wissenschaftler der anorganischen und organischen Chemie.

Ziel ist es, die Daten der synthetischen Chemie dauerhaft und qualitätsgesichert zu speichern. Damit die Forschungsdaten auch zitierfähig und verlinkbar werden, müssen sie durch Persistent Identifier (PI) eindeutig gekennzeichnet werden. Seit 2005 vergibt die TIB DOI-Namen als PI zur Referenzierung von Forschungsdaten. Durch die DOI-Vergabe werden die Forschungsdaten such- und findbar sowie zitierfähig (siehe http://www.tib-hannover.de/de/die-tib/doi-registrierungsagentur/).

Der bisherige Umgang mit Forschungsdaten in der Chemie beinhaltet keine allgemein anerkannten Standards hinsichtlich einer Nachnutzbarkeit oder langfristigen Verfügbarkeit. Überwiegend existiert keine Qualitätssicherung, keine gesicherte Langzeitarchivierung, kein gesicherter Nachweis sowie keine Erschließung der Forschungsdaten und somit keine Datensicherheit.

Im Rahmen dieser Konzeptstudie wird in Kooperation mit allen relevanten Beteiligten der wissenschaftliche Workflow von Forschungsdaten in der Chemie analysiert und die Anforderungen an Prozesse und Strukturen innerhalb einer vernetzten Forschungsdaten-infrastruktur für den Wissenschaftler-Arbeitsplatz in der Chemie identifiziert.

Die vorgelegte Studie identifiziert die konzeptionellen Fragen soweit und zeigt Lösungsansätze auf, so dass in Folgeprojekten die prototypische Realisierung verwirklicht werden kann. Die Konzeptstudie kann hierbei einen Modellcharakter für andere Fachgebiete haben.

Als Grundlage für die Konzeptstudie wurde zusammen mit der Universität Paderborn (Arbeitskreis Prof. Fels) und dem FIZ Chemie ein Fragebogen erstellt und an synthetisch arbeitende Wissenschaftler verschickt, um herauszufinden, ob ein allgemeines Interesse besteht, Daten persistent und für alle verfügbar zu speichern.

Die hier vorliegende Konzeptstudie ist in drei Kapitel aufgeteilt, entsprechend der Struktur des Projektes:

- A Forschungsdaten in wissenschaftlichen Prozessen
- B Metadaten
- C (Zentraler) Datenspeicher

Zusammenfassung

Forschungsdaten in der Chemie

[*… The fabric of science is changing, driven by a revolution in digital technologies that facilitate the acquisition and communication of massive amounts of data. This is changing the nature of collaboration and expanding opportunities to participate in science. If digital technologies are the engine of this revolution, digital data are its fuel. ...*]

Vincent S. Smith, Data publication: towards a database of everything, BMC Research Notes 2009, 2, 113.

Forschungsdaten können als „Artefakte des wissenschaftlichen Prozesses“ verstanden werden. Sie müssen daher immer im fachspezifischen Kontext ihrer Datenentstehung betrachtet werden. Bereits innerhalb einzelner Fachdisziplinen gibt es beispielsweise eine große Heterogenität der Daten und zugehörigen Metadaten. Täglich fallen in der Chemie enorme Mengen an Forschungsdaten bei Experimenten an, beispielsweise in der Analytik (NMR, MS, UV/VIS, IR, X-Ray etc.). Diese Forschungsdaten sind die Grundlage jeglicher wissenschaftlicher Arbeit. Ausgehend vom Experiment durchlaufen Forschungsdaten viele, dem Wissenschaftler bekannte Stadien, die letztendlich als Erkenntnisgewinn in einer wissenschaftlichen Publikation münden. Danach verliert sich der bis dahin so klare Weg der Forschungsdaten, was deren Dokumentation, langfristige Speicherung oder Nachnutzbarkeit für andere Wissenschaftler betrifft. Eine solche (langfristige) Sicht auf Forschungsdaten hat bisher bei Wissenschaftlern oft nur eine untergeordnete Bedeutung. Die Anforderung für einen gesichertern, langfristigen Zugangs zu Forschungsdaten formuliert Vincent S. Smith [1]in drei Sätzen: “*Make it easy – developing a cyberinfrastructure*”, “*Make it citable – motivating data publication through peer recognition*” und “*Make it useful – moving beyond data archival*” mit den Kernaussagen:

- Es muss dem Wissenschaftler so einfach wie möglich gemacht werden, seine Forschungsdaten zu dokumentieren und in eine Infrastruktur zur langfristigen Speicherung einzubringen.
- Forschungsdaten müssen als Anreizsystem für den Wissenschaftler zitierbar sein und somit die Sichtbarkeit seiner Arbeit erhöhen.
- Es müssen Mehrwertdienste für veröffentliche Forschungsdaten zur Verfügung stehen.

Eine öffentliche Bereitstellung von Forschungsdaten oder deren eigenständige Publikation haben in der Chemie bisher noch keine breite Umsetzung gefunden. Eine prominente Ausnahme stellen die Cambridge Structural Database (CSD)[2] und die

Inorganic Crystal Struture Database[3] (ICSD) dar, in denen bereits seit Beginn der siebziger Jahre bibliographische, chemische und kristallographische Daten, die mittels Röntgenstrukturanalyse oder Neutronenbeugung untersucht wurden, erfasst werden. In der Datenbank können Strukturdaten, die zu einer wissenschaftlichen Publikation gehören, aber auch Daten in Form einer „personal communication", hinterlegt werden. In der Datenbank abgelegte Strukturen erhalten eine Deposition Number, die als Identifier in Publikationen genutzt wird. Der Prozess der Datenhinterlegung ist peer-reviewed. Seit Anfang der neunziger Jahre erfolgt die Hinterlegung der Röntgenstrukturdaten im standardisierten CIF-Datei Format. Eine jüngere Entwicklung im Kontext Open Source und Open Data ist die NMR-Spektrendatenbank NMRShift-DB [4] . In dieser Datenbank können chemische, spektroskopische und bibliographische Daten zu einem Molekül hinterlegt werden. Im Gegensatz zu den später diskutierten ursprünglichen Forschungsdaten, werden in dieser Datenbank jedoch nicht die gemessenen Rohdaten hinterlegt, sondern die Daten, die sich aus der Auswertung der Rohdaten ergeben. Für NMR-Daten, beispielsweise ^{13}C-Daten, bedeutet dies ein Speichern von Peaklisten. Ist die Verfügbarmachung solcher Spektrendaten grundsätzlich zu begrüßen, stellt dies doch nur einen ersten Schritt dar, um gemessene, experimentelle Daten einer breiten wissenschaftlichen Öffentlichkeit zur Verfügung zu stellen.

A Forschungsdaten in wissenschaftlichen Prozessen

Die Analyse des wissenschaftlichen Workflows von Forschungsdaten erfolgte exemplarisch mittels einer Umfrage unter Wissenschaftlern der anorganischen, organischen und analytischen Chemie. Auf diese Weise wurden die Anforderungen an Prozesse und Strukturen innerhalb einer vernetzten Forschungsinfrastruktur für den Wissenschaftler-Arbeitsplatz in der Chemie identifiziert. Die Beschränkung auf diese Bereiche der Chemie war vorgenommen worden, da hier vergleichsweise standardisierte Messdaten in Form von spektroskopischen, spektrometrischen und Röntgenstruktur-Daten anfallen. Exemplarisch wurde der Umgang mit NMR, IR, UV/VIS, MS, X-Ray sowie HPLC/GC/GPC abgefragt. Zusätzlich erlaubte der Fragebogen Anregungen und Kommentare, die aufschlussreiche Hinweise zum Meinungsbild der Wissenschaftler lieferten. Insgesamt führte die Umfrage zu 386 Rückläufen aus 106 Arbeitskreisen. Damit haben circa 20% der an deutschen Hochschulen tätigen Arbeitsgruppen der oben erwähnten Fachbereiche teilgenommen. Die ausführliche Beschreibung der Vorgehensweise und der Ergebnisse liefert der Umfragebericht.

Quintessenz der Umfrage

- Während Primärdaten überwiegend als ASCII-Dateien oder in einem proprietären Format vorliegen, werden die analysierten Daten in verschiedenster Form als Daten oder Graphik-Files dezentral elektronisch gespeichert sowie von vielen Befragten auch heute noch ausgedruckt.
- Mit JCAMP-DX liegt ein Datenformat vor, mit dem zumindest für einige der Analysemethoden gemessene Daten archiviert und wiederverwendet werden können, aber weder ist dieses Datenformat bislang allgemein bekannt, noch lässt es sich derzeit routinemäßig fehlerfrei einsetzen.
- Etwa die Hälfte der Befragten wurde bereits mit der Situation konfrontiert, dass Daten nicht zugänglich waren aufgrund von Problemen der Lesbarkeit von Datenformaten oder allgemeinem Datenverlust.
- Die verbesserte Zugangsmöglichkeit zu den eigenen Daten und denen anderer Forschergruppen ist der größte Antrieb einer Referenzierung von Daten mit DOIs zuzustimmen.
- Die große Mehrheit der Befragten befürwortet eine persistente Referenzierung ihrer spektroskopischen und spektrometrischen Daten.
- Es besteht eine gewisse Unsicherheit in Bezug auf Missbrauchmöglichkeiten bei der Veröffentlichung vollständiger Datensätze.
- Daten werden immer in Verbindung mit einer wissenschaftlichen Publikation gesehen. Eine eigenständige Datenpublikation ist derzeit nicht vorstellbar.

Die zusätzlichen Kommentare untermauerten die allgemeinen Aussagen der Umfrage. Diese zeigten auch, dass eine Auseinandersetzung mit dem Thema der Datenreferenzierung stattfindet. Erwähnenswert ist die Befürchtung, dass der erforderliche Aufwand zur Datenablage zu Lasten der eigentlichen wissenschaftlichen Arbeit erfolgt. Interessant ist auch der von Teilnehmern der Umfrage aufgeworfene Aspekt des optimalen Zeitpunkts für die DOI-Vergabe im Lebenszyklus der Daten sowie auch der Vorschlag des Einsatzes von Forschungsdaten in der Lehre.

Erweiterung des Publikationsprozesses

Am 21.4.2010 fand in der TIB im Rahmen des Projektes ein Workshop “Publication of research data“ statt mit Vertetern der Projektpartner, der Arbeitsgruppe Chemie – Information – Computer der GDCh und verschiedener chemierelevanter Verlage (Wiley-VCH, Georg Thieme Verlag und Elsevier), um gemeinsame Optionen zur öffentlichen

Bereitstellung von Forschungsdaten zu diskutieren. Die im experimentellen Teil von Publikationen enthaltenen Daten zur Charakterisierung von Verbindungen stellen üblicherweise nur einen fragmentarischen Auszug aus den Originaldaten bzw. aus den daraus generierten Spektren dar. Zwar werden Publikationen häufig mit sogenannter Supporting Information angereichert. Doch auch dies sind meist nur PDF-Dokumente mit Repräsentationen ausgewerteter Forschungsdaten.

In dem an der TIB Hannover durchgeführten DFG-Projekt „Publikation und Zitierfähigkeit wissenschaftlicher Forschungsdaten“ wurde eine Infrastruktur zur Registrierung von DOI Namen und URNs für wissenschaftliche Datensätze geschaffen und erfolgreich getestet. Mit dem System wurden an der TIB bereits über 650.000 Datensätze aus dem Bereich der Geowissenschaften mit persistenten digitalen Identifikatoren versehen. Seit 2009 existiert eine Kooperation zwischen der TIB und dem Georg Thieme Verlag, in der exemplarisch die Publikation von chemischen Daten, die einem Artikel zugrunde liegen, in den Publikationsworkflow eines Verlages eingebunden wird.[5] Die beispielhafte Umsetzung eines mit Forschungsdaten angereicherten Artikels kann bei der Publikation von K. Jarowicki, C. Kilner, P. J. Kocienski, et.al. (http://dx.doi.org/10.1055/s-2008-1067226) betrachtet werden, (vgl. Abb 16)

Nur in dieser Form sind die zugrundeliegenden Daten einer wissenschaftlichen Publikation im Detail nachvollziehbar und untersützen den peer review Prozess. Die notwendige Erfassung und Bereitstellung von Forschungsdaten kann als Ergebnis der Arbeiten an der Konzeptstudie nur in einer gemeinsamen Anstrengung aller Beteiligter erfolgen: Wissenschaftler, Datenzentren, Fachgesellschaften, Verlage und Bibliotheken.

B Metadaten

Mit der Publikation von Forschungsdaten und deren Zitierbarkeit entstehen neue Anforderungen an Metadatenschemata. Neben den klassischen Metadaten, wie sie in Bibliothekskatalogen angewendet werden, sind weitere Metadaten notwendig, die Forschungsdaten spezifisch beschreiben können. Für Forschungsdaten ist es somit sinnvoll zwischen zwei Arten von Metadaten zu unterscheiden:

1. Metadaten, die Forschungsdaten innerhalb einer Sammlung bzw. eines Kataloges beschreiben. Sie sind am ehesten mit klassischen Katalogdaten zu vergleichen. Diese Metadaten sind die Grundlage, um Forschungsdaten so zu beschreiben, dass sie eindeutig zitiert werden können. Im Rahmen der Konzeptstudie werden solche Daten als *Externe* Metadaten bezeichnet.
2. Metadaten, die die Forschungsdaten auf fachspezifischer Ebene beschreiben. Diese Metadaten enthalten oftmals detaillierte Informationen zu Techniken, Methoden,

Parametern etc., die zum fachlichen Verständnis der Forschungsdaten notwendig sind. Im Rahmen der Konzeptstudie werden solche Daten als *Interne* Metadaten bezeichnet.

In der Konzeptsudie werden verschiedene Ansätze zur Definition von Metadatenschemata für die Beschreibung von Forschungsdaten vorgestellt und analysiert.

Bei der Analyse verschiedener fachspezifischer Metadaten wurde eine erhebliche Komplexität und Diversität festgestellt. Neben unterschiedlichen Formaten und Bezeichnern werden generell je nach Teilgebiet der Chemie sowie verwendeter Technik und Methode unterschiedliche Metadaten erzeugt. Diese Komplexität lassen eine Abstufung bzw. weitere Differenzierung der internen Metadaten für chemische Forschungsdaten ratsam erscheinen.

Grundvoraussetzung für die einfache Archivierung von publizierten Datensätzen ist die weitgehend automatische Extraktion der für eine Archivierung und Wiederverwendung benötigten Daten. Das betrifft in erster Linie die Metadaten der Messung, gekoppelt mit den Daten zur Generierung des Spektrums. Auch wenn mit dem JCAMP-DX-Format bereits seit Längerem ein akzeptiertes Austauschformat für wissenschaftliche Daten gegeben ist, ist dieses Format weder allgemein verbreitet bzw. bekannt noch gibt es in JCAMP-DX Files eine standardisierte Auflistung der Metadaten. Auch der Umfang der Metadaten variiert beispielsweise sowohl hersteller- und geräteabhängig, als auch in ganz besonderem Maße in Abhängigkeit von der Messmethode.

Hier besteht daher primärer Handlungsbedarf, um geräte- und methodenunabhängig die notwendigen Daten und Angaben gewinnen zu können. Darüber hinaus ist eine herstellerunabhängige Wiederverwendung bestehender Datensätze zurzeit ohne Weiteres nicht gegeben, d.h. dass selbst beispielsweise JCAMP-DX Files von NMR-Spektren nicht in allen Fällen herstellerunabhängig gelesen und in das entsprechende Spektrum umgesetzt werden können. Hier wird es daher in Zukunft notwendig sein, verschiedene Szenarien zu erarbeiten und zu erproben, um z.B. die Regenerierung eines Spektrums unabhängig von Gerät und Hersteller auch nach Jahren zu ermöglichen.

C (Zentraler) Datenspeicher

Hierarchisch gesehen gibt es unabhängig von den wissenschaftlichen Verlagen im Bereich der Forschungsdatenpublikation drei Ebenen:

1. Den Wissenschaftler, der für die Datenerhebung zuständig ist
2. Das Datenzentrum als Einrichtung, die die Daten dauerhaft verfügbar macht
3. Die Informationseinrichtungen, die die Daten nachweisen bzw. im DataCite-Kontext durch DOI-Vergabe zitierfähig machen.

Abb. 1: Die Akteure im Bereich Datenpublikation in Ebenen dargestellt

Die entscheidende Lücke stellt in diesem Modell in Deutschland wie auch in anderen Ländern die Ebene der Datenzentren dar. Die Etablierung von Datenzentren wird in den meisten Disziplinen nur auf Basis entsprechender, allgemein gültiger Policies stattfinden können. Bezüglich einer universalen „Forschungsdaten Policy“ sind in den letzten Jahren im Rahmen des Handlungsfeldes „Forschungsprimärdaten" der Schwerpunktinitiative „Digitale Information“ der Allianz der deutschen Wissenschaftsorganisationen und der AG Forschungsdaten der Kommission Zukunft der Informationsinfrastruktur (KII) Konzepte zur Entwicklung einer Deutschen Forschungsdatenpolicy entwickelt worden.

Die vorliegende Konzeptstudie befaßt sich primär mit der technischen Realisierung einer Forschungsdaten-Infrastruktur.

Um die langfristige Verfügbarkeit der Daten sicherzustellen, muss an zwei Stellen angesetzt werden. Zum einen muss der physische Erhalt der gespeicherten Datenobjekte auf einem Speichermedium gesichert werden (Bitstream Preservation). Zum anderen müssen Techniken und Strategien entwickelt werden, die eine dauerhafte Interpretierbarkeit und Nachnutzbarkeit der Archivbestände gewährleisten. Letzteres ist gerade in der Chemie eine anspruchsvolle Aufgabe, da es aufgrund der vielfältigen Messverfahren und der Existenz diverser Gerätehersteller für eine Messmethode ein breites Spektrum an proprietären Dateiformaten gibt. Zudem zeigt sich eine Verfahrensabhängigkeit der für den Erhalt der Dateninterpretierbarkeit wichtigen Metadaten.

Der Aufbau einer technischen Infrastruktur für ein Datenarchiv entsprechend des in der Studie erarbeiteten Konzepts ist mit überschaubarem Aufwand realisierbar. Für die Hardware-Architektur empfiehlt sich eine dezentrale Struktur, um einerseits die Kosten zu reduzieren und andererseits am Knowhow des externen Dienstleisters teilzuhaben und zeitnah mit dem Aufbau beginnen zu können. Im Softwarebereich sollte zwecks Minimierung des eigenen Programmieraufwands weitgehend auf bereits vorhandene Produkte und Applikationen zurückgegriffen werden, zumal es im realen Einsatz des Datenspeichers im wissenschaftlichen Betrieb weiteren Anpassungs- und Optimierungsbedarf geben wird.

Die verhältnismäßig größere Herausforderung ist der Aufbau einer organisatorischen Infrastruktur für die Langzeitarchivierung. Das Archiv sollte auf Basis eines Funktionsmodells wie OAIS aufgebaut werden. Für die Funktionseinheiten Administration, Datenmanagement, Ingest, Access Preservation Planning und die untergeordneten Subprozesse sind Workflows zu entwickeln. Alle Prozesse müssen beschrieben und dokumentiert werden, Policies sind auszuarbeiten.

Das technische Konzept ist weitgehend ausgereift, so dass eine Realisierung der technischen Infrastruktur anhand eines ersten Prototyps direkt im Anschluss an die Konzeptphase möglich wäre. Ausgewählte Arbeitskreise könnten einen Zugang zum Datenspeicher erhalten und erste Forschungsdaten ablegen. Gleichzeitig muss die organisatorische Konzeptionierung mit dem Ziel einer Überführung des Datenspeichers in ein vertrauenswürdiges Langzeitarchiv weiter fortgeführt, verfeinert und schließlich realisiert werden.

Realisierung

Als Ergebnis der Umfrage und des Verlagsworkshops ist zu empfehlen, als Grundlage für eine mögliche Realisierung von einem prozessorientierten Ansatz auszugehen, in dem der klassische Weg der Publikation von wissenschaftlichen Ergebnissen in Aufsätzen durch eine Publikation vollständiger Datensätze ergänzt wird.

Dabei ist es wichtig, die öffentliche Bereitstellung der Daten als Dienstleistung für die Wissenschaftler zu betrachten und nicht als Angebot für die Verlage. Daraus resultiert auch, dass zunächst kein Embargo-Prozess erforderlich ist, da die Daten mit der Publikation veröffentlicht würden.

In Anlehnung an das Beispiel von Thieme sollten Daten im proprietären Format als Beleg ebenso abgelegt werden wie Daten in Austauschformaten wie JCAMP-DX oder CDF damit sie problemlos ohne Zuhilfenahme von Spezialprogrammen geöffnet werden können. Darüber hinaus sollten ohne Limitierung alle weiteren vom Wissenschafter erwünschten Daten uneingeschränkt als supporting information abgelegt werden können.

Das Ablegen von Publikationsdaten im Datenzentrum entbindet den Wissenschaftler zudem von der Aufgabe, selbst seine Daten nach GLP-Richtlinien mind. 10 Jahre aufzubewahren.

Im Umgang mit verschiedenen Disziplinen werden allerdings jeweils ganz individuelle Strukturen und Daten-Philosophien vorgefunden. Gerade in der Chemie erweist sich das Datenmanagement aufgrund besonderer technischer Voraussetzungen als sehr komplex. Es existiert ein breit gefächertes Spektrum an Messmethoden und -verfahren, die unterschiedliche Verbreitung gefunden haben. So gibt es einerseits Messverfahren, wie z. B. die NMR-Spektroskopie, die in nahezu jedem Fachbereich Chemie zur Anwendung kommen. Andererseits finden in einigen Teilgebieten der Chemie eher seltene, sehr spezielle Verfahren Anwendung, so dass nur wenige Institute entsprechende Messgeräte besitzen, z.B. die Partikelgrößenbestimmung mittels einer Scheibenzentrifuge (Biotechnologie, Polymerchemie). Im Extremfall gibt es auch individuelle, im Arbeitskreis einer Universität konzipierte Messapparaturen mit eigenentwickelter Software für die Datenauswertung, z. B. speziell angepasste Kalorimeter in der Polymeren Reaktionstechnik.

Auch ist das Prinzip der Nachnutzung von bereits existierenden Daten in der Chemie nicht sehr verbreitet, vielmehr steht bei der Mehrzahl der chemischen Veröffentlichungen die Reproduzierbarkeit des beschriebenen Experiments an erster Stelle. In der präparativen Chemie erfüllt die Publikation von Daten heute noch in erster Linie eine Belegfunktion im Sinne der Regeln zur Guten Wissenschaftlichen Praxis, was ein Kernargument für die weiteren Arbeiten sein sollte.

Generell kann man nicht davon ausgehen, dass ein System oder Prozess entwickelt werden kann, der trivial auf alle Fachbereiche übertragen werden kann. Vielmehr muss es Ziel sein, in Zusammenarbeit mit den Fachgesellschaften disziplinspezifische Ansätze zu entwickeln, die dann prototypisch realisiert werden können. Dabei wird es Disziplinen geben, die wie die Geowissenschaften eine zentrale Datenzentrenstruktur benötigen, aber auch Disziplinen, die unter Verwendung allgemeingültiger Standards individuelle Lösungen in Form von verteilten Repositorien betreiben.

A - Forschungsdaten in wissenschaftlichen Prozessen

(Arbeitspaket 1)

A 1. Umfrage zu Forschungsdaten unter Wissenschaftlern

Zur Vorbereitung der Arbeiten in diesem Projekt wurde eine Umfrage unter Wissenschaftlern durchgeführt. In erster Linie wurden Wissenschaftler der anorganischen und organischen präparativen Chemie an deutschen Hochschulen befragt, aber es haben einige Wissenschaftler anderer Fachrichtungen sowie 21 Vertreter der Industrie geantwortet. Der Fragebogen dazu wurde vom Arbeitskreis Fels entwickelt, mit den anderen Mitwirkenden der Konzeptstudie abgestimmt und via E-Mail an 58 universitäre Arbeitskreise verschickt, sowie auf dem Wissenschaftsforum der GDCh verteilt. Er enthielt multiple-choice Fragen zum persönlichen Status, zur Erzeugung und Verwaltung von Forschungsdaten, zu Datenbanken und zur Vorgehensweise bei Veröffentlichungen sowie Meinungsumfragen zur Publikation elektronischer Spektren gestellt. Darüber hinaus war es möglich, Anregungen und Kommentare zu verfassen. Insgesamt wurden 386 Fragebögen wieder abgegeben, von denen 328 (85%) vollständig bearbeitet worden sind, bzw. sogar 90%, wenn berücksichtigt wird, dass die Industrievertreter die Statusfrage nicht ausfüllen konnten.

1.1 Resümee der Umfrage

Die allgemeine Meinung der Wissenschaftler zu diesem Projekt reichte von breiter Akzeptanz bis hin zu vollkommener Ablehnung. Es zeigten sich aber auch berechtigte Bedenken der Wissenschaftler hinsichtlich des Mehraufwandes, den ein erweitertes Datenmanagement mit sich führen würde. Die Bedenken richten sich dabei auf den zusätzlichen Zeit- und Verwaltungsaufwand. So sorgte man sich darum, dass durch diesen Mehraufwand die Ergebnisse produzierende Forschung eingeschränkt werden könnte.

Positive Resonanz erreichte man vor allem bei denjenigen, die sich für die Charakterisierung von neuen Substanzen interessieren. Zitat: „Ich halte die Archivierung von spektroskopischen Daten für die Charakterisierung von neuen Verbindungen für unabdingbar. Den als Trend zu verzeichnenden Verzicht auf die Veröffentlichung von Elementaranalysen und spektroskopischen Daten neuer Verbindungen in Publikationen halte ich für ausgesprochen ungut und er kann nicht nur zu Qualitätseinbußen führen, sondern darüber hinaus Betrugsversuchen Vorschub leisten. In diesem Zusammenhang ist z. B. der schwer zu fälschende "Fingerprintbereich" von Infrarotspektren von besonderer Bedeutung.“

Ein Umfrageteilnehmer stellte die Frage, wann ein DOI-Name vergeben werden soll. Damit verbunden ist die Frage, welche Daten zu welchem Zeitpunkt gespeichert und welche Daten überhaupt einen DOI-Namen erhalten sollen. Sinnvoll wäre es nach Meinung des Befragten, die Daten direkt bei der Erzeugung zu hinterlegen. Auf diese Weise dienen sie als

fälschungssicherer Beleg dafür, dass die Messung durchgeführt und nicht manipuliert wurde.

Interessant war auch die Bemerkung von der mangelnden Akzeptanz einiger Verlage, bestimmte „supplemental materials“ zu veröffentlichen, die nicht im „Focus der Journale stehen“ (Zitat).

Als Fazit ergibt sich aus der Umfrage, dass die Mehrheit der Anorganiker und Organiker durchaus eine persistente Referenzierung ihrer spektroskopischen und spektrometrischen Daten genau so befürwortet, wie sie das bereits seit vielen Jahren für Röntgenstrukturdaten akzeptiert hat. Die Abbildungen 2 und 3 machen das deutlich.

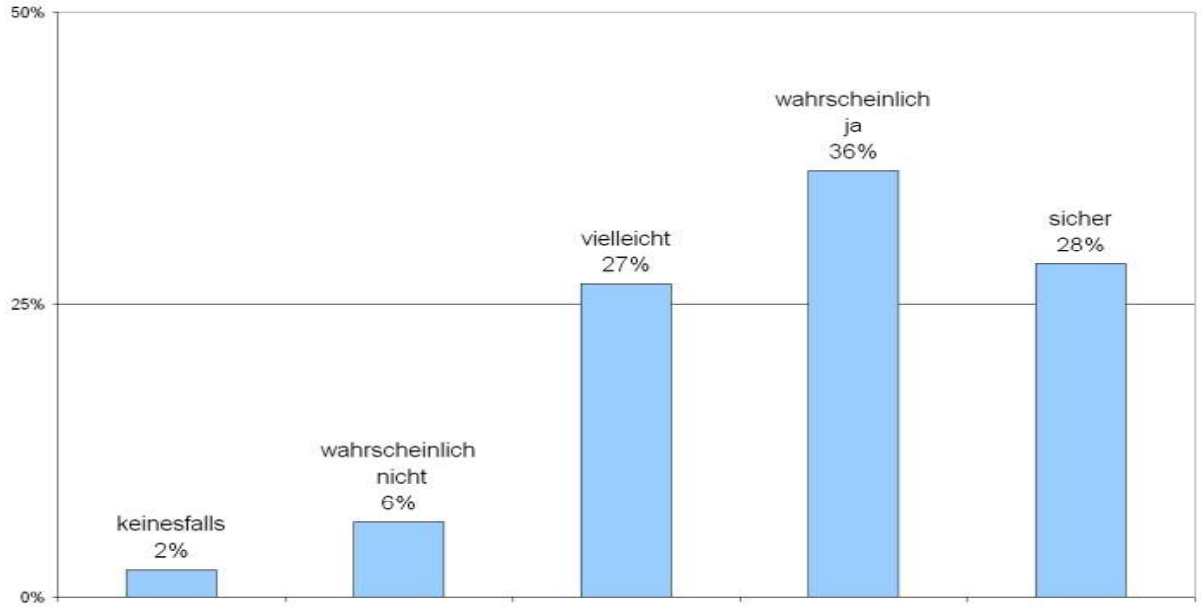


Abb. 2: Würden Sie vollständige, elektronisch zugängliche Spektren/Daten in Publikationen unterstützen?

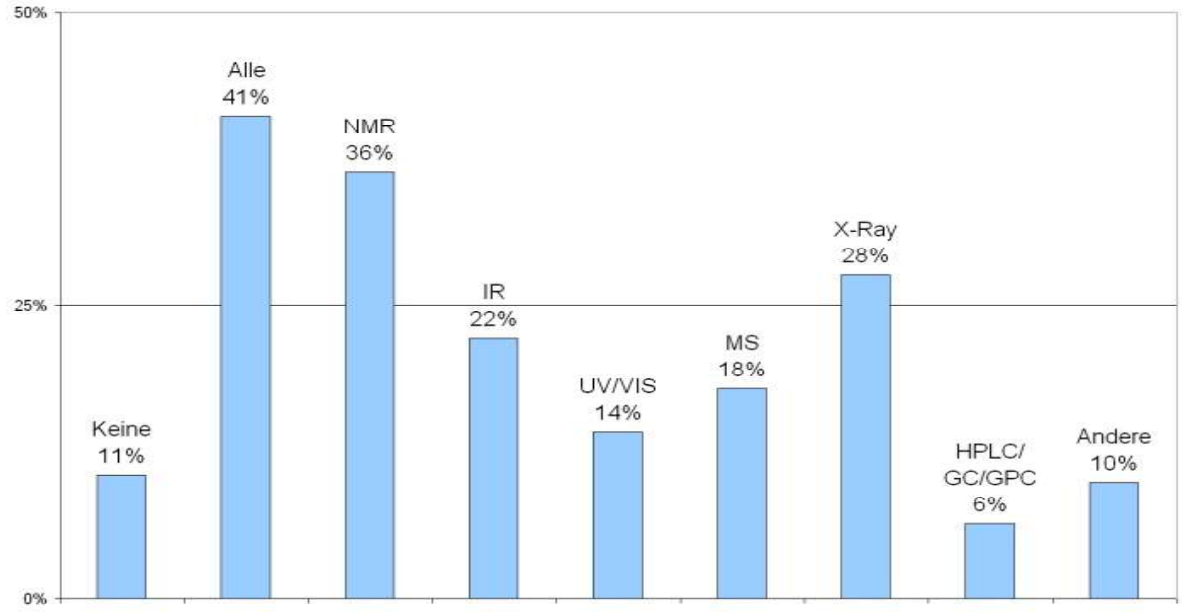


Abb. 3: Welche Messdaten sollten Ihrer Meinung nach mit einem eigenen DOI versehen und dadurch einfach und persistent zugänglich werden?

Einen Embargo-Prozess, d.h. eine Publikation der Daten mit zeitlich verzögerten Zugriffsrechten, halten viele für unnötig (Abb. 5). Ein Großteil der Befragten stellt bereits jetzt zusätzliches Material in Publikationen zur Verfügung und würde dies auch weiterhin tun (Abb. 4). Das Argument des Mehraufwands bei Bereitstellung von Forschungsdaten wir dadurch entkräftet.

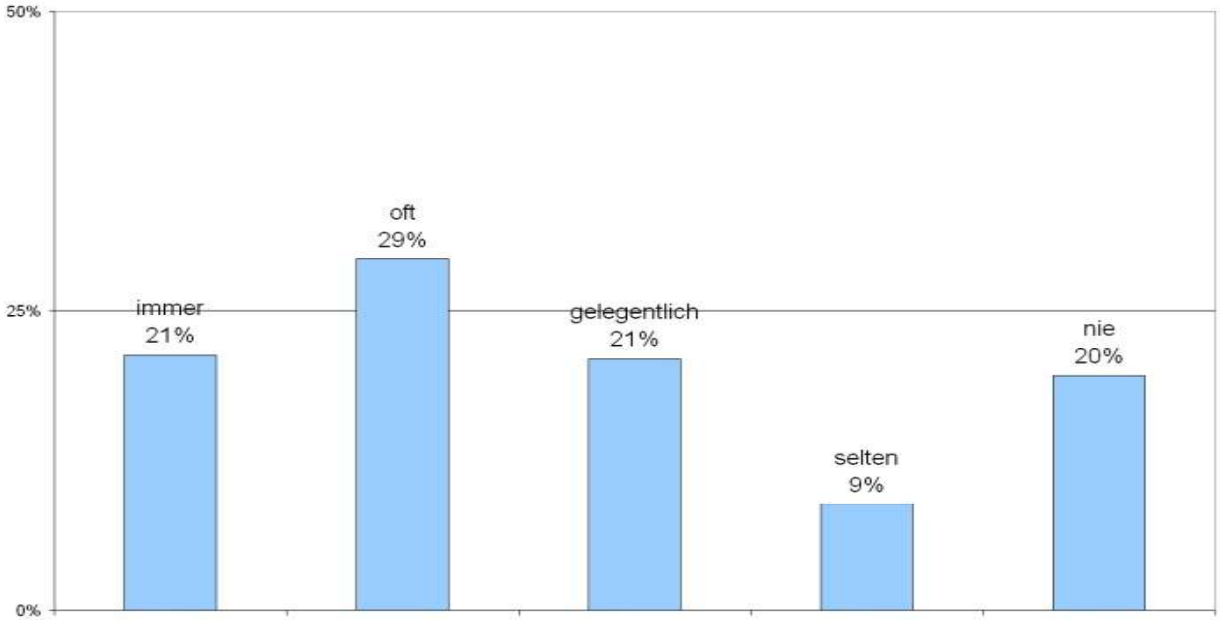


Abb. 4: Ergänzen Sie eine Publikation mit zusätzlichen Informationen oder Daten?

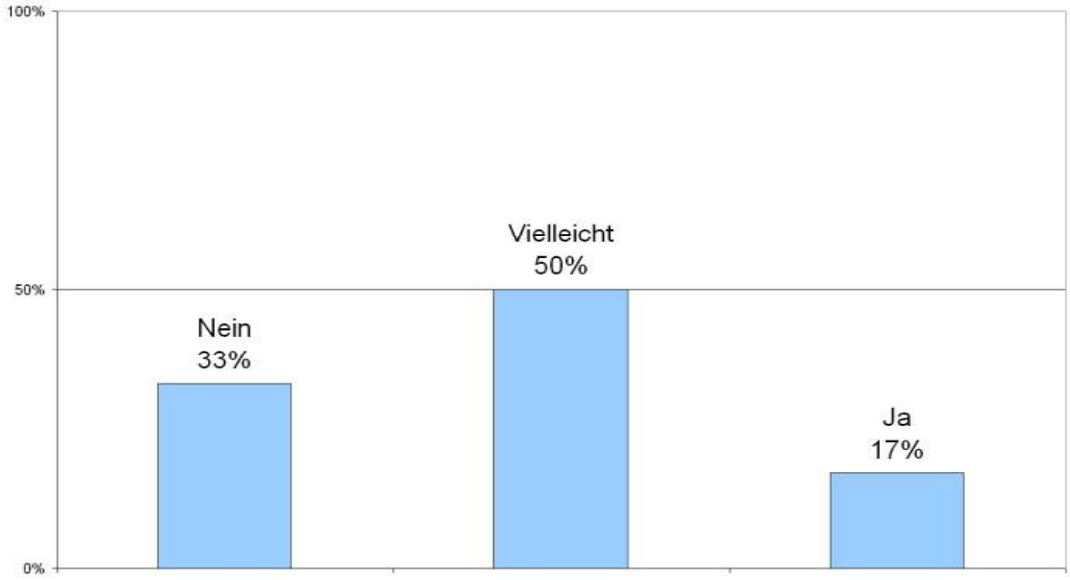


Abb. 5: Halten Sie bei der DOI-Vergabe einen Embargo-Prozess für notwendig?

Bei dem Stellenwert der spektroskopischen und spektrometrischen Daten in der Anorganischen und Organischen Chemie ist es überraschend, dass 40% der Befragten keine Spektrendatenbanken benutzen (Abb.7). Es ist aber darüber hinaus ersichtlich, dass die restlichen 60% nur in sehr begrenztem Umfang (nur je etwa 5%) kommerzielle Spektrendatenbanken verwenden, sondern sich verschiedener frei zugänglicher Datenbanken bedienen, in denen überwiegend anhand von Strukturangaben gesucht wird (Abb.6).

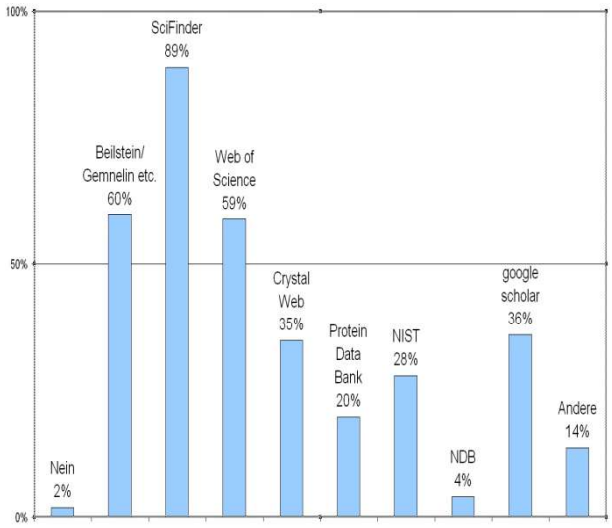


Abb. 6: Nutzen Sie Datenbanken?

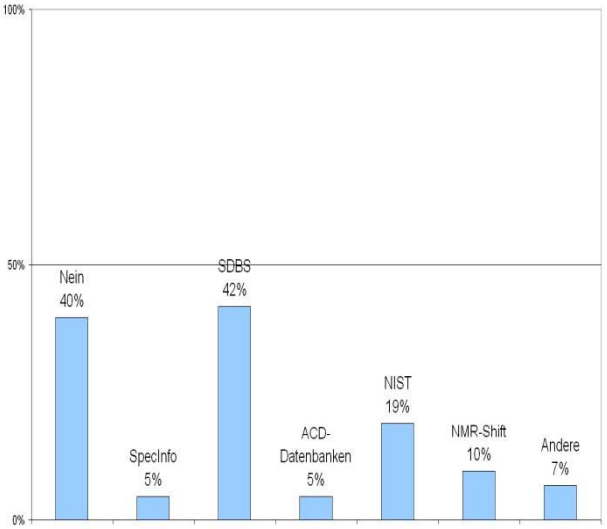


Abb. 7: Arbeiten Sie mit einer Spektren-Datenbank?

Bei der Suche in Datenbanken würden die Wissenschaftler bevorzugt mit Struktur/Substrukturen, Summenformeln, CAS-Nr. und Keywords suchen. Außer der Messmethode spielen andere Suchkriterien eher eine untergeordnete Rolle (Abb. 8).

Problematisch ist jedoch die Tatsache, dass nur wenigen Wissenschaftlern digitale Repostorien, Semantic Web, JCAMP-DX oder andere in diesem Zusammenhang erwähnenswerrte Begriffe bekannt sind (Abb.9).

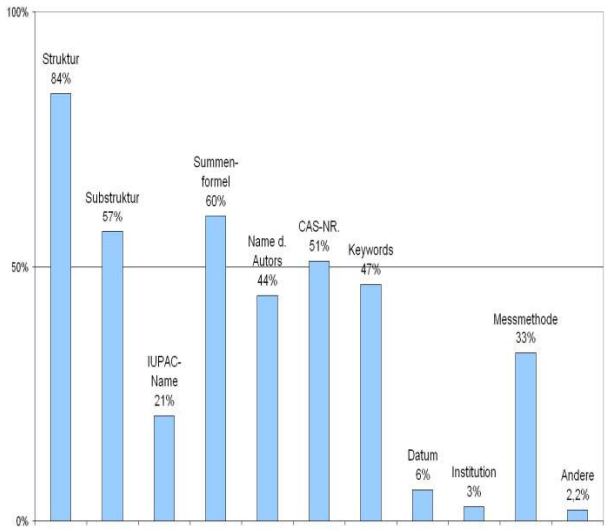


Abb. 8: Mittels welcher Suchkriterien würden Sie bevorzugt in einer Datenbank suchen?

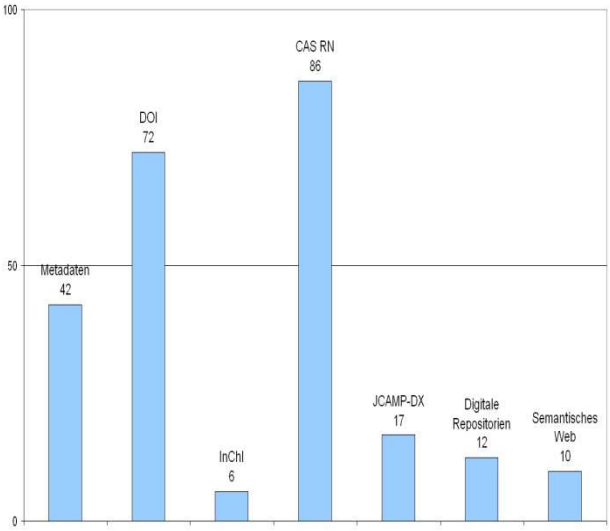


Abb. 9: Welche der folgenden Begriffe sind Ihnen geläufig?

Eine Wiederverwendung von Daten bringt zwei essentielle Erfordernisse mit sich. Zum Einen muss die physikalische Lesbarkeit der Daten garantiert sein, was aber bei professionellen Speichersystemen gegeben sein sollte. Im Vergleich dazu gibt in der Umfrage aber ein relativ großer Anteil (52%) zu erkennen, dass er schon mal mit Problemen bei der Wiederverwendung von gespeicherten Daten konfrontiert war. Hinzu kommt, dass das Spektrum der Formate, in denen zumindest die analysierten Daten abgelegt werden, zu groß ist, als dass man für alle eine Lesbarkeit auch noch in Jahrzehnten garantieren könnte. 18 % der Befragten geben in der Tat an, dass sie bereits Probleme mit Datenformaten hatten. (Abb 10.).

Die Umfrage zeigt, dass die Absicht besteht, Daten länger zu speichern, als es nach der guten Laborpraxis vorgeschrieben ist. Nach Verlassen des Instituts bzw. der Hochschule werden die gewonnenen Daten auf Zentralspeichern hinterlegt und/oder dem Vorgesetzten überreicht (Abb. 11). Häufig mangelt es jedoch an der entsprechenden Infrastruktur, diese Daten auch noch nach Jahren gezielt wiederauffinden zu können. Eine vernetzte Forschungsdateninfrastruktur würde den Wissenschaftler also auch in der Ausübung seiner Belegpflicht unterstützen.

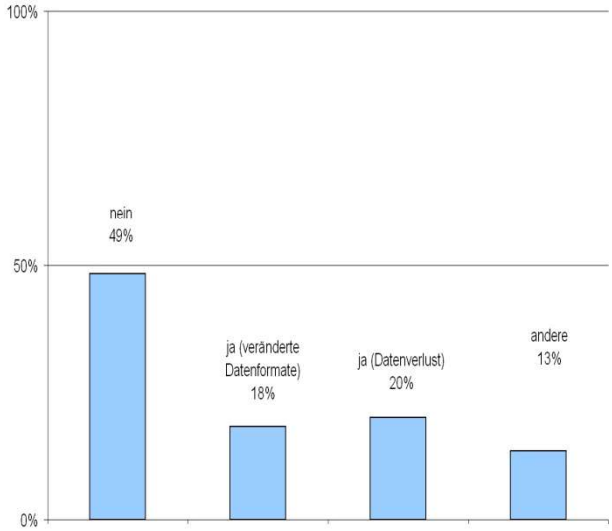


Abb. 10: Gab es jemals Probleme mit der Wiederverwendung von gespeicherten Daten?

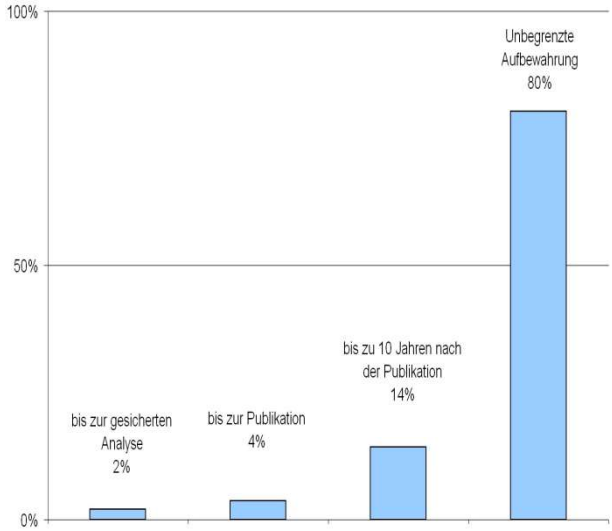


Abb. 11: Wie lange werden Primärdaten aufbewahr

1.2 Analyse der Umfrage

Die Ergebnisse der Umfrage zeigen, dass Wissenschaftler sorgfältig mit Ihren Daten umgehen und die Daten so lange wie möglich speichern. Sie würden auch gerne andere Daten einsehen, um sie mit den eigenen Daten abzugleichen oder auf Fehler zu überprüfen. Darüber hinaus haben nur Wenige Einwände ihre Daten zu teilen, sogar ohne Embargo.

Während Primärdaten überwiegend als ASCII-Dateien oder in einem proprietären Format vorliegen, werden die analysierten Daten in verschiedenster Form als Daten oder Graphik-Files dezentral elektronisch gespeichert sowie von überraschend vielen Befragten auch heute noch ausgedruckt.

Ein kostenfreier Zugang zu den Daten ist aber Voraussetzung für ihren Support.

Zum Zeitpunkt der Umfrage waren einige Begriffewie z. B. Repositorien, Semantisches Web, InChI oder JCAMP-DX den Wissenschaftlern nicht geläufig. In dieser Hinsicht muss Aufklärungsarbeit geleistet werden.

Etwa die Hälfte der Befragten wurde bereits mit der Situation konfrontiert, dass Daten nicht zugänglich waren aufgrund von Problemen der Lesbarkeit von Datenformaten oder allgemeinem Datenverlust.

Die verbesserte Zugangsmöglichkeit zu den eigenen Daten und denen anderer Forschergruppen ist der größte Antrieb einem Daten-DOI-Verfahren zuzustimmen, die große Mehrheit der Befragten befürwortet eine persistente Referenzierung ihrer spektroskopischen und spektrometrischen Daten. Es besteht aber auch eine gewisse Unsicherheit in Bezug auf Missbrauchmöglichkeiten bei der Veröffentlichung vollständiger Datensätze.

Das Interesse der Wissenschaftler an der Umfrage lässt darauf schließen, auch an weiterführenden Untersuchungen, wie z. B. bei der Erprobung eines Prototyps, mitzuarbeiten.

1.3 Vergleich der Nutzerbefragung mit ähnlichen Umfragen im europäischen Ausland

Der permanente Zugang zu wissenschaftlichen Daten wurde in den vergangenen Jahren in verschiedenen Projekten, die auch Umfragen umfassen, thematisiert.

2006 wurde unter der Leitung von Jim Downing und Peter Murray-Rust (University of Cambridge) im Rahmen des Projekts SPECTRa (Submission, Preservation and Exposure of Chemistry Teaching and Research Data) eine Umfrage unter 171 Wissenschaftlern aus den

Bereichen synthetische organische Chemie, Kristallographie und Computational Chemistry durchgeführt. Die Umfrageergebnisse lassen sich wie folgt zusammenfassen:

- Ein großer Teil der Daten (z.B. Angaben aus Laborjournalen, gedruckte Spektren) sind nicht elektronisch gespeichert.
- Digitale Repositorien.sind weitgehend unbekannt.
- Der Bedarf an einem geregelten Zugang zu abgelegten Daten existiert.
- Es liegt eine komplexe Vielfalt an Datenformaten vor.
- Wissenschaftler sind grundsätzlich bereit, zusätzlichen Aufwand zu betreiben, um ausgewählte Daten in JCAMP-DX zu konvertieren und chemische Strukturfiles zu erstellen.
- Eine Suche nach abgelegten Daten sollte als Struktur und/oder Substruktursuche möglich sein.

2009 wurde der Umfragebericht PARSE INSIGHT - into issues of Permanent Access to the Records of Science in Europe - von Tom Kuipers und Jeffrey van der Hoeven (Koninklijke Bibliotheek - Königliche Bibliothek der Niederlande) verfasst. Der Projektbericht beschreibt die Ergebnisse basierend auf einer weltweiten Umfrage mit insgesamt 1840 Antworten, davon 1389 von Wissenschaftlern unterschiedlicher Fachrichtungen (47 % davon naturwissenschaftlich/technisch). 273 Antworten betrafen Daten Manager und 178 Verlage. Die Kernaussagen lauten:

- Die überwiegende Mehrheit (91 %) der Wissenschaftler sieht die Re-Analyse bereits vorhandener Daten als treibenden Motor für die Hinterlegung von wissenschaftlichen Daten an.
- 80 % der Wissenschaftler sehen die Gefahr, dass Informationen unzugänglich werden, weil sie eine langfristige Nutzbarkeit gegebener Hard- und Software nicht für gesichert halten.
- 58 % der Befragten glauben, dass eine internationale Infrastruktur zur Datenarchivierung aufgebaut werden sollte, um den oben genannten Gefahren zu begegnen.
- 25 % der Wissenschaftler machen Ihre Daten für jedermann öffentlich zugänglich.
- Die größte Barriere zur Offenlegung ist die Angst der Wissenschaftler vor Missbrauch ihrer Daten.

Es zeigt sich also, dass sich die Resultate unserer Umfrage mit ähnlichen Recherchen in anderen Nationen decken. Der Stellenwert der Daten, aber auch die Probleme einer

systematischen und langfristig nutzbaren Datenablage werden deutlich. Ebenso offenbart sich eine große Unsicherheit und zum Teil auch Unwissenheit bei der Frage einer Referenzierung und Langzeitarchivierung physikalisch-chemischer Daten.

A 2. Daten in der Chemie

2.1 Datenformate in der Chemie

Die große Herausforderung bei der Diskussion von Forschungsdaten in der Chemie, ihrer Publikation und langfristigen Verfügbarmachung liegt in der enormen Diversifizität der vorhandenen, größtenteils propietären Datenformate. Jedes Meßverfahren, teilweise jedes Meßgerät eines jeden Herstellers hat potentiell ein eigenes Datenformat. Dieser Aspekt ist bei der Diskussion von Forschungsdaten stets im Auge zu behalten. Eine Interoperabilität der Daten läßt sich über Datenaustauschformate realisieren. Hier stehen für spektroskopische Daten Formate wie JCAMP, JCAMP-DX, CML, SpectroML oder AnIML zur Verfügung. Jedoch ist zu berücksichtigen, dass in vielen Fällen bei der Konvertierung in Austauschformate die Informationsdichte der Daten abnehmen kann.

2.2 Werdegang von Forschungsdaten im wissenschaftlichen Prozess

Daten erleben von Ihrer Erzeugung als Forschungsdaten bis hin zu publizierten Daten verschiedene Bearbeitungsschritte durch den Wissenschaftler. In der folgenden Abbildung ist der Werdegang der Daten schematisiert dargestellt.

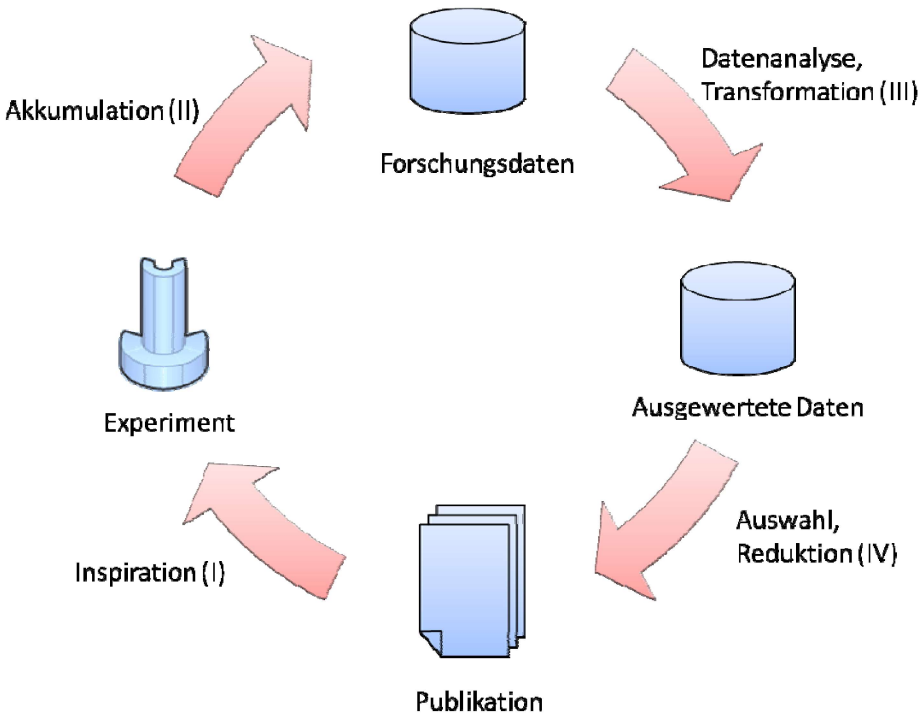


Abb. 12: Datenverlauf vom Experiment bis zur Publikation

Die zyklische Darstellung beginnt mit der Nutzung von „alten“ Forschungsdaten in der Vorbereitungsphase eines neuen Experimentes (*Inspiration I*). In der Vorbereitung eines Experimentes sichtet ein Wissenschaftler in der Regel bisher verfügbare Informationen zu der wissenschaftlichen Fragestellung, mit der sich das Experiment beschäftigt. Dies beinhaltet unter anderem auch die Literaturrecherche in Datenbanken wie Chemical Abstracts, Reaxys oder ähnlichen. Unter Umständen werden zu diesem Zeitpunkt bereits Forschungsdaten gesichtet, die in ihrer bisherigen Form als *Supporting Information* im Anhang von Publikationen zu finden sind. Während des Experiments werden dann in Forschungsdaten unterschiedlichster Art und Umfang erzeugt. Dies kann durch eine einmalige Datenerfassung oder durch die Akkumulation über mehrere Messungen erfolgen. Diese werden sowohl elektronisch erfasst und gespeichert als auch manuell ermittelt und notiert (*Akkumulation II*). So dienen z. B. spektroskopische Messungen in der synthetischen, organischen und anorganischen Chemie nach der Synthese eines Stoffes zu dessen Bestimmung und genauen Beschreibung. Hierbei werden in der Regel verschiedene analytische Verfahren angewendet wie z. B. NMR, MS, IR, UV u.a.

Die im Rahmen solcher analytischen Messungen anfallenden Forschungsdaten werden am häufigsten direkt am Messgerät selber oder einem an das Messgerät angeschlossenen Server gespeichert. Je nach Organisation der Arbeitsabläufe vor Ort ist die Zugänglichkeit solcher Rohdaten in der Regel auf wenige Personen beschränkt. Die Zugänglichkeit reicht vom Experimentator bis zum Arbeitskreis, ist jedoch für außenstehende Personen nicht erlaubt.

Viele der erzeugten Forschungsdaten bedürfen in dieser Rohform noch der Weiterbearbeitung durch den Wissenschaftler. So werden z. B. gemessene FIDs von NMR-Spektren durch eine Software aufbereitet und erst dann in die graphische Repräsentation transferiert (Fourier-Transformation), die zur Auswertung der wissenschaftlichen Ergebnisse genutzt wird (*Datenanalyse/Transformation III*). Die so erzeugten graphischen Darstellungen der ursprünglich gemessenen Forschungsdaten stellen wiederum einen Forschungs-datensatz dar, der sehr häufig auf elektronisch gespeichert und weitergenutzt wird. In diesem speziellen Fall wird die graphische Darstellung des Forschungsdatensatz die Hauptform sein, mit der Wissenschaftler die daran enthaltene Information transportieren und kommunizieren.

Durch die genannten Schritte entstehen Forschungsdaten unterschiedlicher Güte und Inhalts. Mit beginnender Analyse und Interpretation von Daten, werden unter Umständen nicht mehr alle Details eines Rohdatensatzes transportiert. So werden in der Regel in einer wissenschaftlichen Publikation nur noch die Peaklistings eines NMR-Spektrums aufgeführt, die die zur Charakterisierung der Substanz notwendigen Peaks enthalten. Auch werden zusammengeführte Einzelmessungen unter Umständen nur noch als Mittelwert dargestellt,

obwohl ursprünglich eine ganze Reihe von Forschungsdatensätzen erzeugt wurde (*Auswahl/Reduktion IV*).

Einhergehend mit der Bearbeitung von Forschungsdaten steigt die Gefahr von Fehlern und Fehlinterpretationen. Umso komplexer die Experimente, Datenstrukturen und Fragestellungen, desto relevanter wird die Verfügbarkeit von ursprünglichen Forschungsdaten, um Ergebnisse kritisch zu evaluieren. Daher ist der öffentliche Zugang zu den Forschungsdaten im wissenschaftlichen Erkenntnisgewinn eminent.

2.3 Praktisches Fallbeispiel: Prozess der Erzeugung von Forschungsdaten in der organischen Chemie am Beispiel einer Synthese

Das folgende Beispiel macht eine häufig praktizierte Vorgehensweise in der synthetischen Chemie deutlich. Im Zuge der Arbeiten einer Diplomarbeit erfolgte dabei die Synthese einer neuen Substanz, die später im Rahmen einer Dissertation erneut benötigt wurde und daher nochmals synthetisiert werden musste.[6]

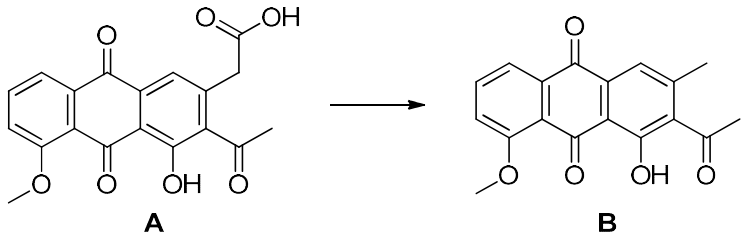


Abb. 13: Darstellung von 2-Acetyl-1-hydroxy-8-methoxy-3-methylanthracen-9,10-dion (B)

Zum Nachweis der Struktur wurde das isolierte Produkt, das in Form oranger Kristalle anfiel, mittels verschiedener analytischer Methoden charakterisiert. Dazu gehörten die Bestimmung des Schmelzpunktes und die Aufnahme von ^{1}H und ^{13}C-NMR-Spektren, IR- und UV-Spektren, eines Massenspektrums unter Elektronenstoßionisierung (70 eV) und eines hochaufgelösten Massenspektrums (EI, 70 eV) als Ersatz für eine Elementaranalyse zur Bestimmung der elementaren Zusammensetzung.

Die Darstellung der Verbindung **B** ist Teil einer komplexen Synthese, die in einem Fachjournal veröffentlicht worden ist. [7] Im Folgenden werden die interpretierten Forschungsdaten in kurzer tabellarischer Form präsentiert.

Auszug aus der Veröffentlichung:

m.p.238 °C

^{1}H NMR (200 MHz, $CDCl_3$): δ = 2.41 (s, 3H, $COCH_3$), 2.65 (s,3H,CH_3), 4.11 (s, 3H, OCH_3), 7.41 (dd, $J_{6,7}$ = 8.3 Hz, $J_{5,7}$ = 1.0 Hz, 1H, 7-H), 7.64 (s, 1H, 4-H), 7.79 (dd, $J_{5,6}$ = 7.8 Hz, $J_{6,7}$ = 8.3 Hz, 1H, 6-H), 8.00 (dd, $J_{5,6}$ = 7.8 Hz, $J_{5,7}$ = 1.0 Hz, 1H, 5-H), 13.30 (s, 1H, OH) ppm. ^{13}C NMR (50 MHz, $CDCl_3$): δ = 20.5 (q, CH_3), 32.3(q, $COCH_3$), 57.1 (q, OCH_3), 115.6 (s, C-9a), 118.8 (d, C-4), 120.7 (d, C-7), 121.2 (d, C-5), 132.6 (s, C-8a), 136.0 (s, C-2), 136.4 (d, C-6), 136.7 (s, C-10a, C-4a), 144.2 (s, C-3), 160.0 (s, C-8), 161.4 (s, C-1), 182.8 (s, C-10), 188.9 (s, C-9), 203.9 (s, CO) ppm. IR (KBr): ν = 3437, 2981, 2927, 2848, 1685, 1631, 1585, 1487, 1468, 1446, 1352, 1286, 1271, 1230, 1184, 1068, 1036, 1012, 962, 849, 750 cm^{-1}. UV (MeOH): λ_{max} (lgε) = 415 (3.59). MS (EI, 70eV): m/z (%) = 310 (25) [M+], 295 (45), 279 (4), 252 (5), 167 (4), 149 (10), 98 (100), 84 (10), 57 (14), 43 (10). HR MS (EI, 70eV): $C_{18}H_{14}O_5$ calcd. for 310.08412; found 310.08414. $C_{18}H_{14}O_5$ (310.30): calcd. C69.67, H 4.55; found C 69.32, H 4.05.

Auf das zusätzliche Einreichen graphischer Darstellungen der Spektren in Form von PDF-Dokumenten oder anderer Daten als *Supporting Information* wurde verzichtet.

Die im Rahmen einer Diplomarbeit durchgeführte Synthese wurde später während einer Dissertation weiter optimiert. Für die Identifizierung der Substanz wurden in dem Falle nicht mehr alle oben aufgeführten analytischen Daten erhoben, sondern es wurden lediglich die ^{1}H NMR und ^{13}C NMR-Spektren zur Kontrolle noch einmal gemessen. Die Ergebnisse wurden unter neuen Dateinamen auf einem Fakultätsserver in einem proprietären Datenformat gespeichert und später in der Dissertation veröffentlicht.[8]

Auszug aus der Dissertation:

Das erhaltene Rohprodukt wird aus Ethanol umkristallisiert und das Methylanthrachinon **B** als oranger, feinkristalliner Feststoff erhalten (419 mg, 1.37 mmol, 90 %, Smp.: 236.8 °C, Lit.:[4] 76 %, Smp.: 238 °C).

^{1}H-NMR (200 MHz, $CDCl_3$): δ [ppm] = 2.41 (s, 3H, $COCH_3$), 2.65 (s, 3H, CH_3), 4.11 (s, 3H, OCH_3), 7.41 (dd, $J_{6',7'}$ = 8.3 Hz, $J_{5',7'}$ = 1.0 Hz, 1H, 7-H), 7.64 (s, 1H, 4-H), 7.79 (dd, $J_{5',6'}$ = 7.8 Hz, $J_{6',7'}$ = 8.3 Hz, 1H, 6-H), 8.00 (dd, $J_{5',6'}$ = 7.8 Hz, $J_{5',7'}$ = 1 Hz, 1H, 5-H), 13.30 (s, 1H, OH).

^{13}C-NMR (50 MHz, $CDCl_3$): δ [ppm] = 20.5 (CH_3), 32.3 ($COCH_3$), 57.1 (OCH_3), 115.6 (C-9a), 118.8 (C-4), 120.7 (C-7), 121.2 (C-5), 132.6 (C-8a), 136.0 (C-2), 136.4 (C-6), 136.7 (C-10a), 136.8 (C-4a), 144.2 (C-3), 160.0 (C-8), 161.4 (s, C-1), 182.8 (s, C-10), 188.9 (s, C-9), 203.9 (CO).

Dieses Beispiel zeigt, dass in der präparativen Chemie das wiederholte Messen einer bekannten Verbindung häufig erforderlich ist. Bereits vorhandene Daten dienen nur dem Vergleich oder als Interpretationshilfe.

Die Reduktion der Spektren auf ausgewählte Werte ermöglicht es den Lesern die Identifizierung nachzuvollziehen. Fakultätsangehörigen ist die Einsicht in die vollständigen NMR-Spektren möglich, sofern der Dateiname bekannt ist, der vom Auftraggeber bestimmt wurde. Ein visueller oder sogar elektronischer Vergleich erleichtert die Beurteilung erheblich und ermöglicht einen direkten Vergleich mit eigenen Ergebnissen.

A 3. Empfehlung speicherungswürdige Daten

Die im Rahmen der Konzeptstudie durchgeführten Untersuchungen und Diskussionen mit Domänexperten haben ein sehr komplexes Bild hinsichtlich der Kriterien für speicherungswürdige Forschungsdaten in der Chemie ergeben. Anders als beispielsweise in den Erd- und Umweltwissenschaften ist die eigenständige Publikation von Forschungsdaten in der Chemie als eher zweitrangig anzusehen. Vielmehr werden Forschungsdaten publiziert, die Teil von wissenschaftlichen Publikationen sind und dort bereits jetzt in reduzierter Form veröffentlicht werden.

Bei Betrachtung der Workflows, bei denen Forschungsdaten entstehen, wurden Forschungsdaten unterschiedlicher Relevanz identifiziert. So wird im Laborbetrieb eine große Menge an Forschungsdaten produziert, die eher in den Bereich der Qualitätskontrolle von laufenden Prozessen fallen und nicht relevant für Publikationen sind. Für solche Forschungsdaten ist eine Speicherung in institutionellen Repositorien vorstellbar. Erst bei der Zusammenfassung von wissenschaftlichen Ergebnissen und deren Aufbereitung für eine Veröffentlichung werden Forschungsdatensätze für die Untermauerung wissenschaftlicher Erkenntnisse und Thesen herangezogen. Solche Forschungsdaten sind von Relevanz für die langfristige Speicherung und öffentliche Zugänglichkeit.

Grundsätzliche Empfehlung sollte sein Forschungsdaten, die Grundlage einer wissenschaftlichen Publikation sind, bei einem Datenzentrum zu hinterlegen und entsprechend zu referenzieren. Selbstverständlich sollte zusätzlich die Option bestehen, Forschungsdaten auch ohne zugehörige Publikation eines Artikels als eigenständige Einheit bei einem Datenzentrum zu hinterlegen. Die Auswahl der zu publizierenden Forschungsdaten obliegt in jedem Fall dem Wissenschaftler.

Als weiterer interessanter Aspekt wurde in Diskussionen mit Wissenschaftlern die Handhabung von substanzbezogenen Forschungsdaten identifiziert. Eine substanzzentrierte Sicht auf Forschungsdaten eines Datenzentrum könnte beispielsweise lauten „Welche Forschungsdaten gibt es zu der Substanz Taxol?“. Bei prominenten Substanzen mit hohem wissenschaftlichen Interesse ist davon auszugehen, dass es eine große Anzahl von Artikeln zu dieser Substanz gibt und folglich auch eine potentiell große Anzahl von Forschungsdaten. Diese Forschungsdaten könnten beispielsweise ^{1}H-NMR Spektren von Taxol sein. Obwohl grundsätzlich alle ein Protonenspektrum der Substanz darstellen, können sich die Forschungsdatensätze hinsichtlich des verwendeten Messgeräts und, was die Qualität der Information betrifft, der Frequenz des Messgerätes unterscheiden. So ist die Informationsdichte eines 800 MHz ^{1}H-NMR Spektrums sicher höher einzustufen als die eines 60 MHz ^{1}H-NMR Spektrums. Die Empfehlung hinsichtlich speicherungswürdiger Datensätze

geht daher dahin, solche „ähnlichen“ Forschungsdatensätze trotzdem zu speichern und Verfahren und Metadatenschemata zu entwickeln, um substanzbezogene Forschungsdaten entsprechend zu clustern.

Hinsichtlich der zu speichernden Datenformate herrscht noch immer großer Diskussionsbedarf, der sich auch in der Komplexität der existierenden Datenformate in den vielen Teilgebieten zu sehen ist. Aus wissenschaftlicher Sicht ist die Erhaltung maximaler Informationsdichte erstrebenswert. Dies würde ein Abspeichern der in erster Instanz erzeugten Rohdaten, in möglicherweise proprietären Datenformaten des verwendeten Messgerätes bedeuten. Hier besteht noch weiterer Gesprächsbedarf, auch mit den Anbietern dieser proprietären Datenformate. Eine Alternative wäre die Verwendung von offenen Austauschformaten wie JCAMP, CML oder AnIML. Hier gibt es jedoch offene Fragen hinsichtlich der Datenverlustes bei der Transformation der Rohdaten in das Austauschformat oder generell der Verfügbarkeit von entsprechender Konvertierungssoftware.

A 4.Analyse des bestehenden Publikationsprozesses in der Chemie

Das nachfolgende Kapitel beleuchtet die bisherige Handhabung von Forschungsdaten in Publikationsprozess, hier insbesondere die Publikation wissenschaftlicher Artikel und von Hochschulschriften.

4.1 Bisherige Publikation von Forschungsdaten

Wie bereits eingangs festgestellt unterliegt die bisherige Handhabung von Forschungsdaten keinen durchgehenden Standards. Forschungsdaten werden täglich im Labor erzeugt und verbleiben dort lange Zeit in der privaten Domäne des Wissenschaftlers.

Unter der *privaten Domäne* (A) sind alle Medien zu verstehen die vom Wissenschaftler genutzt werden, um Daten zu erzeugen und zu bearbeiten. An dieser Stelle werden außer den gerätespezifisch generierten Metadaten nur noch Metadaten erzeugt, die der Wissenschaftler selber für seine persönliche Ordnung benötigt. Eine Nachnutzung dieser Daten kann nur durch direkte Interaktion mit dem Datenproduzenten erfolgen.

Mit der Analyse und Interpretation der Forschungsdaten durch den Erzeuger werden die Daten aufbereitet und in den Kontext der wissenschaftlichen Erkenntnisgewinnung eingebunden. Ein Forschungsdatensatz ist so ein Teil der Beweisführung z. B. in einer Dissertation.

Der Transfer der Daten von der privaten Domäne in die *Gruppendomäne (B)* erfolgt, um den Vorgesetzten und anderen Mitarbeitern der Arbeitsgruppe die Ergebnisse zu präsentieren und zu diskutieren. Der Nutzerkreis ist deutlich erweitert und anonymisiert. Es werden zusätzliche Metadaten erstellt, die auf Grund des Community-Wissens entstehen.

Die Gruppendomäne zeichnet sich durch die arbeitsgruppeninterne Datennutzung aus. Die Gruppe derer, die auf die Daten zugreifen können ist wesentlich größer und anonymer, jedoch besteht immer noch direkter Kontakt zu dem Datenerzeuger. Es erfolgt eine wissenschaftliche Bewertung der Daten und Anreicherung der Metadaten mit Bewertungsergebnissen.

Mit der Publikation eines Artikels werden aktuell die zugrundeliegenden Forschungsdaten in stark reduzierter Form direkt in den Artikel eingebunden. Dies erfolgt in der Regel in dem sogenannten *Experimentellen Teil* einer Publikation, der genau Auskunft über verwendete Methoden, Geräte und Parameter gibt. Weiterhin werden Forschungsdaten in reduzierter Form als Messwerte, Peaklistings oder Kurvenverläufe angegeben. In vielen Fällen werden Artikel mit sogenannten Supporting Information angereichert, einem separatem Dokument,

das detailliertere Informationen zu den gemessenen Daten enthält. Hier ist die Informationsdichte höher als im *Experimentellen Teil* des Artikels. So werden hier beispielsweise graphische Darstellungen von Spektren zur Verfügung gestellt, anstelle der vorher genannten Peaklistings ausgewählter Messwerte. Jedoch ist auch diese Information nicht vergleichbar mit dem ursprünglich gemessenen Forschungsdatensatz. Dieser verbleibt in der Regel beim Erzeuger bzw. Autor, wird dort jedoch nach Erkenntnissen unserer Umfrage nach den Regeln der guten wissenschaftlichen Praxis gespeichert.

Der letzte Schritt stellt somit den Übergang in die *dauerhafte Domäne (C)* mit öffentlichem Zugang und Nachnutzung dar – wobei letztere aktuell stark eingeschränkt ist. Hierfür müssen vollständige Metadaten erzeugt werden, damit eine Suche und Identifizierung der Daten eindeutig gewährleistet ist. In der öffentlichen Domäne sind (theoretisch) die Langzeitarchivierung und alle möglichen Nachnutzer auch außerhalb der wissenschaftlichen Community angesiedelt.

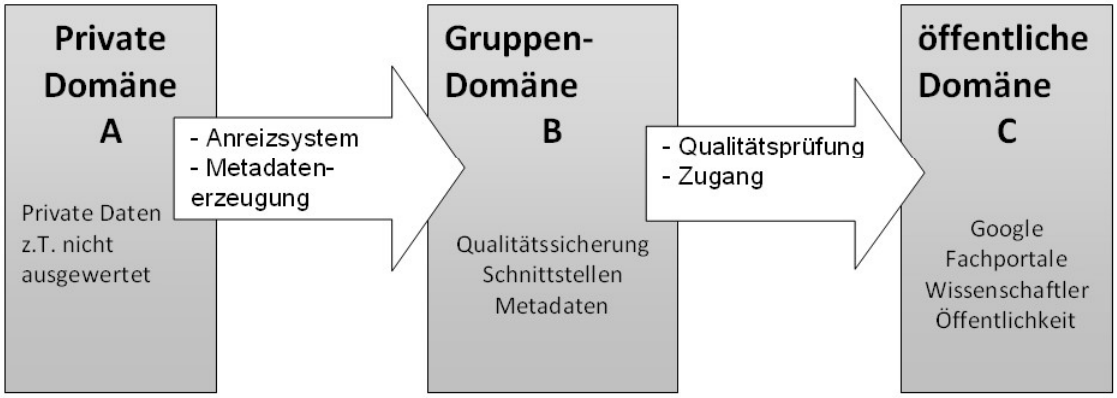


Abb. 14: Das Data Curation Continuum nach A. Treloar[9] beschreibt drei Domänen mit zwei Übergängen. Die drei Domänen sind die private Domäne, die Gruppendomäne und die dauerhafte Domäne

4.2 Ein erweiterter Publikationsprozess

Der bisherige Publikationsprozess in der Chemie enthält keine definierten Verfahren zur Veröffentlichung, Speicherung und öffentlichen Zugänglichkeit der Forschungsdaten, die zu einer Publikation gehören. Der bisherige Publikationsprozess weist somit vor allem Lücken

im Übergang von der Gruppendomäne zur öffentlichen Domäne auf. Es gibt keine Institution, die für die Datenpflege und Datenarchivierung zuständig ist. Diesen Teil sollte eine dauerhafte Domäne übernehmen. Daraus folgt auch, dass es einen Transfer geben muss zwischen der Gruppendomäne zur dauerhaften Domäne und von der dauerhaften Domäne zur öffentlichen Domäne (s. Abb. 15).

Abb. 15: Rollen und Aufgaben bei der Publikation von Forschungsdaten (nach Treloar & Klump)

In der *privaten Domäne (A)* ändert sich kaum etwas. Der Wissenschaftler produziert und verwaltet seine Daten weiterhin eigenständig.

Der Transfer erfolgt dann in die *Gruppendomäne (B)*, wo wieder die Daten für eine Publikation ausgewählt und diskutiert werden. Auch übernimmt der Ersteller der Daten hier die Verantwortung für seine Daten. Hier kann durch bestimmte Vorkehrungen das Qualitätsmanagement (3) für die Daten gesichert werden. Durch die Erstellung eines erweiterten DMP (5) zeigt sich erst welcher Aufwand tatsächlich notwendig war, um die gewünschten Daten zu erzeugen.

Dann folgt der Transfer zu einer *dauerhaften Domäne (C),* die noch geschaffen werden muss. Dieses Aufgabenfeld könnte beispielsweise eine Einrichtung wie das FIZ Chemie übernehmen. Hier werden die Daten in Archiven gespeichert und verwaltet. Dafür müssen aber umfangreiche Metadaten (9) erstellt werden, damit die Daten gesucht werden können.

Deshalb müssen Metadatenstandards eingeführt und eingehalten werden. Die Daten erhalten hier einen persistenten Identifier (10), unter dem sie dauerhaft wiederzufinden und eindeutig identifizierbar sind.

Über diese dauerhafte Domäne kann die *öffentliche Domäne (D)* auf die Daten zugreifen. Dies kann über Fachportale oder durch Suchmaschinen über das Internet geschehen. Außer den Wissenschaftlern können dadurch alle Interessierten auf die Daten zugreifen. Die öffentliche Domäne erweitert den Zugang zu wissenschaftlicher Information erheblich. Die Umfrage hat ergeben, dass Wissenschaftler insbesondere am Zugang zu den einer Publikation zugrundeliegenden Forschungsdaten Interesse haben, da die bisher veröffentlichen Daten für viele Fragestellungen unzureichend sind.

4.2.1 Das Projekt „Publikation und Zitierfähigkeit wissenschaftlicher Primärdaten“

Im wissenschaftlichen Bereich besteht zwar grundsätzlich Bereitschaft, Daten für eine interdisziplinäre Nutzung zur Verfügung zu stellen, aber es ist zur Zeit unüblich, dass die erforderliche Mehrarbeit für Aufbereitung, Kontextdokumentation und Qualitätssicherung im Wissenschaftsbetrieb anerkannt wird. Die klassische Form der Verbreitung wissenschaftlicher Ergebnisse ist ihre Veröffentlichung in Fachzeitschriften, normalerweise ohne Veröffentlichung der zugrunde liegenden Daten. Diese klassische Publikation wird im "Citation Index" erfasst. Dieser Index wird zur Leistungsbewertung von Wissenschaftlern herangezogen. Datenveröffentlichungen werden darin bisher nicht berücksichtigt.

Projektdaten sind breit über Forschungsinstitute verstreut und werden von Wissenschaftlern erhoben und meist selbst verwaltet. Aufgrund der fehlenden Anerkennung der mit der Aufbereitung verbundenen Arbeit sind Projektdaten häufig schlecht dokumentiert und somit schwer zugänglich sowie nicht langfristig gesichert. Große Datenbestände bleiben ungenutzt, da sie nur einen kleinen Kreis von Wissenschaftlern bekannt und zugänglich sind. Viele Forschungsdaten verbleiben als ungenutztes Rohmaterial und sind häufig nach wenigen Jahren verloren.

Die Diskussion um die Fälschung wissenschaftlicher Ergebnisse führte zur Empfehlungen in den „Regeln guter wissenschaftlicher Praxis“ durch die DFG. Die Regeln beinhalten auch Richtlinien für den Datenzugang. Forschungsdaten einer Veröffentlichung müssen mindestens 10 Jahre gespeichert und zugänglich sein, um eine Prüfung der Ergebnisse zu ermöglichen. Zwar werden diese Vorschriften im Regelfall eingehalten, aufgrund der damit verbundenen Zeitbelastung werden Daten aber normalerweise nur in Rohform archiviert und nicht in ihrer Feinstruktur aufgearbeitet, dokumentiert und allgemein verfügbar erschlossen.

Die wichtigsten Ziele eines neuen Umgangs mit Forschungsdaten sind also langfristige und allgemein zugängliche Verfügbarkeit und Speicherung. Durchsetzbar ist dies am besten über eine persönliche Motivation der Wissenschaftler. Dies ließe sich durch die Möglichkeit der Forschungsdatenpublikation fördern:

Daten sollen nicht mehr ausschließlich Teil einer wissenschaftlichen Veröffentlichung sein, sondern eine eigenständige Identität besitzen. Damit würden Forschungsdaten, ähnlich wie Zeitschriftenartikel zitierbar.

Mit der anerkannten Datenpublikation erhält ein Autor also eine zitierfähige Veröffentlichung. Zeitschriftenartikel, welche die Daten verwenden, verweisen auf die Datenpublikation. Umgekehrt kann auch von den publizierten Daten auf Artikel in Zeitschriften verwiesen werden, die den Datensatz verwenden. Die Publikation von Daten kann also sinngemäß in das bestehende System von wissenschaftlichen Veröffentlichungen und deren Zitierbarkeit eingebunden werden.

In dem an der Technischen Informationsbibliothek (TIB) Hannover durchgeführten Projekt „Publikation und Zitierfähigkeit wissenschaftlicher Forschungsdaten“ wurde eine Infrastruktur zur Registrierung von DOI-Namen und URNs für Wissenschaftliche Datensätze geschaffen und erfolgreich getestet. Mit dem System wurden an der TIB bereits über 650.000 Datensätze aus dem Bereich der Geowissenschaften mit persistenten digitalen Identifikatoren versehen. Ein Teil der Datensätze ist über den Online-Katalog der TIB verfügbar, auf die restlichen Datensätze kann offen über die Kataloge der beteiligten Datenarchive als sogenannte Publikationsagenten zugegriffen werden.

Da in den Geowissenschaften/ Erdsystemforschung sowohl die Probleme der Datenvielfalt und -menge auftreten, als auch verschiedenste Formen der Datenerhebung vorkommen (u.a. Experimente, Feldstudien, Simulationen, Monitoring), konnten Informationen gesammelt werden, die eine Beschreibung von Daten ermöglichen.

DataCite

Der im Dezember 2009 gegründete Verein DataCite hat sich zum Ziel gesetzt, Wissenschaftlern den Zugang zu Forschungsdaten über das Internet zu erleichtern, die Akzeptanz von Forschungsdaten als eigenständige, zitierfähige wissenschaftliche Objekte zu steigern und somit die Einhaltung der Regeln guter wissenschaftlicher Praxis zu gewährleisten.

Bis heute haben sich 12 Partner aus 9 Ländern unter dem Dach von DataCite zusammengefunden: die British Library, das französische L’Institut de l’Information Scientifique et Technique (INIST), das Technical Information Center of Denmark, die TU Delft Bibliothek aus den Niederlanden, das Canada Institute for Scientific and Technical

Information (CISTI), die California Digital Library (USA), der Australian National Data Service (ANDS) die Purdue University (USA) und die Eidgenössische Technische Hochschule in Zürich.

Deutsche Mitglieder sind neben der Technischen Informationsbibliothek (TIB), die Goportis Partner Deutsche Zentralbibliothek für Medizin (ZB MED) und voraussichtlich ab dem 01.01.2011 die Deutsche Zentralbibliothek für Wirtschaftswissenschaften, sowie das Leibniz-Institut für Sozialwissenschaften GESIS.

Unter Forschungsdaten werden im DataCite Kontext nicht nur wissenschaftliche Primärdaten verstanden, sondern alle Arten von wissenschaftlicher Information, wie Bilder, Filme, Graue Literatur, die derzeit noch nicht eine eigenständige wissenschaftliche Verlagsveröffentlichung sind.

Andere Identifier-Systeme

Neben der DOI als Persistenten Identifier sind in Deutschland ebenfalls der Uniform Resource Name (URN) und das Handle System verbreitet. Ohne auf die Details der einzelnen Systeme einzugehen, ist für die Wahl des passenden Persistent Identifier (PI)-Systems immer entscheidend, welches Ziel man letztendlich verfolgt. Das Ziel von DataCite ist es, zitierfähige Forschungsdaten zu erhalten.

Für diesen Fall ist aufgrund seiner Verbreitung und Akzeptanz im Verlagswesen der Digital Object Identifier das einzige akzeptable PI-System. Das URN-System bietet keine weltweite universelle Auflösung, das Handle-System hat keine übergeordnete Struktur, die Qualitätssicherheit bietet.

Die Vergabe von URN ist dort sinnvoll, wo auf lokale Dokumente innerhalb eines geschlossenen Systems dauerhaft zugegriffen werden soll, beispielsweise bei Hochschulschriften, die Verwendung von Handle bietet sich an, wo extrem große Mengen von Objekten global persistent identifiziert werden sollen, aber nicht die Zitierung derselben im Vordergrund steht.

Nichtsdestotrotz werden im DataCite Kontext von der TIB in Kooperation mit der Deutschen Nationalbibliothek auch URN-Namen für Forschungsdaten vergeben, wenn das von den Datenzentren so gewünscht wird. Eine Kombination von Handle und DOI-Vergabe wird auch von einigen internationalen DataCite Partnern wie dem ANDS betrieben und ist alleine schon deshalb problemlos möglich, da das DOI-System technisch auf dem Handle System aufbaut.

4.2.2 Eigenständige Forschungsdatenpublikation

Die Vergabe von DOI-Namen für Forschungsdaten ermöglicht die direkte Referenzierung und Verlinkung von Artikel und Forschungsdaten. Somit kann beispielsweise vom Artikel direkt in die Datenbank beim Datenzentrum gesprungen werden.

Die im experimentellen Teil von Publikationen enthaltenen Daten zur Charakterisierung von Verbindungen stellen üblicherweise nur einen fragmentarischen Auszug aus den Originaldaten bzw. aus den daraus generierten Spektren dar. Zwar werden Publikationen häufig mit sogenannter Supporting Information angereichert. Doch auch dies sind meist nur PDF-Dokumente mit Repräsentationen ausgewerteter Forschungsdaten. Seit 2009 existiert eine Kooperation zwischen der TIB und dem Georg Thieme Verlag, in der exemplarisch die Publikation von chemischen Daten, die einem Artikel zugrunde liegen in den Publikationsworkflow eines Verlages eingebunden wird.[10] Die beispielhafte Umsetzung eines mit Forschungsdaten angereicherten Artikels kann bei der Publikation von K. Jarowicki, C. Kilner, P. J. Kocienski, et.al. (http://dx.doi.org/10.1055/s-2008-1067226) betrachtet werden, vgl. Abb. 16.

Nur in dieser Form sind die zugrundeliegenden Daten einer wissenschaftlichen Publikation im Detail nachvollziehbar. Die notwendige Erfassung und Publikation von Forschungsdaten kann als Ergebnis der Arbeiten an der Konzeptstudie nur in einer gemeinsamen Anstrengung aller Beteiligter erfolgen: Datenzentren, Fachgesellschaften, Verlage und Bibliotheken.

Die langfristige Speicherung muss nicht zwingend in einem zentralen Datenzentrum erfolgen. Alternativ könnten Forschungsdaten dezentral am Entstehungsort (Universität, Institut etc.) in einem *institutional repository* gespeichert und durch eine Verknüpfung mit dem zentralen oder vernetzten Datenspeicher verlinkt und suchbar gemacht werden. Dieses Institutional repository kann sowohl eine eigenständige Position sein, die eine vorgegebene Struktur erfüllt, oder aber ein ausgelagertes Repositorium, welches beispielsweise vom FIZ Chemie gestellt werden kann.

Thieme eJournals / Abstract Kontakt

PAPER

Synthesis 2008(17): 2747-2763
DOI: 10.1055/s-2008-1067226

© Georg Thieme Verlag Stuttgart · New York

A Synthesis of 1-Lithiated Glycals and 1-Tributylstannyl Glycals from 1-Phenylsulfinyl Glycals via Sulfoxide-Lithium Ligand Exchange

Krzysztof Jarowicki, Colin Kilner, Philip J. Kocienski*, Zofia Komsta, Jacqueline E. Milne, Anna Wojtasiewicz, Victoria Coombs
Institute of Process Research and Development, School of Chemistry, Leeds University, Leeds, LS2 9JT, UK
e-Mail: p.j.kocienski@leeds.ac.uk;
Received 5 May 2008

Abstract

1-Lithiated glycals generated by reaction of 1-phenylsulfinyl glycals with either *t*-BuLi or PhLi are transformed to 1-tributylstannyl glycals on reaction with tributyltin chloride.

Keywords

lithium - tin - sulfoxides - carbohydrates - glycals

Primary data for this article are available online and can be cited using the following DOI: 10.4125/pd0001th: Primary Data (added August 26th, 2009). FIDs and associated files for the ^{1}H, ^{13}C and DEPT NMR spectra for compounds **14**, (S_S)-**23**, (S_S)-**25**, (R_S)-**26**, **27**, (S_S)-**28**, (R_S,S_S)-**29**, **30**, (R_S)-**36**, (S_S)-**36**, (S_S)-**37**, **38**, (R_S)-**39**, (S_S)-**39**, (S_S)-**44**, (R_S)-**46**, (S_S)-**46**, (R_S)-**48**, (S_S)-**48**, (S_S)-**49**, **52**, (R_S)-**53**, (R_S)-**55**, (R_S)-**57**, (S_S)-**57**, (S_S)-**58**, (R_S)-**61**, (S_S)-**61**, (R_S)-**62**, (S_S)-**62**, (R_S)-**65** and (S_S)-**65** are summarized.

Abb. 16: Verknüpfung eines Artikels mit den dazugehörigen Forschungsprimärdaten

Ein Vorteil zeigt sich darin, dass diese Forschungsdaten in den institutional repositories für institutsinterne Zwecke im Bereich der Verwaltung (Evaluation) benutzt werden können. Nach den Regeln zur „guten wissenschaftlichen Praxis“, welchen die Institute, Universitäten etc. folgen müssen, müssen Forschungsdaten mindestens 10 Jahre lang für die eventuelle Überprüfung gespeichert werden. Durch die institutional repositories würden sowohl die Institutionen als auch die nach Forschungsdaten suchenden Wissenschaftler Vorteile ziehen. Weiterhin unzugänglich blieben die Daten, die in den internen Speichern, auf den Rechnern der Wissenschaftler oder den Messgeräten selber beherbergt sind. So kann der Wissenschaftler eine passende Auswahl treffen, welche Daten publizierbar sein sollen und welche nicht. Der von vielen Wissenschaftlern befürchtete Mehraufwand beim Speichern und Hochladen von Dateien würde sich in Grenzen halten.

Dabei ist es wichtig, die Publikation der Daten als Service für die Wissenschaftler zu betrachten und nicht als Angebot für die Verlage. Alleinstehende Daten, wie sie

beispielsweise in der Umweltchemie oder technischen Chemie anfallen, sollen zunächst nicht berücksichtigt werden. Daraus resultiert natürlich, dass es keinen Embargo-Prozess gibt, da die Daten mit der Publikation veröffentlicht werden.

In Anlehnung an das Beispiel von Thieme sollen Daten im proprietären Format als Beleg ebenso abgelegt werden wie Daten in Austauschformaten wie jcamp-dx oder cdf damit sie problemlos ohne Zuhilfenahme von Spezialprogrammen geöffnet werden können. Darüber hinaus sollen alle weiteren Daten als supplemental material abgelegt werden können, die der Wissenschaftler abzulegen wünscht. Es soll keine Limitation für zusätzliche Daten vorhanden sein.

Grundsätzlich bieten sich zwei Modelle an, wie die Daten zu den Publikationen mit DOI-Namen verknüpft werden sollen. Beide Modelle bieten Vor- und Nachteile.

1. Im ersten Szenario bekommen alle Daten von allen Messungen jeder Substanz einen eigenen DOI-Namen. Dies führt dazu, dass es sehr viele DOI-Namen für die Daten gibt, was sehr unübersichtlich sein kann. Dafür ist es möglich viele Überkreuz-Suchen durchzuführen. So kann beispielsweise in einer Publikation ein DOI-Name auf die Daten verlinken, die Informationen zu einer bestimmten Messung einer Substanz enthalten, und die zu einem Datenpool führt. In diesem Datenpool befinden sich alle meßmethodisch identischen Daten dergleichen Substanz. In diesem Pool wären auch vergleichbare Messung anderer Autoren abgelgt, die dadurch einfach gefunden werden können, und zu deren Publikation man dadurch einfach gelangen könnte. Das bedeutet einen zweifellos einen Mehrwert für den Wissenschaftler.

2. Das zweite Szenario beschreibt ähnlich dem Thieme-Beispiel einen DOI-Namen für alle Messungen einer Substanz, die sowohl proprietäre Daten als auch Daten in Austauschformaten von verschiedenen Messtechniken enthalten soll. Auch hier sind alle zusätzlichen Daten ausdrücklich erwünscht. Diese Art der Datenablage würde die Suche nach bestimmten Daten nicht deutlich verschlechtern aber würde Überkreuz-Suchen deutlich erschweren. Der große Vorteil wäre die Übersichtlichkeit der DOI-Namen und der geringe Mehraufwand für den Wissenschaftler, da nur einmal Metadaten zu den Daten eingegeben werden müssten.

Einem Data-Management-Plan (DMP) einen DOI-Namen zu vergeben wäre ebenfalls zu bedenken. Dieser DMP würde dann quasi als Container für die Daten dienen und würde eine Aussage darüber darstellen, um welche Daten es sich handelt. Dieser Ansatz könnte vor allem mit dem ersten Szenario verbunden werden. Die Eingabe der Metadaten müsste in jedem Fall so einfach wie möglich gehalten werden, um den Mehraufwand für den Wissenschaftler so gering wie möglich zu halten.

Das Ablegen von Daten bei der Publikation im Datenzentrum würde den Wissenschaftler zudem von der Pflicht entbinden, seine Daten nach GLP-Richtlinien mind. 10 Jahre aufzubewahren.

B - Metadaten

(Arbeitspaket 3)

B 1. Einleitung

1.1 Metadaten

Metadaten sind, wenn man es wörtlich aus dem griechischen übersetzt, Daten über den Daten. Bei den beschriebenen Daten handelt es sich um Dokumente, Datenbanken und Dateien, sogenannte elektronische Ressourcen. In der Informatik bedeuten Metadaten, eine Beschreibung der eigentlichen Ressourcen, die kurz Auskunft gibt über Autor, Titel, Format usw. Dieser Begriff der „Metadaten“ ist ein sehr junger Begriff, jedoch gibt es bereits seit langem das Konzept von Daten die Daten beschreiben: in Form von Katalogdaten in Bibliotheken. In den Katalogdaten kann man nach Autoren, Veröffentlichungsdaten, Titeln usw. suchen wie in Metadaten.

Da das Internet in der heutigen Zeit große Möglichkeiten zum Informationsaustausch von elektronischen Dokumenten oder Daten bietet, führt kein Weg daran vorbei vom Modell der Bibliotheken und Archive abzuweichen, um Ressourcen im Internet zu beschreiben.

Die komplexen Regeln, die in Bibliotheken und Archiven angewendet werden, können nicht so einfach auf die chaotischen Verhältnisse im World Wide Web angewendet werden. Deswegen müssen Metadatenschemata an das WWW angepasst werden, um einen optimalen, effektiven und preisgünstigen Zugriff auf Daten zu ermöglichen.

Bei der Speicherung von elektronischen Metadaten gibt es verschiedene Möglichkeiten, wie z. B. die Verwendung von Metatags innerhalb von HTML-Dokumenten (Erscheinungsjahr, Autor etc) oder die Nutzung von Dateiattributen innerhalb von Word- oder PDF-Dateien. Diese Metadaten, wenn sie überhaupt angegeben werden, reichen jedoch in der Regel nicht aus, um Dateien ausreichend zu beschreiben.

Eine beschreibende Suche nach elektronischen Dokumenten, mit den üblichen Mitteln, den Suchmaschinen Google, Yahoo oder anderen erweist sich als äußerst ineffektiv, wenn man nicht den genauen Titel des Dokumentes kennt. Trotz einer hohen Trefferanzahl, zeigt sich, dass nur eine geringe Chance besteht, die gesuchten Daten zu finden. Von Seite der Suchmaschinen besteht allerdings auch kein Interesse diesen Zustand zu ändern, da diese mehr auf die Abdeckung des ganzen Internets ausgerichtet sind und weniger auf die Präzision einzelner Ergebnisse.

Es ist möglich durch die Anwendung von Metadaten, welche mit den Ressourcen verbunden sind, dieses Problem zu lösen. Die Metadaten ermöglichen es erst auf die Inhalte der Ressourcen Einblick zu bekommen. Metadaten bilden eine systematische Methode um Datenquellen zu beschreiben und sie dadurch zugänglich zu machen. Wenn eine

Datenquelle es wert ist zugänglich gemacht zu werden, dann muss sie auch mit Metadaten beschrieben werden um ihre Lokalisierung zu ermöglichen.

Die Metadaten stellen somit die essentielle Verbindung zwischen dem Ersteller der elektronischen Dokumente und dem Nutzer der Informationen dar.

1.2 Bibliothekarische Katalogisierungsstandards

Seit Jahrzehnten gibt es bereits katalogisierende Schemata, die in Archiven und Bibliotheken angewendet wurden. Diese waren nötig, um Publikationen, Texte und Bücher zu beschreiben. Durch diese Katalogisierung konnten Dokumente erst suchbar gemacht werden.

Seit 1969 wurde in den Niederlanden PICA (Project of Integrated Catalogue Automation) ins Leben gerufen, um die Katalogisierung und Automatisierung in Bibliotheken voranzutreiben. Mittlerweile wurde PICA als Teil dem OCLC (Online Computer Library Center) eingegliedert. [11] PICA-Software wird auch in vielen Deutschen Bibliotheken und Bibliotheksverbünden genutzt. Der Grundsätzliche Aufbau von PICA ist der folgende:

0xxx *Kontrollinformationen* - Datum, Zeit, interne Nummer

1xxx *Kodierter Zugang* - Sprache, Staat, physikalische Form, etc.

2xxx *Identifikation* - ISBN, ISSN, etc.

30xx *Personen/ Namen* – Autoren, Herausgeber

31xx *Namen d. Organisationen* – alle beteiligten Organisationen

32xx *Folgenangabefelder* – Details von Teilen etc.

4xxx *Titelbeschreibung* – Titelinformationen einschließlich Fußnoten

5xxx *Klassifikationn* – Inhaltsklassifikation

6xxx *Lokale Daten* – interne Klassifikation

7xxx Signatur/en

8xxx Interne Kennung

9xxx *Zusammenfassung*- Zusammenfassung des Inhalts

Ein weiteres Katalogisierungsformat, welches von der US-amerikanischen Library of Congress ebenfalls Ende der 1960er ins Leben gerufen wurde, heißt MARC (MAchine-

Readable Cataloging). Dieses Format wurde zu MARC 21 weiterentwickelt und ist in Abwandlung in verschiedenen Ländern gebräuchlich.[12] Der Aufbau ist hier der Folgende:

0xx Identifikations- und Kontrollfeld - Sprache, Signatur etc.

1xx *Haupteingabefelder* – Namen, Organisationen, Konferenzen etc.

2xx *Titel und titelbezogene Felder*

3xx *technische Datenfelder* – Größe, physikalisches Medium etc.

4xx *Folgenangabefelder* – Details von Teilen etc.

5xx *Notenfelder* - Fußnoten

6xx *Fachgebietsfelder* - keywords

7xx *Weitere Zugangs- und Personenfelder* – weitere Autoren, andere Titel, etc.

8xx *weitere Folgenfelder* - physikalische and elektronische Lage,etc.

9xx *Lokale Daten* – nicht standardisiert

1.3 Dublin Core

Da es zwischen dem Internet und den Bibliothekskatalogen große Differenzen gibt, die sich vor allem in der Darstellung und den Formaten unterscheiden, war es sinnvoll einen neuen Standard zu definieren, der es erlaubt Objekte zu beschreiben wie in Katalogen, der jedoch auch eine Anpassung an das Internet aufweist.

Während einer Konferenz, dem „OCLC/NCSA Metadata Workshop“ im Jahr 1995, wurde in Dublin, USA die „Dublin Core Metadata Initiative“ (DCMI) gegründet. Die DCMI entwickelte dort eine Sammlung einfacher und standardisierter Konventionen zur Beschreibung von Dokumenten und anderen Objekten im Internet, um diese mit Hilfe von Metadaten einfacher auffindbar zu machen. Diese Sammlung wurde „Dublin Core“ genannt.

Autoren von Webressourcen sollten durch dieses Metadatenschema in die Lage versetzt werden, ihre Ressourcen so zu beschreiben, dass sie etwa von stichwortbasierten Suchmaschinen gefunden werden können. Da das Schema schnell die Aufmerksamkeit von Bibliotheken, Museen usw. auf sich zog, entwickelte sich aus dieser Initiative ein internationales Übereinkommen über eine Kernmenge von Metadaten. Mittlerweile ist Dublin Core seit 2003 ein ISO-Standard, der Standard ISO 15836-2003.

Die DCMI empfiehlt 15 Kernfelder, die so genannten „core elements“ zur sicheren Beschreibung von Daten.[13] Diese core elements enthalten Informationen zu ID, technischen

Daten, Beschreibung des Inhalts, Personen und Rechten, sowie Vernetzung und Lebenszyklus. Auf das Metadatenschema wird im Detail im Vergleich zu den anderen Metadatenschemata noch Stellung genommen.

Als Erweiterung der Core Elements definiert die DCMI als sogenante „DCMI Metadata Terms“ insgesamt 55 weitere Elemente.[14]

B 2. Erweiterte Metadatenschemata für Forschungsdaten

Bei einer eigenständigen Publikation von Forschungsdaten als unabhängige und zitierfähige Objekte, entstehen neue Anforderungen an Metadatenschemata. Neben den klassischen Metadaten, wie sie in Bibliothekskatalogen angewendet werden, sind weitere Metadaten notwendig, die Forschungsdaten spezifisch beschreiben können. Für Forschungsdaten werden somit Metadaten benötigt, die

- die Forschungsdaten innerhalb einer Sammlung bzw. eines Kataloges beschreiben.
  Sie sind am ehesten mit den im vorherigen Kapitel beschriebenen klassischen Katalogdaten zu vergleichen. Diese beschriebenen Metadaten sind die Grundlage, um Forschungsdaten so zu beschreiben, dass sie eindeutig zitiert werden können. Diese Metadaten werden nachfolgend als **Externe Metadaten** bezeichnet.
- die die Forschungsdaten auf fachspezifischer Ebene beschreiben. Diese Metadaten enthalten oftmals detaillierte technische Informationen zu Techniken, Methoden, Parametern etc., die zum fachlichen Verständnis der Forschungsdaten notwendig sind. Diese Metadaten werden nachfolgend als **Interne Metadaten** bezeichnet.

In diesem Kapitel werden verschiedene Ansätze zur Definition von Metadatenschemata für die Beschreibung von Forschungsdaten vorgestellt und analysiert. Die technische Umsetzung der Metadatenschemata ist auf vielfältige Weise möglich, als Stichworte seien hier nur Resource Description Framework (RDF) oder die Extensible Markup Language (XML) genannt. Dieser Aspekt soll hier nicht weiter vertieft werden.

2.1 Externe Metadaten

2.1.1 STD-DOI

In dem an der Technischen Informationsbibliothek (TIB) Hannover durchgeführten Projekt „Publikation und Zitierfähigkeit wissenschaftlicher Forschungsdaten“ wurde eine Infrastruktur zur Registrierung von DOI-Namen und URNs für Wissenschaftliche Datensätze geschaffen und erfolgreich getestet (s. Kapitel A 4.2.1). Mit dem System wurden an der TIB bereits über 650.000 Datensätze aus dem Bereich der Geowissenschaften mit persistenten digitalen Identifikatoren versehen. Ein Teil der Datensätze ist über den Online-Katalog der TIB

verfügbar, auf die restlichen Datensätze kann offen über die Kataloge der beteiligten Datenarchive als sogenannte Publikationsagenten zugegriffen werden.

In diesem Zusammenhang wurde für die Zitierfähigkeit der Daten ein disziplinunabhängiges Metadatenschema entworfen, welches sich stark an der ISO-Norm 690-2 für die Zitierung von Elektronsichen Ressourcen[15] orientiert. Das Schema, genant STD-DOI findet bisher nur Anwendung in den Erd- und Umweltwissenschaften, könnte aber auch in anderen Disziplinen wie beispielsweise in der Chemie genutzt werden. Es enthält sowohl Metadatenfelder, die sich an Dublin Core orientieren, als auch Metadatenfelder, die von der International DOI Foundation (IDF) verlangt werden, und umfasst 25 beschreibende Eingabefelder.[16]

Im Kontext von DataCite wird auf Basis dieses gerade ein eigenes Metadatenschema für Forschungsdaten entwickelt, welches als erster Draft im September 2010 vorliegen wird.

2.1.2 Metadatenschema nach Altman und King

Im April 2007 veröffentlichten Altman und King im D-LIB Magazin einen Entwurf für ein Metadatenschema für die Zitierung von Wissenschaftlichen Daten „*A Proposed Standard for the Scholarly Citation of Quantitative Data*“.[17] Sowohl Micah Altman als auch Gary King sind Vorsitzende des Institute for Quantitative Social Science (IQSS). Dieses Institut ist teilweise eine eigenständige Forschungseinrichtung und teilweise ein integrierter Teil der Harvard Universität in Cambridge MA, USA. Die Aufgabe dieses Institutes ist es für die Sozialwissenschaften und die Gesundheitswissenschaften statistische und analytische Werkzeuge zu schaffen und diese weitläufig verfügbar zu machen. Diese Werkzeuge sollen der Lösung gravierender Probleme dienen, die die Gesellschaft und das Wohlergehen der Bevölkerung betreffen.

Für eine Zusammenarbeit und den Wettbewerb unter Wissenschaftlern halten sie es für nötig einen einfachen Standard zu nutzen, der die Vorteile der gedruckten Zitierung mit denen der elektronischen Daten miteinander verbindet. Der von Altmann und King vorgeschlagene Standard basiert größtenteils auf Dublin Core und stellt eine minimale Menge von Attributen dar.

2.1.3 Metadatenschema des OECD-Verlags

Die OECD ist eine internationale Organisation mit 30 Mitgliedsländern, die sich Demokratie und Marktwirtschaft verpflichtet fühlen. Der ökonomische Nutzen von Bildung für den

Einzelnen und die Gesellschaft sowie Chancengleichheit im Bildungssystem stehen in der bildungspolitischen Arbeit im Vordergrund. In jährlich erscheinende Publikationen veröffentlicht die OECD vergleichende Statistiken und Indikatoren zum Ressourceneinsatz in Form von Finanzmitteln oder Personalausstattung in nationalen Bildungssystemen und analysiert, wie sich Bildung auf Innovationskraft und Arbeitsmarkt auswirken. Um diese Publikationen der OECD suchbar zu machen schlug die Gruppe um Toby Green, den Leiter des OECD-Verlags, im Jahr 2008 ebenfalls ein Metadatenschema vor, das für die Zitierung von wissenschaftlichen Daten von Bedeutung ist.[18] Dieses Schema basiert auf dem Entwurf von Altman und King und besteht aus 27 Elementen.

2.1.4 Metadatenschema DANS (Data Archiving and Networked Services)

Das DANS steht unter der Schirmherrschaft der königlichen niederländischen Akademie für Kunst und Wissenschaft und wird ebenfalls von der niederländischen Organisation für wissenschaftliche Forschung gefördert. Das DANS ist für die Speicherung und Verfügbarmachung von wissenschaftlichen Daten in den Geisteswissenschaften, Sozialwissenschaften und in der Kunst zuständig. Das DANS entwickelt permanent archivierende Dienstleistungen, animiert andere ihrem Beispiel zu folgen und arbeitet eng zusammen mit Datenmanagern, um sicherzustellen, dass ein Maximum an Daten frei für wissenschaftliche Zwecke zur Verfügung steht. Zur Indexierung ihrer Daten werde in DANS 15 DC Terms Elemente verwendet, die ungefähr einer Verfeinerung der Core Elements entsprechen.

2.1.5 Metadatenschema ANDS (Australian National Data Service)

Das ANDS wurde vom Australian Commonwealth Government's Department of Innovation, Industry, Science and Research gegründet und hat sich zum Ziel gemacht die nationale Politik der australischen Forschungsgemeinschaft im Bereich des Datenmanagements zu beeinflussen, über die optimale Handhabung von Daten zu informieren und die weit verteilten Sammlungen an wissenschaftlichen Daten in Australien in eine zusammenhängende Kollektion von wissenschaftlichen Ressourcen umzuwandeln. Für die Verfügbarmachung seiner Daten hat das ANDS ein eigenes Metadatenschema entwickelt. Die Metadaten werden hier in vier verscheidene Gruppen eingeteilt, die in verscheidenen Beziehungen zueinander stehen können: Collection, Service, Party und Activity.

2.2 Detaillierte Analyse der Metadatenschemata

Für die einfachere Gegenüberstellung der Metadatenschemata und den besseren Überblick der einzelnen Metadatenelemente können die Metadatenelemente allgemein in sechs Blöcke zusammengefasst werden. Diese Gruppierung basiert auf der Struktur, die die DCMI für Dublin Core etabbliert hat. Die Blöcke umfassen die Themen identifier, technische Daten, Beschreibung des Inhalts, Personen und Rechte sowie Vernetzung und Lebenszyklus.

ID:

- identifier: eindeutige Identifizierung der Daten durch einen passenden Katalog wie DOI, URN, URL/PURL oder ISBN/ISSN.

Technische Daten:

- format: Die Formatangabe zeigt womit die Dateien geöffnet bzw. weiterverarbeitet werden können.

- type: Der Typ gibt an um was für eine Art Daten es sich handelt. Hierbei unterscheidet man:

- collection: Beschreibung von Unterordnern mit jeweils eigenen Metadaten

- sound: Sprachaufnahmen oder Klänge

- image: Fotos, Drucke, Landkarten und Diagramme aber auch Videos

- **Text:** Schriftliche Dokumente

- physical objects: also reale Objekte aber nicht deren Bilder oder Beschreibungen (Text)!

- interactive document: Dokument, welches Benutzereingaben erfordert wie z. B. ein Formular

- software: ausführbare Dateien

- dataset: der interessanteste Typ für unsere Analyse

- **service:** z. B. ein web-server

- **size:** Größe und Größeneinheit der Daten

- **mode:** Modus in dem die primären erhalten wurden

- language: Sprache des Dateninhalts. Meist Sprachkürzel nach ISO 639 oder DIN EN

ISO 3166

Beschreibung des Inhalts:

- **title:** Titel des Dokuments unter dem es formell publiziert wird
- **subtitle:** näher beschreibender Titel
- **subject:** Thema des Inhalts in suchbaren Schlagwörtern/ keywords
- **edition:** bei geänderten Ressourcen die Auflage der Ressource
- **coverage:** Limits des vom Dokument abgedeckten Bereichs (temporal, spatial)
- **descripion:** Zusammenfassung des Inhalts; abstract; table of contents
- **classification/ dicipline**: Einordnung der Ressource, Fachrichtung

Personen und Rechte:

- **creator:** verantwortlicher Verfasser oder Urheber der Datei
- **publisher:** Verleger oder Herausgeber
- **contributor:** Name je einer Person, die einen Beitrag zu den Daten geleistet hat
- **rightsholder:** Person/ Organisation, die Eigner/ Verwerter der Daten
- **rights:** Information zur Klarstellung der Rechte; Lizenz/ license
- **provenance:** Angaben zur Echtheit der Daten
- **registration agency:** Agentur, die die Registrierung vorgenommen hat

Vernetzung:

- **source:** Verweist auf Daten, von denen diese abgeleitet wurden
- **relation:** Verweist auf Daten, die mit diesen in Beziehung stehen
- **audience:** Klassifizierung der Zielgruppe der Daten
- **external links:** Links zu Daten, Personen, Instituten etc.
- Parent-child linking/ belonging: Stammdaten oder Datenauszug

Lebenszyklus:

- date: relevantes Datum z.B: copyright, veröffentlicht am, erstellt am

- next publication date: Bei sich fortsetzenden Daten

- peridicity: Periode der sich fortsetzenden Daten

- period covered start-end: Bei Veröffntlichungszeiträumen Anfang und Ende

- numeric fingerprint: Andere nicht-identifier Erkennungsnummer

- publication place: Ort der Publikation

- address: Standort der Daten

- embargo: Zeitraum für den der Inhalt gesperrt ist

Hierbei ist anzumerken, dass die einzelnen Elemente möglicherweise auch in anderen Blöcken gepasst hätten. Zum Teil wurden unterschiedliche Begriffe mit gleichem oder ähnlichem Inhalt zusammengefasst unter einem Begriff.

Das Element „embargo“ beinhaltet einen Zeitraum für den der Inhalt gesperrt wäre. Dieser Zeitraum erlaubt nur dem Erzeuger und den vom Erzeuger benannten zugriffsberechtigten Personen den Zugriff auf die Daten. Dieser Embargozeitraum sollte limitiert und nicht uneingeschränkt sein. Die gespeicherten Daten sollten nur solange eingeschränkt zugänglich sein, solange die Daten für Publikationen benötigt werden und nicht darüber hinaus, da die Daten auch für andere Personen suchbar und einlesbar sein sollen. Bei eingeschränkt zugänglichen Daten stellt sich die Frage, ob diese Daten auch externe Metadaten erhalten sollen. Bei einem durch Metadaten beschriebenen Datensatz würden andere User zwar sehen, dass es bereits Daten gibt, und bräuchten diese nicht wiederholt erzeugen, da sie nicht mehr publizierbar sind. Andererseits könnten Nutzer, die die selben Daten erzeugt haben, diese nicht mehr anderen zur Verfügung stellen. Es ist umstritten ein solches Element einzuführen.

publishing metadata field	Simple Dublin Core	Altman & King	DANS	ANDS	STD-DOI	OECD/Toby Green
persistent identifiers:	<persistent identifiers>	<unique,persistent, global identifier>	<persistent identifiers>	<key>	<Recourceidentifier>	<unique,persistent, global identifier>
DOI	1	1	1	1	<DOI>	1
other identifiers	1	1	5	5	optional	1
technische Daten:						
format	<format>	-	<format>	-	<format>	-
language	<language>	-	<language>	-	<language>	optional
type	<type>	-	-	<electronic>	optional	-
size	-	-	-	-	<size>	optional
mode	-	-	-	-	optional	-
Beschreibung des Inhalts:						
title	<title>	<main title>	<title>	<name>	<title>	<main title>
subtitle	-	-	-	<namepart>	optional	optional
edition	-	-	-	<edition>	<edition>	-
subject/ keywords	<subject>	-	<subject>	<subject>	2	optional

coverage	<coverage>	-	<spatial coverage>	<spatial>	-	<countries covered/optional>
description/ abstract	<description>	-	<description>	<description>	<description>	<abstract>*
classification/ discipline	-	-	-	-	<discipline>	optional
Personen und Rechte:						
creator	<creator>	<authors>	<creator>	<party>	<creator>	<authors>
publisher	<publisher>	-	-	<party>	<publisher>	optional
contributor	<contributor>	<authors> **	<contributor>	<party>	<contributor>	<authors> **
rightsholder	<rightsholder>	-	-	-	-	<is copyrighted by>
rights	<rights>	-	<access rights>	<access policy>	-	-
provenance	<provenance>	-	-	-	-	-
registration agency	-	-	-	-	<registration agency>	-
Vernetzung:						
source	<source>	-	-	<originating>	-	-

relation	<relation>	-	<relation>	<related object>	<related identifiers/ relationtype>	optional
audience	<audience>	-	<audience>	-	3	-
external links	-	-	-	<external links>	-	optional
parent-child relationship/ belongs to	-	-	-	-	-	optional
Lebenszyklus:						
date	<date>	<publication date>	<date created/ date availlable>	<date accessioned>	<date> ***	<publication date>
next publication date	-	-	-	-	4	<next publication date>
periodicity	-	-	-	-	-	<peridicity>
period covered start-end	-	-	-	-	<startpublicationdate>	optional
numeric fingerprint	-	<numeric fingerprint>	-	-	-	-
publication place	-	-	-	-	<publication place>	-
address	-	-	-	<address>	-	-

Tabelle 1: Gegenüberstellung der einzelnen Metadatenschemata für Forschungsdaten

* = Toby Green unterscheidet in Unterpunkten long abstract und short abstract.

** = Toby Green und Altman-King unterscheiden nicht separat creator und contributor sonder fassen beide Punkte unter authors zusammen.

*** = Das STD-DOI-Schema unterscheidet in Subpunkten noch issue date, creation date, publication date

1 = Bis auf das STD-DOI unterscheiden die anderen Schemata die PI nicht.

2 = Dieser Teil ist integriert unter dem Begriff description

3 = Die Zielgruppe wird impliziert durch den Begriff diszipline mitbeschrieben, jedoch nicht ausdrücklich erwähnt.

4 = Dieser Begriff wird z.T. durch andere Begriffe unter date mitbestimmt.

5 = Diese identifier sind nicht zwangsläufig persistent.

ID:

Jedes dieser Schemata besitzt mindestens ein Element zur Bestimmung der ID. Hierbei unterscheidet keines dieser Schemata außer STD-DOI die verschiedenen persistenten Identifier wie URN, DOI usw. Das STD-DOI hingegen unterscheidet zwischen DOI und anderen persistenten Identifiern.

Die Schemata von DANS und ANDS hingegen haben zusätzlich zu den persistenten Identifiern noch Identifier, die für die interne Nutzung vorhanden sind.

Technische Daten:

Bei den technischen Daten, die *format, language, type* und *size* enthalten, gibt es große Unterschiede zwischen den Schemata, während das STD-DOI zumindest optional alle Felder abdeckt, benötigt das Altman-King-Schema (AKS) keines dieser Elemente. Das Element *language* wird in den anderen Schemata mindestens optional angeboten außer bei ANDS, die ein ausschließlich nationales Institut sind. Auf die Elementen *type* und *format* wird in allen Schemata außer dem AKS und dem OECD-Schema Bezug genommen.

Beschreibung des Inhalts:

Zu diesem Block gehören die Elemente *title, subtitle, edition, subject/ key words, coverage, description/ abstract* und *classification/ discipline*. Alle Schemata enthalten das Element title, welches den Titel des Dokuments/ der Daten beschreibt. Das AKS begnügt sich mit diesem einen Element zur Beschreibung des Inhalts, wohingegen die anderen Schemata auch die Elemente *subject/ keywords* und *description/ abstract* anführen. Im STD-DOI wird jedoch das Element *subject/ keywords* nicht explizit genannt, aber in *description* als sonstige Angabe geführt. Andere Elemente sind nur vereinzelt von wenigen Schemata geführt.

Personen und Rechte:

Die hier enthaltenen Elemente *creator* und *contributor* werden von allen Schemata unterstützt, auch wenn das AKS und Toby Green *creator* und *contributor* unter *authors* zusammenfassen. Das Element *publisher* kommt noch in vier der sechs Schemata vor. Die Elemente *rights* und *rightsholder* spielen eine untergeordnete Rolle, in den Schemata von DC, DANS, ANDS und OECD werden diese genannt. Hierbei können sich auch beide Begriffe überschneiden.

Vernetzung:

Bei dem Block Vernetzung sind Elemente aufgeführt die eine Verbindung oder einen Bezug zu anderen Daten, Personen oder Quellen haben. Die Elemente hier sind *source*, *relation, audience, external links* und *parent-child linking/ belonging*. Das AKS nennt keine Elemente in diesem Block, wohingegen die anderen Schemata zumindest den Begriff *relation*

anführen. Dieser Begriff ist hierbei sehr weitläufig und kann andere Elemente dieses Blocks mit einbeziehen. Die Schemata von DC und DANS sowie implizit das STD-DOI durch das Element *discipline* richten die Daten an eine Zielgruppe (*audience*).

Lebenszyklus:

In diesem Block sind die verschiedene relevante Daten und Plätze der Datensätze aufgeführt. Jedes Schemata hat ein Element *date*, das vor allem das Datum der Veröffentlichung oder der Verfügbarkeit beschreibt. Außer dem STD-DOI und dem ANDS wird in keinem Schema Wert auf Orte gelegt, wo Daten gespeichert oder veröffentlicht werden.

2.3 Zitierung von Forschungsdaten

Für die Zitierung von Forschungsdaten ist es sinnvoll, wie auch von Toby Green und Altmann & King vorgeschlagen und von der TIB, bzw. DataCite praktiziert, sich prinzipiell an Zitierungsformen für wissenschaftliche Artikel zu orientieren. Die meisten Zitierungsformen orientieren sich am Standard ISO 690-2.

Hierbei werden beispielsweise folgende Metadatenelemente, die auch in allen untersuchten Schemata vorhanden sind, verwendet (in DCMI Elementen):

<creator> (<data>)
<title>
<publisher>, <publicationPlace>
<identifier>

Die gleiche Struktur und Syntax und die Verwendung von DOI-Namen als Identifier ermöglicht eine elegante Verlinkung zwischen einem wissenschaftlichen Artikel und den im Artikel analysierten Forschungsdaten. Artikel und Datensatz sind durch ihre jeweiligen DOI Namen in gleicher Weise eigenständig zitierbar. Diese Form der Zitierung und Verlinkung bietet sich insbesondere bei Forschungsdaten an, die in direkter Beziehung zu Wissenschaftlichen Artikeln stehen, sogenannte „suppplementary data“.

So wird beispielsweise der Datensatz

Kuhlmann, H et al. (2009):
Age models, iron intensity, magnetic susceptibility records and dry bulk density of sediment cores from around the Canary Islands.
PANGAEA, Bremen
doi:10.1594/PANGAEA.727522

in folgendem Artikel verwendet:

Kuhlmann, Holger; Freudenthal, Tim; Helmke, Peer; Meggers, Helge (2004): Reconstruction of paleoceanography off NW Africa during the last 40,000 years: influence of local and regional factors on sediment accumulation.
Marine Geology, 207(1-4), 209-224,
doi:10.1016/j.margeo.2004.03.017

Diese Verlinkung wird auch bei der Darstellung des Artikels über das Portal „ScienceDirect“ dargestellt (Abb 17.). Durch eine Kooperation des Datenzentrums „Publishing Network for Geoscientific & Environmental Data (PANGAEA)“ mit Elsevier wird bei jedem Artikel, der in ScienceDirect angezeigt wird automatisch geprüft, ob für diesen Artikel Forschungsdaten verfügbar sind, die mit einer DOI registriert wurden, und ggf. ein Verweis direkt auf die Vorschauseite des Artikels platziert.

Abb. 17: Anzeige eines Artikels in ScienceDirect mit Verweis auf die verfügbaren Forschungsdaten (Supplementary Data)

2.4 Interne Metadaten

2.4.1 Einleitung - Fachspezifisches Metadatenschema für Forschungsdaten aus der Chemie

Neben den externen Metadaten sind für die fachspezifische Beschreibung von Forschungsdaten aus der Chemie die internen Metadaten von herausragender Bedeutung. Diese internen Metadaten liefern notwendige Informationen, um einen Forschungsdatensatz aus inhaltlicher Sicht zu verstehen und nachvollziehen zu können.

Es ist an dieser Stelle anzumerken, dass in der Chemie bei der Generierung von Datensätzen in Austauschformaten sowohl die Daten als auch ein Teil der zugehörigen Metadaten in einer gemeinsamen Datei abgespeichert wird. Dieser Umstand ist Ansatzpunkt der nachfolgenden Überlegungen für Metadatenschemata interner Metadaten für Forschungsdaten in der Chemie. Insbesondere soll hier das JCAMP-DX Austauschformat näher untersucht werden.

Die präparative Chemie ist bis heute ein elementarer Schwerpunkt der chemischen Forschung. Sie steht am Beginn einer Prozesskette vom Stoff mit einer erwünschten Eigenschaft bis hin zum fertigen Produkt. Die Herstellung und Charakterisierung neuer Substanzen mit gezielter Struktur und Reaktivität wird begleitet von einer umfassenden Reihe analytischer Methoden. Im Verlauf dieses Prozesses werden im Rahmen der Dokumentation und Spezifikation verschiedenste Forschungsdaten in vielfältigen Datenformaten erzeugt, erfasst und archiviert. Ebenso spezifisch wie die Forschungsdaten sind die affiliierten Metadaten, die eingesetzt werden, um diese Forschungsdaten zu beschreiben und dadurch besser in Datenbanken auffindbar zu machen.

Zusätzlich zu ***externen Metadaten***, die der Beschreibung von Forschungsdaten innerhalb eines Kataloges oder einer Datenbank dienen, ist die Erfassung ***interner Metadaten*** von Forschungsdaten im Focus der Konzeptstudie.

Unter diesen Metadaten versteht man die anwendungsbezogenen physikalischen und technischen Parameter, die sich unmittelbar auf die Probe und die Erzeugung chemischer Forschungsdaten beziehen. Die Beschaffenheit und die Anzahl interner Metadaten sind abhängig von der angewendeten Untersuchungsmethode.

Im Rahmen der Konzeptstudie wurden die gängigsten analytischen Methoden der organischen Chemie näher betrachtet. Hierzu zählen die Kernresonanz-, Infrarot- und Ultraviolett-Spektroskopie, Massenspektrometrie, als chromatographische Verfahren HPLC und GPC und die Röntgenstrukturanalyse.

Zu den ausgewählten Methoden wurden Beispieldateien (siehe anliegender Datenträger) erfasst und die verwendeten Datenformate im Hinblick auf ihre Metadaten analysiert. Desweiteren wurde die Auswertungssoftware insbesondere unter dem Aspekt der Möglichkeit des Datenexports im JCAMP-DX-Format betrachtet. JCAMP-DX (**J**oint **C**ommittee on **A**tomic and **M**olecular Physical **D**ata E**X**change) ist ein universelles Austauschformat, dass für verschiedene spektroskopische Methoden einsetzbar ist.[19]

Da die Datenformate nicht nur von der Untersuchungsmethode, sondern insbesondere auch vom Hersteller bzw. Entwickler des verwendeten Gerätes abhängig sind, wurden erste Kontakte zu Vertretern von marktführenden Firmen hergestellt, die auch einige Beispielspektren zur Verfügung stellten.

2.4.2 JCAMP-DX

1988 wurde JCAMP-DX 4.24 als standardisiertes Datenformat zum Austausch von (zunächst) IR-Spektren und verwandten chemischen und physikalischen Informationen verschiedener Datensysteme unterschiedlicher Gerätehersteller entwickelt.[20] [21] Es basiert darauf, dass alle Daten unter verschiedenen Feldern (***L**abelled **D**ata **R**ecords – LDRs)* variabler Länge mit ASCII Zeichen gespeichert werden. Daraus entsteht ein Textfile, das mit einem Editor bearbeitet werden kann. Verschiedene JCAMP-DX Strukturen sind möglich. Eine einfache Datei besteht aus einem einzigen Block, der folgendermaßen aufgebaut ist:

```
CORE FIXED HEADER
CORE VARIABLE HEADER
NOTES
CORE DATA
END
```

Gibt es mehrere Datensätze, werden diese in mehreren (voneinander unabhängigen) Blöcken in einer Datei gespeichert, z. B.

```
LINK BLOCK (Informationen zum gesamten Datensatz sowie Projektbeschreibung)
BLOCK2 (z. B. IR-Spektrum)
BLOCK3 (z. B. NMR-Spektrum)
BLOCK4 (z. B. Massenspektrum)
BLOCK5 (z. B. Struktur)
END.
```

Mehrdimensionale Daten werden unter Verwendung von NTUPLES und PAGES gespeichert. Dadurch können verschiedene Blöcke, z. B. für den Total Ion Current (TIC) und Single Ion Current (SIC) eines LC/MS-Spektrums verknüpft werden.

JCAMP-DX-Dateien beinhalten also Metadaten eingeteilt in CORE (maschinenlesbar, essentiell) und NOTES (für Menschen lesbar, optional).

Der CORE besteht aus vier Teilen:

Core Fixed Header Informationen = generelle LDRs, erforderlich für alle JCAMP-DX-Files
Core Variable Header Information = methodenspezifische LDRs, spezielle LDRs für spezielle JCAMP-DX-Formate
Core Data = relevante Parameter
Core Data Table = XY Daten, können komprimiert vorliegen (squeezed, diffdup…)

NOTES vervollständigen den CORE und beschreiben ein Experiment detailliert. NOTES können Informationen beinhalten, die nicht in den Forschungsdaten enthalten sind. Der Inhalt ist abhängig vom Benutzer und von der Messtechnik. NOTES können daher stark variieren.

JCAMP-DX Protokolle sind erweiterbar und wurden auch an weitere Messmethoden (Raman, UV, NMR, MS, X-Ray, Chromatogramme, Thermogramme) angepasst.

2002 erschien die Weiterentwicklung von JCAMP-DX V.5.00 für eindimensionale Datensätze zu JCAMP-DX V.6.00 für mehrdimensionale Datensätze, wie sie z. B. bei zweidimensionalen NMR-Methoden vorliegen.[22]

2.4.3 Software

Die im Rahmen der Konzeptstudie gesichtete Software zeigt nur einen kleinen Ausschnitt aus der Vielfalt der verwendeten Datenverarbeitungssysteme, die mit der Methode, dem Hersteller und auch dem Alter der verschiedenen Instrumente variieren.

Im Falle der Kernresonanz wurden zunächst TopSpin™-Spektren der im Hause verwendeten Bruker Maschinen und Delta™-Spektren eines Jeol Spektrometers (des Pharmazeutischen Instituts der Universität Marburg) eingehender untersucht. Beide Applikationen können NMR-Spektren als JCAMP-DX Dateien exportieren. In TopSpin™ sind JCAMP-DX Formate der Version 6.0 in unterschiedlichen Kompressionsmodi (z. B. packed, squeezed, diffdup) auswählbar. Zusätzlich kann man festlegen, welche Daten in der JCAMP-DX Datei enthalten sein sollen (z. B. Realspektrum, Imaginärspektrum, FID). Davon hängt u. a. der Umfang der JCAMP-DX Dateien ab.

Auf dem Gebiet der Massenspektrometrie liegen uns momentan nur zu MASPECII detaillierte Informationen vor. MASPECII kann Formate verschiedener Hersteller im- und exportieren. Als Standardformat kann ANDI/MS (*.cdf) bzw. Text ausgewählt werden.

Finnigan bietet die Xcalibur-Software für massenspektrometrische Anwendungen an. Mass Frontier ist ein hierin enthaltenes Softwarepaket zur Verwaltung und Interpretation von Massenspektren und Chromatogrammen. Mass Frontier unterstützt JCAMP-DX.

Opus ist die von Bruker Optics entwickelte Software für IR-Spektren. Ebenso wie TopSpin™ kann auch Opus JCAMP-DX Dateien öffnen, bearbeiten und generieren. Mit Opus können beispielsweise auch IR-Spektren von Nicolet und Perkin Elmer eingelesen werden. Der OPUS Viewer ist ein frei erhältliches Programm, das alle Arten von Bruker OPUS, JCAMP-DX und Galactic Grams Dateien (Thermo Fisher Scientific) öffnen kann. Er erlaubt die Darstellung spektraler Daten (Spektren, Interferogramme etc.) ebenso wie weitere mit dem Spektrum gespeicherte Daten wie Evaluationsberichte, Messgrößen, Parameter, Protokolle, Signaturen etc.

JASCO bietet als Softwarelösung den Spectra Manager II™ an. Er unterstützt standardisierte Datenformate wie ASCII, TXT und JCAMP-DX, um einen Austausch zu anderen Datenverarbeitungssystemen und Softwarelösungen wie z. B. Macintosh zu gewährleisten.

UV WinLab (aktuelle Version 2.8) unterstützt die Lambda-Platform von Perkin Elmer. Mit UV WinLab ist das Öffnen und Speichern im JCAMP-DX Format mit der Extension DX möglich.

Abgesehen von den proprietären Softwareformaten gibt es auch sogenannte Konverter, also universell einsetzbare Programme zur Visualisierung, Verarbeitung und Verwaltung spektroskopischer Daten.[23] [24]

2007 wurde JSpecView, ein Open Source Java Viewer und Konverter für JCAMP-DX und XML Spektraldaten, veröffentlicht.[25] [26] JSpecView kann eine Reihe von JCAMP-DX und AnIML/CML Formaten und Protokollen darstellen und ermöglicht außerdem die Bearbeitung und Auswertung der Daten z. B. durch graphische Überlagerung mehrerer Spektren.[27] Eine Zuordnung von Signal und entsprechender Struktur durch Kombination von JSpec und Jmol für IR, NMR, MS ist ebenfalls machbar. JSpecView erlaubt den Export zu verschiedenen JCAMP-DX Komprimierungsformaten und in Formate wie AnIML, CML, JPG, PNG oder SVG.

Verschiedene weitere z. T. frei erhältliche Konverter sind im WWW zu finden.[28] FIDtoJCAMP konvertiert ursprüngliche Bruker MALDI TOFMS FID Formate so, dass sie mit JSpecView oder anderen Viewern dargestellt werden können.[29]

Spekwin32 1.71.3 ist eine freie Spektroskopie-Software für UV-VIS, NIR, IR, Raman und Fluoreszenz. [30] Sie ermöglicht die simultane Darstellung und Bearbeitung zahlreicher Spektrenformate: ASCII, JCAMP-DX, THERMO Galactic GRAMS SPC, Varian Cary 50, Perkin Elmer, Avantes Avasoft, Roper Scientific, Scinco Neosys. Ebenso ist der Datenexport als ASCII, Galactic SPC oder binäre Daten sowie graphische Dateien (WMF, GIF, PNG, TIFF, BMP) möglich.

Abgesehen von den Open Source Softwarelösungen sind kommerzielle Konverter erhältlich. Als ein Beispiel wäre hier Grams/AI zu nennen.[31] Grams arbeitet mit nahezu allen Arten von Analyseinstrumenten der verschiedensten Methoden. Die GRAMS-Softwaresuite erkennt offenbar automatisch Hunderte verschiedener Dateiformate wie z. B. Agilent/HP, Beckman, Bio-Rad, Bruker, Gilson, Hitachi, PerkinElmer, Shimadzu, Varian und Waters/Micromass. Es werden aber auch zahlreiche universelle Datenformate unterstützt wie SPC, ASCII, JCAMP und AnDI/NetCDF. Andere Beispiele sind bei ACD (Advanced Chemistry Development) zu finden.[32] Hier werden der ACD/MS Processor, der ACD/NMR Processor und der ACD/UV-IR Processor angeboten. Diese Programme unterstützen Datenformate zahlreicher Anbieter und Open Source Formate wie ASCII, JCAMP und netCDF.

Da die genannten Softwarelösungen bisher nicht getestet wurden, konnten die Angaben dazu bisher nicht nachvollzogen bzw. bewerten werden.

2.4.4 Beispiele

Zu den im Rahmen der Konzeptstudie ausgewählten Messmethoden (NMR, MS, IR, UV, HPLC/GPC und Röntgenstrukturanalyse) wurden Beispieldatensätze beliebiger Proben erfasst und Auszüge daraus dargestellt. Eine Auflistung der Dateien befindet sich im Dateiverzeichnis. Dem Dokument ist eine CD zugefügt, auf der die Beispieldateien abgelegt sind.

Kernresonanz (NMR)

Die Kernresonanz ist in der präparativen und analytischen Chemie eine der gebräuchlichsten analytischen Methoden. In der Forschung gehört sie zur täglichen Routine. Die Firma Bruker BioSpin ist einer der marktführenden Hersteller von NMR-Technologie. Alternativ werden Geräte der Firmen Varian und Jeol häufig eingesetzt. Jede dieser Firmen bietet eine eigene Software zur Messung, Bearbeitung, Analyse und Darstellung der NMR-Daten an. Das hat zur Folge, dass die Daten in den Instituten in proprietären Datenformaten vorliegen. Die aktuelle NMR-Software der Firma Bruker ist TopSpin™. Ein mit TopSpin™ generierter NMR-Datensatz besteht aus einem Ordner mit Unterordnern in mehreren Hierarchien, die wiederum diverse Dateien umfassen. Dazu gehören u. a. Textdateien wie das Peaklisting

(peak.txt) oder eine Parameter Datei (parm.txt) die wichtige interne Metadaten der Messung enthält.

Als Austauschformat bietet Bruker JCAMP-DX an. Eine JCAMP-DX Datei ist eine Textdatei, die mit einem ASCII Texteditor eingesehen, bearbeitet und kommentiert werden kann. So besteht z. B. die Datei des Protonenresonanzspektrums von 1,3-Propandiol aus 5535 Zeilen, die sich zusammensetzten aus Abschnitten mit unterschiedlichen Metadaten, Parametern und Messdaten.

Den Kopfeintrag bilden rund 40 Zeilen, die wichtige Angaben zur Probe, zur Messung, zum Messgerät und zum Datenformat enthalten. Diese Daten könnten auch bei einer Datenbankrecherche zum Auffinden des Spektrums nützlich sein, wobei der Titel von der messenden Person (Operator) eingegeben wird und frei wählbar ist. Häufig wird an dieser Stelle eine individuelle Kennung gewählt, die nicht zwingend einer genormten Probenbezeichnung entspricht.

Auszug aus einem mit TopSpin™ erzeugten JCAMP-DX File:

```
##TITLE=1,3-Propandiol  1H  CDCl3 05.06.07
##JCAMPDX= 6.0          $$ Bruker NMR JCAMP-DX V2.0
##DATA TYPE= NMR FID
##DATA CLASS= NTUPLES
##ORIGIN= Bruker BioSpin GmbH
##OWNER= ac
$$ 1.72 TOPSPIN                Version 2.1
$$ 2009-12-03 15:22:27.475 +0100  Helch@AC17
$$ Compression mode = diff/dup
##.OBSERVE FREQUENCY= 500.13350091
##.OBSERVE NUCLEUS= ^1H
##.DELAY= (7, 7)
##.ACQUISITION MODE= SIMULTANEOUS (DQD)
##.ACQUISITION SCHEME= undefined
##.AVERAGES= 16
##.DIGITISER RES= 18
##SPECTROMETER/DATA SYSTEM= av500
##.PULSE SEQUENCE= zg30
##.SOLVENT NAME= CDCl3
##.SHIFT REFERENCE= INTERNAL, CDCl3, 1, 14.98474
##AUDIT TRAIL= $$ (NUMBER, WHEN, WHO, WHERE, PROCESS,
```

Danach folgen Abschnitte, die spezielle Parameter enthalten, wie z. B.:

```
$BRUKER FILE EXP=uxnmr.par
$BRUKER FILE PROC=integrals.txt
$BRUKER FILE PROC=intrng
$BRUKER FILE PROC=parm.txt
```

Ab Zeile 1504 bis zum Ende wird der FID in Form einer komprimierten Datentabelle aufgelistet.

```
##DATA TABLE= (X++(R..R)), XYDATA
0@@@@@@@@@@@@@Ac2C2c2C2f4F4g6I6a19A28a60A92b24B58c04C
46d00D63e28E95
33f76G66h61I75a095A218a358A518a687A875b080B298b540B806c096C411
```

Die NMR-Software Delta™ der Firma Jeol ermöglicht ebenso den Import und Export von Daten anderer Spektrometer-Hersteller. Sie beinhaltet die Option zur Konvertierung der proprietären Daten in JCAMP-DX 6.0. Die so erzeugten Dateien tragen die Extension JDX. Auch hier findet man einen Kopfeintrag mit wichtigen Metadaten. Er umfasst jedoch nur 21 Zeilen. Danach folgen rund 400 Zeilen Jeol spezifische Parameter und anschließend die eigentlichen Messdaten (insgesamt ca. 8800 Zeilen).

Auszug aus einem mit Delta™ erzeugten JCAMP-DX File:

```
##TITLE= y/y633r
##JCAMP-DX= 6.00 $$ JEOL NMR v1.10
##DATA TYPE= NMR SPECTRUM
##DATA CLASS= NTUPLES
##NUM DIM= 1
##ORIGIN= DELTA2_NMR
##OWNER= AUTO_ECX400
##LONG DATE= 2009/10/ 6 15:46:52
##SPECTROMETER/DATA SYSTEM= DELTA2_NMR
##.OBSERVE FREQUENCY= 3.99787870642448E+02
##.OBSERVE NUCLEUS= ^1H
##.DELAY= (1.614E+01,1.614E+01)
##.ACQUISITION MODE= SIMULTANEOUS
##.ACQUISITION SCHEME= PHASE SENSITIVE
##.PULSE SEQUENCE= single_pulse.ex2
##.SOLVENT NAME= CHLOROFORM-D
##.FIELD= 9.3899
##.FILTER WIDTH= 1.80443647311872E+01
##.ACQUISITION TIME= 4.36731904
##.SPINNING RATE= 16
##.OBSERVE 90= 12
$$ ----------------------
$$ JEOL SPECIFIC PARAMETERS
```

Auch wenn der Aufbau augenscheinlich vom ersten Beispiel abweicht, die grundlegende Struktur des JCAMP-DX Files ist gleich und im Protokoll festgelegt.

Ein Datenaustausch zwischen Jeol und Bruker NMR-Geräten sollte demnach problemlos möglich sein. Erste Tests zeigten jedoch, dass die gegenseitige Lesbarkeit nicht ohne Schwierigkeiten verläuft. Ein von einem Bruker-Gerät erzeugtes Spektrum konnte von Delta™ problemlos eingelesen werden (Abb. 18). Umgekehrt war eine Lesbarkeit der Delta™-JCAMP-DX Daten in TopSpin™ zwar möglich, es konnte aber kein sinnvolles Spektrum erzeugt werden. Abb. 19 zeigt ein Spektrum, das mit einem Jeol-Gerät gemessen wurde.

Abb. 20 zeigt die graphische Darstellung des Jeol JCAMP-DX-Spektrums mit TopSpin ™. Von dieser Messung lag neben dem Spektrum-JCAMP-DX-File auch FID-JCAMP-DX-File vor. Dieser konnte transformiert werden, das resultierende „Spektrum“ ist in Abb. 21 dargestellt.

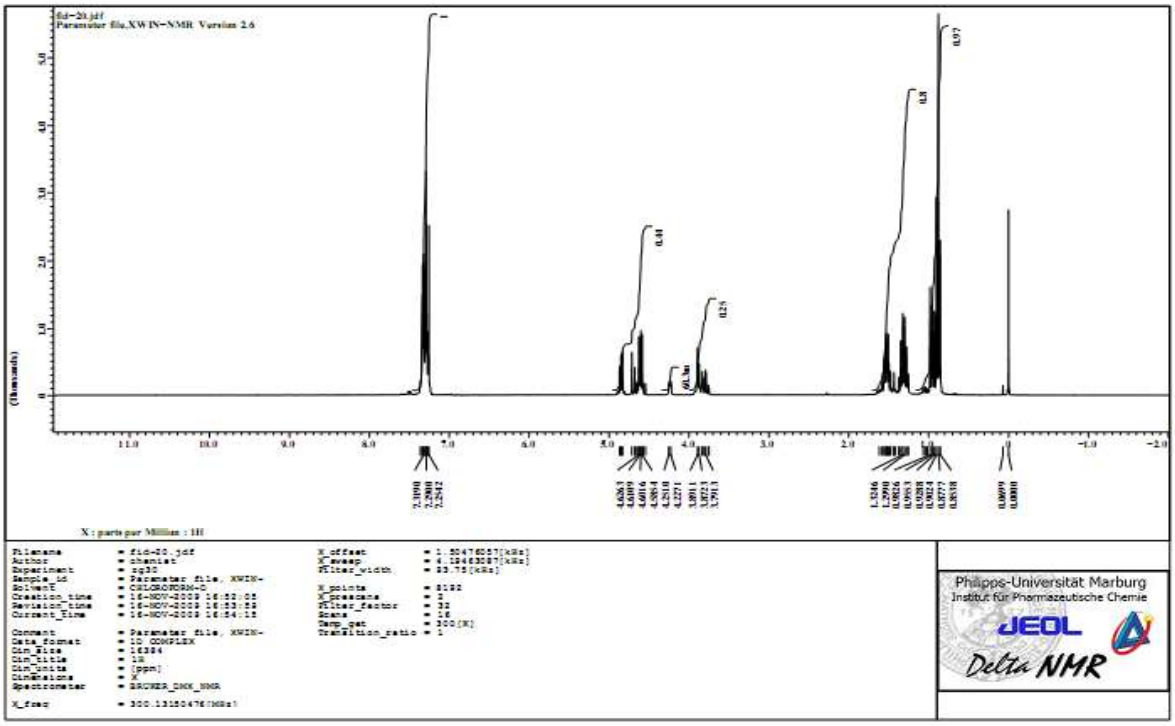


Abb. 18: Darstellung eines Bruker-Spektrums mit Jeol Delta™

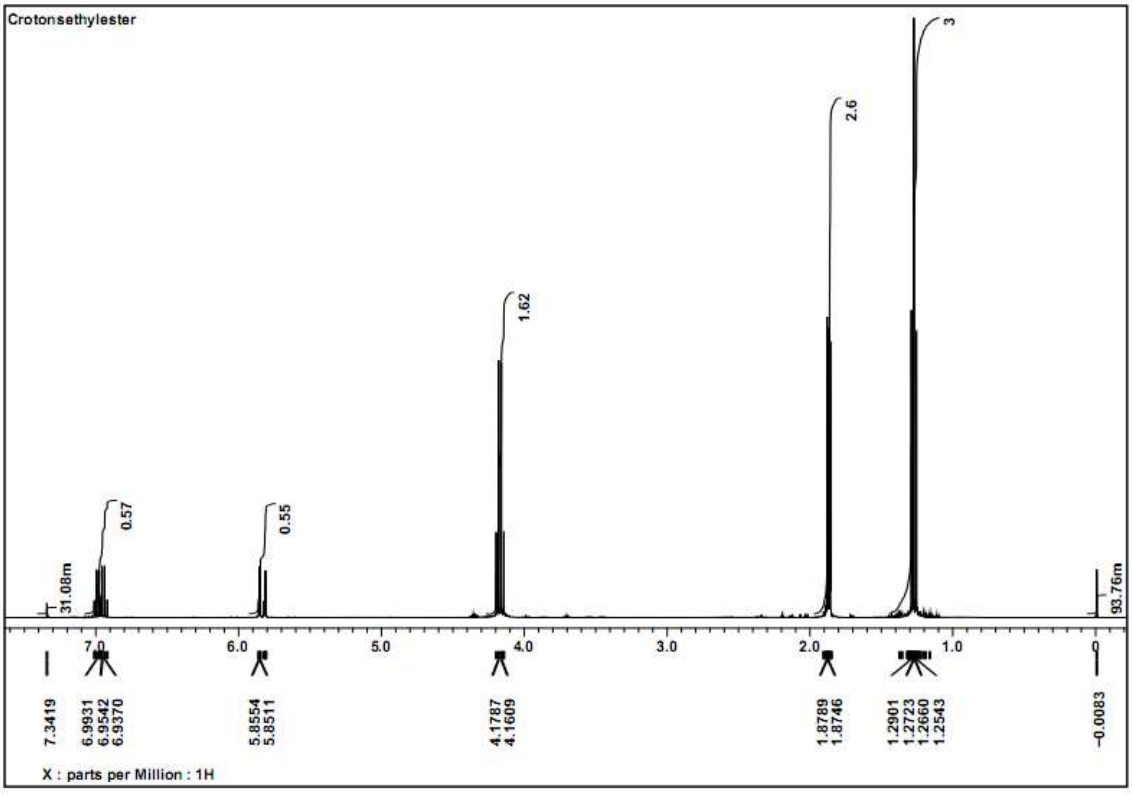


Abb. 19: Crotonsethylester_1H.pdf, Probe mit Delta™ gemessen und visualisiert

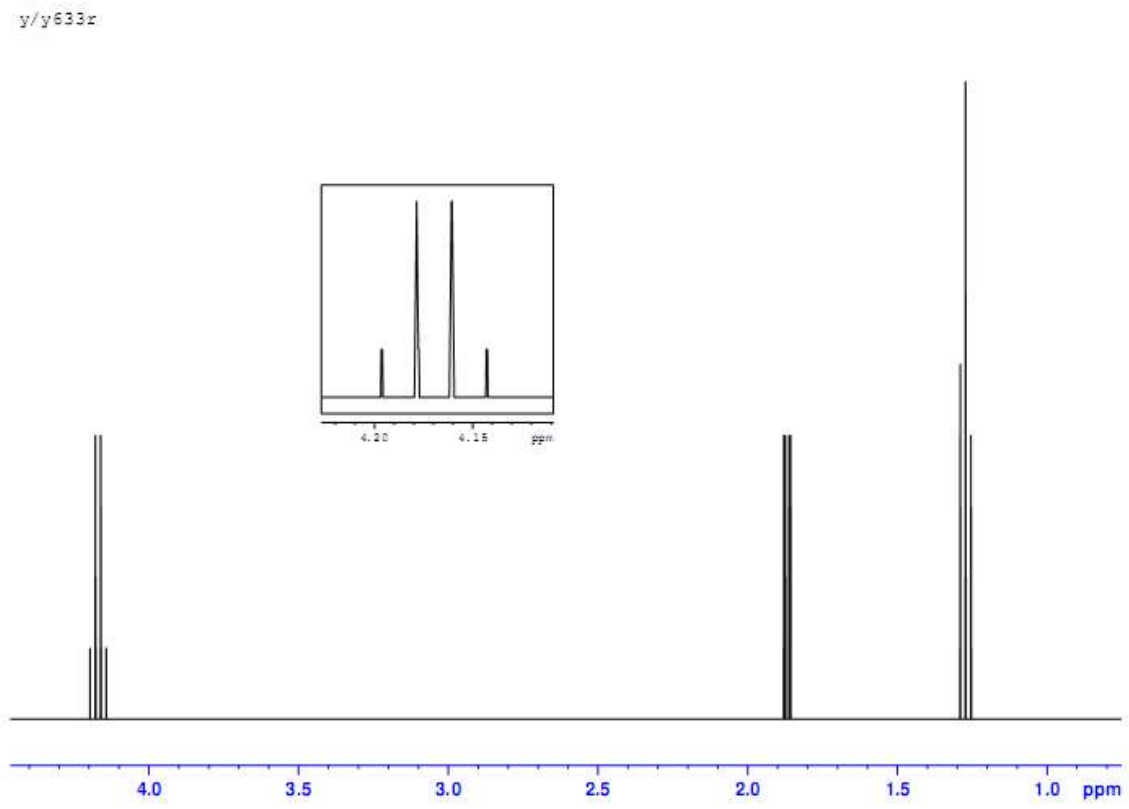


Abb. 20: Graphische Darstellung mit TopSpin™ des Jeol-Files CrotonsethylesterSpectrum_1H_JCAMP-DX6.jdx

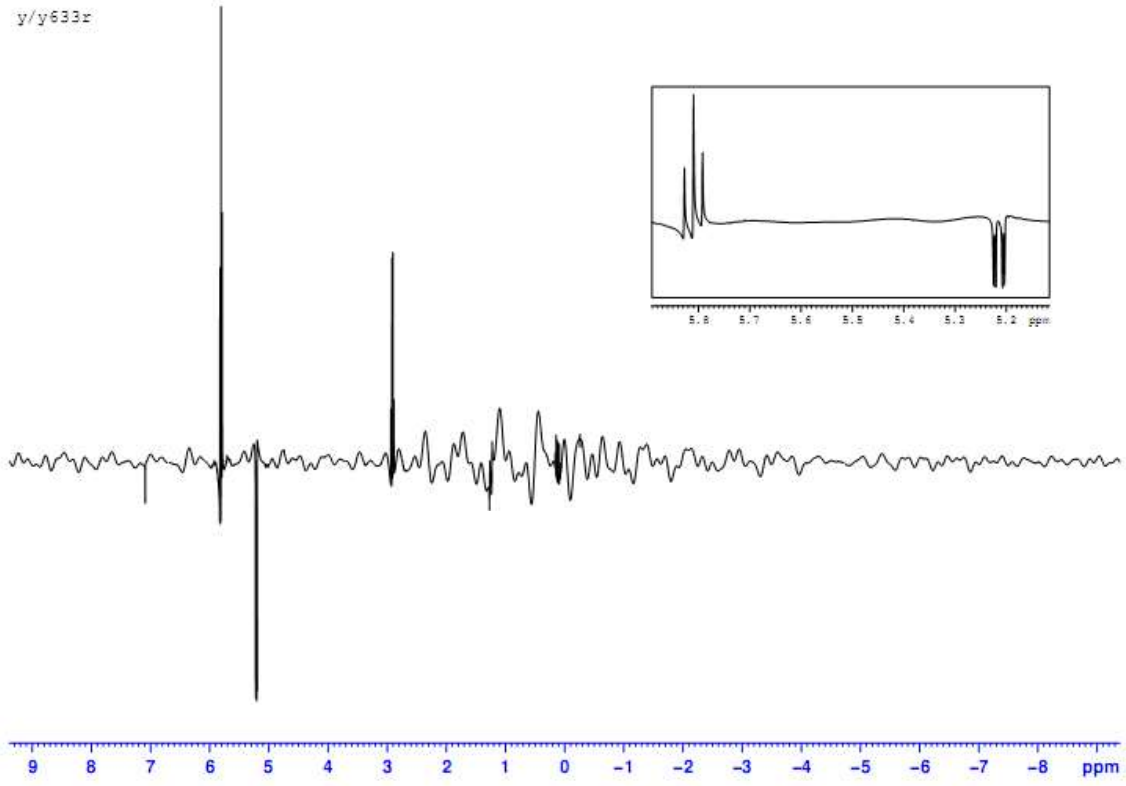


Abb. 21: Aus dem FID des Jeol-Files CrotonsethylesterFID_1H_JCAMP-DX6.jdx mit TopSpin™ transformiertes Spektrum

Offenbar sind hier beim Konvertieren der Daten Informationen verloren gegangen. Die genaue Ursache dieses Effekts ist gegenwärtig nicht bekannt. Weitere Testmessungen mit unterschiedlichen JCAMP-DX Outputs sind dazu erforderlich.

NMR-Geräte von Varian werden mit der Software VnmrJ 3.0 betrieben. Entsprechende Beispielspektren und nähere Informationen zu Austauschformaten liegen bisher nicht vor.

Massenspektrometrie (MS)

Die in der Universität Paderborn gemessenen Massenspektren der Finnigan MAT Geräte werden als proprietäres MS2 Format gespeichert. Diese Dateien können mit der in der Fakultät verwendeten MASPECII 32 offline-Software ausgewertet, bearbeitet und als ANDI/netCDF Format exportiert werden. ANDI/netCDF (Network **C**ommon **D**ata **F**ormat) ist ein gängiges Austauschformat für Massenspektren. Umgekehrt können mit MASPECII Massenspektren anderer Gerätehersteller in Form von CDF-Dateien importiert werden. Die meisten MS-Geräte und MS-Softwareprodukte weiterer Anbieter unterstützen dieses Format zum Austausch von Dateien zwischen verschiedenen Datenverarbeitungssystemen. Mit Hilfe von MASPECII können File-Informationen eingesehen und zusammen mit dem Massenspektrum in einem PDF-File gespeichert werden.

File-Informationen aus MASPECII:

```
File Details.
File
Name:D:\dohmeier\Primärdatenprojekt\Metadaten\4_Ms\Weber\jab325c1.ms2
File Date/Time:29.10.2009 at 10:04:00
File Type:Lo-Res Mass/Int data
File Source:Imported from ICIS II format
file D:\DIENST DATEIEN\MASSENSPEKTROSKOPIE\DATA\NEU\jab325c1.dat
File Title:jab325c1
Operator:Zukowski
Instrument:MAT95
```

Eine Beispielmessung von Jeol wurde an einem GC/MS-Gerät des Typs JMS-T100GC aufgenommen. Die Ausgabe der Firma Jeol war eine JPF und eine JSP Datei. Beide Dateien sind mit einem Editor lesbar. Aus dem Kopfeintrag wird ersichtlich, dass es sich bei beiden Dateien um eine Art JCAMP-DX (JEOL-DX=2.00) Format handelt. Das Spektrum mit der Extension JPF kann dem Totalionenstroms zugeordnet werden (##DATA TYPE=Mass Profile Spectrum), bei dem zweiten Spektrum JSP handelt es sich um das eigentliche Massenspektrum (##DATA TYPE=Mass Spectrum).

Auszug aus der Jeol-Datei eines Totalionenstroms:

```
##JEOL-DX=2.00
##DATA TYPE=Mass Profile Spectrum
##DATA PROCESSING=unrounded
##SPECTROMETER=JMS-T100GC
##DATA SYSTEM=JEOL MassCenter
##MEASUREMENT DATE=2004/07/05,13:46
##SAMPLE NAME=
##NOTE=
##INLET=GC
##IONIZATION MODE=EI
##IONIZATION POLARITY=+
##ION SPECIES=Regular
##SCAN RANGE=35.000000,600.000000
##REPETITION RATE=0.400000 seconds
##RETENTION TIME=0.379750 min
##RETENSION TIME=0.379750 min
```

Auszug aus einem Jeol-Massenspektrum:

```
##JEOL-DX=2.00
##DATA TYPE=Mass Spectrum
##DATA PROCESSING=unrounded
##SPECTROMETER=JMS-T100GC
##DATA SYSTEM=JEOL MassCenter
##MEASUREMENT DATE=2004/07/05,13:46
##SAMPLE NAME=
##NOTE=
##INLET=GC
##IONIZATION MODE=EI
##IONIZATION POLARITY=+
##ION SPECIES=Regular
##SCAN RANGE=35.000000,600.000000
##REPETITION RATE=0.400000 seconds
##RETENTION TIME=0.379750 min
##RETENSION TIME=0.379750 min
##SCAN
NUMBER=MDT[CTR[50.0000..50.0000,80,Center,30,2.0,Area]]
##SCAN LAW=TOF
##DATA COMPONENT=mass        intensity
##DATA UNITS=m/z counts
##DATA MAXIMUM=592.953334    82292.216940
##DATA MINIMUM=100.175068    12.299416
##NPOINTS=162
##DATA=
100.175068    89.190710
…
```

Beide Dateien enthalten spezifische Metadaten, die detaillierte Parameter zu den Messbedingungen angeben.

Zusätzlich zu diesen standardisierten Formaten übergab Jeol eine CDF-Datei, die mit der uns zur Verfügung stehenden MS Software MASPECII in eine MS2-Datei konvertiert und bearbeitet werden kann.

Die Firma Shimadzu stellte von einer Messung Dateien der Formate CDF, PDF sowie eine Textdatei zur Verfügung. Wie oben erläutert, kann die CDF-Datei als gängiges Austauschformat mit MASPECII eingelesen, bearbeitet und in verschiedene Ausgabeformate exportiert werden.

Der Analysis Report des Esquire 3000 der Massenspektrometrischen Abteilung der Universität Bielefeld gibt verschiedene Metadaten aus. Unter Analysis Info sind Angaben zur Datei, Methode, Arbeitskreis, Operator und ein Kommentar vorhanden, zusätzlich ist eine Reihe von Aufnahmeparametern ausgedruckt.

Auszug aus den Metadaten eines Massenspektrums eines Esquire 3000:

```
Analysis Info
AC2_Kloesener_0503_JK2R4_03.d
Analysis Name
MethodTune-nan.MS
WorkgroupAC II

CommentKloesener, JK2R4, CH2Cl2

S.Heitkamp Operator
Acquisition Date10.5.2010 11:58:27
09:21:35 Print Date18.5.2010

Acquisition Parameter

Mode NanoESI, off-line Ion Source Type
Positive   Ion Polarity
Std/Normal  Mass Range Mode
50 m/z Scan Begin 1500 m/z Scan End
20 Spectra Averages
on   Rolling
…
```

Außerdem liegen Massenspektren im JCAMP-DX V 4.1 von 1999 vor. Sie zeigen ebenfalls einen Kopfeintrag, der aber nur unzureichend ausgefüllt wurde (##CAS NAME=Holger). Aus dem Spektrum geht auch nicht hervor, mit welchem Gerät oder welcher Software das Spektrum aufgenommen wurde.

Auszug eines Beispiels eines Massenspektrums im JCAMP-DX Format:

```
##TITLE=Library Entry 1 in C:\DATABASE\test01.l
##JCAMPDX=Revision 4.10
##DATA TYPE=MASS SPECTRUM
##SAMPLE DESCRIPTION=
##NAMES=
##CAS NAME=Holger
##MOLFORM=C20H40
##CAS REGISTRY NO=000000-00-0
##MP= -300
##BP= -300
##MW= 280.312
##$RETENTION INDEX=0
##$CONDENSED SPECTRUM=NO
##NPOINTS= 76
##XYDATA=(XY..XY)
   37       521
   39     76432
…
```

Infrarotspektroskopie (IR)

Bruker Optics ist ein führender Hersteller von IR-Geräten. Die vorliegenden Beispiele wurden an einem ALPHA FT-IR Spektrometer unter Verwendung der Software Opus (Version 6.5) aufgenommen, bearbeitet und gespeichert. Opus ermöglicht einen Datentransfer über verschiedene Austauschformate, wie z. B. Datenpunkttabelle.dpt, JCAMP-DX (Version 4.24), Pirouetta.dat, Galactic.spc.

JCAMP-DX 4.24. Dateien sind entsprechend dem JCAMP Protokoll aufgebaut. In den ersten Zeilen stehen auch hier Metadaten, die aufgrund der einfacheren Messtechnik gegenüber den NMR-JCAMP-DX-Dateien weniger umfangreich sind (18 Zeilen).

Auszug aus einem mit OPUS erzeugten JCAMP-DX 4.24 Format:

```
##TITLE=BDG
##JCAMP-DX=4.24
##DATA TYPE=INFRARED SPECTRUM
##DATE=5/5/2010
##SAMPLING PROCEDURE=ATR platinum Diamond 1 Refl
##ORIGIN=Administrator
##XUNITS=1/CM
##YUNITS=TRANSMITTANCE
##RESOLUTION=4
##FIRSTX=3997.7786
##LASTX=373.99488
##DELTAX=-1.4166473
##MAXY=1.0042931
##MINY=0.47493482
##XFACTOR=1
##YFACTOR=9.3532082e-010
##NPOINTS=2559
##FIRSTY=0.99623656
##XYDATA=(X++(Y..Y))
3998+1065128178+1065065790+1065100011+1065219179+1065348225
+1065429732….
```

Ultraviolettspektroskopie (UV)

Beispiele von UV-Spektren liegen von den Geräteherstellern der Firmen JASCO, Perkin Elmer und Shimadzu vor. Mit dem Gerät der Firma JASCO werden die Formate JWS, JCAMP-DX und TXT erzeugt. Vom Perkin Elmer-Gerät existieren SP und DAT Daten und von Shimadzu liegen TXT und PDF-Daten vor.

Auszug aus einem UV Spektrum im JCAMP-DX Format eines JASCO UV-Gerätes:

```
##TITLE= ascolactone-5_cmAcN
##JCAMP-DX= 4.24
##DATA TYPE=
##ORIGIN= JASCO
##OWNER= University of Debrecen
##DATE= 00/06/17
##TIME= 04:42:05
##SPECTROMETER/DATA SYSTEM=
##RESOLUTION=
##DELTAX= -0.50000000
##XUNITS= NANOMETERS
##YUNITS= ABSORBANCE
##XFACTOR= 1.00000
##YFACTOR= 1E-04
##FIRSTX= 350.0000
##LASTX= 190.0000
##NPOINTS= 321
##FIRSTY= -0.26508
##MAXY= 500.00000
##MINY= -0.27380
##XYDATA= (X++(Y..Y))
350.0000 -2651 -2537 -2572 -2738 -2656 -2154 -977 -238 1130 1840
2614 2970
9…
```

Auszug aus einem TXT UV Spektrum eines JASCO IR-Gerätes:

```
TITLE  ascolactone-5_cmAcN
DATA TYPE
ORIGIN        JASCO
OWNER
DATE  00/06/17
TIME   04:42:05
SPECTROMETER/DATA SYSTEM  JASCO Corp., J-810, Rev. 1.00
RESOLUTION
DELTAX        -0.5
XUNITS        NANOMETERS
YUNITS        ABSORBANCE
FIRSTX        350.0000
LASTX190.0000
NPOINTS       321
FIRSTY        -0.26508
MAXY 1436.15540
MINY  -0.27380
XYDATA
350.0000      -0.265083
349.5000      -0.25373
…
```

Beide Dateitypen enthalten umfangreiche Metadaten und sind ähnlich aufgebaut. Im Gegensatz dazu sind die Perkin-Elmer UV-Daten, die im ASCII-Format zu Verfügung stehen, nicht mit Metadaten versehen. Ebenso wenig informativ sind die verfügbaren UV-Daten von Shimadzu.

Hochdruckflüssigchromatographie/Gel-Permeations-Chromatographie (HPLC/GPC)

Die zur Verfügung gestellten Beispiel HPLC- Daten von Shimadzu liegen ebenfalls in einem CDF-Format vor. Darüber hinaus wurden Shimadzu HPLC-Daten als PDF- und Text-Datei gesammelt, die ebenfalls Metadaten enthalten.

Zu Varian liegen verschiedene Ausgabeformate vor (CDF, TXT, XLS, RTF). Es gibt eine zusätzliche Textdatei, an deren Ende sich etwa 60 Zeilen mit Metadaten befinden. Diese Datenausgabe ist offensichtlich optional.

Auszug aus einer Varian Textdatei mit Metadaten:

```
ACQMETHODNAME=
ACQMETHODVERSION=
ACQTIME=    17.04.2009 11:14:05
BATCHCOUNT=     -1
BATCHNAME=      0241
BATCHPOS= -1
CALDATE=    N.A.
CALNAME=
CALTYPE=    0
CHANNELNUMBER=          1
CHROMATONAME= 0241_10
DATAPATH=  LC Gruppe\Laptopdemo\Demolabor-2009-04-17\
…
NPEAKS=     21
NPOINTS=    3001
OPERATOR= Sebastian Krahe
PROCESSDATE=    01.10.2009 14:23:23
PROJECTNAME=    Demolabor
RACKNUMBER=     1
REPORTNAME=     SKR style
RMSNOISE=  0
RUNINFO=    N.A.
RUNNAME=   0241_10
RUNTIME=    5.00
SAMPLENAME=     N.A.
```

In den Shimadzu-Textdateien befinden sich die Metadaten im Kopfeintrag. Sie sind gegliedert in Header, File Information und Sample Information.

Auszug aus einer Shimadzu Textdatei:

```
[Header]
Application Name      LCsolution
Version         1.24
Data                                File                                Name
        C:\LabSolutions\LCsolution\Sample\PDA_Demo_Data-003.lcd
Output Date   05.10.2009
Output Time   10:53:26

[File Information]
Type   Data File
Generated      16.01.2003 14:34:58
Generated by SHIMADZU
Modified       25.12.2006 13:00:34
Modified by    Admin

[Sample Information]
Operator Name        SHIMADZU
Acquired       16.01.2003 14:35:27
Sample Type  1:Standard
Level  3
Sample Name STD
Sample ID      3
ISTD Amount 1        1
ISTD Amount 2        1
```

GPC Daten werden in Paderborn als Textfile gespeichert. Sie enthalten ebenfalls einen Kopfeintrag mit Metadaten.

Auszug aus GPC Textdaten:

```
Sample :        mom160
Inject date :   Donnerstag   12/11/09   11:37:45
Inject volume :          100.000        ul
Concentration :           6.000 g/l
Project :       D:\GPC User\Momen\momen
Calibration :   PMMA_THF_2009-05-06.CAL
Method :        c:\pss_wingpc7\THF-Anlage 1.MET
Export file :   D:\GPC User\GPC ASCII Report.TXT

Operator :      Momen
Account :

Calibration MH-K :      1.000000      ml/g    univ. Calibration MH-K :
0.000000      ml/g
Calibration MH-A :      0.000000      -       univ. Calibration MH-A :
0.000000      -
Internal Standard Calibration :         23.586         ml
Internal Standard Acquisition :         23.606         ml
Data Interval :  1.000 s

Eluent :        THF
Flow :   1.000 ml/min

Column 1 :      PSS-SDV guard 5m  Temperature :  0.000 C
Column 2 :      PSS-SDV 10e5A, 5 m        Temperature :  0.000 C
Column 3 :      PSS-SDV 10e3A, 5 m        Temperature :  0.000 C

Detector 1 :    Knauer RI (THF)       Delay :  0.000 min

Baseline from :         18.100        to :      22.567        ml
Integration from :      18.085        to :      22.548        ml

          Knauer RI (THF)
Mn:      4.203E+3      g/mol
Mw:      6.091E+3      g/mol
```

Röntgenstrukturanalyse

Ebenso wie bei den anderen analytischen Methoden sind bei der Datensammlung Datenformat und Umfang der Dateien geräteabhängig. In der Zentralen Analytik der Universität Paderborn werden Röntgenstrukturanalysen mit einem Bruker SMART APEX CCDC durchgeführt.

Bei einer Messung mit diesem Gerät erhält man ***Frames*** (1. Stufe der Primärdaten); die Anzahl der zu messenden Frames (z. B. 1800) wird vom Kristallographen vorgegeben. Sie beinhalten Intensitäten im Hexadezimal-Code. Die Gesamtheit dieser Frames wird als Messung oder Datensatz bezeichnet. Die Bezeichnung der Frames erfolgt manuell und wird so gewählt, dass sie Rückschlüsse auf den Auftraggeber der Messung erlaubt. Die anschließende notwendige, weitgehend automatisierte Datenreduktion der Frames führt zum ***RAW-File*** (2. Stufe der Primärdaten). Die RAW-Datei enthält die Messdaten (hkl-Indices, F_o^2, $\sigma(F_o^2)$, Richtungscosini) als ASCII-File. Eine weitere Bearbeitung der Daten durch den Kristallographen schließt die Absorptions-Korrektur ein und führt zum ***HKL-File*** (3. Stufe der Primärdaten – enthält wieder Miller Indices und F_o^2-Daten (o = observed)). Die HKL-Datei ist der Ausgangspunkt der Strukturanalyse. Sie wird intern abgelegt. Eine Software zur Strukturlösung (z. B. ShelXS) greift auf die HKL-Daten und ein ***INS-File*** (Instruction-File) mit Zellparametern, Raumgruppe, wahrscheinlich vorliegenden Atomsorten etc. zu und findet (meistens) ein mehr oder weniger vollständiges ***Modell*** der (Molekül-)Struktur. In weiteren Schritten wird dieses (rudimentäre) Modell verfeinert, d.h. komplettiert und die zunächst nur grob bestimmten Atompositionen sowie die anisotropen Auslenkungsparameter (adp) werden exakt bestimmt. Bei diesem Prozess werden auch F_c^2 –Daten (c = calculated) berechnet, die zur Berechnung des sog. R-Wertes, einem Gütekriterium für die Richtigkeit des Strukturmodells, mit den gemessenen F_o^2 Daten verglichen werden. Das Ergebnis dieser Verfeinerung ist das ***RES-File*** (Instruction-File), das die Zelle, Atomkoordinaten und weitere Strukturparameter enthält. Außerdem wird im letzten Verfeinerungszyklus ein ***CIF-File*** (Crystallographic Information File) erzeugt.[33] Das CIF-File enthält auch Zellparameter, Raumgruppe, Atomparameter, adp's, sowie berechnete Parameter (Bindungslängen, -winkel, Torsionswinkel, Wasserstoffbrücken etc.), Mess-Informationen und R-Werte (aus der Verfeinerung). Es wurde 1991 als standardisiertes Datenformat für die Archivierung und Verbreitung von kristallographischen Informationen eingeführt.[34] Es ist seit vielen Jahren etabliert und wird zur Veröffentlichung von Kristallstrukturanalysen verwendet.

Beim *Cambridge Crystallographic Data Centre* (*CCDC*) oder der *Inorganic Crystal Structure Database (ICSD)* müssen CIF-File und F_oF_c-File im Rahmen einer Publikation eingereicht werden.

Auszug aus einem CIF-File eines Supplemental Materials des *Journals of the American Chemical Society:*

```
data_09082

_audit_creation_method            SHELXL-97
_chemical_name_systematic ;  ? ;
_chemical_name_common             ?
_chemical_melting_point           ?
_chemical_formula_moiety          ?
_chemical_formula_sum   'C25 H22 Br N'
_chemical_formula_weight          416.35
_chemical_absolute_configuration  ad
loop_
 _atom_type_symbol
 _atom_type_description
 _atom_type_scat_dispersion_real
 _atom_type_scat_dispersion_imag
 _atom_type_scat_source
 'C'  'C'   0.0033   0.0016
 'International Tables Vol C Tables 4.2.6.8 and 6.1.1.4'
 'H'  'H'   0.0000   0.0000
 'International Tables Vol C Tables 4.2.6.8 and 6.1.1.4'
 'N'  'N'   0.0061   0.0033
 'International Tables Vol C Tables 4.2.6.8 and 6.1.1.4'
 'Br'  'Br'  -0.2901   2.4595
 'International Tables Vol C Tables 4.2.6.8 and 6.1.1.4'

_symmetry_cell_setting            Orthorhombic
_symmetry_space_group_name_H-M    P2(1)2(1)2(1)
  loop_
 _symmetry_equiv_pos_as_xyz
 'x, y, z'
 '-x+1/2, -y, z+1/2'
 '-x, y+1/2, -z+1/2'
 'x+1/2, -y+1/2, -z'
  _cell_length_a                  9.6774(7)
```

2.4.5 Analyse

Wie anhand der Beispiele deutlich wird, ist die Metadatenausgabe sehr unterschiedlich in Form, Art und Weise, Ausführlichkeit und Menge abhängig von der Methode und vom Gerätehersteller.

Ausführlich und übersichtlich ist die Darstellung der Metadaten in den JCAMP-DX Files, die bisher für NMR, IR- und UV-Spektren vorliegen. In den JCAMP-DX Protokollen werden ***L**abelled **D**ata **R**ecords (LDRs)* definiert. LDRs entsprechen Metadaten. Im *Core* müssen maschinenlesbare Metadaten angegeben werden, *Notes* sind dagegen optional und daher in Art und Umfang Benutzer- und methodenabhängig. Da *Notes* vornehmlich aus Text bestehen, sind sie menschenlesbar. In der *Core Fixed Header Information* sind Metadaten wie TITLE (Text), JCAMP-DX (String), DATA TYPE (String), ORIGIN (Text) und OWNER (Text) angegeben. Die *Core Variable Header Information* enthält methodenspezifische Metadaten, wie z. B. XUNITS (String) und YUNITS (String). In den *Notes* können Informationen zu *Global Notes*, z B. LONG DATE (Text), CAS NAME (String) oder SPECTROMETER/DATASYSTEM (Text) und *Data-Type Specific Notes*, z. B. IONIZATION ENERHGY (AFFN) oder TOTAL ION CURRENT (AFFN, String) enthalten sein.

Bei den MASPECII MS2-Dateien gibt es eine Übersicht, die sich *File Details* nennt und im PDF der Messung ausgedruckt werden kann. Sie gibt Parameter wie z. B. FILE NAME, FILE TITLE, OPERATOR, INSTRUMENT, NOTES (EI, 70 eV…) und NUMBER OF SCANS an.

Unsere Beispieldaten (GC/MS, HPLC) der Shimadzu Text-Files sind ebenfalls strukturiert. Hier wird differenziert zwischen *Header*, der die Metadaten APPLICATION NAME, DATA FILE NAME umfasst, *File Information* mit den Metadaten TYPE, GENERATED und *Sample Information* mit OPERATOR NAME, SAMPLE TYPE als Metadaten.

Die in der Röntgenstrukturanalyse verwendeten CIF-Files sind ASCII. Sie sind sehr umfangreich und beschreiben messtechnische und auswertungstechnische Metadaten sehr ausführlich. Im Kopfeintrag stehen chemische Metadaten, wie z. B _chemical_name_systematic ; ? oder _chemical_name_common ?. Sind Daten nicht zutreffend oder unbekannt, benötigen sie einen Platzhalter "." oder "?". Im vorliegenden Beispiel ist der Name der Verbindung also nicht eingegeben worden. Die Angabe chemischer Metadaten, die für eine Archivierung relevant sind, ist anscheinend fakultativ.

2.4.6 Erkenntnisse

Es gibt keine Konsistenz der abgelegten Metadaten, weder von einer Messmethode zur nächsten noch zwischen scheinbar analogen Metadaten-Outputs einer gegebenen Messung von unterschiedlichen Herstellern. Als Beispiel sind hier eine NMR-JCAMP-DX-Datei von Bruker und eine von Jeol zu nennen. Die gegenseitige Lesbarkeit ist nicht ohne Einschränkung gegeben. So trägt beispielsweise die Delta Datei (Jeol, NMR) auch abweichend die Extension JDX, was möglicherweise ein Hinweis auf eine andere Variante des Formats ist. TopSpin™-Files von Bruker (FID und JCAMP-DX File) können von der Delta -Software gelesen und korrekt in ein Spektrum umgesetzt werden. Umgekehrt kann TopSpin™ nur die JCAMP-DX, nicht jedoch die von Delta™ (Jeol) erzeugten JDF-Files einlesen.

Es gibt unterschiedliche JCAMP-DX Formate. Man nutzt parallel verschiedene Versionen für verschiedene Methoden. So findet in der IR- und UV-Spektroskopie nach wie vor die Version 4.24 Anwendung. In der ein- und zweidimensionalen NMR-Spektroskopie wird Version 6.0 verwendet. Bei dieser JCAMP-DX Version hängt die Größe der Outputs von der Art der Messung und deren Darstellung ab. Maßgeblich ist dabei die Art der Komprimierung (z. B. diff-dup-, packed- oder squeeze-Files) und der Ausgabeoutput (z. B. fid, fid + rspec + ispec, fid+all procnos).

Von Jeol ist bekannt, dass bei GC/MS ein eigenes, scheinbar JCAMP-DX analoges Format JEOL-DX 2.0 als Austauschformat mit den Extensionen JSP (Spektrum) und JPF (TIC) erzeugt werden kann.

Erwartungsgemäß stimmen Metadaten der *Core Fixed Header Information* bei verschiedenen Methoden und Instrumenten überein. Die in der *Core Variable Header Information* enthaltenen Metadaten sind methodenspezifische Angaben und variieren entsprechend. JCAMP-DX Files verschiedener Protokolle und Messmethoden stimmen also zu einem gewissen Teil in ihren Metadaten überein, enthalten darüber hinaus auch spezifische Parameter.

In Form von Beispieldatensätzen stehen uns aktuell JCAMP-DX Files im Zusammenhang mit NMR und IR-und UV-Messungen und ältere Massenspektren zur Verfügung. Informationen zu kommerziellen Softwarelösungen anderer Messmethoden geben aber Hinweise darauf, dass darüber hinaus bei weiteren Methoden ein Datentransfer über JCAMP-DX machbar ist. Gespräche mit Wissenschaftlern machen jedoch deutlich, dass JCAMP-DX in der Praxis nicht verwendet wird, was auch durch die Umfrage belegt werden kann.[35] JCAMP-DX war 83 % der Befragten kein Begriff. Allgemein lässt sich festhalten, dass die meisten Hersteller von Analytik-Applikationen gewisse Austauschformate anbieten. Die meisten Nutzer verwenden aber allenfalls nicht standardisierte Textformate. Messungen werden als

proprietäre Daten, ASCII-Daten ohne Metadaten, PDF oder Bilddateien gespeichert. Dieser Eindruck wird ebenfalls durch das Ergebnis der Umfrage bestätigt. Bei der Frage nach dem Speicherformat überwiegen proprietäre Formate deutlich Nichtproprietäre.

Keines der Beispiele enthält ausreichende Angaben (IUPAC-Name, InChI, CAS-Nr. oder Ähnliches), um die Messung durch eine Datenbanksuche auffinden zu können, obwohl in einigen Fällen (z. B. bei JCAMP-DX Files) unter ##TITLE zumindest ein frei wählbarer Name als Probenbezeichnung enthalten ist. In den Notes könnten solche Angaben allerdings ergänzt werden.

Hersteller und Erfinder der Probe sind in der Regel unklar, da bei den Messungen meist nur ein OPERATOR eingegeben wird, der die Messung durchführt.

Das Problem der Rückverfolgung vom Dateinamen zum Hersteller der Probe und zur standardisierten Probenbezeichnung ist ebenfalls aus der Umfrage bekannt. Ein interessanter Ansatz, der diese Problematik angeht, sind die maschinenlesbaren MS- und NMR-Anträge der MS- und NMR-Abteilungen der Universität Bielefeld. Hier wird der Dateiname vom System vorgegeben. Die systematische Probenbezeichnung ermöglicht langfristig Rückschlüsse auf den Auftraggeber der Messung. Dem Wissenschaftler wird der langfristige Beleg seiner Untersuchungsergebnisse erleichtert.

Ein maschinenlesbares Auftragsformular könnte auch als erster Arbeitsschritt zur Anreicherung der Messdaten mit Metadaten, die die Software nicht umfasst, genutzt werden. Eine Verknüpfung solch einer Automatisierung mit der Datenausgabe in einem standardisierten Datenformat könnte den Upload in einem Datenzentrum vereinfachen.

Da die JCAMP-DX Daten einfach in einem Editor bearbeitet werden können, besteht theoretisch das Problem der Manipulierbarkeit. Es scheint möglich zu sein, das Spektrum zu verändern und mit Hilfe spezieller Software das JCAMP-File durch eine inverse Fourier-Transformation in ein FID-File umzuwandeln.

B 3. Empfehlung eines fachspezifischen Metadatenschemas

Die vorliegenden Erkenntnisse zeigen die Komplexität der fachspezifischen Metadaten auf. Neben unterschiedlichen Formaten und Bezeichnern werden generell je nach Teilgebiet der Chemie sowie verwendeter Technik und Methode unterschiedliche Metadaten erzeugt. Diese Komplexität lässt eine Abstufung bzw. weitere Differenzierung der internen Metadaten für chemische Forschungsdaten ratsam erscheinen. Die enorme Diversifizität von Forschungsdaten in der Chemie legt ferner zum jetzigen Zeitpunkt einen allgemein ausgerichteten Ansatz nahe. So sollen die internen Metadaten unterteilt werden in technische bzw. methodenspezifische und chemische Metadaten. Chemische Metadaten sollen einen Datensatz inhaltlich unter Anwendung von Metadatenfeldern wie Chemischer Name, Summenformel oder chemischer Strukturformeln beschreiben. Die technischen Metadaten sollen relevante Informationen aufnehmen, um einen Datensatz aus technischer und inhaltlicher Sicht nachvollziehen und nutzen zu können. Die technischen Metadaten weisen eine hohe Diversifizität auf, daher sollen hier in einem ersten Ansatz Metadatencontainer definiert werden, die erst später in ihrer Granularität weiter ausgearbeitet werden. Dies erfolgt exemplarisch für die in der Konzeptstudie bereits mehrfach genannten Forschungsdaten aus den Bereichen der spektrometrischen Methoden NMR, MS, IR, UV und HPLC/GPC).

In Anlehnung an das Shell-Concept von Davies et. al. könnte ein Forschungsdatensatz in der Chemie durch unterschiedliche Metadaten-Ebenen beschrieben werden. Diese Ebenen könnten unterschiedliche Metadaten-Container enthalten, die es ermöglichen der Diversifizität chemischer Forschungsdaten Rechnung zu tragen und flexibel darauf zu reagieren. Wie in Abb. 22 dargestellt ummanteln die Metadaten-Ebenen den Kern, der die Daten des Forschungsdatensatzes darstellt. Die äußerste Hülle stellen die externen Metadaten dar, die in Katalogsystemen ähnlich bibliographischer Metadaten Anwendung finden und zur Zitierung eines Forschungsdatensatzes dienen. Zwischen Hülle und Kern liegen die internen (fachspezifischen) Metadaten, die den Datensatz umso spezifischer beschreiben, je näher sie am Kern liegen.

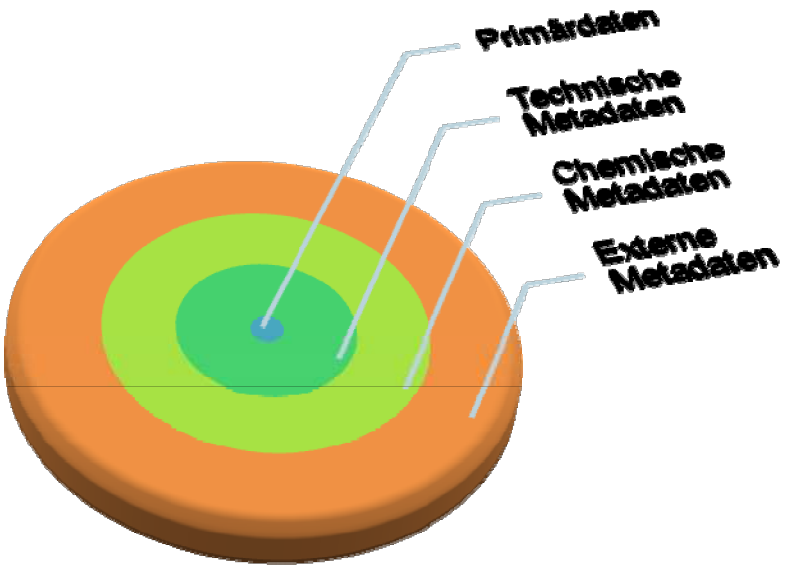


Abb. 22: Shell- Konzept der Metadaten-Ebenen

3.1 Chemische Metadaten

Liegt ein Forschungsdatensatz vor, der im weitesten Sinne die Daten über eine chemische Probe beinhaltet, so kann diese in den chemischen Metadaten beschrieben werden. Beispielhaft sei hier ein Forschungsdatensatz zu einem NMR-Spektrum einer chemischen Substanz genannt. In den chemischen Metadaten läßt sich dieser Datensatz durch unterschiedliche Namen (IUPAC-Name, Trivialname, INN), Summenformel, chemische Strukturformel, Elementarzusammensetzung etc. beschreiben. Diese Metadaten stellen wichtige Elemente für eine fachspezifische Suche dar. Aufgrund der Heterogenität chemischer Forschungsdaten ist davon auszugehen, dass nicht in allen Fällen chemische Metadaten im gleichen Maße definiert werden können. Basiert der Forschungsdatensatz nicht direkt auf einer chemischen Substanz oder auf einem Substanzgemisch, so können diese Metadaten unter Umständen nicht angegeben werden.

3.2 Technische Metadaten

Die Ebene der technischen Metadaten liegt noch näher am Kern des Shell-Konzeptes und dient zu einer spezifischen Beschreibung eines Forschungsdatensatzes, um diesen im Detail verstehen und nutzen zu können. Sie umfassen technische Infomationen, wie beispielsweise den Dateinamen, das Dateiformat, die Dateigröße, Hashwerte zur Validierung der

archivierten Daten und Angaben zur Software. Diese können in der Regel automatisch ermittelt werden.

Zusätzlich benötigte Metadaten zur Langzeitarchivierung werden ausführlich in Kapitel C 6.1.6 beschrieben.

C - Technische Aspekte, Datenspeicher

(Arbeitspaket 2)

C 1. Konzeptstudie „Langzeitarchivierung – Anforderungen an Hardware, Software und Datenorganisation“

Der Schaffung eines nachhaltigen Zugangs zu wissenschaftlichen Forschungsdaten wird national und international ein großer Stellenwert eingeräumt. Die wichtigsten Ziele dabei sind eine langfristige Speicherung der Daten sowie deren dauerhafte Verfügbarkeit. Die Daten sollen nicht mehr ausschließlich Teil einer wissenschaftlichen Publikation sein, sondern eine eigene Identität besitzen und durch DOI-Vergabe zitierfähig sein.

In diesem Kapitel „Technische Aspekte, Datenspeicher“ wird eine vernetzte Infrastruktur erarbeitet, mit der diese Ziele erreicht werden können. Die Infrastruktur kann für alle wissenschaftlichen Daten verwendet werden, so dass das vorgeschlagene Konzept auch Modellcharakter für andere Fachgebiete besitzt. Die Einhaltung der speziellen Anforderungen der einzelnen Disziplinen wird durch eine flexible softwaretechnische Verarbeitung gewährleistet.

Die Aufgabe des FIZ CHEMIE in dieser Konzeptstudie besteht in der Konzipierung eines von Universitäten und anderen Instituten nutzbaren Archivs für chemische Forschungsdaten. Das Archiv dient dabei nicht nur der Bewahrung der digitalen Daten, sondern stellt vielmehr ein Zusammenspiel einer technischen und organisatorischen Infrastruktur dar, die vor allem die langfristige Verfügbarkeit der Daten zum Ziel hat. Um dies zu erreichen, muss an zwei Stellen angesetzt werden. Zum einen muss der physische Erhalt der gespeicherten Datenobjekte auf einem entsprechenden Speichermedium gesichert werden (Bitstream Preservation). Zum anderen müssen Techniken und Verfahren entwickelt werden, die eine dauerhafte Interpretierbarkeit der Archivbestände gewährleisten. Hierzu gehören beispielsweise das Erstellen und Verwalten von Metadaten als Informationsgrundlage für spätere Maßnahmen zum Erhalt der Datenlesbarkeit.

Die Archivierung umfasst folgende Aufgaben:

- **Übernahme und Validierung:** Die Daten müssen aus unterschiedlichen Quellen und in verschiedenen Formaten übernommen, geprüft und in das Archiv integriert werden.
- **Dauerhafte Aufbewahrung und Verfügbarkeit:** Die Archivdaten müssen jederzeit zugänglich (lesbar) sein. Dabei müssen technologische Generationenwechsel durch die permanente Pflege von Migrationsketten überwunden werden (Datenträger, Datenformate, Software).
- **Verständlichkeit und Authentizität:** Die Archivdaten müssen in ihrem ursprünglichen Entstehungs- und Nutzungszusammenhang dauerhaft verstehbar und authentisch gehalten werden.

- **Nachnutzbarkeit:** Die Archivdaten müssen jederzeit für die ursprünglichen Datenproduzenten, Datennutzer sowie für Forschung und Öffentlichkeit in einer bedarfsgerechten Form zur Benutzung bereitgestellt werden.

Das zu konzipierende Archiv für chemische Forschungsdaten soll keine individuelle Insellösung sein, sondern in eine nationale, vernetzte Forschungsdateninfrastruktur eingebettet werden können.

Um die Bereitschaft der Wissenschaftler für die Langzeitarchivierung ihrer Forschungsdaten zu erhöhen, müssen geeignete technische Lösungen zur Verfügung gestellt werden. Der Prozess der Datensicherung sollte sich möglichst nahtlos in die wissenschaftlichen Arbeitsabläufe einfügen, damit sich der Wissenschaftler auch weiterhin seinem eigentlichen Tätigkeitsschwerpunkt, der Forschung, widmen kann.

1.1 Problemstellung

Wissenschaftliche Daten sind geprägt durch ihre Herkunft aus experimentellem Vorgehen, denn sie stammen aus Arbeitsabläufen, die immer wieder an die untersuchte Fragestellung angepasst werden. Häufig durchlaufen diese Daten von ihrer Erzeugung bis zur Publikation der darauf basierenden Forschungsergebnisse eine Reihe von Bearbeitungsschritten. Es lässt sich kaum universell festlegen, in welchem Stadium die Daten in ihrem Lebenszyklus erhaltenswert sind. In der Regel existieren keine Formatvorgaben, so dass Forschungsdaten in einer Vielzahl von Dateiformaten erzeugt werden, die semantisch selten einheitlich strukturiert und nur lückenhaft mit Metadaten beschrieben sind. Faktoren wie diese stellen für die digitale Langzeitarchivierung von Forschungsdaten oft eine größere Herausforderung dar als die Datenmenge selbst.

In Deutschland werden derzeit Forschungsdaten in den meisten Disziplinen nicht systematisch archiviert. Es fehlen allgemein anerkannte Standards hinsichtlich einer Nachnutzbarkeit und langfristigen Verfügbarkeit der Daten. Es existieren in der Regel keine Qualitätssicherung, kein gesicherter Nachweis und keine Datensicherheit. Zwar gibt es die Empfehlungen der DFG zur guten wissenschaftlichen Praxis[36], die unter anderem fordern, dass Forschungsdaten, die Basis einer Publikation sind, mindestens zehn Jahre auf gesicherten Datenträgern aufbewahrt werden sollen. Die experimentellen Daten liegen meist auf einem Messrechner bzw. auf einem arbeitsgruppeninternen Server und werden in der Regel nach Projektende bzw. nach Ausscheiden des Wissenschaftlers aus der Arbeitsgruppe für eine längerfristige Speicherung auf ein Speichermedium abgelegt. Häufig wissen nur der Forscher selbst sowie der Arbeitsgruppenleiter um deren Existenz bzw. Bedeutung. Eine allgemeine Zugänglichkeit der Daten zwecks Nachvollziehbarkeit von

Publikationsergebnissen oder eine Zitierbarkeit der Daten ist selbstverständlich nicht gegeben.

Bei dieser herkömmlichen Umgangsweise mit Forschungsdaten gibt es immer wieder viele technische und organisatorische Schwierigkeiten, die letztendlich dazu führen, dass wichtige Daten verlorengehen.

- Die Haltbarkeit von Datenträgern ist aufgrund von Alterung oder Verschleiß begrenzt. Ist der Datenträger nicht mehr auslesbar, sind die Daten verloren.
- Zum Auslesen der Daten wird in der Regel ein entsprechendes Lesegerät benötigt. Durch den technologischen Wandel und die damit verbundene Weiterentwicklung der Datenträgertechnologie kann es passieren, dass die notwendigen Lesegeräte nicht mehr verfügbar sind.
- Auch können Datenträger verlorengehen, versehentlich gelöscht oder gar bewusst entsorgt werden. So hat die NASA durch die Speicherung auf nicht oder nur wenig beschrifteten Magnetbändern und durch unsachgemäße Lagerung Daten aus drei Jahrzehnten verloren[37].
- Software-Umgebungen veralten und werden möglicherweise nicht weiterentwickelt. So ist bei proprietären Formaten die Datenlesbarkeit gefährdet. Verschärft wird dieses Problem durch die Tatsache, dass oft neue Programmversionen mit veränderten Datenformaten veröffentlicht werden, die ältere Datenformate gar nicht oder nur unvollständig nutzen können.

Auch das Fach Chemie ist signifikant von der Problematik des Datenverlustes betroffen. So hat die Umfrage von Fels und Dohmeier-Fischer [38] im Rahmen der Konzeptstudie nachgewiesen, dass bei 52 % aller befragten Wissenschaftler bereits Probleme bei der Wiederverwendung von gespeicherten Daten aufgetreten sind. Bei einem Anteil von 18 % der Befragten waren die Schwierigkeiten auf veränderte Dateiformate zurückzuführen, 20 % haben einen Datenverlust erlitten und 14 % gaben diverse Gründe an.

Gerade in der Chemie erweist sich das Datenmanagement aufgrund besonderer technischer Voraussetzungen als sehr komplex. Es existiert ein breit gefächertes Spektrum an Messmethoden und -verfahren, die unterschiedliche Verbreitung gefunden haben. So gibt es einerseits Messverfahren, wie z. B. die NMR-Spektroskopie zur Strukturanalyse, die in nahezu jedem Fachbereich Chemie zur Anwendung kommen. Andererseits kommen in einigen Teilgebieten der Chemie eher seltene bzw. sehr spezielle Verfahren zum Einsatz, so dass nur wenige Institute entsprechende Messgeräte besitzen, z. B. Partikelgrößenbestimmung mittels einer Scheibenzentrifuge (Biotechnologie, Polymerchemie). Im Extremfall gibt es auch individuelle, im Arbeitskreis einer Universität

konzipierte Messapparaturen mit eigens entwickelter Software für die Datenauswertung, z. B. speziell angepasste Kalorimeter in der Polymeren Reaktionstechnik.

Für ein Messgerät gibt es in der Regel unterschiedliche Hersteller, die ihre eigenen Datenformate entwickelt haben. Die im Einsatz befindlichen Geräte können unterschiedlichen Gerätegenerationen angehören. Entsprechend groß ist die Formatvielfalt der zu archivierenden Forschungsdaten. Oft sind die Formate proprietär und ihr langfristiger Erhalt nur in Zusammenarbeit mit den Geräteherstellern möglich. Ist ein Gerätehersteller später einmal nicht mehr existent, können seine Formate möglicherweise nur mittels Emulation erhalten werden. Emulation ist ein teils aufwendiges Verfahren, bei dem die Funktion eines älteren, nicht mehr verfügbaren Computersystems durch ein leistungsstärkeres, aktuelles System nachgebildet wird.

Diese nichtbearbeiteten Formate aber – als Beispiel sei hier das NMR-FID aus dem Bereich der NMR-Spektroskopie genannt – können oft nicht ohne spezielle Software visualisiert werden. Aus ihnen resultierende nachbearbeitete Formate, wie z. B. JCAMP-DX, enthalten nur reduzierte Informationen. JCAMP-DX ist ein in der Chemie etabliertes, standardisiertes Austauschformat, welches eine Visualisierung von Spektren ermöglicht.

Die beschreibenden Metadaten, die in der Chemie benötigt werden, enthalten Strukturinformationen. Struktur-Files aber sind ohne spezielle Retrievalsysteme nicht suchbar, so dass ebenfalls chemische Identifier wie der InChI[39] zur Charakterisierung von Strukturen benötigt werden.

1.2 Existierende Projekte zur Langzeitarchivierung

Die allgemeine Problematik der Zitierbarkeit und Langzeitarchivierung von wissenschaftlichen Daten ist bereits ausführlich durch nationale und internationale Forschung beleuchtet worden. Ein großes nationales Projekt ist nestor[40] (Network of Expertise in Long-term Storage of Digital Resources), das deutsche Kompetenznetzwerk zur Langzeitarchivierung und -verfügbarkeit. nestor vereinigt Kooperationspartner aus den Bereichen Bibliotheken, Museen und Archive und wird seit dem Jahr 2009 nach Auslauf der Projektförderung von den Kooperationspartnern getragen. nestor kooperiert mit europäischen Partnern, bündelt Standardisierungsaktivitäten und dient als Anlaufstelle für Fragen zur digitalen Langzeitarchivierung. Das aus dem Projekt hervorgegangene nestor-Handbuch[41] beschäftigt sich mit nahezu allen Aspekten der digitalen Langzeitarchivierung.

Das Projekt PLANETS[42] (Preservation and Long-term Access through Networked Services) ist ein von der EU von 2004 bis 2010 gefördertes Projekt mit dem Ziel, Services und Tools zu entwickeln, die Gedächtnisinstitutionen bei der Sicherstellung der langfristigen

Zugänglichkeit von digitalen kulturellen und wissenschaftlichen Ressourcen unterstützen. Im Projekt haben Partner aus Nationalbibliotheken, Archiven und universitären Forschergruppen mit Firmen wie IBM und Microsoft zusammengearbeitet. Aus dem Projekt ist 2010 die Open Planets Foundation [43] hervorgegangen, die die Entwicklung von Technologien für die Langzeitarchivierung weiter vorantreiben will.

Tatsächlich ist die Langzeitarchivierung noch ein recht junges Themenfeld und wird deshalb erst von relativ wenigen Instituten praktiziert.

Die grundlegende Voraussetzung für die Langzeitarchivierung von Forschungsdaten ist, dass es vertrauenswürdige Archive gibt, die diese Aufgabe übernehmen können. Insbesondere in den Bio-, Geo-, Klima- und Sozialwissenschaften haben sich bereits seit längerer Zeit Institutionen etabliert, die für das Forschungsdatenmanagement zuständig sind. In einigen Disziplinen wird diese Aufgabe von so genannten Datenzentren übernommen, die zumeist mit der Motivation entstanden, wissenschaftliche Daten, die regelmäßig in einem institutionellen oder projektgebundenen Rahmen erhoben werden, langzeitlich zu sichern. Auch die Weltdatenzentren des International Council for Science[44] (ICSU WDCs) haben sich dieser Aufgabe verpflichtet.

PANGAEA / WDC-MARE

Um Daten aus der Klima- und Umweltforschung allgemein verfügbar und einer weiteren wissenschaftlichen Auswertung zugänglich zu machen, hat das Alfred-Wegener-Institut für Polar- und Meeresforschung [45] (AWI) gemeinsam mit dem Zentrum für Marine Umweltwissenschaften [46] (MARUM) der Universität Bremen ein Informationssystem zur Archivierung, Publikation und Verarbeitung von Daten aufgebaut. In dem als Langzeitarchiv betriebenen System PANGAEA[47] werden Milliarden von Messwerten aus dem Bereich der Erdsystemforschung gespeichert, die der Wissenschaft über Internet-Portale für weiterführende Analysen zur Verfügung gestellt werden.

PANGAEA hat ursprünglich mit den Daten des Alfred-Wegener-Instituts begonnen und sich mittlerweile zu einer zentralen Bibliothek für eine Vielzahl geowissenschaftlicher Disziplinen entwickelt. Das System wird von verschiedenen internationalen Forschungsprojekten und vom World Data Center for Marine Environmental Sciences[48] (WDC-MARE) als Datenarchiv genutzt.

Die technische Grundlage dieses Konzepts bildet die relationale Datenbank Sybase IQ, die sich durch hohe Komprimierung und automatische Indexierung auszeichnet. Das Datenmodell wurde so generisch und offen gehalten, dass es jederzeit um neue Parameter erweitert werden und sich damit neuen wissenschaftlichen Entwicklungen anpassen kann. Die Daten werden georeferenziert in Zeit und Raum abgelegt.

In langjähriger Arbeit ist es gelungen, mit PANGAEA eine zuverlässige Langzeitverfügbarkeit von vielfältigsten Daten zu gewährleisten. Das Datenarchiv dient gleichzeitig als Publikationssystem, welches die Datenpublikation in den etablierten Prozess der wissenschaftlichen Veröffentlichung integriert und somit ein Anreizsystem für den Datenproduzenten schafft.

Scientific Drilling Database

Auch das Geoforschungszentrum Potsdam[49] (GFZ) veröffentlicht Forschungsdaten im Sinne des Open Access. Die Scientific Drilling Database [50, 51] (SDDB) ist ein Archiv für Forschungsdaten aus dem International Continental Scientific Drilling Program (ICDP) und assoziierten Projekten. Das Archiv wird vom Geoforschungszentrum Potsdam und der Operational Support Group ICDP[52] (OSG) betrieben.

Die Daten in der Scientific Drilling Database werden unter Creative-Commons-Lizenzen[53] publiziert. Die Entwicklung des Archivs ist noch nicht abgeschlossen. Derzeit ist unter anderem eine Navigation über Autoren, Research-Programme oder DOI möglich.

Die Daten in PANGAEA und der Scientific Drilling Database werden durch DOI-Registrierung bei der TIB Hannover – seit 2010 über das internationale Konsortium DataCite[54] – zitierfähig (s. Kap. 6.1.5).

kopal

kopal[55] (Kooperativer Aufbau eines Langzeitarchivs digitaler Information) stellt ein von 2004 bis 2007 vom BMBF gefördertes Verbundprojekt unter Leitung der Deutschen Nationalbibliothek [56] dar, dessen Ziel der Aufbau eines kooperativ entwickelten und betriebenen Archivsystems zur Sicherung der Langzeitverfügbarkeit digitaler Dokumente war. Das Archivsystem gewährleistet nicht nur den physischen Erhalt der Daten, sondern stellt auch deren künftige Interpretierbarkeit sicher. Der technische Betrieb des Langzeitarchivs ist bei dem Rechenzentrum GWDG[57] (Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen) angesiedelt. Auf Server-Seite wurde das Archiv auf Basis des DIAS-Systems[58] von IBM[59] entwickelt, während auf Client-Seite die im kopal-Projekt entwickelte Open Source Software koLibRI [60] (kopal Library for Retrieval and Ingest) eingesetzt wird, die das Erstellen, Einspielen und Abfragen von Archivpaketen ermöglicht. kopal ist mandantenfähig und bietet Institutionen verschiedene Nutzungsmodelle an. Das Archivsystem wird unter Berücksichtigung internationaler Standards für Langzeitarchivierung und Metadaten sowie im Rahmen des Referenzmodells OAIS (Open Archival Information System, s. Kap. 6.1.1) implementiert.

Als Mandanten nutzen derzeit die SUB Göttingen[61] und die Deutsche Nationalbibliothek das Archivsystem.

Im nachfolgenden, von der DFG geförderten Projekt KoLaWiss [62] (Kooperative Langzeitarchivierung für Wissenschaftsstandorte) wurden unter der Projektleitung der GWDG die Anforderungen der wissenschaftlichen Community an digitale Langzeitarchivierungsmaßnahmen evaluiert. Unter anderem wurde auch eine Gap-Analyse für das kopal-Langzeitarchiv durchgeführt.

BABS

Das Bibliothekarische Archivierungs- und Bereitstellungssystem BABS[63] entstand 2005 bis 2007 im Rahmen eines von der DFG geförderten Projekts, an dem die Bayerische Staatsbibliothek[64] und das Leibniz-Rechenzentrum[65] beteiligt waren. In diesem Projekt wurde exemplarisch eine organisatorische und technische Infrastruktur für die Langzeitarchivierung von Netzpublikationen einer Universalbibliothek aufgebaut. BABS vereint eine heterogene organisatorische und technische Infrastruktur für die Langzeitarchivierung mit der Bereitstellung von elektronischen Publikationen unterschiedlicher Art. Mit Ende der Projektlaufzeit hat BABS den Projektstatus verlassen und ist in den Produktionsbetrieb übergegangen.

Die vom OAIS-Referenzmodell (s. Kap. 6.1.1) definierten Funktionalitäten Ingest, Data Management und Access werden einerseits von dem am Münchener Digitalisierungszentrum (MDZ) entwickelten Electronic Publishing System ZEND [66] (Zentrale Erfassungs- und Nachweisdatenbank) für Retrodigitalisate, andererseits von dem Digital Asset Managementsystem DigiTool [67] der Firma Ex Libris für elektronische Publikationen bereitgestellt. Die Aufgabe des Archival Storage übernimmt das robotergesteuerte Archiv- und Backupsystem mit dem Softwarepaket Tivoli Storage Manager[68] der Firma IBM am Leibniz-Rechenzentrum.

In dem Nachfolgeprojekt BABS2[69] wird im Zeitraum von 2008 bis 2010 die bestehende Infrastruktur zu einem vertrauenswürdigen und skalierbaren digitalen Langzeitarchiv ausgebaut.

Digitales Archiv

Das Digitale Archiv[70] ist die Archivierungslösung des Bundesarchivs, potenzielle Nachnutzer sind alle Bundesbehörden. Der Produktivbetrieb wurde 2008 aufgenommen. Das Digitale Archiv basiert auf einem Speichersystem von Hewlett-Packard und auf Software-Komponenten von SER. Mit dem Digitalen Archiv können digitale Objekte und Metadaten aus disparaten Systemen der Behörden kontrolliert, fehlerfrei und effizient archivtauglich aufbereitet sowie in das Bundesarchiv überführt werden. Das System soll die dauerhafte Sicherung und Authentizität dieser Daten gewährleisten. Der Gesamtprozess von der abgebenden Stelle bis in das Storage-System orientiert sich strikt am OAIS-Referenzmodell (s. Kap. 6.1.1) und verwendet offene Standards.

Das Digitale Archiv entlastet die Behörden von nicht mehr laufend benötigten Unterlagen, stellt eine zentrale Infrastruktur für eine langfristige Speicherung zur Verfügung und bietet zudem komfortable Zugriffsmöglichkeiten auf archivierte Unterlagen.

1.3 Zielsetzung

In vielen Fachgebieten existieren noch keine Institutionen bzw. Datenzentren, die die Aufgabe des Datenmanagements übernehmen. Entsprechend fehlen bislang in diesen Disziplinen Konzepte für eine Langzeitarchivierung von Forschungsdaten. Diese Situation trifft auch auf die Chemie zu, und deshalb könnte das FIZ CHEMIE die Institution sein, die in Deutschland die Aufgabe der Langzeitarchivierung chemischer Forschungsdaten übernimmt. Ein vom FIZ CHEMIE betriebenes Langzeitarchiv müsste sich nahtlos in die Forschungslandschaft einbetten und für alle wichtigen, erhaltenswerten Daten aus dem Bereich der Chemie offen sein.

Die Konzeptstudie leistet die für das Erreichen dieses Ziels notwendige Vorarbeit und entwickelt ein realisierbares, technisches Konzept für eine Forschungsdateninfrastruktur in Deutschland. Diese Infrastruktur soll den Datenproduzenten, -manager und -nutzer miteinander vernetzen. Die Rolle des Datenmanagements sowie die Verantwortung für die chemischen Forschungsdaten und deren Langzeitarchivierung übernimmt in diesem Konzept das FIZ CHEMIE.

Um dieser Aufgabe gerecht zu werden, muss das FIZ CHEMIE ein geeignetes Langzeitarchiv konzipieren, in welchem wichtige und erhaltenswürdige Forschungsdaten aus dem Bereich der Chemie abgelegt werden. Dazu müssen zum einen die hard- und softwaretechnischen Anforderungen geprüft und definiert werden, um in einem ersten Schritt eine Basis-Infrastruktur für ein Archiv zu entwickeln, dessen Hauptziel zunächst der langfristige Erhalt der Daten (Bitstream Preservation) ist. Die eigentlichen Herausforderungen aber liegen weniger beim Aufbau und Betrieb der Hard- und Softwarearchitektur, sondern vielmehr in der Gewährleistung der Langzeitverfügbarkeit und Nachnutzbarkeit von Daten in unterschiedlichsten Dateiformaten sowie in der Organisation des Langzeitarchivs. Alle Prozesse, die die Daten im Archiv durchlaufen, müssen definiert und beschrieben sein. Das Management der Forschungsdaten setzt eine organisierte Struktur voraus, so dass ein Langzeitarchiv als ein Zusammenspiel von Mensch und Technik beschrieben werden kann. Es sind Kriterien und Standards für die Langzeitarchivierung und Nachnutzbarkeit zu entwickeln bzw. vorhandene Lösungen auf die Archivierung chemischer Forschungsdaten zu übertragen.

Die Existenz von Langzeitarchiven sowie geeignete organisatorische Rahmenbedingungen in den Instituten und bei der Forschungsförderung sind notwendige Voraussetzungen für die digitale Langzeitarchivierung von wissenschaftlichen Forschungsdaten. Sie müssen aber auch durch technische Lösungen unterstützt werden, die die Mitwirkung der Wissenschaftler an der digitalen Langzeitarchivierung ihrer Forschungsdaten so einfach wie möglich gestalten. Ein Beispiel dafür ist die Beschreibung der Forschungsdaten durch Metadaten. Die Generierung und Pflege der Metadaten stellt für den Wissenschaftler oft eine enorme Hürde dar, weil die Metadatenschemata meist sehr komplex sind und eine rein manuelle Metadatenerzeugung sehr aufwändig ist. In der Praxis zeigt sich, dass die Bereitschaft zur Datenablage in ein Repository deutlich höher ist, wenn die Erstellung und Pflege von Metadaten so weit wie möglich automatisiert wird. Ein hoher Technisierungsgrad des Datenmanagements erlaubt es den Wissenschaftlern, sich auch weiterhin vorrangig ihrer Forschung zu widmen. Ein weiterer wichtiger Schwerpunkt dieser Studie ist deshalb die Analyse der Anforderungen, die der Chemiker an Tools und Werkzeuge stellt, die ihn in der Handhabung und Verwaltung seiner Daten und Metadaten unterstützen.

Viele Aspekte zur Langzeitarchivierung und Nachnutzbarkeit von Forschungsdaten sind bereits in nationalen und internationalen Projekten bearbeitet worden. Für viele der genannten Punkte wurden bereits Empfehlungen ausgesprochen oder prototypische Lösungen entwickelt. Es existieren eine Reihe international anerkannter Standards, Lösungsansätze und Werkzeuge, die in der Konzeptstudie aufgegriffen und für den Aufbau einer Infrastruktur für chemische Forschungsdaten zusammengeführt werden. Auf diese Weise ist es möglich, auch die besondere Problematik im Fach Chemie effizient zu bearbeiten und zu lösen. Im Fokus steht dabei stets die Anpassung des zu konzipierenden Systems an die spezifischen Bedürfnisse des Chemikers, über die der Umfragebericht zum Fragebogen von Fels und Dohmeier-Fischer[38] wichtige Informationen liefert.

C 2. Konzept eines vernetzten Langzeitarchivs

2.1 Erweiterter Publikationsprozess

Der bisherige Publikationsprozess soll durch die Installation einer Institution, die für die Datenpflege und -archivierung zuständig ist, erweitert werden. Dieser erweiterte Publikationsprozess lässt sich auf Basis eines Modells beschreiben, welches auf dem von Treloar [71] entwickelten Data Curation Continuum beruht: In verschiedenen Stufen im Lebenszyklus der Forschungsdaten werden unterschiedlichste Anforderungen an die Eigenschaften der Datenobjekte, ihre Beschreibung, Persistenz und an die Datenverfügbarkeit gestellt.

In dem erweiterten Publikationsprozess, der bereits in Kapitel A.4.2 beschrieben ist, durchlaufen die Forschungsdaten vier Domänen, die über Transferprozesse miteinander verknüpft sind.

Die Erweiterung des Publikationsprozesses durch eine Dauerhafte Domäne zielt auf den Aufbau eines Datenzentrums, welches die langfristige Speicherung und Pflege der Daten übernimmt. Dies wird gegenüber dem Wissenschaftler mittels Trusted Archive Policies garantiert. Weiterhin bietet das Datenzentrum Schnittstellen unterschiedlichster Art zum Upload und Download von Forschungsdaten an. Eingebettet in die Dauerhafte Domäne ist auch die DOI-Registrierung von Forschungsdaten. Die Daten erhalten hier einen DOI-Namen oder einen sonstigen persistenten Identifier, unter dem sie dauerhaft wiederzufinden und eindeutig identifizierbar sind.

2.2 Vernetzte Forschungsdaten-Infrastruktur in Deutschland

Auf Grundlage dieses theoretischen Ansatzes lässt sich ein Konzept für eine vernetzte Forschungsdaten-Infrastruktur auf nationaler Ebene entwickeln, welches in der Abb. 23 dargestellt ist.

Die äußere Schicht des Netzwerks spiegelt die *Private Domäne* wider. Diese setzt sich aus einer großen Zahl an *Wissenschaftlern W* zusammen, die in verschiedenen Forschungseinrichtungen im Rahmen von Projekten Daten unterschiedlichster Art erzeugen und analysieren. Der einzelne Forscher gehört jeweils einer Arbeitsgruppe an und stellt die Messdaten bzw. analysierten Daten seinen Kollegen zur Verfügung, indem er sie auf den Arbeitsgruppenserver legt. Alle *Arbeitsgruppenserver AS*, in der Abbildung durch die rote Schicht dargestellt, bilden die *Gruppendomäne*.

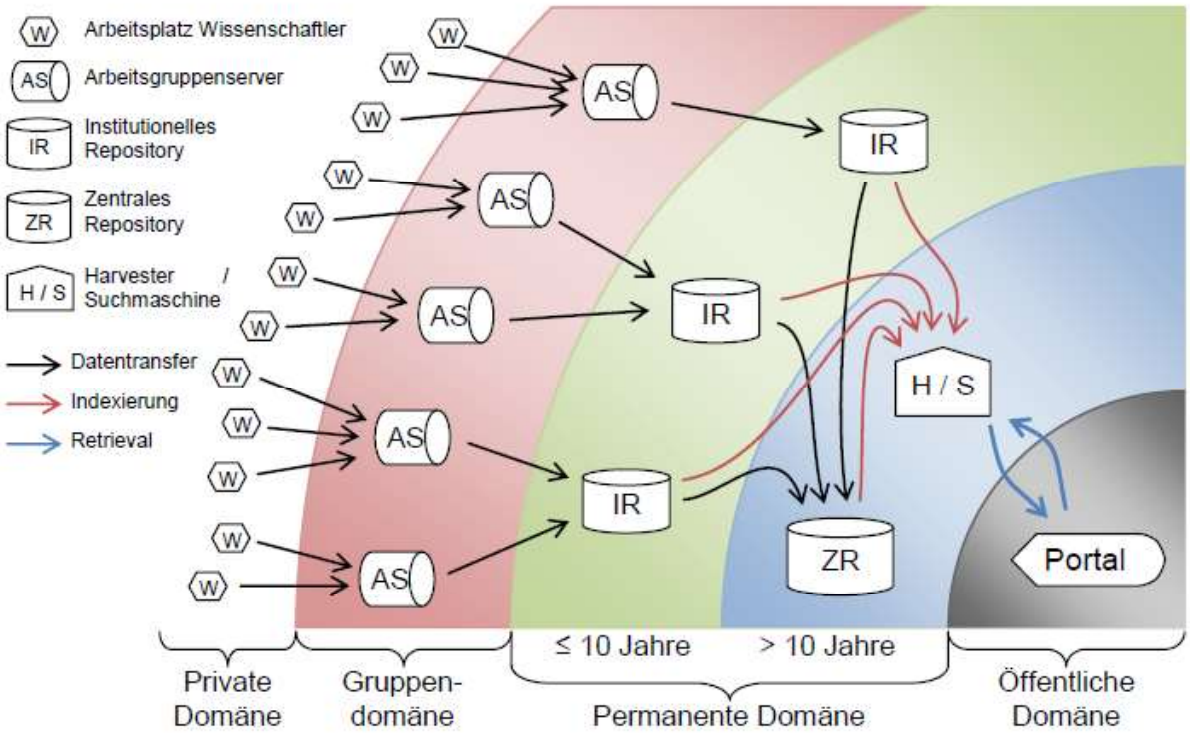


Abb. 23: Vernetzte, nationale Forschungsdaten-Infrastruktur

Vom Arbeitsgruppenserver fließen die Daten in ein *Institutionelles Repository IR*. Ein Institutionelles Repository wird idealerweise von jedem Forschungsinstitut bereitgestellt, um die Daten dezentral an ihrem Entstehungsort zu speichern (grüne Schicht). Das Institutionelle Repository bietet ein Zugriffsmanagement, so dass die Forschungsdaten je nach Bedarf entweder nur einem ausgewählten Personenkreis oder auch bereits der breiten Öffentlichkeit zur Verfügung gestellt werden können. Im Institutionellen Repository, welches für verschiedene Disziplinen offen ist, sollten die Daten zur Einhaltung der Empfehlungen der DFG zur guten wissenschaftlichen Praxis[36] mindestens zehn Jahre aufbewahrt werden. Datensätze, die langfristig, also länger als zehn Jahre aufbewahrt werden müssen, weil sie z. B. einmalig, schwer reproduzierbar oder ein wichtiger Bestandteil einer Publikation sind, werden in das Langzeitarchiv übertragen. Das Langzeitarchiv wird durch das *Zentrale Repository ZR* repräsentiert, welches im Gegensatz zu den Institutionellen Repositories eine zertifizierte Langzeitstrategie mit einer Archivadministration entsprechend dem OAIS-Referenzmodell (s. Kap. 6.1.1) sowie eine Gewährleistung der dauerhaften Nachnutzbarkeit aller Datenformate nachweisen kann und seine Vertrauenswürdigkeit über entsprechende Zertifikate belegt. Das Zentrale Repository steht im Mittelpunkt dieser vernetzten Infrastruktur, alle in ihm archivierten Datensätze müssen über einen DOI-Namen registriert sein. Die *Dauerhafte Domäne* setzt sich aus dem Zentralarchiv sowie den Institutionellen

Repositories zusammen. Alle Elemente der Dauerhaften Domäne werden durch einen so genannten *Harvester H*, der regelmäßig die Metadaten sämtlicher Repositories einsammelt, vernetzt. Diese Funktionalität des Harvesters erfolgt über eine standardisierte Kommunikationsschnittstelle. Hier hat sich das Protocol für Metadata Harvesting der Open Archives Initiative OAI-PMH [72] durchgesetzt. Eine Vernetzung aller Repositories ist nur möglich, wenn die Betreiber ausnahmslos diesen gemeinsamen Standard verwenden. Gleichzeitig wird damit auch der Weg für eine zukünftige interoperable Einbindung der aufzubauenden nationalen Forschungsdaten-Infrastruktur in internationale sowie interdisziplinäre Netzwerke geebnet. Der Harvester liefert die Daten für die zentrale *Suchfunktion S*, mit der der Datennutzer eine Suche über die publizierten Daten aller Repositories sowie des Langzeitarchivs durchführen kann.

Auf dieser Basis kann dem Datenproduzenten bzw. -nutzer ein zentrales *Forschungsdatenportal FP* angeboten werden, von welchem er Zugriff auf alle Komponenten der vernetzten Infrastruktur erhält und welches ihm die für eine komfortable Handhabung seiner Daten benötigten Werkzeuge zur Verfügung stellt.

Die Installation einer vernetzten Infrastruktur mit einem Zentralen Langzeitarchiv und lokal angesiedelten Institutionellen Repositories als Subsystemen bietet den Vorteil, dass den Wissenschaftlern vor Ort die notwendige Unterstützung für die systematische Ablage ihrer Forschungsdaten zukommen kann. Die Betreiber der Institutionellen Repositories können aufgrund ihrer Nähe zu den Forschern eine viel bessere Autorenbetreuung, z. B. bei der Metadatenanreicherung und dem Upload-Prozess, leisten als das zentrale Langzeitarchiv, da dessen räumlicher Abstand zu den Wissenschaftlern in der Regel groß ist. Das Langzeitarchiv bündelt Objekte aus verschiedensten Quellen und steht damit einer sehr viel größeren Masse an Objekten gegenüber, die nur durch automatisierte Verarbeitung sinnvoll zu bewältigen ist. Im Idealfall konzentriert sich das das Archiv auf die Erfüllung seiner primären Aufgaben, der Langzeitarchivierung, und zieht keine Kompetenzen an sich, die die Subsysteme viel besser und effektiver wahrnehmen können.

2.3 Perspektive des Datennutzers

Es wird angenommen, dass ein Wissenschaftler in der Rolle des Datennutzers spezielle Daten für seine Forschung benötigt. Die

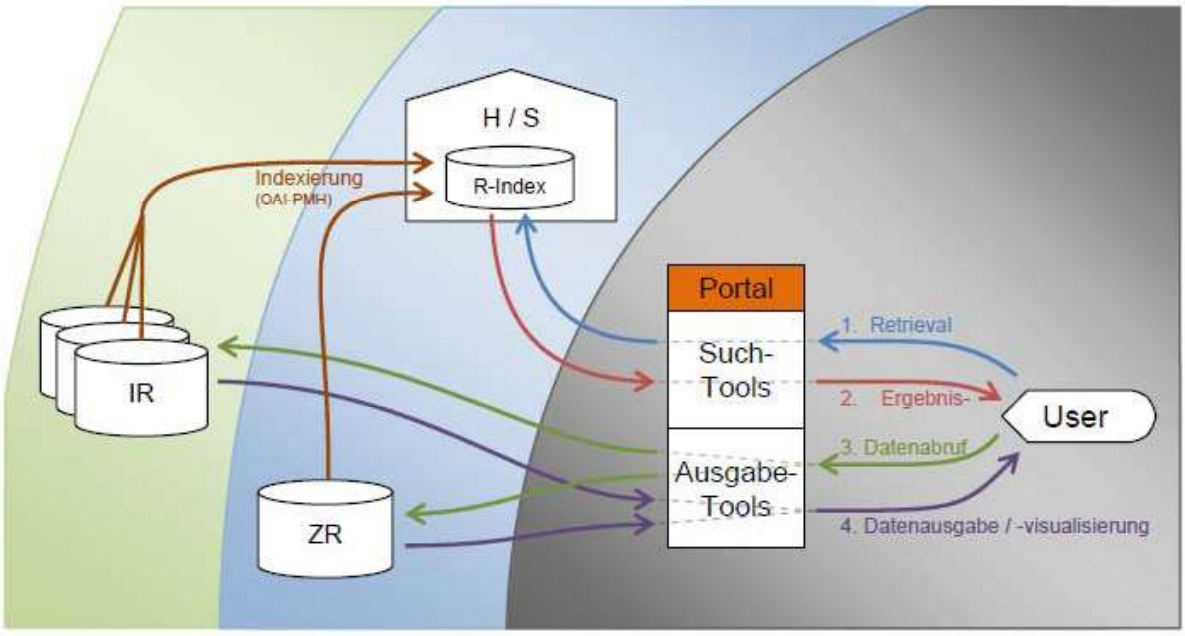


Abb. 24 zeigt die vernetzte Infrastruktur einschließlich der Datentransferprozesse aus der Perspektive des Datennutzers.

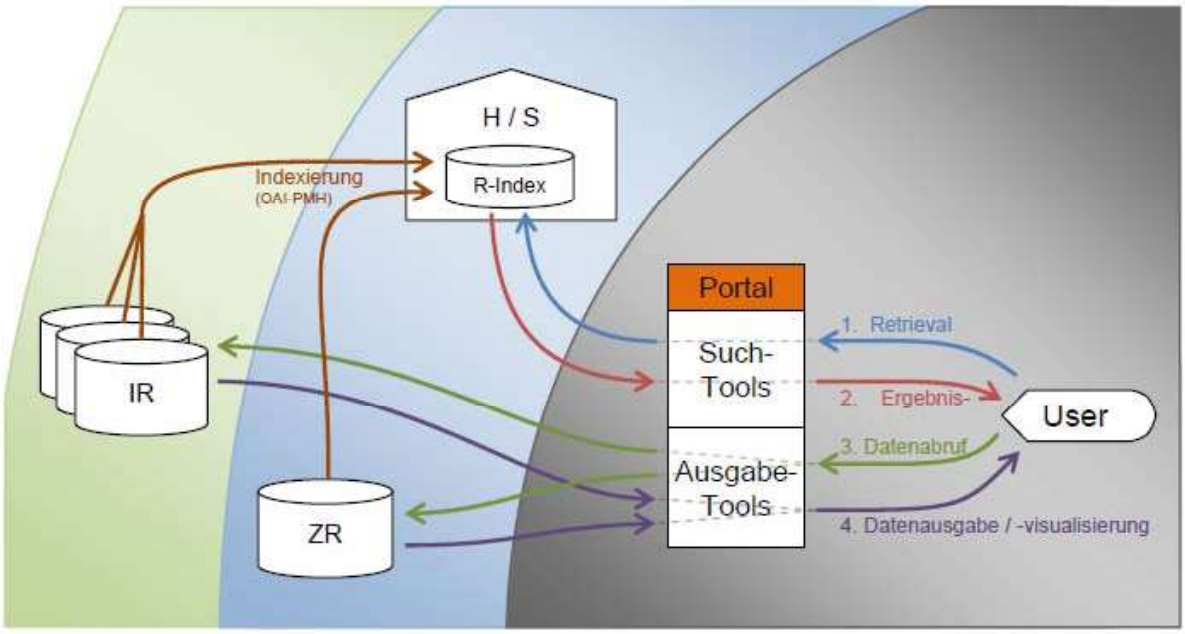


Abb. 24: Vernetzte Infrastruktur und Datentransferprozesse aus der Perspektive des Datennutzers

Der Datennutzer hat Zugriff auf ein zentrales Forschungsdatenportal mit vielen Funktionalitäten zur Datenverwaltung, Retrieval und Präsentation. Er kann über spezielle, auf ihn zugeschnittene Suchtools – wie z. B. einer Struktursuche – die vom Harvester zur Verfügung gestellten Datenbestände aus verschiedenen Quellen durchsuchen. Über die ausgelieferte Ergebnisliste erhält er gleichzeitig Zugriff auf die Datenbestände im Zentralen Repository ZR und auf die in den Institutionellen Repositories IR. Die den Datennutzer interessierenden Datensätze werden ihm entweder direkt vom Repository ausgeliefert oder durch einen nachgeschalteten Filter im Portal aufbereitet zur Verfügung gestellt. So werden dem Nutzer z. B. die von ihm gesuchten NMR-Daten nicht nur im Rohdatenformat ausgeliefert, sondern ihm wird zusätzlich eine Visualisierung der Daten auf Basis des Formats JCAMP-DX angeboten, welches ihm sogleich eine Interpretation und Einschätzung des Datensatzes ermöglicht.

2.4 Perspektive des Datenproduzenten

Der Wissenschaftler ist in der Regel nicht nur Datennutzer, sondern auch Datenproduzent. Die Abb. 25 zeigt die vernetzte Infrastruktur und die Datentransferprozesse aus der Perspektive des Datenproduzenten.

Ein Wissenschaftler erzeugt im Rahmen seiner Forschungsarbeit, oft bezogen auf ein Projekt, an diversen Geräten experimentelle Daten. Bei der Datenerzeugung müssen alle Geräte- und messspezifischen Parameter – wie z. B. Temperatur, Druck, Geräte-Hersteller, Geräte-Software, Geräte-Parameter, Datum, Operator oder auch das Pulsprogramm bei der NMR-Spektroskopie – als Metadaten in der Messdatendatei abgelegt werden. Zu einem späteren Zeitpunkt ist dies nicht mehr möglich und nur so ist gewährleistet, dass diese für die Dateninterpretation zwingend notwendigen Metadaten erhalten bleiben. Der Forscher legt seine Messdaten zunächst auf den Arbeitsplatzrechner bzw. Arbeitsgruppenserver, bearbeitet sie und wertet sie aus. Dann entscheidet er, welche weiteren Experimente er durchführen muss. Im Verlauf seiner Forschungen werden Datensätze zusammengefügt und verglichen.

Wichtige Daten, auf die sich seine Forschungsergebnisse stützen, legt er über das zentrale Forschungsdatenportal in das Institutionelle Repository seiner Universität bzw. Forschungseinrichtung. Bei der Datenablage im Institutionellen Repository werden die bereits existierenden Metadaten (Mess- und Geräteparameter) mit weiteren beschreibenden Metadaten ergänzt. Damit die Daten für seine Kollegen bzw. auch für eine breite Öffentlichkeit verständlich sind, müssen beispielsweise Strukturinformationen, der Name des Autors, Informationen zum wissenschaftlichen Kontext und Ziel des Experiments, zur

verwendeten Technik usw. hinterlegt werden. Gegebenenfalls können auch mehrere zusammengehörige Datensätze in Form eines Datencontainers zusammengefasst werden, wobei dann zusätzlich weitere Metadaten über den Container erforderlich sind.

Die Erfassung und Hinterlegung von Metadaten sollte so weit wie möglich automatisiert erfolgen, um die Bereitschaft des Wissenschaftlers für diese wichtige Aufgabe zu erhöhen. Datensätze, die Bestandteil einer Publikation sind, muss der Datenproduzent zur dauerhaften Speicherung vom Institutionellen Repository, gegebenenfalls auch vom Arbeitsgruppenserver, über das zentrale Forschungsdatenportal im Langzeitarchiv ablegen. Wenn sich die Metadatenschemata von Institutionellem Repository und Archiv unterscheiden, müssen die beschreibenden Metadaten möglicherweise vervollständigt werden. Sind die im Institutionellen Repository abgelegten Daten kein Bestandteil einer Publikation, ist zu Projektende bzw. mit Ablauf von zehn Jahren die Notwendigkeit einer dauerhaften Aufbewahrung zu prüfen.

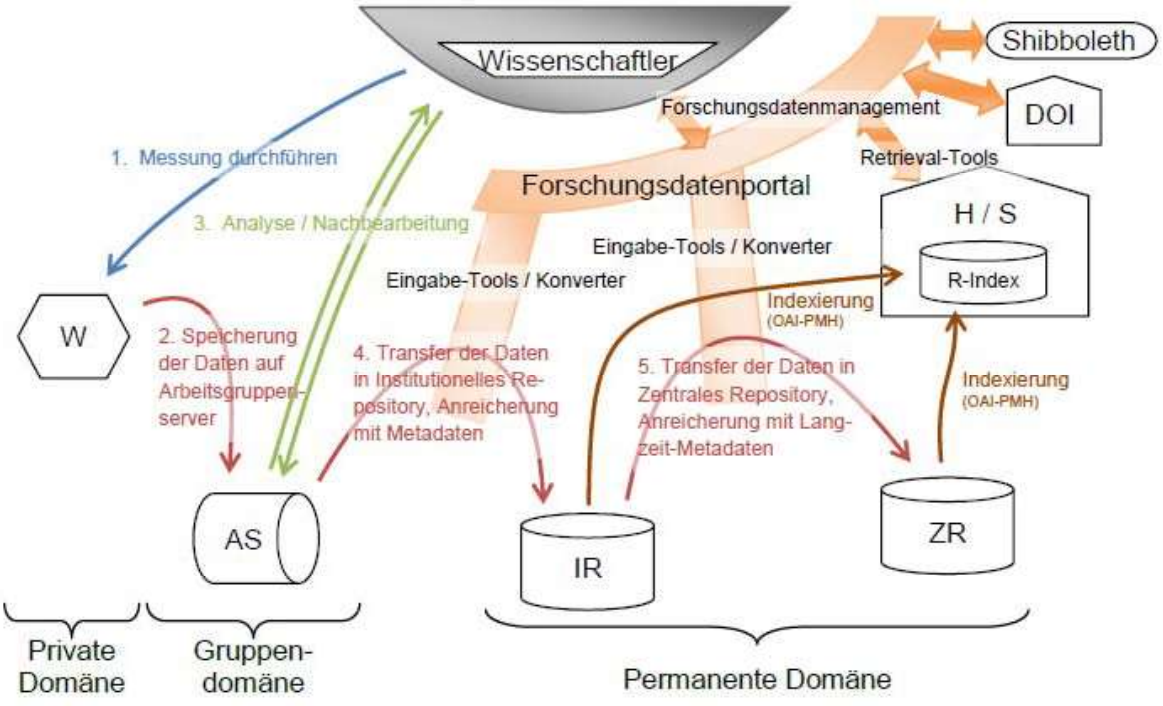


Abb. 25: Vernetzte Infrastruktur und Datentransferprozesse aus der Perspektive des Datenproduzenten

Mit der Ablage von Daten in das Institutionelle Repository kann der Wissenschaftler seine Daten bequem von einer Stelle aus verwalten: Das Forschungsdatenportal, welches die permanente Domäne verbindet, bietet ihm einen Repository-übergreifenden Zugriff auf seine

Daten und gibt ihm Werkzeuge zur Hand die ihm die Datenorganisation in Zukunft deutlich vereinfachen.

2.5 Institutionelle Repositories

Ein Repository ist ein an einer Universität oder sonstigen Forschungseinrichtung betriebener Dokumentenserver, auf dem wissenschaftliche Materialien digital publiziert und weltweit zugänglich gemacht werden. In den letzten Jahren haben sie sich im Rahmen der Open Access Initiative[73] an vielen Instituten etabliert. Dabei wird zwischen Institutionellen und Disziplinären Repositories unterschieden. Institutionelle Repositories werden innerhalb einer Institution, häufig von der Universitätsbibliothek, betrieben. Sie sind disziplinübergreifend und ermöglichen den Wissenschaftlern eine digitale Publikation ihrer Forschungsergebnisse und deren Archivierung. Disziplinäre Repositorien hingegen sind institutionsübergreifend und konzentrieren sich auf eine Fachdisziplin. Die Gestaltungsformen von Repositories sind reichhaltig, z. B. unterscheiden sie sich durch die Art des dargebotenen Inhalts oder durch ihre Mehrwertdienste. Es gibt eine Vielzahl existierender Software für die Installation eines Repositories, darunter auch viele Open Source-Produkte. In der Regel sind in der Software von Haus aus standardisierte Schnittstellen wie OAI-PMH[72] integriert, über die sich eine Vernetzung von Repositories realisieren lässt.

Es ist sinnvoll, für ein Netz von Institutionellen Daten-Repositories einheitliche Richtlinien zu definieren. Ein erster Ansatz könnte folgendermaßen aussehen:

- Jedes einzelne Repository muss die von der Open Archive Initiative entwickelte Schnittstelle OAI-PMH besitzen. Über diese Schnittstelle werden die Metadaten der gespeicherten Datensätze mit Hilfe des Harvesters eingesammelt und nutzbar gemacht. Über eine im Forschungsdatenportal angebotene Suche können alle Repositories gleichzeitig durchsucht werden. Die Repository-übergreifende Suche ist bereits von OAI-Serviceprovidern wie DRIVER[74], OAIster[75] und BASE[76] realisiert worden.
- Es sollte ein einheitlicher Basis-Metadatensatz definiert werden, den alle Institutionellen Repositories implementieren müssen. Ein Basis-Datensatz erleichtert die Vernetzung der Repositories und stellt die Interoperabilität mit anderen Schemata sicher.
- Soweit wie möglich sollten die Policies für die Institutionellen Repositories disziplinübergreifend, also in einer Art nationaler Gesamtstrategie, vorgegeben werden. Auch für die Organisation der Institutionellen Repositories ist eine Anlehnung an das OAIS-Referenzmodell (s. Kap. 6.1.1) denkbar. Ebenso hat DINI[77], die

Deutsche Initiative für Netzwerkinformation, eine Reihe von Kriterien entwickelt, die als Mindeststandard für ein qualitätsgesichertes Repository angesetzt werden könnten.

- Viele Aspekte, wie z. B. die Art der verwendeten Repository-Software, werden aber auch weiterhin der Betreiberorganisation obliegen.
- Das Institutionelle Repository soll die Einhaltung der Empfehlungen der DFG zur Guten Wissenschaftlichen Praxis[36] garantieren und deshalb für mindestens zehn Jahre die Daten speichern. Nach dieser Zeit sollte eine Löschung oder Übertragung in ein Langzeitarchiv überprüft werden (s. Kap. 6.2.1).
- Zwischen dem Datenproduzenten und dem Betreiber des Repositories werden Vereinbarungen für die Datenablage abgeschlossen.
- Für jedes Institutionelle Repository sollte als Qualitätsnachweis eine Zertifizierung, z. B. über DINI[78], angestrebt werden.

Die Realisierung von konkreten Vorgaben für Institutionelle Daten-Repositories ist nur durch eine nationale Initiative möglich.

2.6 Datentransfer vom Institutionellen Repository zum Archiv

Sollen Datensätze langfristig und qualitätsgesichert gespeichert werden sowie langzeitverfügbar sein, ist ein Datentransfer vom Institutionellen Repository in das Langzeitarchiv unerlässlich. Die technisch einfachste Lösung ist der manuelle Download der Daten aus dem Institutionellen Repository und der nachfolgende Upload in das Langzeitarchiv. Dieser Weg ist aber für den Datenproduzenten aufwändig und die Wahrscheinlichkeit einer Akzeptanz sehr gering.

Viel sinnvoller wäre eine automatisierte Schnittstelle, auf dessen Basis der Nutzer einen Datensatz im Repository einschließlich seiner Metadaten ohne erneutes Hochladen, also quasi auf Knopfdruck, in das Langzeitarchiv transferieren kann. Der Nutzer könnte so seine gesamten, von ihm erzeugten Daten in einem einzigen, zentralen Forschungsdatenportal verwalten und in aller Ruhe entscheiden, welche Datensätze dauerhaft aufbewahrt werden sollen. Mit steigender Anzahl an Institutionellen Repositories wird aber die Realisierung einer automatisierten Schnittstelle schwieriger, weil umso mehr verschiedene Software-Lösungen zu berücksichtigen sind. Zudem gibt es individuelle Installationen und spezielle Lösungen, so dass eine derartige Schnittstelle möglicherweise nicht für alle existierenden Institutionellen Repositories realisierbar ist.

Der Datentransfer von Institutionellem Repository in das Langzeitarchiv ist charakterisiert durch eine Metadatenanreicherung. Der Datenproduzent ist verantwortlich für einen vollständigen Satz an beschreibenden Metadaten, während der Archivbetreiber das abzulegende Datenpaket mit administrativen Metadaten einschließlich spezieller Metadaten zur Langzeitarchivierung versehen muss.

Der Datentransfer stellt aber nicht nur einen technischen Prozess zwischen zwei Systemen dar, sondern ist auch mit vielen organisatorischen Anforderungen verbunden. Die Voraussetzung für die Übernahme von Daten in das Langzeitarchiv ist immer der Abschluss einer Datenübergabevereinbarung zwischen dem Datenproduzenten bzw. dem datenproduzierenden Institut und dem Archivbetreiber. Darin sind u. a. die Rechteproblematik, die Archivierungsdauer, die Art der zu archivierenden Objekte und die Verantwortlichkeiten geregelt.

2.7 Langzeitarchiv und Harvester

Im Zentrum der vernetzten Infrastruktur für chemische Forschungsdaten steht der Verbund aus Langzeitarchiv und Harvester, der in diesem Konzept vom FIZ CHEMIE als Archivbetreiber zur Verfügung gestellt wird. Das Langzeitarchiv nimmt nicht nur aufgrund dieser strukturellen Anordnung eine zentrale Position ein, sondern auch wegen seiner Langzeitstrategie. Nur das Langzeitarchiv kann eine Verfügbarkeit der in vielfältigen Formaten vorhandenen Forschungsdaten auf lange Sicht garantieren. Die dauerhafte Archivierung, insbesondere die Sicherstellung der Nachnutzbarkeit, ist eine mit Kosten verbundene, verantwortungsvolle Aufgabe, die von den Betreibern der Institutionellen Repositories in der Regel nicht geleistet werden kann. So bedarf es entsprechender finanzieller und personeller Kapazitäten sowie umfassender Expertise im Bereich Datenmanagement, um dem komplexen Gebiet der Langzeitarchivierung gerecht zu werden.

2.8 Zentrales Forschungsdatenportal

In den vorhergehenden Kapiteln wurden die Komponenten einer vernetzten Forschungsdaten-Infrastruktur und die darin stattfindenden Datentransfers beschrieben sowie die Perspektiven von Datenproduzent und -nutzer betrachtet. Eine vernetzte Infrastruktur mit den vorgestellten Funktionalitäten kann aber nur funktionieren, wenn ein übergeordnetes zentrales Forschungsdatenportal die einzelnen Komponenten und Strukturen in ein einheitlich bedienbares Gesamtsystem überführt.

Das Forschungsdatenportal stellt die zentrale Anlaufstelle und Arbeitsplattform sowohl für den Datennutzer als auch den Datenproduzenten dar. Es umspannt die Permanente Domäne, bestehend aus Institutionellen Repositories, Langzeitarchiv und Harvester, und ermöglicht dem Datenproduzenten eine Repository-übergreifende Datenorganisation, indem es einen zentralen Zugang zu den drei genannten Komponenten anbietet.

Das Forschungsdatenportal besitzt viele nützliche Features und vereint alle Tools, die dem Wissenschaftler die Handhabung und Organisation seiner Daten erleichtern. Nur durch die Installation eines zentralen Forschungsdatenportals können die zu entwickelnden Werkzeuge für die Datenorganisation die Wissenschaftler großflächig erreichen, unabhängig davon, in welchem Institut sie forschen und welches Repository ihnen zur Verfügung steht. So werden dem Forscher die komfortablen Upload-Tools nicht erst bei der Datenablage in das Langzeitarchiv angeboten, sondern er kann diese bereits einen Schritt früher, also für den Datentransfer in das Institutionelle Repository seines Instituts einsetzen.

Der experimentierende Wissenschaftler erhält beispielsweise Werkzeuge, mit denen Metadaten aus der Messwertedatei ausgelesen werden, und über ein Composer-Tool kann er zusammengehörige Datensätze zu einem Datenpaket zusammenfassen. Ein Datenpaket sollte ganz einfach per Knopfdruck verschoben werden können, um dieses, oder alternativ auch eine Kopie davon, im Langzeitarchiv abzulegen. Bei diesem Prozess sollte eine DOI-Registrierung des Datensatzes erfolgen.

Ausgehend vom zentralen Portal startet auch die Repository-übergreifende Suche auf Basis der vom Harvester eingesammelten Metadaten. Die bei einer Suche gefundenen Datensätze werden angezeigt, visualisiert, vernetzt und zum Download angeboten.

Für den Zugang zum Forschungsdatenportal sollte ein Authentifizierungs- und Rechteverwaltungssystem aufgebaut werden, welches die Nutzung beliebig verteilter Ressourcen mit einem einzigen Account ermöglicht.

Unter Berücksichtigung dieser beschriebenen vernetzten Infrastruktur ist in dieser Studie ein Langzeitarchiv zu konzipieren, welches in eine Landschaft diverser, auf unterschiedlicher Software beruhender Institutioneller Repositories nahtlos eingebettet werden kann. Deshalb muss das Archiv zusätzlich mit einem Harvester verknüpft werden, auf dessen Basis eine vernetzte Architektur realisiert werden kann. Komfortable, in ein Forschungsdatenportal integrierbare Tools sollen dem Wissenschaftler die Datenhandhabung und -organisation deutlich erleichtern.

C 3. Digitale Erhaltungsstrategien

3.1 Bitstream Preservation

Digitale Informationen werden als Bits, also als Informationseinheiten mit den Werten „0“ oder „1“, auf Datenträgern gespeichert. Datenträger aber sind nicht unendlich lange haltbar, also nach gewisser Zeit möglicherweise unlesbar, so dass Strategien entwickelt werden müssen, um diesen so genannten Bitstream zu erhalten (Bitstream Preservation). Weiterhin bedarf es entsprechender Hardware, wie z. B. Laufwerken, um die Daten von den Medien überhaupt lesen zu können, und zudem werden Computer benötigt, um die Daten zu verarbeiten.

Die Sicherstellung des physischen Erhalts der Datensätze sowie deren Lesbarkeit ist eine grundlegende Voraussetzung für die digitale Langzeitarchivierung. Gemäß nestor[79] gibt es folgende Erhaltungsstrategien:

- **Regelmäßige Medienmigration**: Die verwendeten Speichertechniken bzw. Datenträger müssen regelmäßig durch neue ersetzt werden.
- **Redundante Datenhaltung**: Die Daten sollten in mehrfacher Kopie vorliegen. Zur Sicherung gegen äußere Einflüsse empfiehlt sich auch eine räumlich getrennte Aufbewahrung der unterschiedlichen Kopien.
- **Diversität der eingesetzten Speichertechnik**: Die Daten sollten auf mindestens zwei unterschiedlichen Datenträgertypen gesichert werden.
- **Standards**: Die verwendeten Speichermedien sollten internationalen Standards entsprechen und auf dem Markt eine weite Verbreitung aufweisen.

Zuverlässige Strategien berücksichtigen nicht nur ein Refreshment der Datenträger innerhalb einer Speichertechnik, sondern darüber hinaus auch die Aktualisierung ganzer Speichertechniken. Der Bitstream-Erhalt kann über den Aufbau einer geeigneten Hardware-Architektur sichergestellt werden. Ein entsprechendes Konzept wird in Deliverable AP2.1.1 (Kap. 4) entworfen.

3.2 Migration und Emulation

Aber Bitstream Preservation allein kann nicht die Nachnutzbarkeit von digitalen Daten sichern. So muss auch die passende Software existieren, die die gespeicherten und lesbaren Daten anschließend interpretiert und die digitalen Objekte erst wieder nutzbar macht. Auch der Gefahr, dass das Wissen um die korrekte Interpretation der Daten verloren geht, muss mit geeigneten Strategien entgegengewirkt werden. Lösungsansätze, die die Interpretierbarkeit der Daten über lange Zeit garantieren sollen, sind die Migration von Formaten (davon zu unterscheiden ist die Datenträgermigration bei der Bitstream Preservation) und die Emulation.

Bei der Migration werden die Daten einem neuen Umfeld angepasst, indem sie von einem Datenformat in ein aktuelleres, möglichst standardisiertes und offen gelegtes Format überführt werden. Dies muss frühzeitig geschehen, wenn die Gefahr der Veralterung eines Formates droht. Das Ziel ist eine möglichst authentische Weitergabe der Daten zur Sicherstellung der dauerhaften Nutzbarkeit. Für eine verlustfreie Migration müssen Original- und Zielformat eindeutig und in offener Form spezifiziert sein, einfache Formate sind im Allgemeinen von Vorteil.

Vorteile der Migration:

- technisch (verglichen mit Emulation) gut realisierbar
- Möglichkeit der Automatisierung
- Aufbewahrung der originalen Objekte

Nachteile der Migration:

- jedes Objekt nur einzeln migrierbar
- Gefahr von Datenverlust bzw. -veränderung besonders bei mehreren Migrationsschritten
- höherer Speicherbedarf aufgrund der Aufbewahrung von Vorgänger-Formaten

Bei der Emulation wird die originale Umgebung der archivierten digitalen Objekte simuliert, um z. B. drohende Verluste einer Datenformatmigration zu umgehen. Emulation kann auf der Ebene des Betriebssystems oder auf der Ebene der Hardware-Plattform erfolgen.

Vorteile der Emulation:

- Originalobjekte bleiben unverändert, keine Konvertierung
- weniger Speicherplatz als bei Migration

Nachteile der Emulation

- technisch teils schwer zu implementieren
- hoher Aufwand pro Hardware-Generationenwechsel, für jede Plattform neue Emulatoren
- Spezifikationen für zu emulierende Objekte/Systeme nicht immer hinreichend bekannt

Es ist notwendig, dass der Archivbetreiber laufend die Formate der in seinem Archiv aufbewahrten Objekte beobachtet. Droht für ein Format die Gefahr einer zukünftigen Nichtinterpretierbarkeit, müssen frühzeitig Maßnahmen ergriffen werden. In einem ersten Schritt ist der Hersteller anzusprechen. Dieser sollte an der Interoperabilität der Daten interessiert sein und über das nötige Knowhow für die Bereitstellung einer Lösung verfügen. In vielen Fällen stehen rechtzeitig Konverter bereit, über die eine Formatkonvertierung im Batchbetrieb realisiert werden kann. Bietet der Hersteller keine Lösung, weil er z. B. nicht mehr existiert, können auch andere Anbieter diese Lücke schließen. So stellt beispielsweise ACD/Labs eigene Konverter und Batchprozessoren zur Verfügung.

Sollte auch über einen Drittanbieter keine Lösung realisierbar sein, ist zunächst der Aufbau einer Emulationsumgebung zu prüfen. Ist es möglich, die Emulation auf Betriebssystemebene durchzuführen, können aktuelle Systeme relativ problemlos in Virtualisierungslösungen wie VMWare langfristig weiter betrieben werden. So unterstützen die aktuellen Systeme von VMWare auch noch den Betrieb von MS-DOS[80].

Auch die Emulation spezieller Hardware ist heutzutage machbar. So können OpenVMS-Systeme auf Alpha-CPU-Basis im Emulator unter Windows betrieben werden[81].

Ist auch die Emulation nicht realisierbar, bleibt nur der Versuch, mittels eines Programm-Herstellers – oder wie im Fall des FIZ CHEMIE als Archivbetreiber mit der eigenen Softwareentwicklung – eine spezielle Lösung zu schaffen.

Diese Erhaltungsstrategien müssen in der Dokumentation des Archivs (Deliverable AP2.2, Kap. 6.2.2) definiert und festgelegt werden. Auch gibt es administrative Metadaten, die die Durchführung dieser Erhaltungsmaßnahmen unterstützen, indem sie z. B. Migrationszyklen festlegen. Letztendlich muss jede mit den Daten vorgenommene Maßnahme wiederum in den Provenienz-Metadaten, die die Historie des archivierten Datensatzes beschreiben, erfasst werden. Auch auf die rechtlichen Metadaten können diese Maßnahmen Einfluss nehmen, da in ihnen gegebenenfalls Einschränkungen für die Migration und Emulation festgelegt sind.

C 4. Hardware-Architektur für ein Langzeitarchiv

Der Bitstream-Erhalt fällt in den Bereich der Hardware-Planung und kann durch die Konzipierung einer geeigneten Hardware-Architektur als Kern des Archivs zur Verfügung gestellt werden. Das nachfolgend vorgestellte Hardware-Konzept berücksichtigt die dauerhafte Aufbewahrung der Daten z. B. durch rechtzeitige Migration auf neue Datenträger bzw. Datenträgertechnologien, sichert aber keinesfalls deren Nachnutzbarkeit.

Der Hardware-technische Teil des Projekts erfordert den Aufbau eines zentralen Datenspeichers für die Forschungsdaten, eines Datenbanksystems für den Harvester sowie des Frontendservers mit Tools für eine Validierung, Konvertierung und Visualisierung. In einer späteren Phase werden zudem Universal Virtual Computer (UVC) benötigt, in denen Applikationen, die unter aktuellen Betriebssystemen nicht mehr funktionsfähig sind, betrieben werden können.

Die Anforderungen an die Hardware des zentralen Datenspeichers sind abhängig von der Art der Datenspeicherung. Für das Archivieren von Daten bestehen zunächst zwei klassische Möglichkeiten:

- das Speichern in einer Datenbank
- das Speichern in einem Filesystem

Bei beiden Möglichkeiten werden im Normalfall die zugehörigen Metadaten zusammen mit einer Referenzierung zum Datensatz in einer Datenbank gespeichert und indexiert. Der Nachteil der klassischen Datenspeicherung besteht in der Tatsache, dass eine Sicherung der Datenbestände nur mittels eines klassischen Backups erfolgen kann. Mit ansteigendem Datenvolumen aber erhöht sich proportional der Sicherungsaufwand. Die Sicherung von mehreren Terabyte an Daten kann unter Umständen einige Tage dauern. Hinzu kommt, dass bei klassischen Verfahren alle Datenbestände auf Festplatten verfügbar gehalten werden müssen. Dies ist sowohl hinsichtlich der Investitionskosten als auch der Betriebskosten unvorteilhaft. Berücksichtigt man zudem ökologische Aspekte, so ist die Frage zu stellen, ob die Speicherung sehr alter oder sehr wenig genutzter Datenbestände kontinuierlich Strom verbrauchen sollte. Daraus resultiert das Fazit, dass eine klassische Datenspeicherung für den Betrieb eines Langzeitarchivs nicht empfehlenswert ist.

Aufgrund der Tatsache, dass nur ca. 20 % der gespeicherten Daten aktive Daten sind (Quelle: IDC[82]), hat sich das Hierarchische Speichermanagement (HSM) als Alternative zur klassischen Datenspeicherung etabliert. Hierbei werden Daten, auf welche längere Zeit nicht zugegriffen wurde, auf alternative Speichermedien wie Magnetbänder ausgelagert. Diese Speichermedien sind in der Regel preiswerter, haben jedoch den Nachteil, dass die

Zugriffszeiten auf die Daten verlängert werden. Aber gerade bei sehr großen Datenbeständen kann durch den Einsatz von Magnetbändern eine erhebliche Energieeinsparung erzielt werden.

Für die Speicherung von Daten in einem HSM-System werden Policies definiert, die bestimmen, unter welchen Bedingungen die Daten wie, wo und wie oft gespeichert werden, aber auch, wann sie von einer Speicherhierarchiestufe zu einer anderen verschoben werden. Durch die Möglichkeit, mehrere Instanzen eines Datums auf unterschiedlichen Medien zu speichern, entfällt das klassische Backup. Auch ein Austausch veralteter Medientechnologien ist ohne Unterbrechung des Betriebs und ohne eine verringerte Datensicherheit möglich. Bekannte Vertreter einer HSM-Lösung sind der Tivoli Storage Manager (TSM)[68] von IBM und der Storage- und Archive-Manager (SAM-FS/QFS)[83] von SUN.

Sowohl die klassische als auch die hierarchische Datenspeicherung benötigen im Prinzip die gleichen Hardwarekomponenten, jedoch in unterschiedlichen Verhältnissen. Bei der klassischen Datenspeicherung wird angenommen, dass für die bereitgestellte Festplattenspeicherkapazität ca. die dreifache Kapazität an Magnetbändern zur Datensicherung benötigt wird. Bei der hierarchischen Datenspeicherung hingegen wird davon ausgegangen, dass nur für die aktiven Datenbestände Festplattenkapazitäten bereitgestellt werden müssen. Dies sind ca. 20 % der geplanten Archivkapazität. Geht man von der Speicherung von drei Instanzen je Datum aus, entsprechen die benötigten Magnetbandkapazitäten denen der klassischen Speicherung.

Diese Systeme sehen aus Sicherheitsgründen von Haus aus die Erweiterbarkeit auf einen zweiten Standort vor. Aufgrund der Tatsache, dass das Ziel des Projekts der Langzeiterhalt von Forschungsdaten ist, sollte diese Möglichkeit, die Sicherheit der Datenbestände zu erhöhen, nicht außer Acht gelassen werden und als Anforderung an die Hardwarelösung eingeplant werden.

Mit einer professionellen HSM-Lösung sind folglich die von nestor postulierten vier Strategien für den Bitstream-Erhalt (s. Kap. 3.1) bereits erfüllt bzw. problemlos realisierbar.

Für den Aufbau des Datenbanksystems und des Frontendservers kann aufgrund der hohen Leistungsfähigkeit moderner Computer auf eine virtualisierte Infrastruktur (z. B. VMWare) zurückgegriffen werden. Diese bietet den Vorteil, dass sie transparent erweitert werden kann und schon jetzt alle Voraussetzungen für den Aufbau der Emulationsumgebung mittels UVCs liefert.

4.1 Modell I: Aufbau eines Langzeitarchivs im FIZ CHEMIE

In Modell I ist das Langzeitarchiv vollständig im FIZ CHEMIE lokalisiert. Damit ist das FIZ CHEMIE sowohl für den Bitstream-Erhalt zuständig als auch für die Entwicklung von Archivierungsstrategien sowie für die Sicherstellung der Zugänglichkeit und Verfügbarkeit der Forschungsdaten.

Die im FIZ CHEMIE vorhandene Server- und Storage-Umgebung konzentriert sich auf wenige Hersteller. Im Storage-Umfeld sind dies HP und SUN, im Serverumfeld Dell, HP und Supermicro. Für die Integration eines Archivs im FIZ CHEMIE ist eine Infrastruktur von SUN oder HP zu bevorzugen, da einerseits diese Systeme seit vielen Jahren erfolgreich betrieben werden und die entsprechenden Erfahrungen vorhanden sind und andererseits auch teilweise auf die Ressourcen vorhandener Systeme zurückgegriffen werden kann.

Geht man davon aus, dass für die ersten Phasen des Projekts, in denen die reine Bitstream-Archivierung im Vordergrund steht (s. Kap. 3.1), mit einem Datenvolumen von ca. 50 Terabyte zu rechnen ist, belaufen sich die benötigten Festplattenkapazitäten auf 10 Terabyte und die Magnetbandkapazitäten auf 150 Terabyte. Mit den Produktpaletten der genannten Hersteller kann heutzutage ein solches Datenvolumen problemlos verarbeitet werden. Ein kleines Festplattensystem (z. B. HP MSA1000 oder SUN StorageTek 6140) und ein modularer LTO-4-Taperoboter (z. B. HP MSL6000 oder SUN StorageTek SL500) erfüllen bereits die prinzipiellen Voraussetzungen. Um eine Ausfallsicherheit zu gewährleisten und eine erhöhte Datensicherheit realisieren zu können, sollte der Taperoboter, idealerweise auch das Festplattensystem, redundant eingeplant werden. Zusätzlich zu den Speichersystemen sind zwei geclusterte Serversysteme (z. B. HP ProLiant DL38x oder SUN Fire X4270) zum Betrieb der HSM-Lösung und der Archivlösung einzuplanen. Die folgende Abbildung zeigt einen möglichen Hardware-Aufbau eines Archivs.

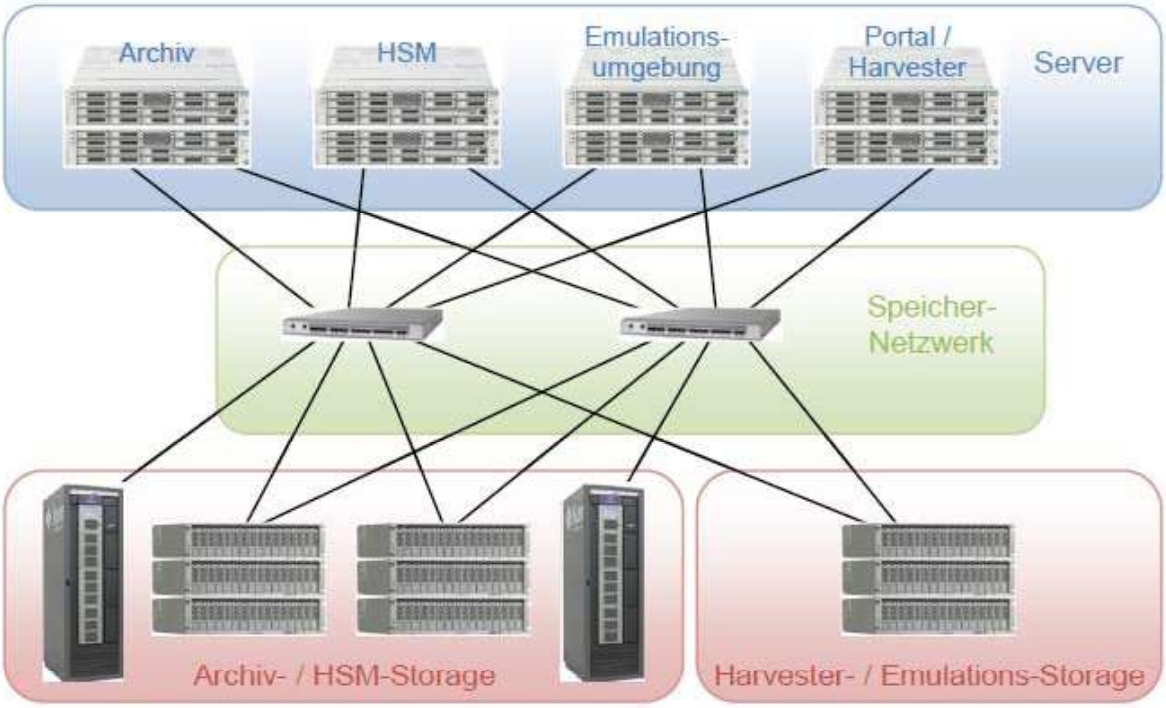


Abb. 26: Hardware-Aufbau eines Archivs im FIZ CHEMIE

Da die Verfügbarkeit eines Archivs für Forschungsdaten ein sehr wichtiges Kriterium für seine Akzeptanz ist, sollte die gesamte Hardwareumgebung so ausgelegt werden, dass sich alle Komponenten im laufenden Betrieb ergänzen und austauschen lassen. Aber auch eine Erweiterbarkeit auf einen zweiten Standort sollte realisierbar bleiben.

4.2 Modell II: Langzeitarchiv in Kooperation mit Archivdienstleister

Eine interessante Alternative zum Aufbau eines im FIZ CHEMIE lokalisierten Archivs ist eine dezentrale Struktur, die auf einer Kooperation des FIZ CHEMIE mit einer bereits ein Archiv betreibenden Organisation basiert. Prinzipiell sind für eine Kooperation alle Organisationen geeignet, deren Archive die technischen Voraussetzungen für einen dauerhaften Bitstream-Erhalt – entsprechend der von nestor vorgegebenen Erhaltungsstrategien[79] – erfüllen.

In Modell II werden dementsprechend die Verantwortlichkeiten zwischen zwei Partnern aufgeteilt: Für den Bitstream-Erhalt ist der Archivdienstleister zuständig, während das FIZ CHEMIE für die Entwicklung von Archivierungsstrategien und die Sicherung der dauerhaften Verfügbarkeit der Forschungsdaten verantwortlich ist.

Dieses Modell bietet den Vorteil, dass der Hardware-Aufwand im FIZ CHEMIE sehr gering gehalten werden kann und der geplante Aufbau eines Archiv-Prototyps deshalb ohne wesentliche Verzögerungen möglich ist. In dieser Konstellation wird im FIZ CHEMIE nur der

Frontend-Server für die Archivierung aufgebaut, während die Speicherung der eigentlichen Forschungsdaten über ein ausgelagertes Backendsystem, in der Abb. 27 grau hinterlegt, erfolgt.

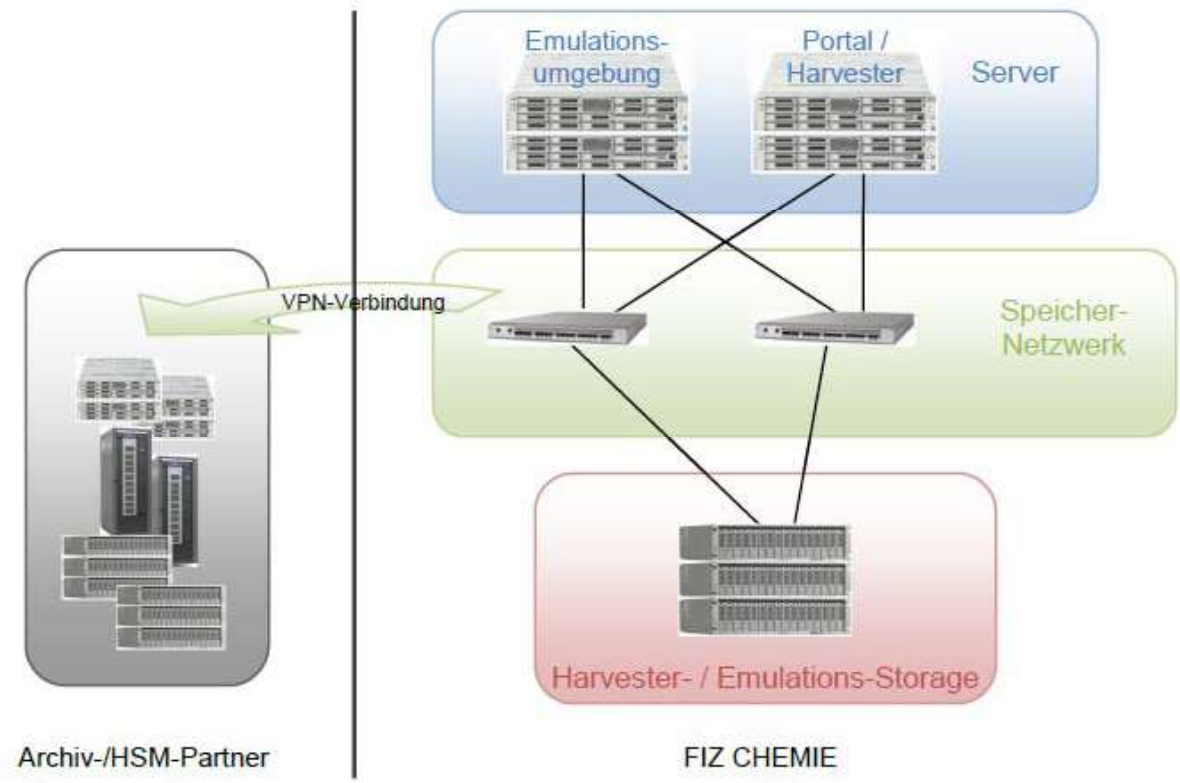


Abb. 27: Archiv in Kooperation mit einem Archivdienstleister

Eine für diese Architektur besonders geeignete HSM-Lösung ist ein Archiv, welchem die Software SUN SAM-FS/QFS zugrunde liegt. Hierbei handelt es sich um ein policy-basierendes, selbstsicherndes, virtuell unlimitiertes Filesystem. Es ermöglicht die transparente, gleichzeitige Speicherung von Files an verschiedenen Stellen (Festplatten-Arrays und Tape-Libraries) und unterschiedlichen Standorten. Durch den Einsatz eines Festplattencaches skaliert die Performance des Filesystems mit seinem Wachstum. Das Verfahren, bis zu vier gleichzeitige Kopien speichern zu können, löst zusätzlich das Problem der Sicherung großer Multi-Terabyte-Datenmengen, da nur noch Filesystemtabellen gesichert werden müssen. Im Falle eines Desaster-Recoveries reduziert sich die Wiederherstellung des Systems maximal auf wenige Stunden, da gegebenenfalls nur die Server neu installiert und die Filesystemtabelle neu eingelesen werden müssen, jedoch nicht der gesamte Datenbestand des Archivs. Ein weiterer Vorteil dieser Lösung besteht darin, dass ein Technologie-Wechsel für den Anwender vollständig transparent durchgeführt werden kann. Beim Wechsel der Tape-Technologie müssen nur die neue Anlage in Betrieb

genommen, eine neue Kopie der Files auf der neuen Anlage erzeugt und abschießend die alte Anlage abgeschaltet werden. Für dieses System steht auch eine zusätzliche Lösung zur Datenträgerüberwachung zur Verfügung, welche einen potentiellen zukünftigen Ausfall eines Bandes erkennen lässt und somit ein rechtzeitiges Kopieren des Bandes initiiert.

Die Anbindung zwischen den Frontend-Servern und dem SAM-FS/QFS-Backend-Servern kann direkt via eines Network Files System (NFS) oder über einen eigenen QFS-Client in den Frontend-Servern erfolgen. Die Netzwerkverbindung zwischen den Rechenzentren sollte über eine VPN-Lösung abhörsicher eingerichtet werden.

Da in der Entwicklungsphase nur geringe Ressourcen sowohl bezüglich Server-Kapazitäten wie auch Storage-Kapazitäten benötigt werden, besteht der besondere Charme dieses Konzepts darin, dass man mit einfachen, ja sogar nur mit virtuellen Server-Systemen als Frontend-Server starten kann und dass die Storage-Volumina und die Netzwerk-Bandbreite zwischen den Rechenzentren sehr gering sind. Für den Archivdienstleister würde das erforderliche Storage-Volumen nahezu keine Kosten verursachen.

Werden in einer späteren Phase des Projekts höhere Storage-Volumina und höhere Zugriffsgeschwindigkeiten erforderlich, ist es aufgrund der Architektur von SAM-FS/QFS problemlos möglich, unterbrechungsfrei weitere Storage-Kapazitäten bereitzustellen. Zur Erhöhung der Performance könnten später sogar der Festplattencache und auch die primäre Arbeitskopie unterbrechungsfrei zum FIZ CHEMIE verschoben werden, wobei dann nur noch Sicherungskopien beim Archivbetreiber vorgehalten werden. Dieses Vorgehen wäre ein logischer Schritt, zumal er zusätzlich die Daten auf zwei Standorte verteilt und somit die Datensicherheit erhöht.

Dieser in Modell II beschriebene Ansatz reduziert die benötigten IT-Ressourcen und die Investitionskosten für eine Realisierung des Datenarchivs signifikant. Während der Entwicklungsphase sind nur minimale Kosten für die virtuellen Server und ein Storage-Volumen von weniger als 1 TB erforderlich. Für die Realisierungsphase beschränken sich die Investitionen am FIZ CHEMIE nur auf vier Server mit Betriebssystem und Applikationen sowie einige Infrastrukturgeräte. Die sonst übliche Anschaffung erweiterbarer Backupsysteme sowie die nötige Backupsoftware entfallen.

4.3 Virtualisierungsumgebung für Forschungsdatenportal, Harvester und Emulation

Da für die später benötigten UVCs eine Virtualisierungsumgebung, wie z. B. VMWare Sphere, benötigt wird, ist es nur logisch, diese von Anfang an als technische Basis auch für

den Forschungsdatenportal-Server inklusive der Umgebung für die Tools einzuplanen. Ein weiterer Vorteil besteht darin, dass später sehr einfach zusätzliche Ressourcen in das System eingebracht werden können. Das FIZ CHEMIE betreibt seit fünf Jahren sehr erfolgreich einen solchen VMWare Sphere-Cluster, der Virtualisierungsgrad der Serverinfrastruktur beträgt inzwischen über 60 %. Aufgrund der vorliegenden Erfahrungen mit diesen Systemen ist der Aufbau des Forschungsportal-Servers inklusive der Umgebung der für die Tools benötigen Systeme als unproblematisch anzusehen. Auch der Aufbau der Harvester-Datenbank ist auf einer virtualisierten Umgebung realisierbar, zumal sie nur für die Indexierung der Metadaten, welche nur einen Bruchteil der zu erwartenden Gesamtdatenmenge ausmachen, benötigt wird.

Basierend auf einem zentralen Storage-System empfiehlt sich der Aufbau eines Virtualiserungsclusters mit vier leistungsstarken Servern. Stand der Technik für diese Systeme sind Dual-Hex-Core-CPUs, min. 128 GB Arbeitsspeicher und 10 GBit FCoE-Adaptern. Vertreter dieser Systeme sind z. B. HP ProLiant DL38x oder SUN Fire X4270. Dieser Hardwareausbau gewährleistet, dass auch sehr leistungsstarke VM-Server für die Indexierung, Konvertierung und Visualisierung bei hohen IO-Werten bereitgestellt werden können. Durch den Einsatz des Virtualisierungclusters kann auf weitere High-Availability-Techniken verzichtet werden. Dies reduziert den Wartungsaufwand der gesamten Umgebung erheblich.

4.4 Zusammenfassung Hardware-Lastenheft

Die beschriebenen Anforderungen lassen sich für beide Modelle zusammenfassen und durch weitere konkrete Anforderungen ergänzen.

4.4.1 Hardware-Lastenheft für Modell I: Aufbau Langzeitarchiv im FIZ CHEMIE

Allgemeine Voraussetzungen:

- bevorzugte Hardware-Hersteller: HP oder SUN, ggf. IBM
- Hardware Linux RedHat oder Solaris (ggf. IBM AIX) kompatibel
- Management und Monitoring unter Linux oder Solaris (ggf. IBM AIX)
- Dimensionierung der Hardware-Komponenten für den Betrieb eines HSM-Speichermanagement-Systems (20 % Archivkapazität für Festplatten-Systeme, 300 % Archivkapazität für Magnetband-Systeme)

- Online-Integration neuer Speichersysteme (Festplatten- bzw. Magnetband-Systeme)
- Medientechnologiewechsel online, d. h. Online Integration eines neuen Storage und Migration der Daten im Hintergrund
- Hochverfügbarkeit der einzelnen Storage-Systeme
- Erweiterbarkeit um einen 2. Standort (asynchroner Betrieb)
- Hardware-Komponenten und HSM-Lösung aus einer Hand

Hardware für HSM-Lösung:

- 2 x 10 TB Festplattensystem mit redundantem Controller, redundanter Stromversorgung, Multipath-Unterstützung, Raid-Level 1,5,6,51,60, modulare Erweiterbarkeit, alle Komponenten Hot-Plug-fähig, Kapazität online erweiterbar
- 2 x 100 TB Bandlibrary mit 2 LTO-4 Laufwerken, erweiterbar auf 4 LTO-4 Laufwerke, 150-200 Bandstellplätze, redundanter Controller, modular, Hot-Plug-fähig, redundante Stromversorgung
- redundantes Speichernetzwerk, basierend auf 8 GBit FC-SAN Technologie, alternativ 10 GBit FCoE-Netzwerk
- 4 x Server mit Dual-Quad-Core-CPU (AMD/Intel), 128 GB Ram, 130 GB Festplatten-raid (Raid-Level 1), 2 x 8 GBit FC-Hostbus-Adapter bzw. 2 x 10 GBit FCoE-Adapter

Hardware für Forschungsdatenportal, Harvester- und Emulationsumgebung:

- 1 x 10 TB Festplattensystem mit redundantem Controller, redundanter Stromversorgung, Multipath-Unterstützung, Raid-Level 1,5,6,51,60, modulare Erweiterbarkeit, alle Komponenten Hot-Plug-fähig, Kapazität online erweiterbar
- 4 x Server mit Dual-Hex-Core-CPU (AMD/Intel), 128 GB Ram, 130 GB Festplatten-raid (Raid-Level 1), 2 x 8 GBit FC-Hostbus-Adapter bzw. 2 x 10 GBit FCoE-Adapter

4.4.2 Hardware-Lastenheft für Modell II: Aufbau Langzeitarchiv mit Kooperationspartner

Allgemeine Voraussetzungen:

- bevorzugte Hardware-Hersteller: HP oder SUN, ggf. IBM
- Hardware Linux RedHat oder Solaris (ggf. IBM AIX) kompatibel
- Management und Monitoring unter Linux oder Solaris (ggf. IBM AIX)

- Hochverfügbarkeit der einzelnen Storage-Systeme
- Hardware-Komponenten und HSM-Lösung aus einer Hand

Hardware für Forschungsdatenportal, Harvester- und Emulationsumgebung:

- redundantes Speichernetzwerk basierend auf 8 GBit FC-SAN Technologie, alternativ 10 GBit FCoE-Netzwerk
- 1 x 10 TB Festplattensystem mit redundantem Controller, redundanter Stromversorgung, Multipath-Unterstützung, Raid-Level 1,5,6, modulare Erweiterbarkeit, alle Komponenten Hot-Plug-fähig
- 4 x Server mit Dual-Hex-Core-CPU (AMD/Intel), 128 GB Ram, 130 GB Festplatten-raid (Raid-Level 1), 2 x 8 GBit FC-Hostbus-Adapter bzw. 2 x 10 GBit FCoE-Adapter

C 5. Software

Die Analyse der Software-Anforderungen für eine Langzeitarchivierungsinfrastruktur kann in zwei Hauptkomponenten aufgeteilt werden: das Langzeitarchiv und das Forschungsdatenportal.

Die Anforderungen an das Langzeitarchiv können weiter in die Schwerpunkte

- Backend-Storage-System,
- Archiv-Software und
- Emulationsumgebung

gegliedert werden.

Für das Forschungsdatenportal liegt der Fokus auf

- den Schnittstellen: DOI-Registrierung, Authentifizierung,
- den Tools für die Chemiker: Suche, Struktursuche, Eingabe-Interfaces, Datenmanagement, Visualisierung sowie auf
- den internen Tools: Konvertierung, Validierung.

5.1 Beschreibung des Backend-Storage-Systems

Für den Aufbau eines Langzeitarchivs wird ein Backend-Storage-System benötigt, welches primär drei Anforderungen erfüllt:

- **Linear skalierende Leistung mit steigendem Datenvolumen**
  Nur durch ein System, dessen Leistung mit dem Datenvolumen linear skaliert, wird es möglich sein, auch nach fünf bis zehn Jahren Betriebszeit ohne einen Wechsel des Backend-Storage-Systems akzeptable Antwortzeiten für den User garantieren zu können.
- **Rapid Desaster Recovery**
  Eine weitere Anforderung ist die Zeit, die für mögliches Desaster-Recovery benötigt wird. Normalerweise skaliert das Daten-Recovery mit dem Datenvolumen linear. Dies wäre schon nach einer Betriebszeit von fünf Jahren fatal, wenn berücksichtigt wird, dass für den Restore von 10 TByte Daten schon weit über einen Tag eingeplant werden muss. Dieses würde im Desaster-Fall zu einem inakzeptablen, mehrtägigen Ausfall der Infrastruktur führen.

- **Verwaltung infiniter Datenmengen**
  Der Aufbau einer Infrastruktur zur Langzeitarchivierung muss berücksichtigen, dass das System für eine Zeitdauer von weit über zehn Jahren betrieben wird. Da zur heutigen Zeit nicht abgeschätzt werden kann, mit welchen Datenmengen in der Zukunft gerechnet werden muss, ist die Verwendung eines Systems, welches prinzipiell unbegrenzte Datenmengen verwalten kann, sehr sinnvoll.

Wie schon in Kap. 4 zur Hardware beschrieben, bildet ein Backend-Storage-System, basierend auf einer HSM-Lösung, den heutigen Stand der Technik für die Speicherung sehr großer Datenmengen ab. Aktuelle HSM-Lösungen (IBM TSM oder SUN SAM-QFS) erfüllen diese primären Anforderungen bedenkenlos. Beide Systeme sind für einen Langzeitbetrieb ausgelegt und bieten zusätzlich Features wie die Online-Integration von neuen Storage-Komponenten, eine automatische Datenmigration sowie eine Erweiterbarkeit auf einen zweiten Standort.

Ein weiterer Vorteil einer HSM-Lösung ist der Verzicht auf eine zusätzliche Backup-Lösung, da HSM-Lösungen prinzipiell mehrere Kopien eines Datums ablegen und sich automatisch um den Erhalt der Lesbarkeit der Datenträger kümmern. Weiterhin organisieren diese HSM-Lösungen die Speicherung je nach Anforderung des einzelnen Datums nach vordefinierten Regeln automatisch. So ist es nicht nötig, durch regelmäßige Prozesse quasi nicht mehr abgerufene Daten auf preisgünstigere bzw. alternative Medien zu verschieben, da HSM-Lösungen über eine automatische Archivierung und ein automatisches Staging (Zurückholen archivierter Daten in den Onlinecache) verfügen.

5.2 Archiv-Software

Für den Aufbau des Archivs für chemische Forschungsdaten soll auf eine gängige und bewährte Archiv-Software zurückgegriffen werden. Bei der Auswahl der Software steht im Vordergrund, den planbaren Aufwand für die eigenen Entwicklungsarbeiten der benötigten Erweiterungen möglichst überschaubar zu halten. Daher soll ein System zum Einsatz kommen, welches den gewünschten, in der nachfolgenden Funktionsbetrachtung beschriebenen Features möglichst nah kommt, um sich auf die Kernaufgaben – die Erweiterungen für den chemischen Content und die Anpassungen für die Usability – konzentrieren zu können. Zudem sollte die Software international etablierte Standards berücksichtigen (s. Kap. 6.1).

Aus den grundlegenden organisatorischen Bedürfnissen eines Langzeitarchivs und der vom FIZ CHEMIE geplanten Infrastruktur ergeben sich für die benötigte Archiv-Software die folgenden Voraussetzungen:

- OAIS-Konformität
- Berücksichtigung von Langzeitarchivierungsaspekten
- erweiterbares Webinterface mit allen grundlegenden Funktionen
- OAI-PHM-Schnittstelle
- Unterstützung der benötigten Metadatenstandards
- Unterstützung von Handle-Systemen (wie DOI)
- Workflow
- Rechteverwaltung
- APIs zur Integration von Erweiterungen wie Datenvalidierung und -konvertierung
- Metadatenspeicherung in Datenbank (z. B. Oracle, PostgreSQL)
- Forschungsdatenspeicherung im Filesystem
- Open Source

Im Auftrag des nestor-Kompetenznetzwerks zur Langzeitarchivierung wurde von Prof. U. M. Borghoff am Institut für Softwaretechnologie der Universität der Bundeswehr München eine sehr umfangreiche Produktanalyse von 65 Archivierungssystemen durchgeführt. Das resultierende nestor-Dokument „Vergleich bestehender Archivierungssysteme“[84] liefert für sechs Systeme mit hoher Relevanz eine Expertise, auf deren Basis die projekteigenen Parameter abgeglichen und eine Auswahl geeigneter Systeme genannt werden können:

DIAS

DIAS ist eine kommerzielle Lösung, die von IBM[59] in Zusammenarbeit mit der Niederländischen Nationalbibliothek[85] mit dem Fokus auf Langzeitarchivierung entwickelt wurde. DIAS ist konform zu gängigen Standards wie OAIS (ISO 14721), METS sowie DC (Dublin Core) und bietet Konzepte zur Langzeitarchivierung wie das Preservation Layer Model (PLM) und Universal Virtual Computer (UVC).

Als Schwachpunkte der Lösung können das fehlende Rechtemanagement, Authentizität, Verwaltung von Urheber- und Lizenzrechten sowie die fehlenden offenen Schnittstellen für Eigenentwicklungen genannt werden. Für die nötigen Entwicklungen zur Integration der chemischen Informationen wäre somit immer die Kooperation mit der IBM notwendig.

Da die Lösung von IBM angeboten wird, ist eine Fokussierung auf IBM Hightech Hardware gegeben. Da das FIZ CHEMIE Berlin keinerlei IBM-Systeme betreibt, wäre diese Lösung mit einem Fremdkörper in der vorhandenen IT-Infrastruktur vergleichbar. Ferner ist die Idee, den Storage über einen Partner bereitzustellen, mit dieser Lösung nicht realisierbar.

Weltweit bekannte Installationen sind kopal[55] und die Niederländische Nationalbibliothek.

DigiTool / Rosetta

Die ebenfalls kommerziellen Lösungen DigiTool[67] und Rosetta[86] werden von Ex Libris angeboten. Sie sind auch an die Standards OAIS, METS sowie DC angelehnt und enthalten ein Digital Preservation System für die Langzeitarchivierung.

Der Schwachpunkt beider Lösungen sind die fehlenden offenen Schnittstellen für Eigenentwicklungen, sodass auch bei dieser Archiv-Software immer der Hersteller bei Erweiterungen einbezogen werden muss.

Technisch können die Lösungen in die bestehende Infrastruktur des FIZ CHEMIE integriert werden, da sie Hardwarehersteller-unabhängig sind und mit gängigen Betriebssystemen und Datenbanken arbeiten. Eine Auslagerung des Storages zu einem Partner ist mit DigiTool und Rosetta problemlos möglich, da die Daten im Filesystem liegen und auch NFS unterstützt wird.

Eine Referenzanwendung ist beispielsweise die British Library[87].

DSpace

DSpace[88] ist ein Open Source-out of box-Repository, dessen Entwicklung durch das MIT und HP initiiert wurde. Bis Juli 2009 koordinierte die DSpace Federation die Planung und Entwicklung, mittlerweile wurde das Projekt von DuraSpace[89] übernommen.

DSpace folgt den Standards OAIS und METS. Die Software ist in Java entwickelt worden und kann durch eigene Patches erweitert werden. Verfügbare Erweiterungen sind zum Beispiel ein Embargo-Mechanismus, eine semantische Suche oder das Dublin Core Meta Toolkit. Für die Verwendung als Langzeitarchiv müssen Mechanismen zur Digital Preservation implementiert werden.

Da DSpace hardwareunabhängig ist, kann diese Lösung analog zu DigiTools in die bestehende Infrastruktur am FIZ CHEMIE integriert werden. Da auch hier die Daten im Filesystem liegen, ist eine Storage-Auslagerung problemlos durchführbar.

Weltweit bekannte Anwender sind z. B. die Smithsonian Institution[90], Pandektis at National Research Foundation of Greece[91] sowie die Texas Digital Library[92].

EPrints

EPrints[93] ist von der Universität Southampton als OpenSource Publikationsmanagement-System entwickelt worden. EPrints orientiert sich an den Standards OAIS und METS, ist jedoch eher für kleinere Installationen ausgelegt.

EPrints wurde primär in Perl entwickelt und besitzt Schnittstellen für eigene Erweiterungen. Eine Erweiterung zur Digital Preservation ist in Arbeit.

Da EPrints auf konventionelle Standardtechniken (z. B. Apache-Webserver, MySQL-DB, etc.) ausgerichtet ist, erscheint eine Integration in die IT-Infrastruktur des FIZ CHEMIE und eine Auslagerung des Storages unproblematisch.

Eine bekannte Anwendung ist das dänisch-deutsche Gemeinschaftsprojekt Organic EPrints[94].

Fedora

Fedora[95] ist ein OpenSource-Content-Management-System und wurde ursprünglich an der Cornell University entwickelt. Inzwischen wird Fedora von DuraSpace[89] verwaltet.

Fedora ist in Java entwickelt worden. Für die Erweiterung stehen umfangreiche APIs, wie z. B. eine Web-API, zur Verfügung. Fedora ist für die Archivierung von sehr großen Datenmengen ausgelegt, besitzt eine Volltextsuche und stellt zudem eine RDF-Suche zur Verfügung. Eine Integration in die IT-Infrastruktur des FIZ CHEMIE ist problemlos möglich.

Eine weltweit bekannte Anwendung ist z. B. eSciDoc[96]. Hervorzuheben ist auch das Projekt RODA[97], welches ein OAIS-konformes Archiv, basierend auf den Standards EAD, METS und PREMIS, darstellt.

MyCoRe

Das ursprünglich von den Universitäten Essen und Bonn entwickelte MyCoRe[98] ist ein OpenSource-System zur Entwicklung von Archivanwendungen und Repositorien. Inzwischen wurde eine Geschäftsstelle in Hamburg gegründet, die die Arbeiten um MyCoRe organisiert.

Das in Java entwickelte MyCore berücksichtigt die Standards OAIS und METS, legt aber den Schwerpunkt auf Publikationen und Sammlungen.

Auch bei MyCoRe sind eine Integration in die IT-Infrastuktur des FIZ CHEMIE sowie eine Storage-Auslagerung möglich.

Eine Referenzanwendung stellt beispielsweise das Papyrus-Projekt Halle-Jena-Leipzig[99] dar.

Nach dieser Analyse zeigen sich die Software-Pakete DSpace und Fedora für den Aufbau eines Archivs für chemische Forschungsdaten als grundsätzlich geeignet. Dabei erscheint nach aktuellen Betrachtungen die Verwendung von Fedora bzw. einem Fedora-Ableger, wie z. B. eSciDoc, der erfolgversprechendere Ansatz zu sein, da die vorhandenen Fedora-APIs eine Integrationen eigener Web-Interfaces und die Anbindung von Validierern und Konvertern mit überschaubarem Aufwand ermöglicht. Auch gibt es bezüglich der Integration

von gängigen Archivsystemen (z. B. SamFS/QFS) in Fedora-Umgebungen über das Projekt eSciDoc positive Erfahrungen.

Alle Systeme besitzen von Haus aus keine Datenstrukturen, die den Bedürfnissen für chemischen Content gerecht werden. Sie müssen um die Möglichkeit der Ablage von Strukturinformationen in den Metadaten (idealerweise InChI-Strings/-Keys) und einer chemischen Struktur- bzw. Substruktur-Suche erweitert werden.

5.3 Emulationsumgebung

Aufgrund der Tatsache, dass gerade im Umfeld der Emulation und der meist damit verbundenen Virtualisierung in den nächsten Jahren sehr viele Entwicklungen zu erwarten sind, soll zu diesem Punkt an dieser Stelle nur eine mögliche Perspektive angerissen werden.

Für den Betrieb einer Emulationsumgebung ist es erforderlich, dass zeitnah eine frisch installierte, personalisierte UVC mit einer speziell benötigten Softwareumgebung bereitgestellt wird. Diese muss aus einem Pool von vielen im Laufe des Betriebs der Forschungsdateninfrastruktur entstandenen Varianten möglichst automatisch erstellt werden. Wenn die UVC vom Anwender nicht mehr benötigt wird bzw. eine bestimmte Laufzeit erreicht, muss sie von der Emulationsumgebung wieder automatisch entfernt werden.

Schon heute kann mit dem Produkt VMWare Views eine solche Funktionalität zur Verfügung gestellt werden[100].

5.4 Zusammenfassung Software-Lastenheft

Speicher-Backend-System:

- linear skalierende, hochverfügbare HSM-Lösung, z. B. IBM TSM oder SUN SAM-FS/QFS
- Rapid Desaster Recovery
- keine zusätzliche Backup-Applikation für die Sicherstellung der Datenbestände
- Online-Integration neuer Storage-Systeme und Migration der Daten im Hintergrund
- Erweiterbarkeit um einen 2. Standort (asynchroner Betrieb)
- HSM-Lösung aus gleicher Hand wie Hardware-Komponenten
- Verwaltung infiniter Datenmengen

- automatische Archivierung, automatisches Staging
- Betrieb unter Linux RedHat oder Solaris (ggf. IBM AIX)

Archivsoftware:

- OAIS-Konformität
- Berücksichtigung von Langzeitarchivierungsaspekten
- erweiterbares Webinterface mit allen grundlegenden Funktionen
- OAI-PHM-Schnittstelle
- Unterstützung der benötigten Metadatenstandards
- Unterstützung von Handle-Systemen (wie DOI)
- Workflow
- Rechteverwaltung
- APIs zur Integration von Erweiterungen wie Datenvalidierung und -konvertierung
- Metadatenspeicherung in Datenbank (z. B. Oracle, PostgreSQL)
- Forschungsdatenspeicherung in Filesystem
- Open Source

Emulationsumgebung:

- automatisches Ausrollen und Löschen von UVCs
- zentrale Verwaltung von UVCs
- UVCs Zugriff via Remote-Desktop, VNC oder PC-over-Ethernet

5.5 Forschungsdatenportal und Tools

Disziplinen wie beispielsweise die Klimaforschung oder die Geowissenschaften sind auf dem Gebiet der Archivierung von Forschungsdaten schon sehr weit entwickelt. Dennoch wurde festgestellt, dass es für den Wissenschaftler immer noch zu komplex ist, Daten in ein Repository hochzuladen[101]. Stolpersteine sind häufig die Eingabe von Metadaten und die Semantik. Eine einheitliche Benutzeroberfläche mit besseren Tools muss dementsprechend im Mittelpunkt der Entwicklung stehen, um Abhilfe für dieses Problem zu schaffen. So müssen zum Beispiel Tools bereitgestellt werden, die für den Wissenschaftler die Extraktion von geräte- und messungsspezifischen Metadaten aus seinen Messdaten automatisch übernehmen.

Für das Fach Chemie kommt erschwerend hinzu, dass die Kommunikation zwischen Chemikern über die textuelle Sprache hinaus geht, sie ist durch chemische Strukturzeichnungen, also Grafiken von Molekülen, geprägt. Damit das Forschungsdatenportal diesem Bedürfnis Rechnung tragen kann, bedarf es einer besonderen Unterstützung des Chemikers durch spezielle Tools zur Strukturerfassung und Speicherung, zum Strukturretrieval und zur Strukturvisualisierung.

Um einen Anforderungskatalog für die zu entwickelnden Tools erstellen zu können ist zunächst ein detaillierterer Blick auf die Funktionalität des Forschungsdatenportals und den Umgang des Users mit selbigem nötig.

Datennutzer

Aus der Sicht des Datennutzers liegt der Fokus des Interesses primär in dem Aufspüren von Forschungsdaten, die seine eigene Arbeit unterstützten oder bestätigen. Zu diesem Zweck benötigt er im Portal zunächst ausgefeilte Retrieval-Funktionen, die ihm die Suche nach den passenden Daten erleichtern. Die Suche selbst kann nur in den gespeicherten Metadaten erfolgen, da die Forschungsdaten selbst meist nur aus Zahlenkolonnen bestehen. Dem Datennutzer muss daher zunächst ein Volltext-Retrieval mit semantischer Unterstützung zur Verfügung stehen. Die semantische Suche sollte einerseits die grundlegenden chemischen Kenntnisse zu Verbindungen und Verbindungsklassen, aber auch zu speziellen Messmethoden und alternativen Verfahren oder Bezeichnungen beherrschen. Neben der Volltext-Suche muss dem Chemiker das Suchwerkzeug seiner Sprache, d. h. eine Struktur- und Substruktur-Suche, bereitgestellt werden. Die nach erfolgreicher Suche erhaltene Trefferliste sollte zusätzlich durch eine Facettierung einschränkbar gemacht werden. Der Vorteil dieser Faceted Search liegt darin, dass die erhaltene Trefferliste nachträglich durch zusätzliche Kriterien (z. B. Messmethode, Messgerät, o. ä.) weiter eingeschränkt werden kann. Die Trefferliste selbst liefert eine Übersicht über die gespeicherten Metadaten und einen Link zu dem eigentlichen Forschungsdatensatz. Über diesen Link wird dem Anwender eine visualisierte Fassung des Datensatzes angeboten. Die Ausgabe kann je nach Art der vorliegenden Daten entweder von einem Spektren-Viewer, einem Datenplotter oder bei unbekannten Datenformaten gegebenenfalls von einem Text-Viewer erfolgen. Zusätzlich wird dem Datennutzer eine Download-Möglichkeit für den Datensatz inklusive Metadatensatz angeboten. Zusammenfassend muss festgestellt werden, dass der Datennutzer folgende Tools benötigt:

- Volltextsuche
- semantische Suche mit Fokus auf der Chemie
- Struktur- und Substruktur-Suche

- Faceted Search unterstütze Trefferliste
- Spektren-Viewer
- Datenplotter

Datenproduzent

Für den Datenproduzenten, der das Forschungsdatenportal für die Ablage seiner Daten verwenden möchte, ist der benötigte Funktionsumfang ungleich größer.

In einem **Szenarium A** beginnt der erste Kontakt des Wissenschaftlers mit der Authentifizierung an dem System. Dies ist nötig, um nachweisen zu können, wer die Daten abgelegt hat und wer der Rechteinhaber ist. Gefolgt wird dieser Schritt von dem Zugriff auf eine Verwaltungsplattform für seine Daten. Von dort ist er in der Lage, sowohl auf alle seine Datensätze, mögliche konvertierte oder migrierte Versionen seiner Daten und die zugehörigen Metadaten, als auch auf den Status, auf die Preservation-Informationen sowie auch auf Nutzungsstatistiken zuzugreifen. Auch der Datenproduzent benötigt für die Übersicht über seine Daten eine Suchfunktion und eine Faceted Search, um bestimmte Datensätze effektiv wiederzufinden.

Des Weiteren ist aus Sicht des Datenproduzenten für die Speicherung seiner Daten ein Wizard unterstütztes Upload-Tool notwendig, welches ihm neben dem geführten Upload eines oder mehrerer Datensätze in das ihm zugehörige Institutionelle Repository bzw. in das Langzeitarchiv auch die Erfassung und Prüfung der Metadaten ermöglicht. Das Upload-Tool muss im Hintergrund weitere komplexe Aufgaben wie eine Fileformat- und Metadaten-Validierung, Daten-Plausibilitätsprüfung, Datenkonvertierung, Prüfsummenberechnung zur Sicherstellung der Authentizität der Daten und eine SIP-Paketerstellung (Submission Information Package, s. Kap. 6.1.1) erfüllen.

Das Forschungsdatenportal ermöglicht es dem Datenproduzenten schließlich, seine Daten für Veröffentlichungen mit einem DOI zu ergänzen und bei Bedarf für eine dauerhafte Verfügbarkeit in das Zentrale Repository mit Langzeitarchivierungsstrategie zu verschieben.

Neben dem beschriebenem Szenarium A, welches vollständig webbasiert zur Verfügung gestellt wird, sollte dem Anwender in einem **Szenarium B** langfristig eine lokale, plattformunabhängige Applikation (Java, idealerweise Webstart zur Vermeidung komplizierter Installationen) zur Verfügung gestellt werden. Diese Applikation stellt zunächst nur einen SIP-Paket-Composer dar, mit dessen Hilfe der Datenproduzent SIP-Pakete mit mehr als einem Datensatz erzeugen kann. Ähnlich einem Warenkorb können die einzelnen Datensätze zusammengefasst und mit Metadaten ergänzt werden. Der Vorteil des Composers liegt darin, dass Metadaten, die sonst für jeden einzelnen Datensatz erfasst werden müssen (z. B. Autor, etc.), nur einmal eingegeben werden brauchen. Auch eine

Metadatenübernahme aus zuvor erfassten Datensätzen ist möglich. Somit wird der Aufwand zur Beschreibung der Forschungsdatensätze minimiert und infolgedessen die Anwenderakzeptanz erhöht. Nach Fertigstellung des SIP-Pakets erfolgt eine erste Datenvalidierung beim Datenproduzenten, anschließend wird im Falle einer erfolgreichen Validierung das Gesamtpaket zum Institutionellen Repository bzw. Archiv übertragen. Später soll diese Applikation dem Datenproduzenten eine analoge Umgebung wie die Webbrowser-Version des Szenariums A bieten. Die Entwicklung dieser Applikation stellt nur eine verhältnismäßig geringfügige Zusatzaufgabe dar, da die Mehrzahl der benötigen Komponenten schon für das Szenarium A entwickelt worden sind. Zusammengefasst benötigt der Datenproduzent zusätzlich folgende Tools:

- Authentifizierungssystem
- Verwaltungsplattform
- Upload-Tool inklusive Metadaten-Editor, Metadaten-Validator, Fileformat-Validator, Daten-Plausibilitätsprüfung, Datenkonvertierung, Prüfsummenberechnung, SIP-Paketerstellung
- DOI-Registrierung
- Datenmigrator in Zentrales Repository
- SIP-Paket-Composer

Die benötigten Tools können in verschiedene Gruppen unterteilt werden:

- Retrieval-Tools
- Visualisierungstools
- Tools zur Daten-Organisation
- externe Schnittstellen

Für viele Aufgaben stehen bereits geeignet erscheinende Tools zur Verfügung. Diese müssen im Vorfeld eingehend auf ihre Verwendbarkeit geprüft werden, um doppelte Entwicklungen zu vermeiden.

5.5.1 Tools zur Daten-Organisation

Für den Datenproduzenten steht die Speicherung und Verwaltung seiner Forschungsdaten im Mittelpunkt des Interesses. Hierbei ist für ihn wichtig, mit möglichst geringem Aufwand seine Daten in dem System speichern und verwalten zu können.

In dem Forschungsdatenportal wird dem Datenproduzenten eine Verwaltungsoberfläche zur Verfügung gestellt, die ihn beim Daten-Upload unterstützt, die nachträgliche Bearbeitung der Metadaten ermöglicht und die Beantragung und Verwaltung von DOIs erlaubt.

Das erste Tool, das dem Datenproduzenten zur Seite gestellt werden muss, ist ein leicht bedienbares Upload-Tool mit integriertem Metadaten-Editor, in dessen Hintergrund ein Metadaten-Validator arbeitet, der die Eingabe auf Gültigkeit und Vollständigkeit prüft. Für den eigentlichen Daten-Upload werden zusätzlich ein Prüfsummen-Generator und ein Datenkonvertierer mit integriertem Format-Validator benötigt.

Mit der Akzeptanz des Web-basierten Upload-Tools steht und fällt die Akzeptanz des gesamten Forschungsdatenportals. Das zu entwickelnde User-Interface zum Daten-Upload sollte daher analog zu einem Wizard funktionieren. Es leitet den Anwender Schritt für Schritt durch die Metadateneingabe, berücksichtigt dabei, dass Teile der Metadaten abhängig von den Messverfahren sind, prüft die Plausibilität der Eingaben und gibt umfangreiche Hilfestellungen. Die Validierung der Metadaten wird während der Realisierungsphase auf eine technische Validierung gegen ein XML-Schema und für bestimmte Daten auf eine Plausibilitätsprüfung beschränkt. (Eine Realisierung des Projekts wird zeigen, an welchen Stellen und in welcher Form die Metadatenvalidierung optimiert werden kann. Dazu sollte gegen Ende der Realisierungsphase eine intensive Betrachtung der Metadaten durch das Projektteam erfolgen.)

Abschließend unterstützt der Wizard den Daten-Upload. Ist der Datensatz in das System übertragen, müssen nun automatisch eine technische Format-Prüfung und eine technische Daten-Validierung erfolgen. Die technische Format-Prüfung stellt eine sehr große Herausforderung dar. Schon das JCAMP-DX-Format bietet eine Vielzahl von Variationen in Abhängigkeit von Gerätehersteller und Messverfahren (s. Ergebnisse Arbeitspaket 3). Die Prüfung der Datenformate sowie eine gegebenenfalls nötige Konvertierung kann folglich nicht ohne Unterstützung durch externe, kommerzielle Produkte und ohne die Hilfe der Hersteller erfolgen, da eine Neuentwicklung von Validatoren und Konvertern bei der Vielzahl der Datenformate einen unverhältnismäßig hohen Aufwand darstellt und zu viele Ressourcen des Projekts binden würde. Ein mögliches Produkt für diese Aufgabe könnte zum Beispiel der ACD/Automation Server von ADC/Labs [102] sein, welcher automatisiert eine Daten-prozessierung und -konvertierung durchführen kann. Aber auch dieses Tool ist nur für Daten handelsüblicher Messgeräte einsetzbar. Für die Daten aus sehr speziellen Messgeräten oder gar Eigenentwicklungen der Datenproduzenten ist eine automatische Formatprüfung bzw. Datenvalidierung nicht möglich. Nach erfolgreicher Datenvalidierung ist der Datensatz mit einer Prüfsumme und einer digitalen Signatur zum Nachweis der Datenintegrität erforderlich. Der Upload-Prozess ist nun abgeschlossen.

Um die Durchführung des Uploads insbesondere für den Transfer größerer Datenmengen zu verbessern ist es sinnvoll, in einem zweiten Entwicklungsschritt dem Datenproduzenten eine plattform-unabhängige Applikation zur Verfügung zu stellen. Der Vorteil einer solchen Applikation ist, dass sie direkt beim Anwender ausgeführt wird und somit die Datenformat-Validierung und Prüfsummen-Generierung vor dem eigentlichen Upload durchführen kann. Dadurch können fehlerhafte Angaben und defekte Dateien vor dem eigentlichen Upload erkannt und korrigiert werden und müssen nicht mehrfach übertragen werden. Zusätzlich können die Daten direkt beim Anwender digital signiert werden, um die Authentizität sicherzustellen. Analog zum web-basierten Upload erfolgt die Metadateneingabe über einen Wizard. Nach Auswahl und Validierung der Forschungsdatendatei(en) für den Upload wird ein Daten-Paket (SIP-Paket) erzeugt, welches neben den eigentlichen Forschungsdaten auch die sie beschreibenden Metadaten in Form eines METS-Files (s. Kap. 6.1.7) enthält. Dieses SIP-Paket kann prinzipiell mindestens ein und maximal beliebig viele Datensätze enthalten. Nach dem Upload in das System erfolgt eine Prüfung der Prüfsummen und digitalen Signaturen. Ist der Nachweis für die Datenintegrität erfolgt, ist der Upload-Prozess abgeschlossen.

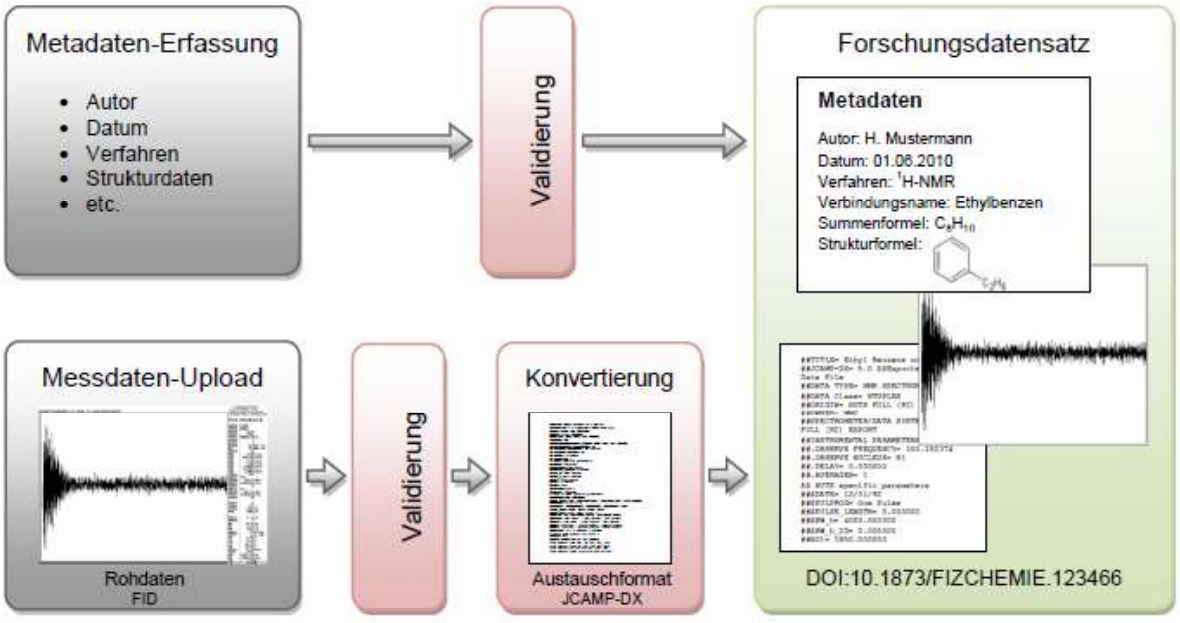


Abb. 28: Prozesse beim Upload der Rohdaten einer ^{1}H-NMR-Messung von Ethylbenzen

Nach erfolgreichem Abschluss des Upload-Prozesses steht dem Datenproduzenten eine Übersicht über die eigenen Forschungsdaten in einer Verwaltungsoberfläche zur Verfügung. Diese kann er nutzen, um für den hochgeladenen Datensatz einen DOI und gegebenenfalls Sub-DOIs zu beantragen, aber auch, um die eingegebenen Metadaten zu korrigieren oder

im Bedarfsfall die Datensätze mit alternativen Darstellungsformaten (z. B. JCAMP-DX) zu ergänzen. Auch hat er dort Zugang zu Statistiken, wie oft seine Daten abgerufen wurden, und zu administrativen Informationen, ob und wie seine Daten im Rahmen von Preservation-Maßnahmen konvertiert oder bearbeitet wurden inklusive einer zugehörigen History.

Zusammenfassung der verfügbaren Tools:

- Fileformat-Validator
- Messdaten-File-Konvertierer

Zusammenfassung der Entwicklungsaufgaben:

- Verwaltungsoberfläche für Datennutzer
- Web-basierendes Upload-Tool
- Metadaten-Editor, -Validator
- Tools zur Datenintegrität, d. h. Prüfsummen und digitale Signatur
- Upload-Applikation, die alle Tools zum Daten-Upload vereinigt

5.5.2 Retrieval-Tools

Das Retrieval stellt den zentralen Arbeitsschritt zur Daten-Akquise sowohl für den Datenproduzenten als auch den Datennutzer dar. Dementsprechend muss der Entwicklung geeigneter Retrieval-Tools eine primäre Rolle zufallen.

Die textorientierte Suche bildet die erste Komponente der benötigten Retrieval-Möglichkeiten. Sie umfasst zunächst die reine Volltextsuche über die gespeicherten Metadaten. Die auf diesem Wege erhaltenen Ergebnislisten zeigen nur exakte Treffer. Sie leiden aber häufig unter dem Problem, entweder viel zu lang zu sein, weil die Suche zu unspezifisch ist, oder zu kurz zu sein, weil die Suche zu spezifisch ist. Zu umfangreiche Trefferlisten lassen sich durch spezifischere Suchen oder durch die Anwendung von Facettierungen reduzieren. Die Erweiterung zu kurzer Trefferlisten ist nur durch eine Erweiterung mit ähnlichen Treffern möglich. Diese Erweiterung lässt sich durch semantische Werkzeuge erreichen. So ist es möglich, mittels Thesauri oder Ontologien z. B. ähnliche Verbindungen oder Messverfahren zu identifizieren und diese als erweiterte Begriffe in die Suche zu integrieren. Durch solche semantischen Verfahren ist eine Erweiterung der Ergebnisliste auf ähnliche Treffer möglich. Für die Bereitstellung der textuellen Suche liegt

somit der Fokus auf der Entwicklung einer Volltextsuche mit semantischer Erweiterung, also auf der Integration geeigneter Thesauri und Ontologien, wie z. B. des Mesh-Thesaurus[103] von der U.S. National Library of Medicine und der RSC Ontologies[104] von der Royal Society of Chemistry.

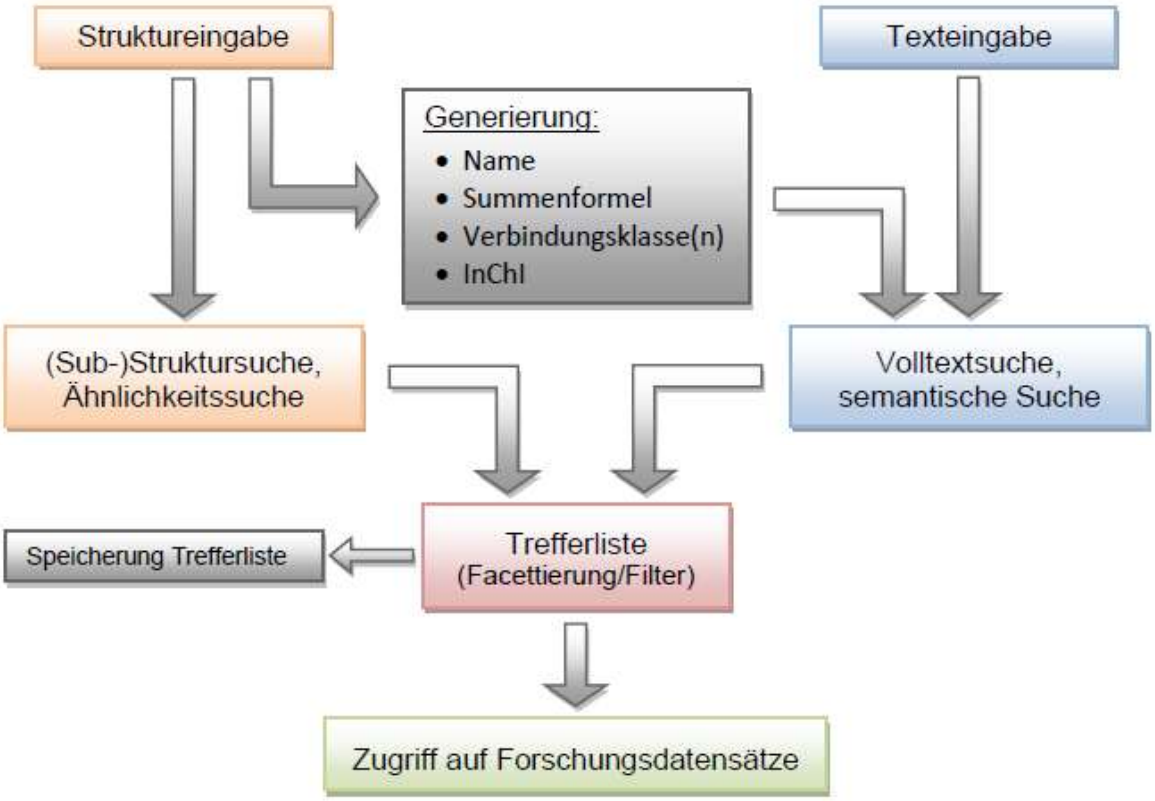


Abb. 29: Vielfalt an Retrieval-Tools für chemische Forschungsdaten

Neben der textorientierten Suche ist die grafische Suche, d. h. die Struktursuche, gesondert hervorzuheben. Für eine Struktursuche ist zunächst die Eingabe einer chemischen Struktur mittels eines Struktur-Editors erforderlich. In den letzten Jahren sind einige OpenSource-Struktur-Editoren veröffentlicht worden. Das Java-Applet JChemPaint [105] ist einer der bekanntesten und verbreitetsten Vertreter. JChemPaint bietet alle Features, die zur Erfassung auch komplexer Verbindungen inklusive stereochemischer Aspekte notwendig sind, die Bedienung ist für den Chemiker intuitiv. Der Datenaustausch von und zu JChemPaint erfolgt in den Standard-Formaten SMILES, Molfile oder CML (Chemical Markup Language).

Die erfasste Struktur kann nun auf zwei Wegen für eine Suche verwendet werden. Der erste Weg ist eine grafenorientierte Struktur- bzw. Substruktur-Suche. Sie entspricht der klassischen chemischen Struktur-Suche, wie sie seit vielen Jahren bei großen Anbietern

chemischer Datenbanken Standard ist. Die Bereitstellung einer (Sub-)Struktur-Suche ist für das Forschungsdatenprojekt unerlässlich. Die Entwicklung einer eigenen grafenorientieren (Sub-) Struktur-Suche wäre jedoch unverhältnismäßig, zumal geeignete OpenSource-Tools wie OrChem[106] zur Verfügung stehen. OrChem liefert alle nötigen Features zur chemischen Struktur-, Substruktur- und auch Ähnlichkeitssuche. Der zweite Weg ist eine textuelle Suche. Hierfür wird aus der Struktur zunächst ihr Verbindungsname, ihre Summenformel und der zugehörige InChI-String generiert. Mit Hilfe spezieller Algorithmen zur Erkennung funktioneller Gruppen, wie sie z. B. in checkmol[107] Verwendung finden, können des Weiteren der Struktur auch Verbindungsklassen zugeordnet werden. Über diese textuellen Bezeichner der eingegebenen Struktur kann nun wie oben beschrieben zusätzlich eine Volltextsuche bzw. eine semantische Suche durchgeführt werden. Für die grafenorientierte Struktursuche und einer damit verknüpfbaren Volltextsuche ist somit die Integration eines Struktur-Editors und grafenorientierter Suchwerkzeuge in das Portal notwendig. Zusätzliche Tools zur Generierung von Verbindungsnamen und Verbindungsklassen ergänzen die Entwicklungsaufgabe.

Die aus der einzelnen Suche bzw. der Kombination von grafischer und textueller Suche erhalte Trefferliste kann mitunter sehr lang und unübersichtlich werden. Daher ist es unerlässlich, die Trefferliste mittels einer Facettierung einschränkbar zu gestalten. Die Facettierung sollte im Minimum den Zeitraum der Messung und die Messtechnik umfassen. Mit zunehmender Trefferzahl sollten weitere Facetten hinzukommen. Aus technischer Sicht bietet die Facettierung den Vorteil, dass für die weitere Einschränkung der Trefferliste keine neue Suche über den gesamten Datenbestand benötigt wird, sondern nur eine Suche über den deutlich geringeren Datenbestand der Trefferliste erfolgen muss.

Zusätzlich sollte der Anwender die Möglichkeit bekommen, die erhaltenen Suchen und Trefferlisten speichern zu können. Die Abspeicherung der Suche ermöglicht dem Anwender zu einer späteren Zeit, selektiv neu hinzugekommene, auf die Suche passende Datensätze zu identifizieren und zu begutachten. Die Speicherung der Trefferliste kann dem Anwender dazu dienen, in Ruhe Datensatz für Datensatz abzurufen und die Verwendbarkeit für seine eigene Arbeit zu beurteilen.

<u>Zusammenfassung der verfügbaren OpenSource-Tools:</u>

- Struktur-Editor => JChemPaint
- Struktur-Suche => OrChem
- Erkennung funktioneller Gruppen => checkmol
- semantische Tools => Mesh-Thesaurus / RSC Ontologies

Zusammenfassung der Entwicklungsaufgaben:

- Volltextsuche
- semantische Suche, Integration verfügbarer semantischer Tools
- Integration von Struktur-Editor, Struktur-Suche
- Trefferliste mit Facettierung und Speicheroption

5.5.3 Visualisierungstools

Mit Erhalt einer Trefferliste rückt für den Datennutzer die Analyse der erhaltenen Treffer in den Vordergrund. Mit der Auswahl eines Datensatzes erhält der Anwender eine Übersichtsseite mit allen Metadaten, die den Datensatz, dessen Herkunft und Historie beschreiben. Sofern bei der Ablage der Daten ein Preview erzeugt wurde, wird dieser zusätzlich zur Verfügung gestellt. Die Bereitstellung eines Previews ist abhängig von der Art der gespeicherten Messdaten. Bei der Betrachtung von chemischen Forschungsdaten sind zwei Gruppen von Daten zu unterscheiden:

Die erste Gruppe stammt aus technischen Quellen, die die Daten in einem bekannten, wohldefinierten Datenformat speichern, bzw. Konverter zur Verfügung stellen, um die Daten in ein solches Datenformat überführen zu können. Zu dieser Gruppe gehören zum Beispiel NMR-, IR- und UV-Spektren. Die Daten dieser Beispiele werden oft schon von dem jeweiligen Messgerät als JCAMP-DX-Files abgespeichert. Dieses Format lässt sich mittels JSpecView[108] direkt im Browser visualisieren und dem Anwender bereitstellen.

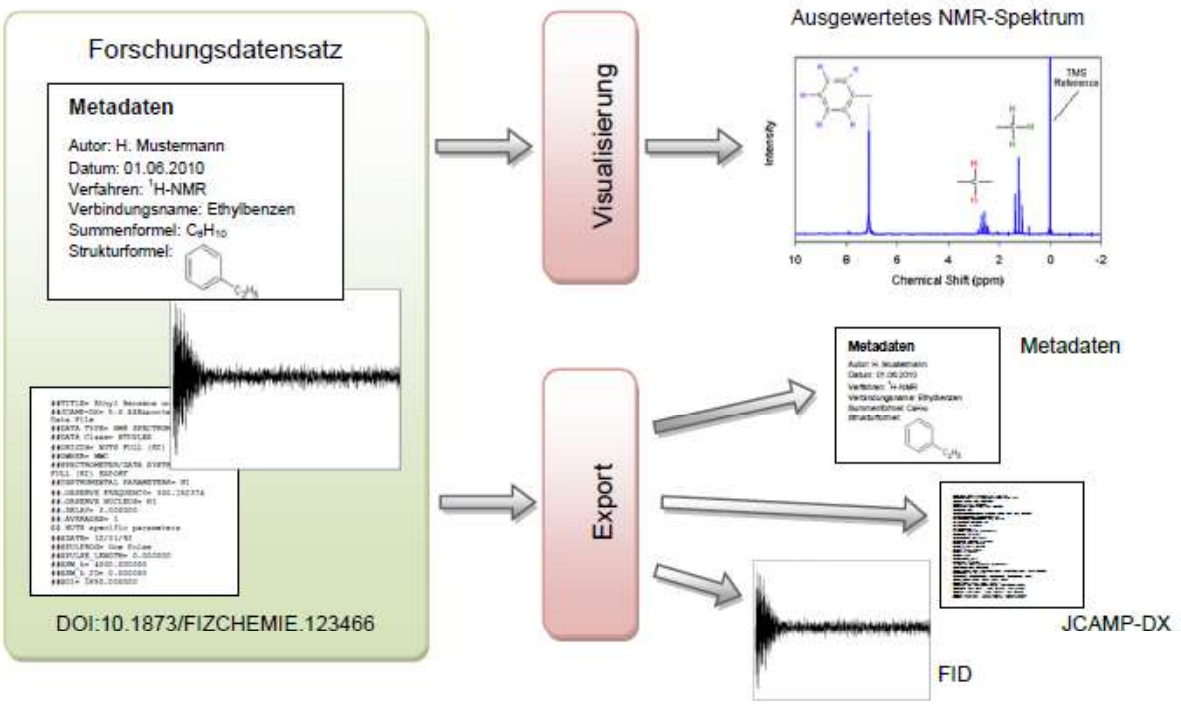


Abb. 30: Ausgabe eines archivierten Datenpakets mit Visualisierung des ^{1}H-NMR-Spektrums und Export von Rohdaten, Austauschformat und Metadaten

Die Daten der zweiten Gruppe sind in unbekannten Formaten gespeichert. Dies können z. B. einfache Textfiles mit tabellarischen gespeicherten Daten sein, wo jedoch die Bedeutung der einzelnen Zahlenkolonnen nicht bekannt oder dokumentiert ist. Diese Daten stammen häufig aus speziellen Messgeräten, die möglicherweise der publizierende Wissenschaftler selbst entwickelt hat. Viele dieser Daten können daher nicht automatisch visualisiert werden. Für diesen Fall stehen nur zwei Möglichkeiten zur Verfügung: Die Ausgabe über einen allgemeinen Daten-Plotter wie ImageJ [109], in dem der Datennutzer gegebenenfalls die Darstellung und die Skalierungen verändern kann, oder die rein textuelle Ausgabe.

Die Visualisierung der Forschungsdaten ist demzufolge ein weiterer Entwicklungsschwerpunkt des Projekts. Es sind Tools zu entwickeln, die die verfügbaren Spektrenviewer und Datenplotter im Forschungsdatenportal zur Verfügung stellen. Analog zum Datenimport in das Portal liegt die Herausforderung dieser Aufgabe darin, dass eine Vielzahl von unterschiedlichsten Datentypen verlässlich erkannt und passend visualisiert werden müssen.

Neben der Visualisierung der Forschungsdaten ist es für den Datennutzer wichtig, dass er die Daten aus dem Portal herunterladen kann. Er erhält somit die Möglichkeit, selber alternative Datenauswertungen durchzuführen oder die Messdaten als Basis von Simulationen zu verwenden. Der Export eines Datensatzes muss alle relevanten Metadaten zur Nachnutzung enthalten. Die Tools zum Datenexport müssen dem Datennutzer neben

dem Download der Messdaten-Dateien auch den Download der Metadaten bzw. idealerweise eines Gesamtdatenpakets ermöglichen, in dem Metadaten und Messdaten enthalten sind.

Zusammenfassung der verfügbaren OpenSource-Tools:

- Spektren-Viewer => JSpecView
- Daten-Plotter => ImageJ und andere

Zusammenfassung der Entwicklungsaufgaben:

- Datenausgabe / -visualisierung
- Datenexport

5.5.4 Externe Schnittstellen

Für den Betrieb eines Forschungsdatenportals ist die Ansteuerung mehrerer externer Infrastrukturen erforderlich. Für diesen Zweck ist die Entwicklung von Schnittstellen bzw. Tools für das Ansteuern externer Schnittstellen notwendig.

User-Authentifizierung

Um die Voraussetzung der Authentizität der Datenquelle (s. Kap. 6.2.4) erfüllen zu können, ist es erforderlich, dass sich jeder Datenproduzent an dem Forschungsdatenportal identifiziert. Dies kann zum Beispiel durch eine Nutzerdatenbank in dem Portal erfolgen. Es wäre jedoch zur Erfüllung der Authentizitätskriterien in einem zweiten Schritt notwendig, die Nutzerangaben auf Korrektheit zu prüfen, da es nicht passieren darf, dass sich ein „Nutzer“ unter falschen Namen anmeldet und das System unkontrolliert mit fehlerhaften Inhalten verunreinigt. Dieser Ansatz ist möglich, erfordert aber den Einsatz von Personalkapazitäten. Auch die Akzeptanz beim Datenproduzenten würde leiden, da er einerseits warten muss, bis sein Account freigegeben wurde und er sich andererseits wieder einmal eine neue Benutzerkennung und Password merken muss.

Eine Alternative stellen zentrale Identity-Management-Systeme dar, die heutzutage schon in vielen Universitäten und Forschungseinrichtungen betrieben werden. Über diese zentralen Einrichtungen wird auch ein institutsinternes Accounting bereitgestellt, mit dem sich die Institutsmitarbeiter an der Infrastruktur der eigenen Einrichtung anmelden können. Viele dieser Systeme sind bereits über das Projekt DFN-AAI[110] miteinander vernetzt, sodass es dem Forscher möglich ist, basierend auf einem einheitlichen Benutzernamen und seinem

Password über sein Institut seine Identität zu übermitteln. DFN-AAI wird seit einigen Jahren vom Deutschen Forschungsnetz (DFN) entwickelt und aufgebaut. Den DFN-AAI-Dienst nutzen inzwischen auch eine große Zahl der wissenschaftlichen Verlage (z. B. De Gruyter, Elsevier und Springer). Technisch erfolgt die Authentifizierung mittels Shibboleth[111]. Die nötigen Tools inklusive dem technischem Support werden vom DFN zur Verfügung gestellt. Es muss somit nur eine Integration dieser Tools in das Forschungsdatenportal durchgeführt werden.

DOI-Registierung

Zur Bereitstellung eines dauerhaften Identifiers ist die Registrierung eines DOI-Namen[112] bei der TIB Hannover bzw. DataCite[54] (s. Kap. A 4.2.1) erforderlich. Da die TIB Projektpartner ist, stehen dem Projekt alle nötigen Tools inklusive des zugehörigen Erfahrungsschatzes zur Verfügung, so dass sich die Entwicklung der DOI-Registrierungskomponente im Wesentlichen auf die Integration der zur Verfügung stehenden Tools in das Forschungsdatenportal und das nötige Mapping der Metadaten konzentriert. Das Metadaten-Mapping stellt im Fall von chemischen Daten eine besondere Herausforderung dar, da die Beschreibung eines Datensatzes häufig Strukturdaten oder gar Reaktionsdaten enthält. Die Aufgabe liegt darin, diese nichttextuellen chemischen Inhalte möglichst vollständig in die textuelle Beschreibung des Datensatz zu integrieren, z. B. mittels InChI-Codes. Auch Geräte- und Messparameter müssen in die beschreibenden Metadaten einfließen. Der Aufwand des Metadaten-Mappings ist erforderlich, da der Forschungsdaten-Suchindex der TIB bzw. DataCite eine der zentralen Anlaufstellen des Datennutzers ist. Die beschreibenden Metadaten sind demnach so zu kapseln, dass sie in die Datenbank von der TIB bzw. DataCite integriert und suchbar gemacht werden können.

OAI-PMH-Schnittstelle

Die Bereitstellung einer OAI-PMH-Schnittstelle für die projekteigenen und externen Harvester kann durch die Verwendung einer geeigneten Repository-Software, wie sie in Kap. 5.2 vorgestellt wurde, auf die Definition eines Metadaten-Mappings reduziert werden. In diesem Metadaten-Mapping ist festzulegen, welche Metadaten bei eingehenden Anfragen geliefert werden sollen. Da ein analoges Metadaten-Mapping schon für die DOI-Registrierung erfolgt, ist es sinnvoll, dieses im Wesentlichen zu übernehmen.

<u>Zusammenfassung der benötigten externen Schnittstellen:</u>

- Authentifizierung Shibboleth
- DOI-Registrierung

Zusammenfassung der bereitgestellten Schnittstellen:

- OAI-PMH-Schnittstelle

Zusammenfassung der Entwicklungsaufgaben:

- Integration DFN-AAI Schnittstelle
- Entwicklung der DOI-Registrierung mit Metadaten-Mapping
- Metadaten-Mapping für OAI-PMH-Export

C 6. Qualitätssicherung

Die Basis für eine Datenpublikation ist die Bereitstellung von Forschungsdaten durch den Wissenschaftler einschließlich einer hinreichenden Beschreibung durch Metadaten (1). Zudem muss eine an die Bedürfnisse der Wissenschaftler angepasste, digitale Infrastruktur zur Verfügung gestellt werden, die nicht nur die Ablage und Bereitstellung von Forschungsdaten ermöglicht, sondern auch deren Langzeitarchivierung und Nachnutzbarkeit sichert (2).

Abb. 31: Voraussetzungen für die Publikation von Forschungsdaten

Die Hardware-Anforderungen einer für einen dauerhaften Bitstream-Erhalt geeigneten Infrastruktur sind ausführlich in Deliverable AP2.1.1 erörtert worden. In weitergehenden Betrachtungen sind bereits viele Aspekte zur Sicherstellung der Nachnutzbarkeit der Forschungsdaten analysiert worden. Dies sind u. a. Maßnahmen gegen die Alterung von Dateiformaten wie Migration oder Emulation. Die Definition der Software-Anforderungen in Deliverable AP2.1.2 berücksichtigt nicht nur eine geeignete Archiv-Software, sondern beinhaltet ebenso die Planung komfortabler Werkzeuge für die Wissenschaftler zur Datenablage und -bereitstellung.

Die Langzeitarchivierung digitaler Objekte setzt aber auch eine Organisation der Datenhaltung im Archiv sowie die Entwicklung einer Langzeitarchivierungsstrategie voraus.

Alle Maßnahmen und Prozesse müssen ausreichend dokumentiert und beschrieben werden. Für diesen Zweck ist der ISO-Standard 14721:2003 OAIS entwickelt worden, auf den in Kap. 6.1.1 eingegangen wird (3).

Die dokumentierte Organisation des Archivs ist aber auch gleichzeitig ein wichtiges Qualitätsmerkmal. Weitere, im Folgenden beschriebene Maßnahmen sind vorzunehmen, um die Qualität und Vertrauenswürdigkeit sowohl für das Archiv als auch für die darin enthaltenen Daten sicherzustellen (5).

Über interoperable Schnittstellen wird ein persistenter Zugang zu den qualitätsgesicherten Forschungsdaten ermöglicht (4).

In den nachfolgenden Kapiteln sollen Möglichkeiten für eine Qualitätssicherung des konzipierten Langzeitarchivs erarbeitet werden. Im Mittelpunkt stehen dabei folgende Aspekte:

- Standards und Interoperabilität
- Organisation der Datenhaltung
- Datenauswahlkriterien
- Policies
- Technische Qualitätskontrolle
- Inhaltliche Qualitätskontrolle
- Informationssicherheit und Vertrauenswürdigkeit
- Qualitätsnachweis durch Zertifizierung

6.1 Standards und Formate

Für den Erfolg der digitalen Langzeitarchivierung bilden Standards eine unabdingbare Voraussetzung für kompatible und interoperable Systeme aller Art. Standards werden sowohl für organisatorische als auch für technische Aspekte der digitalen Langzeitarchivierung benötigt. Sie fördern nicht nur die Austauschbarkeit von Komponenten, sondern gewähren auch verlässliche Vorgaben für die Produktentwickler.

Für die Konzipierung des Archivs für chemische Forschungsdaten sollen etablierte und international anerkannte Standards berücksichtigt werden. Diese erhöhen nicht nur die Wahrscheinlichkeit für eine Akzeptanz und erfolgreiche Etablierung des Archivs in den wissenschaftlichen Betrieb, sondern dienen ebenso als Basis für eine Qualitätssicherung und Zertifizierung. Nur durch die Verwendung von bereits etablierten Standards ist es möglich,

das Archiv in nationale und internationale Netzwerke interoperabel einzubinden sowie auch interdisziplinäre Grenzen zu überschreiten.

6.1.1 OAIS-Referenzmodell

Das Referenzmodell OAIS-RM (Open Archival Information System)[113] wurde seit 1998 von der NASA gemeinsam mit rund 50 Organisationen aus der Luft- und Raumfahrtindustrie sowie 150 Nationalarchiven und -bibliotheken entwickelt und im Jahr 2003 als internationaler Standard ISO 14721 verabschiedet. Es stellt heute das Referenzmodell im Bereich der digitalen Archivierung dar, es gibt keine anderen Modelle mit einem ähnlichen Anspruch und einer vergleichbaren weltweiten Akzeptanz.

OAIS ist ein rein funktionales und generisches Modell. Es macht keine Vorgaben zur technischen Implementierung und ist nicht auf bestimmte Systemarchitekturen oder Datenformate beschränkt. Das Referenzmodell versteht sich als gültig für jede Organisation, die langfristig Daten aufbewahren und die im Standard beschriebenen Verantwortlichkeiten für die Langzeitarchivierung von Informationsobjekten für eine bestimmte Nutzergruppe übernimmt. Dabei ist OAIS neutral gegenüber unterschiedlichen Archivierungstechniken und ermöglicht aufgrund seiner Containerstruktur eine dezentrale Implementierung.

Das Referenzmodell OAIS beschreibt ein Archiv als Organisation, in dem Menschen und Systeme mit der Aufgabenstellung zusammenwirken, Informationen zu erhalten und einer definierten Nutzerschaft verfügbar zu machen. OAIS identifiziert die zentralen Funktionen und Abläufe eines Archivsystems und legt fest, wie die elektronische Information in das Archivsystem gelangt, welche Bearbeitungsschritte für die Archivierung notwendig sind und wie auf die gespeicherte Information zurückgegriffen werden kann.

Die Umgebung eines OAIS-Archivs besteht aus Datenproduzent, Datenmanager und Datennutzer.

Es wird zwischen drei so genannten Informationsobjekten unterschieden, die sich aufeinander beziehen:

- Submission Information Packages (SIP) sind die digitalen Ressourcen, die die aufbewahrenden Institutionen übernehmen.
- Archival Information Packages (AIP) sind SIPs, die vom Archiv durch Metainformationen ergänzt wurden. In Form von AIPs werden die digitalen Ressourcen langfristig aufbewahrt.

- Dissemination Information Packages (DIP) heißen die digitalen Ressourcen, die entsprechend den Bedürfnissen bestimmter Nutzergruppen generiert und zielgruppenorientiert zur Verfügung gestellt werden.

Es werden Prozesse für die sechs Hauptfunktionen Ingest, Archival Storage, Data Management, Administration, Preservation Planning und Access definiert.

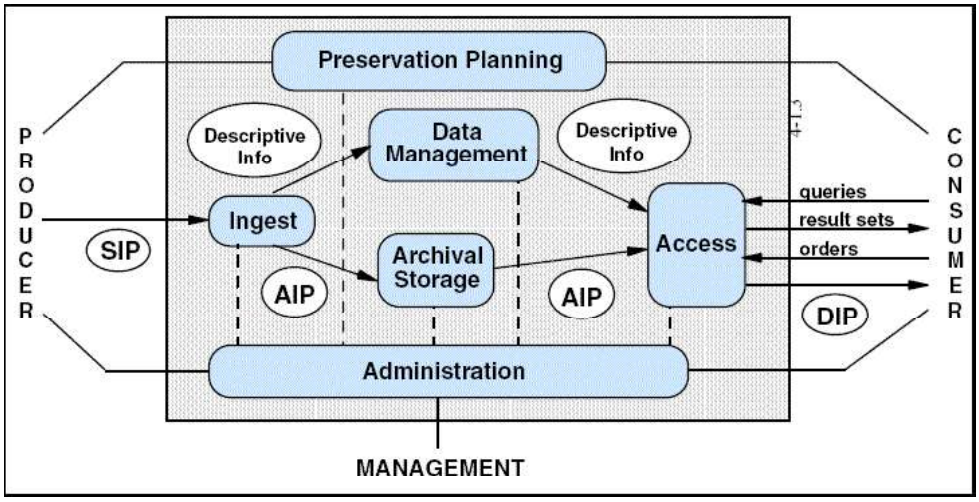


Abb. 32: Funktionseinheiten des OAIS-Referenzmodells

- **Ingest** (Datenempfang): stellt die Dienste und Funktionen zur Verfügung, um die von den Datenproduzenten übermittelten Submission Information Packages (SIPs) entgegenzunehmen und für die Langzeitarchivierung als Archival Information Packages (AIPs) aufzubereiten.
- **Archival Storage** (Archivierung): stellt Dienste und Funktionen bereit, um die AIPs physisch auf Datenträgern zu speichern, zu pflegen und von dort wieder zu lesen.
- **Data Management** (Datenverwaltung): umfasst alle Dienste und Funktionen, die benötigt werden, um die in Archival Storage gespeicherten AIPs technisch und inhaltlich zu identifizieren, zu beschreiben und langfristig zu verwalten.
- **Administration** (Verwaltung): beinhaltet die Bereitstellung und operative Leitung aller administrativen, planerischen und organisatorischen Prozesse, um das Archiv als Gesamtsystem fachlich und technisch nachhaltig zu betreiben und strategische Vorgaben (Management) umzusetzen.
- **Preservation Planning** (Konservierungsstrategie): stellt Dienstleistungen und Funktionen für die Entwicklung von Methoden, Standards und Strategien für die Langzeitarchivierung zur Verfügung, um externe Veränderungen zu antizipieren und

es der Administration zu ermöglichen, die Langzeitarchivierung der im OAIS archivierten Informationen kontinuierlich zu garantieren.

- **Access** (Zugang): liefert Dienste und Funktionen, welche es dem Endnutzer erlauben, durch Abfragen die Existenz, Beschreibung und Verfügbarkeit von Archivdaten festzustellen und diese als Dissemination Information Packages (DIPs) anzufordern und in bedarfsgerechter Form einzusehen.

Ein OAIS-konformes Archiv bietet eine Reihe von Vorteilen:

- Das OAIS-RM trägt zur Rationalisierung der digitalen Archivierung bei, indem es eine international standardisierte Kodifizierung der Verantwortlichkeiten und Voraussetzungen eines digitalen Langzeitarchivs liefert. Es legt hinsichtlich Archivplanung einen definierten Startpunkt fest und hilft, Fehlplanungen zu vermeiden.
- Es führt eine gemeinsame Terminologie zur Beschreibung und Vergleichbarkeit von digitalen Archivlösungen ein. Dies fördert die effiziente Kommunikation zwischen Nutzern, Betreibern und Herstellern.
- OAIS ermöglicht die Implementierung von standardisierten und zertifizierten Lösungen, was die Vertrauenswürdigkeit und rechtliche Absicherung fördert.
- Es vereinfacht die Anforderungsanalyse und -definition in Projekten und eignet sich als Grundlage für die Erstellung von Pflichtenheften.
- Auf Basis dieses Modells sind Kriterienkataloge entstanden, die der Überprüfung der Konformität einer Archivlösung mit ISO 14721 und als Grundlage einer Archiv-Zertifizierung dienen sollen (s. Kap. 6.2).
- Anwendungsbeispiele für OAIS-konforme Langzeitarchive in Deutschland sind das aus dem Projekt kopal[55] hervorgegangene, gleichnamige System, das Bibliothekarische Archivierungs- und Bereitstellungssystem[63] der Bayerischen Staatsbibliothek (BABS) sowie das Digitale Archiv[70].

PAIMAS

Als Ergänzung zum OAIS-RM wurde 2006 der internationale Standard ISO 20652 PAIMAS[114] (Producer Archive Interface – Methodology Abstract Standard) verabschiedet, der das Vorgehen bei der Spezifikation von Schnittstellen zwischen Produzentensystemen und OAIS-konformen digitalen Archivsystemen standardisiert. Dieser Standard wurde entwickelt, weil die Beziehung zwischen Datenproduktion und Archivbetrieb aufgrund ihrer Komplexität definierter Regelungen bedarf. Probleme können z. B. auftreten aufgrund von

- Nichtkonformität der Archivierungsdaten,
- Erkennung von Fehlern in archivierten Daten und
- steigender Diversität der Datenproduzenten und wachsender Datenkomplexität.

So muss im Allgemeinen die Aufnahme von Informationen in ein Langzeitarchiv durch Gesetze und Übernahmevereinbarungen geregelt werden. Dabei sind auch rechtliche Aspekte zu berücksichtigen, weil mit der Datenannahme gleichzeitig eine Übertragung bestimmter Rechte und Pflichten auf den Archivbetreiber erfolgt. Alle Parameter des Datentransferprozesses sowie der Aufbewahrung der Daten im Archiv sind festzulegen.[115]

=> Konzeptstudie: OAIS-konformes Langzeitarchiv

Der Aufbau und die Organisation des zu konzipierenden Archivs für chemische Forschungsdaten sollen auf dem OAIS-Referenzmodell basieren. Dabei sollten, wie in dem an der Niederländischen Nationalbibliothek[56] entwickelten und vom OAIS-RM abgeleiteten Prozessmodell DSEP-PM [116] (Process Model for a Deposit System for Electronic Publications) bereits praktiziert, zusätzlich die Prozesse Delivery & Capture und Packaging & Delivery berücksichtigt werden. Weiterhin werden in Analogie zum Projekt KoLaWiss[117] auch die Datenproduktion und die Datennutzung als weitere Prozesse in das Modell aufgenommen. Die Betrachtung dieser vor- und nachgeschalteten Prozesse ist wichtig, da die Bedürfnisse von Datenproduzenten und Datennutzern mit in die Archivplanung einfließen müssen, um z. B. deren Motivation für ihre Mitwirkung an der Langzeitarchivierung von Forschungsdaten zu steigern. Gerade bei der Konzipierung eines Archivs, welches für diverse Universitäten und unterschiedlichste Arbeitsgruppen Daten speichern soll, muss die Datenproduktion einer besonderen Betrachtung unterzogen werden, um verschiedene Bedürfnisse bedienen zu können. Während hinsichtlich der Datennutzung auch spätere Verbesserungen problemlos möglich sind, sollten Datenproduktion und Ingest möglichst frühzeitig optimiert werden.

Die Tabelle 1 listet zusammenfassend die wichtigsten Subprozesse der einzelnen Funktionseinheiten in einem Langzeitarchiv für chemische Forschungsdaten auf. Viele der genannten Subprozesse sind in dem vorliegenden Dokument bereits bearbeitet bzw. angerissen worden.

Top Level Prozesse	Wichtige Subprozesse	Bearbeitung
Datenproduktion	Datenauswahl entsprechend Auswahlkriterien	Kap. 6.2.1
	Formatempfehlungen	
	inhaltliche Qualitätskontrolle	Kap. 6.2.3
	Tools für Wissenschaftler: Authentifizierung, Verwaltungsplattform, Daten-Upload-Tool	Kap. 5.5
	automatische und manuelle Generierung beschreibender und technischer Metadaten	Kap. 5.5, 6.1.6
Delivery & Capture	Datenübernahmevereinbarung	Kap. 6.2
	Annahme von Daten und Metadaten	Kap. 5.5
	technische Qualitäts-, Integritäts- und Authentizitätskontrollen	Kap. 5.5
	Konvertierung, Validierung	Kap. 5.5.3
	Generierung von administrativen und preservation Metadaten	Kap. 6.1.6
	Packen von SIPs mit Composer-Tool	Kap. 5.5.2
Ingest	Annahme der SIPs	Kap. 5.5
	Formatnormalisierung	Kap. 5.5
	DOI-Registrierung	Kap. 5.5.6
Data Management	Einstellen, Verändern und Löschen von Daten	Kap. 5.5.2
	Bearbeitung von Suchanfragen	Kap. 5.5.4
	Berichte und Statistiken	Kap. 5.5.2
Archival Storage	Handhabung der AIPs	Kap. 4
	Bitstream Preservation	
	Speichermedienerneuerung	
	Desaster Recovery	
	Backup	

Administration	Datenübergabevereinbarungen Archivstandards und -policies Warten von Hard- und Softwaresystemen Optimierung von Archivprozessen Migrationen bzw. Updates der Archivinhalte, Standards und Policies Support	Kap. 6.2 Kap. 6.2
Preservation Planning	Beobachtung und Bewertung der Archivinhalte Empfehlungen für Standards und Policies Migrations- und Emulationspläne	 Kap. 3.2
Access	Authentifizierungssystem Nutzerverwaltung Zugrifffsgestaltung Erstellen und Ausliefern von DIPs Nutzersupport	Kap. 5.5.6 Kap. 5.5.6 Kap. 5.5.3
Packaging & Delivery	Aufbereitung und Übergabe der Archivinhalte an Datennutzer	Kap. 5.5.3
Datennutzung	Werkzeuge für Wissenschaftler, wie Suchtools Visualisierung von Daten	Kap. 5.5.4 Kap. 5.5.5

Tabelle 1: Funktionseinheiten und wichtige Subprozesse im OAIS-Referenzmodell

Das in Kap. 4.2 beschriebene Modell II, bei dem das Archiv in Kooperation mit einem Archivdienstleister aufgebaut wird, stellt eine dezentrale Implementierung des OAIS-RM dar, bei der die Funktionseinheit Archival Storage zum Kooperationspartner ausgelagert wird. Ein professionelles HSM-System beinhaltet bereits Lösungen für alle Subprozesse des Archival Storage, wie Bitstream Preservation, Speichermedienerneuerung und Desaster Recovery.

Wesentliche Funktionsmodule des OAIS lassen sich auch weitgehend problemlos auf die Struktur eines Institutionellen Repositories abbilden, ausschließlich das Preservation Planning kann für ein Repository innerhalb der vernetzten Infrastruktur (s. Kap. 2.5) unberücksichtigt bleiben. Ein Institutionelles Repository muss ebenfalls organisiert,

dokumentiert und durch Policies beschrieben werden, um eine Zertifizierung zum Nachweis seines qualitätsgesicherten Betriebs zu erhalten. Die Abb. 33 zeigt ein OAIS-konformes Modell eines kooperativ arbeitenden Systems aus Langzeitarchiv und Institutionellem Repository. Beide Komponenten erfahren in dem Konzept der vernetzten Infrastruktur bereits eine enge Kopplung durch das übergeordnete Forschungsdatenportal, welches sowohl für das Archiv als auch für das Repository dieselben Werkzeuge für Datenproduzent und -nutzer zur Verfügung stellt und eine Datenorganisation an übergeordneter Stelle ermöglicht. Der Output des Institutionellen Repositories kann in das Langzeitarchiv eingespeist werden. Ein ähnliches Modell verteilter Dienste wurde bereits in dem Projekt SHERPA DB[118] entwickelt.

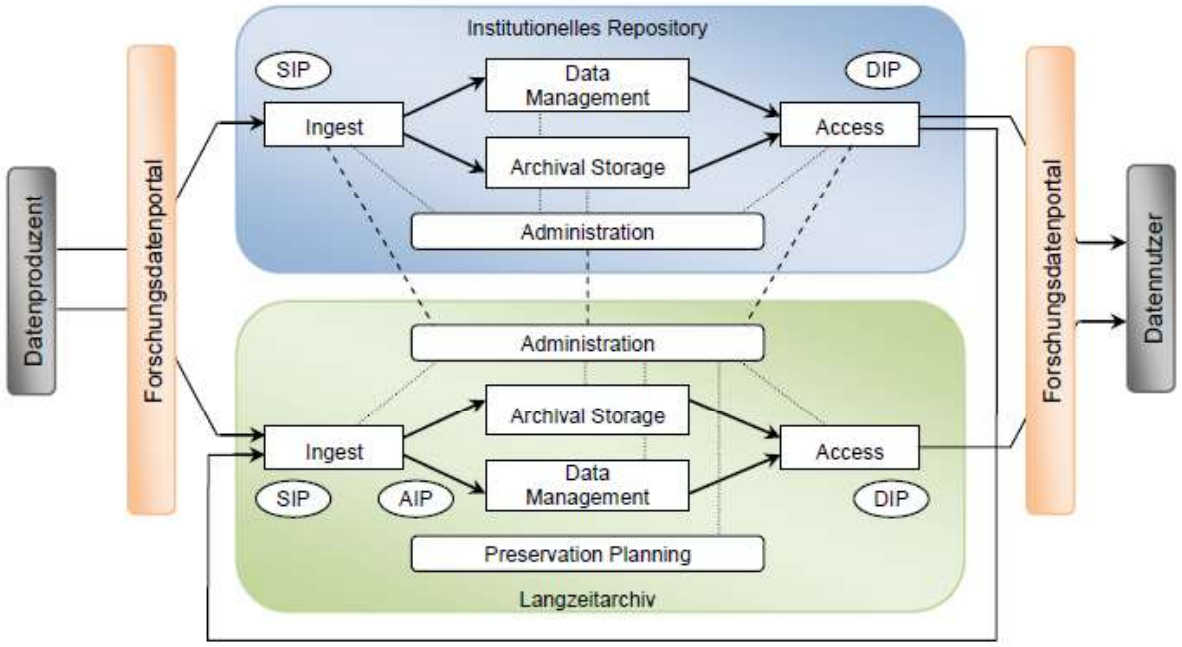


Abb. 33: OAIS-konformes Modell eines kooperativen Systems aus Langzeitarchiv und Repository

6.1.2 Digitale Identifikation

Um die Verfügbarkeit und Authentizität von Informationen zu garantieren, müssen diese eindeutig und permanent identifiziert werden können. Eine stabile und weltweit eindeutige elektronische Kennzeichnung online gespeicherter Dokumente ermöglichen so genannte persistente Identifier. Ein persistenter Identifier besteht aus einer eindeutigen Zeichenkette, über die die Auffindbarkeit der referenzierten Daten – unabhängig von deren Lagerungsort – sichergestellt wird. Dazu existiert bei einem vertrauenswürdigen Dienstleister ein Verzeichnis, welches den Identifier auf den Lagerort der Daten auflöst. So bietet ein

persistenter Identifier auch bei sich ändernden URLs (Uniform Resource Locator) eine Zitationssicherheit und ist damit für ein digitales Langzeitarchiv unumgänglich.

Es gibt verschiedene Arten von persistenten Identifiern, ein wichtiger ist der 1997 eingeführte und technisch auf dem Handle-System[119] aufbauende DOI[112]. Die Betreiberin des DOI-Systems ist die 1998 gegründete International DOI Foundation (IDF)[120]. Dem IDF-System sind eine Reihe von Registrierungsagenturen zugeordnet, die für eine Gemeinschaft mit definiertem Interesse die Aufgabe der Referenzierung digitaler Objekte übernehmen.

=> Konzeptstudie: DOI als persistenter Identifier

Die Daten im Langzeitarchiv für chemische Forschungsdaten sollen durch DOI-Namen persistent identifizierbar sein. Speziell für die Synthesechemie werden von den Projektpartnern der Konzeptstudie unterschiedliche Szenarien der DOI-Vergabe diskutiert:

- DOI für jeden einzelnen Datensatz (in einer Publikation)
- DOI für einen Container, in dem alle Datensätze für eine chemische Struktur enthalten sind
- DOI für Container, in dem alle Datensätze einer Publikation vorhanden sind

Ein DOI-Name besteht aus einer eindeutigen alphanumerischen Zeichenfolge, die in zwei Teile, ein Präfix und ein Suffix, gegliedert ist. Das Präfix wird von der DOI-Agentur, also von DataCite, vergeben, das Suffix vom Archivbetreiber. Über das Suffix kann der Archivbetreiber die Organisation und Erfassung zusammengehöriger Datensätze realisieren.

Die mögliche hierarchische Strukturierung des Suffix erlaubt eine Gruppierung von Datensätzen. Dabei werden Suffix-Nodes verwendet, um hierarchische Informationen oder Granularitätsebenen widerzuspiegeln. Das Suffix sollte ein konsistentes, logisches System reflektieren, dass gut dokumentiert werden kann[121].

Beispiel: In einer Publikation, die durch einen DOI identifiziert ist, können weitere Sub-DOIs wiederum Teile identifizieren und ansprechen, z. B. eine darin enthaltene Abbildung. DOI-Suffixe sind damit auch erweiterbar.

Auf Basis dieses Ansatzes kann der angesprochene Container für eine Sammlung von Datensätzen (Szenarien 2 und 3) realisiert werden. Es würde einen übergeordneten DOI geben, der den Container charakterisiert, und untergeordnete Sub-DOIs, welche die einzelnen, im Container enthaltenen Datensätze ansprechen. Dem den Container beschreibenden DOI sind damit Sub-DOIs zugeordnet. Dieses Datensatz-Containermodell lässt sich ideal mit Hilfe des im nachfolgenden Kapitel behandelten Metadatenstandards METS beschreiben.

6.1.3 Metadaten

Metadaten sind strukturierte Daten, mit denen eine Informationsressource beschrieben wird. Sie dienen nicht nur dem Auffinden von Informationsobjekten und einer besseren Recherche, sondern enthalten im Bereich der Langzeitarchivierung auch wichtige Informationen, die für die Verwaltung, Verarbeitung, Nutzung und Pflege der Objekte notwendig sind.

Auch im Bereich der Archivierung von Forschungsdaten ist eine ausführliche Erfassung von Metadaten eine notwendige Voraussetzung, um deren technische sowie inhaltliche Nachnutzbarkeit sicherzustellen. Beispielsweise muss festgehalten werden, was die Daten bedeuten und warum sie generiert wurden, in welchem wissenschaftlichen Kontext sie stehen oder auch welchen Prozessen bzw. Konvertierungen sie unterzogen wurden. Damit sind Metadaten nicht statisch, sondern erfassen auch Veränderungen nach dem Daten-Upload. Je umfassender und differenzierter ein Datensatz mit Metadaten beschrieben ist, umso höher ist im Allgemeinen sein wissenschaftlicher Wert.

Für die Beschreibung von digitalen Ressourcen in einem Langzeitarchiv werden zu den in Kapitel B erwähnten Metadaten noch 3 weitere Gruppen von Metadaten benötigt: strukturelle und administrative Metadaten sowie Preservation Metadaten. Für die Struktur dieser Metadaten existieren eine Reihe internationaler Standards bzw. Schemata. Metadaten können Ressourcen auf verschiedenen Aggregationsebenen beschreiben. So können sie sich auf eine einzelne Ressource, eine ganze Sammlung oder auch nur einen Teil einer größeren Ressource beziehen.

Strukturelle Metadaten

Die strukturellen Metadaten erfassen den Zusammenhang eines Objekts mit anderen Objekten im Langzeitbewahrungssystem. Beispielsweise besteht im Bereich der Forschungsdaten die Notwendigkeit, verschiedene Messdaten einer chemischen Struktur oder die Daten einer zusammengehörigen Versuchsserie in Beziehung zu setzen. Dieses lässt sich über den Metadata Encoding & Transmission Standard METS[122] realisieren, in den sich externe Metadatenschemata für die anderen Metadatentypen einbinden lassen.

Administrative Metadaten

Die administrativen Metadaten liefern Informationen, die zur Verwaltung der archivierten Objekte notwendig sind. Dazu gehören rechtliche Aspekte wie Verwertungs- und Urhebernachweise, Zugriffssteuerung, aber ebenso Provenienzangaben.

Preservation Metadaten

Zur Sicherstellung der Langzeitbewahrung und vor allem der -verfügbarkeit sind besondere Informationen notwendig, wie z. B. Hinweise zur Migration und zur technischen Umgebung, die notwendig ist, um das Objekt verfügbar zu halten. Die Provenienz-Metadaten erfassen die Historie eines Objekts wie dessen Herkunft sowie alle Maßnahmen, die unternommen wurden, um die Langzeitverfügbarkeit zu erhalten. Weiterhin umfasst dieser Datentyp Informationen zur Sicherung der Authentizität sowie Rechte-Informationen hinsichtlich der auf die Daten anwendbaren Prozesse. Bekannte Standards für die Langzeitarchivierungsmetadaten sind PREMIS[123] und LMER[124].

PREMIS (PREservation Metadate: Implementation Strategies) ist eine Initiative, welche die Entwicklung und Pflege des international anerkannten gleichnamigen PREMIS-Langzeitarchivierungsmetadatenstandards verantwortet. Sie wurde 2003 von der Research Libraries Group (RLG) und dem Online Computer Library Center (OCLC) ins Leben gerufen. Das Datenmodell von PREMIS definiert fünf grundlegende Einheiten, die als Entities bezeichnet werden: Intellectual Entity, Object Entitiy, Event Entity, Rights Entity und Agent Entity. Die Eigenschaften der Entities werden durch so genannte Semantic Units, die die für die Langzeitarchivierung relevanten Eigenschaften beinhalten, näher beschrieben[125].

Die Langzeitarchivierungsmetadaten für elektronische Ressourcen LMER wurden im Jahr 2003 von der Deutschen Bibliothek entwickelt. Das Objektmodell basiert auf dem Preservation Metadata Implementation Schema der Nationalbibliothek von Neuseeland. Das Metadatenschema LMER dient vorrangig als Austauschformat in kooperativen Archivierungsumgebungen. Es ermöglicht die Erfassung von technischen Informationen sowie der Veränderungshistorie eines Objekts und beschränkt sich damit auf Angaben, die größtenteils automatisch generiert werden können. Zum einen enthält LMER Kernelemente, die für alle Dateikategorien und jedes Dateiformat gültig sind, zum anderen gibt es auch einen flexiblen Teil für spezifische Metadaten. Das Objektmodell umfasst die Abschnitte LMER-Objekt, LMER-Datei, LMER-Prozess und LMER-Modifikation, die jeweils durch ein eigenes XML-Schema beschrieben werden.

Metadaten im LMER-Format können im Wege der Konkordanz auf das Datenmodell von PREMIS gemappt werden.

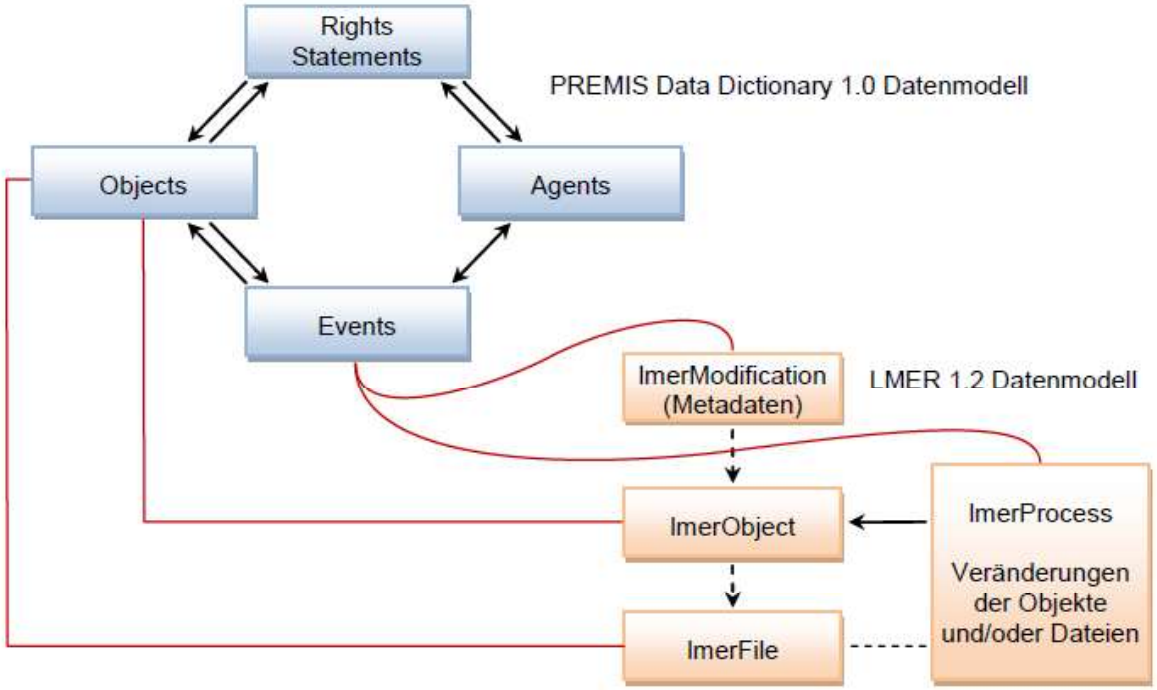


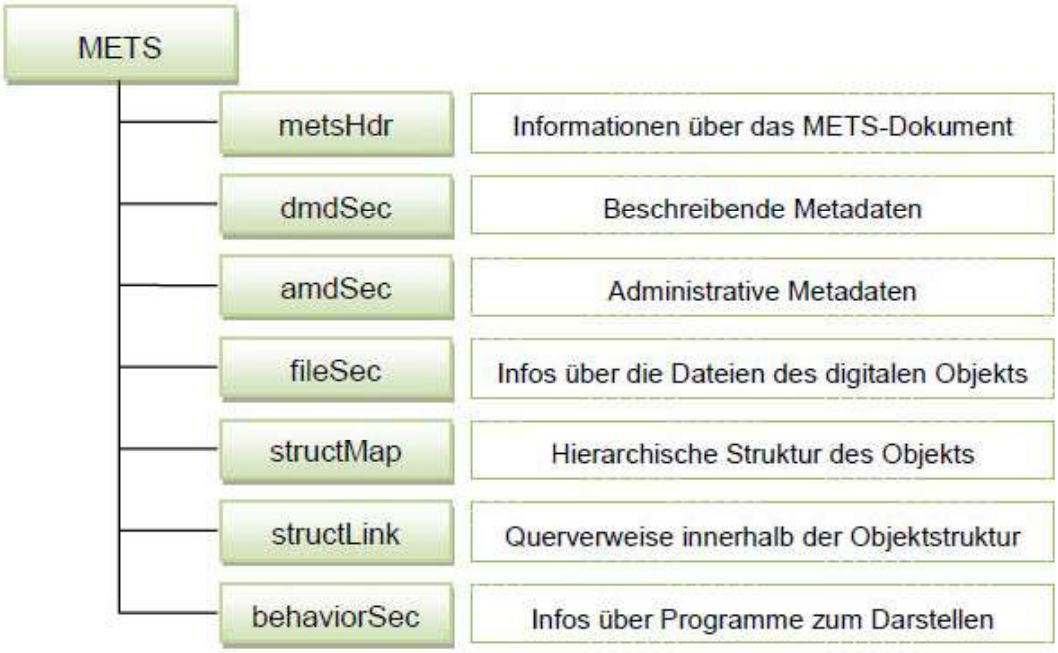
Abb. 34: Konkordanz LMER – PREMIS

METS

Der Standard METS[122] wurde auf Initiative der Digital Library Federation entwickelt und wird heute von der Library of Congress betreut. METS dient zur Beschreibung von digitalen Objekten mit Metadaten und bietet eine kohärente, übergeordnete Gesamtstruktur, die alle genannten Typen von Metadaten (beschreibende, strukturelle und administrative Metadaten) in einem XML-Dokument vereint. Die Metadaten sind entweder eingebettet in die METS-eigene Struktur oder sie werden in einer eigenen, externen Datei erfasst, auf die referenziert wird. METS kann in seiner Eigenschaft als Containerfomat beliebige XML-Schemata (so genannte Extension Schemas) integrieren und ist damit sehr flexibel. Der auf XML basierende Standard METS ist Plattform- und Software-unabhängig sowie problemlos mit anderen Schemata austauschbar. Obwohl METS noch recht jung ist, stellt dieses Format bereits jetzt den ersten, weitläufig akzeptierten Standard für die Beschreibung archivierter, digitaler Objekte dar.

Als XML-Anwendungen können METS-Dateien durch viele gängige Software-Tools erstellt und verwaltet werden. Derzeit werden eine Reihe von Software-Anwendungen entwickelt, die die Verwaltung von METS-Dateien erleichtern.

Abb. 35: Struktur eines METS-Dokuments



Wie die Abb. 35 zeigt, besteht ein METS-Dokument aus sieben Hauptabschnitten:

- **METS Header**
  Der Kopfteil enthält Metadaten, die das jeweilige METS-Dokument selbst beschreiben, einschließlich der Angaben zum Bearbeiter bzw. Produzenten des METS-Dokuments.

- **Beschreibende Metadaten (Descriptive Metadata)**
  Der Abschnitt für die beschreibenden Metadaten kann sowohl Verweise auf ein externes Dokument, wie z. B. einen Dublin Core Datensatz enthalten, wie auch in das METS-Dokument eingebettete Angaben oder auch beides miteinander kombiniert. Ebenso ist es möglich, mehrere externe und interne Erschließungspakete in diesen Abschnitt zu integrieren.

- **Administrative Metadaten (Administrative Metadata)**
  Der Abschnitt für die administrativen Metadaten liefert Informationen über die Herstellung und Speicherung von Dateien, über Urheberrechte und über die digitalisierte Vorlage. Außerdem werden hier Angaben zur Herkunft der Digitalisate erfasst. Ähnlich wie die Erschließungsangaben können diese Metadaten extern oder in dem METS-Dokument integriert vorliegen.

- **Dateienabschnitt (File Section)**
  Im Dateienabschnitt werden alle Dateien mit Inhalten, aus denen das digitale Objekt besteht, aufgelistet.

- **Strukturbeschreibung (Structural Map)**
  Die Strukturbeschreibung ist der zentrale Bestandteil jedes METS-Dokuments. Sie bildet den inneren Aufbau des digitalen Objektes ab und verknüpft die Elemente der Struktur mit den Dateien, aus denen der Inhalt des digitalen Objekts besteht, sowie mit deren Metadaten.

- **Strukturverknüpfung (Structural Links)**
  Der Abschnitt zur Strukturverknüpfung ermöglicht es den Erstellern von METS-Dokumenten, vorhandene Hyperlinks zwischen einzelnen Knoten des im Strukturabschnitt dargestellten hierarchischen Aufbaus des digitalen Objekts zu beschreiben.

- **Verhalten (Behavior)**
  Ein Abschnitt zum Verhalten kann eingefügt werden, um ausführbare Anweisungen für das Verhalten mit den Inhalten in METS-Objekten zu verknüpfen. So kann z. B. für die Anzeige einer bestimmten Bilddatei ein Viewer mit Parametern für die Steuerung der Darstellung aufgerufen werden.

=> Konzeptstudie: METS als Metadatencontainer

Das Arbeitspaket 3 beschäftigte sich detailliert mit der Analyse bestehender Metadatenstandards und definierte spezifische für die Beschreibung chemischer Forschungsdaten notwendige Metadaten. Es wurde festgestellt, dass es spezieller messtechnischer Metadaten bedarf, die die Geräteparameter sowie die Messbedingungen erfassen. Weiterhin müssen chemische Metadaten definiert werden, die z. B. im Bereich der Synthesechemie Substanzinformationen wie Struktur- und Summenformel enthalten. Der Standard METS kann in seiner Eigenschaft als Containerformat flexibel sowohl bereits existierende Metadatenstandards als auch die speziell zur Beschreibung chemischer Forschungsdaten entwickelten Metadatenschemata integrieren. Deshalb ist er hervorragend für die Beschreibung von chemischen Forschungsdaten in einem Repository bzw. Langzeitarchiv geeignet.

Im Kontext des OAIS-Referenzmodells kann ein METS-Dokument als Submission Information Package (SIP), als Archival Information Package (AIP) oder Dissemination Information Package (DIP) eingesetzt werden. Der Datenproduzent gibt für den zu publizierenden Datensatz bzw. für den Container, der mehrere zusammengehörige Datensätze enthält, beschreibende und strukturelle Metadaten ein. Andere Metadaten werden automatisch erfasst. Die Daten inklusive der in einer METS-Datei befindlichen Metadaten werden zum SIP zusammengeschnürt und hochgeladen. Die METS Datei wird vom Archivbetreiber durch Hinzufügen weiterer administrativen Metadaten inklusive Preservation Metadaten erweitert und als AIP archiviert. Jedes archivierte Datenpaket

enthält somit einen oder mehrere Datensätze und wird immer durch genau ein METS Dokument beschrieben.

Die Erfassung der Metadaten soll so weit wie möglich automatisiert erfolgen. Viele der technischen Metadaten wie Dateiformate und -versionen, Datenumfang, Messparameter, Zeitstempel usw. können durch entsprechende Werkzeuge automatisch ermittelt werden.

Beschreibende Metadaten hingegen müssen in der Regel durch den Wissenschaftler erzeugt werden. Dazu zählen einerseits die Beschreibung des Experiments sowie dessen thematische Einordnung, andererseits die chemischen Metadaten wie Struktur-, Summenformel und Verbindungsname, die sich oft erst aus der Dateninterpretation selbst, teils aus der Analyse unterschiedlicher Datensätze, ergeben. Ebenso müssen auch Teile der administrativen Metadaten, wie rechtliche Informationen über Urheber- und Verwertungsrechte, manuell erstellt werden.

Die nachfolgende Tabelle greift die in Kap. 2.4, Abb. 24 dargestellte Struktur eines vernetzten Langzeitarchivs für chemische Forschungsdaten wieder auf und zeigt beispielhaft, welche Metadaten in welchem Stadium des Daten-Publikationsprozesses entstehen. Die grün dargestellten Datentypen lassen sich automatisch generieren, während die rot markierten Informationen manuell vom Wissenschaftler zu erstellen sind.

Schritt im Daten-Publi-kationsprozess	**Erfassung Metadatentyp (automatisch / manuell)**
1. Messung	Zeitstempel der Messung Messparameter: Typenbezeichnung Messgerät, Geräteparameter, Messbedingungen Operator Name/Bezeichnung der Messung
2. Speicherung 3. Nachbearbeitung auf Arbeitsgruppenserver	Zeitstempel der Speicherung

4. Transfer in Institutionelles Repository	Zeitstempel des Transfers Autorendaten Beschreibung der Messung, Ziel, wissenschaftlicher Kontext, Keywords, Kommentierungen Chemische Metadaten: Struktur- und Summenformel, Name, Chemischer Identifier Rechtliche Informationen: Urheberrecht, Verwertungsrecht Dateiformatspezifische Informationen: Fileformat-Version, kompatible Visualisierungstools, Checksumme
5. Transfer in Langzeit-archiv	Zeitstempel des Transfers Quell-Repository Langzeitinformationen: Format-Lebensdauer, Historie, Erhaltungsstrategie, Rechte-Informationen

Tabelle 2: Metadaten in verschiedenen Stadien des Daten-Publikationsprozesses

=> Konzeptstudie: Interoperabilität durch OAI-PMA

Wie bereits zuvor beschrieben, strebt die Open Archive Initiative an, dass alle Datenprovider ihre individuellen Metadatensätze in einen gemeinsamen, definierten Core Set konvertieren, um diesen über die OAI-PMH-Schnittstelle für ein Harvesting zur Verfügung zu stellen. Auf dieser Technik basierend lassen sich Daten-Archive miteinander vernetzen. Das Archiv für chemische Forschungsdaten kann auf diese Weise sowohl in nationale und internationale Repository-Netzwerke eingebunden werden als auch interdisziplinär, also mit existierenden Archiven anderer Disziplinen, vernetzt werden.

6.2 Policies für das Langzeitarchiv

6.2.1 Auswahl der Daten

Nicht alle von den Wissenschaftlern erzeugten Datensätze sollten in das Langzeitarchiv aufgenommen werden. Einerseits würde dies früher oder später zu Kapazitätsproblemen führen, andererseits muss der Wissenschaftler vor einer intellektuell nicht mehr zu bewältigenden Datenflut geschützt werden. Damit müssen Auswahlkriterien entwickelt werden, die letztendlich auch ein weiterer Aspekt einer Qualitätssicherung des Archivs und des Archivguts sind, da mit der Aufnahme in das Langzeitarchiv gleichzeitig eine Verdichtung relevanter Information erfolgt. Archivare greifen bei dieser Aufgabe auf einen Erfahrungsschatz von über 100 Jahren zurück und geben wichtige Anhaltspunkte für eine Auswahl.

Zur Ableitung von Kriterien verwenden Archivare ein Modell, welches sich in den letzten Jahren etabliert hat. Dabei wird zwischen Primärnutzen und Sekundärnutzen unterschieden. Daten besitzen zunächst vorrangig einen primären Wert für den Datenproduzenten, gleichzusetzen mit dem Wissenschaftler selbst. Dieser entscheidet, welche Daten er generiert, welche Ergebnisse er aus ihnen ableitet und an wen er die Daten weitergibt. Der Primärwert verringert sich im Laufe der Zeit, bis er nach etwa zehn Jahren so gering ist, dass der Wissenschaftler selbst die Daten nicht mehr benötigt, weil z. B. diese speziellen Forschungsarbeiten abgeschlossen und die Ergebnisse publiziert sind.

Für die Gesellschaft hingegen können diese Daten auf langfristige Sicht einen Sekundärwert besitzen und ein wichtiges Kulturgut darstellen. In diesem Fall dürfen sie keinesfalls verloren gehen, sondern müssen dauerhaft archiviert werden.

Speziell im Bereich e-Science muss dieses herkömmliche Modell um eine weitere Rolle erweitert werden. Zusätzlich ist der Fachnutzer, auch Sekundärwissenschaftler genannt, mit seinen individuellen Interessen zu berücksichtigen. Dieser gehört nicht zum Kreis der Datenproduzenten, sondern möchte bestehende Daten im Rahmen seiner eigenen, speziellen Forschung nachnutzen. Bei der Ableitung von Auswahlkriterien sind alle drei Kategorien zu berücksichtigen.

<u>Eindeutige Auswahlkriterien:</u>

- Alle publizierten Forschungsdaten müssen dauerhaft aufbewahrt werden.
- Auch Forschungsdaten, die Basis einer Publikation sind, sollten langzeitarchiviert werden.

Alle anderen Forschungsdaten sind nach etwa zehn Jahren hinsichtlich einer Langzeitarchivierung zu prüfen. Dieser Zeitpunkt erscheint sinnvoll, weil zum einen der Primärwert der Daten nach dieser Zeit signifikant gesunken ist und zum anderen eine Einhaltung der Empfehlungen der DFG zur Guten Wissenschaftlichen Praxis nachgewiesen werden kann. Bei der Prüfung muss festgestellt werden, ob die Daten für Sekundärwissenschaftler von Interesse sind oder ob sie einen Sekundärwert für die Gesellschaft besitzen. Ist dies der Fall, sind die Daten – bezugnehmend auf die in Kap. 2 entworfene vernetzte Infrastruktur – aus dem Institutionellen Repository in das Langzeitarchiv zu transferieren.

<u>Weitere Auswahlkriterien:</u>

- Wert der Daten für einen Fachnutzer (Nachnutzung) bzw. Wert der Daten für die Gesellschaft (Kulturerbe)
- Häufigkeit der Nachfrage
- Möglichkeit der Datenreproduktion (Einmaligkeit der Daten, hohe Kosten des Experiments)
- Auswahl des Besonderen
- Auswahl des Gewöhnlichen, Querschnitt, Längsschnitt

Die Kriterien müssen auf die Disziplin zugeschnitten und nachvollziehbar sein. Eine Kontinuität der Auswahlkriterien ist notwendig, um eine Vergleichbarkeit der Datenarchivierung zu gewährleisten. Gehen Forschungsdaten in die dauerhafte Aufbewahrung über, unterliegen sie als Archivgut den Archivgesetzen.

6.2.2 Organisatorisches Konzept

Die Planung digitaler Langzeitarchivierungsmaßnahmen und deren Dokumentation, wie im Referenzmodell OAIS vorgesehen und von Zertifizierungsinitiativen vorgeschrieben, stellen einen relativ komplexen und aufwändigen Prozess dar. Unterstützung beim Planungsprozess liefert das Planning Tool Plato [126], welches im Rahmen des EU-Projekts PLANETS[42] entwickelt wurde. Plato ist als Web-Applikation frei verfügbar und führt den Anwender durch die einzelnen Schritte des Workflows zur Erstellung eines Langzeitarchivierungsplans.

Die nachfolgende Übersicht zeigt beispielhaft, welche Aspekte bei der Planung eines OAIS-konformen Langzeitarchivs für chemische Forschungsdaten berücksichtigt und dokumentiert werden müssen. Sie basiert auf dem Dokument “Policy-making for Research Data in Repositories: A Guide“[127] und liefert bereits erste Lösungsansätze für den Planungsprozess.

Inhaltliche Abdeckung

Fachzuordnung/Themenfelder

- berücksichtigte bzw. ausgeschlossene Fächer/Themengebiete

Konzeptstudie: Es erfolgt eine Beschränkung auf die organische und anorganische Synthesechemie.

Sprache

- Vorgaben für die Sprache
- Notwendigkeit von Textübersetzungen, wie z. B. der Metadaten

Konzeptstudie: Der Schwerpunkt liegt auf Englisch als wissenschaftliche Publikationssprache, gegebenenfalls kann auch eine Mehrsprachigkeit berücksichtigt werden.

Art der Forschungsdaten

- Daten aus wissenschaftlichen Experimenten
- Daten aus Modellierungen und Simulationen
- abgeleitete Daten aus dem Prozessieren oder Kombinieren von Rohdaten oder anderen Daten

Konzeptstudie: Es werden zunächst experimentelle Daten der analytischen Standardverfahren NMR, IR, UV/VIS, MS, X-Ray und HPLC/GC/GPC betrachtet.

Status der Forschungsdaten

Forschungsdaten können im Data Lifecycle unterschiedliche Stadien annehmen:

- Rohdaten
- bearbeitete Daten, akkumulierte Daten
- analysierte, interpretierte Daten
- zusammengefasste bzw. tabellarische Daten

Konzeptstudie: Die Rohdaten sind gemeinsam mit den ausgewerteten und möglichst auch visualisierbaren Daten zu speichern.

Dateiformate

- Festlegung bevorzugter, archivierungsfähiger Dateiformate

Konzeptstudie:

Es sollten Empfehlungen für archivfähige Formate erarbeitet werden. Das Archiv sollte aber letztendlich Messdaten in jedem beliebigen Format annehmen. Jeder Rohdatensatz muss zusätzlich in ein visualisierbares und archivfähiges Austauschformat konvertiert werden. (Beispiel NMR: FID-File + JCAMP-DX).

Bei der Beschränkung auf die Synthesechemie ist die Anzahl der Formate überschaubar. Soll das Archiv für alle Teilgebiete der Chemie offen sein, sind auch viele individuelle Formate zu berücksichtigen, deren umfassende Planung schwierig ist.

Volumen- und Größenbeschränkungen

- Einschränkungen für Dateigröße oder Zahl separater Dateien
- Anwendung von Komprimierungssoftware

Konzeptstudie: Bei Forschungsdaten darf es hinsichtlich Dateigröße keine Beschränkungen geben.

Metadaten

Metadatenkategorien

Zur Beschreibung digitaler Objekte in einem Archiv werden im Allgemeinen beschreibende, strukturelle und administrative Metadaten benötigt. Es ist im Detail zu definieren, welche untergeordneten Metadatenkategorien es gibt, wann im Publikationsprozess sie generiert werden und wer dafür verantwortlich ist.

Konzeptstudie:

Metadatenkategorie	**Inhalt**
Messtechnische Metadaten	Typbezeichnung Messgerät, Geräteparameter, Messbedingungen

Chemische Metadaten	Summenformel, Strukturformel, Verbindungsname, chemischer Identifier
Bibliographische Metadaten	beschreibende Metadaten, technische Metadaten, administrative Metadaten (DataCite)
Preservation Metadaten	History, Format-Lebensdauer, Archivierungsstrategie
Strukturelle Metadaten	Relationen zwischen Datensätzen

Tabelle 3: Metadatenkategorien für chemische Forschungsdaten

Der Metadatenstandard METS wird als Containerformat für externe – bereits existierende bzw. in der Konzeptstudie entwickelte – Schemata für die einzelnen Metadatenkategorien verwendet.

Zugang zu Metadaten

Kosten und Zugriffskontrolle für Metadaten

Konzeptstudie:

Die beschreibenden Metadaten sind frei zugänglich.

Die administrativen und technischen Metadaten dienen vorrangig internen Zwecken. Wichtige administrative Informationen sollten aber auch dem Datenautor als Urheber der Daten über das Forschungsdatenportal zur Verfügung gestellt werden.

Reuse von Metadaten

Definition von Nutzungsszenarien für die Metadaten

Konzeptstudie:

Das Archiv besitzt eine OAI-PMH-Schnittstelle für das Harvesting von Metadaten.

Die bibliographischen Metadaten werden zusätzlich bei der TIB Hannover im Rahmen der DOI-Registrierung gespeichert.

Ingest

Zugangsberechtigung

- Bestimmung der für einen Daten-Upload berechtigten Personen
- nur Repository-Betreiber
- registrierte Studenten, Lehrpersonal, akkreditierte Mitglieder
- Datenproduzenten

Konzeptstudie: Datensätze dürfen durch den Datenproduzenten oder dessen Bevollmächtigten in das Repository geladen werden.

Inhaltliche Einschränkungen

- Festlegung von Vorgaben für Datenproduzenten

Konzeptstudie:

Die Metadaten müssen vollständig sein.

Das Hochladen kompletter Datensätze erfolgt entsprechend der Vorgaben des Archivbetreibers.

Vorgaben für Repository

- Datenprüfung zur Sicherstellung der Integrität, Validität, Vollständigkeit beim Transferprozess
- Stichproben oder alle Datenpakete
- Prüfung der Metadaten

Konzeptstudie: Datenprüfungen zur Integrität, Validität und Vollständigkeit der Datenpakete einschließlich Metadaten durch den Archivbetreiber erfolgen in dem Maße, wie sie automatisierbar sind (technische Qualitätskontrolle, s. Kap. 6.2.3). Der Einsatz von Personal ist unwahrscheinlich.

Vergabe eines persistenten Identifiers

Konzeptstudie: Alle hochgeladenen Datensätze werden bei DataCite registriert und mit einem DOI versehen.

Weitere Kriterien zur Prüfung der Archivobjekte durch Archivbetreiber:

- Berechtigung des Datenproduzenten zur Datenablage
- Relevanz für den Geltungsbereich des Archivs
- Validität von Formaten
- Ausschluss von Spam
- Festlegung der Verantwortlichkeiten

Konzeptstudie: Die Verantwortlichkeiten müssen klar definiert und vertraglich geregelt werden. Der Datenproduzent ist verantwortlich für die Qualität, Validität und Authentizität seiner Forschungsdaten. Der Archivbetreiber muss für die Qualität der Speicherung und für die dauerhafte Datenverfügbarkeit garantieren.

Qualitätsabschätzung

- Merkmale zur Einschätzung und Sicherstellung der Datenqualität:
- inhaltliche Qualitätssicherung der Daten durch Experten und Kollegen vor dem Daten-Upload in das Langzeitarchiv
- Reputation des Datenproduzenten
- Sekundärwert der Daten

Konzeptstudie: Ein Workflow für eine inhaltliche Qualitätskontrolle für chemische Forschungsdaten ist auszuarbeiten (s. Kap. 6.2.3).

Wird der Datensatz im Rahmen einer Veröffentlichung publiziert, ist die Wahrscheinlichkeit hoch, dass der Datensatz qualitativ in Ordnung ist.

Wird der Datensatz unabhängig publiziert, ist ein Review-Prozess zwingend notwendig, z. B. Installation eines Editorial Bords oder Community-basierte Qualitätsprüfung.

Embargoprozess

Konzeptstudie: Die Berücksichtigung eines Embargoprozesses ist – beispielsweise aufgrund von Vorgaben der Verlage für die Veröffentlichung – gegebenenfalls notwendig.

Rechte und Besitzverhältnisse

Es ist der Abschluss eines Lizenzvertrags zwischen Datenlieferant und Archivbetreiber notwendig, in dem Rechte und Pflichten in Form einer klar und präzise definierten Datenübergabevereinbarung festgelegt sind. Der Archivbetreiber archiviert die Daten, macht sie zugänglich und erhält die für die Datenpflege notwendigen Rechte. Im Gegenzug muss der Datenproduzent auf entsprechende Rechte verzichten.

Konzeptstudie:

Der Lizenzvertrag regelt Rechte und Pflichten des Archivbetreibers:

Konvertieren, Kopieren, Umorganisieren von Datensätzen

Migration von Datensätzen in ein anderes Repository

Haftung des Archivbetreibers für Schäden und Verlust von Datensätzen

Die Pflichten des Datenlieferanten sind:

keine Verletzung von Rechten Dritter

bei Upload von Fremddaten Vorlage einer Einverständniserklärung des Rechteinhabers

bei gesponserten Forschungsarbeiten Erfüllung aller Verpflichtungen gegenüber der Förderorganisation

Access und Reuse

Zugang zu den Datenobjekten

- Datenzugang: Open Access, Controlled Access, Restricted Access
- Registrierung

Konzeptstudie:

Der Zugang zu den Datensätzen soll offen sein für alle Wissenschaftler und Interessenten (Open Data). Ein kontrollierter Zugang sollte über einen anerkannten Authentifizierungsdienst wie DFN-AAI (Shibboleth[111]) realisiert werden.

Die chemischen Forschungsdaten sollen leicht auffindbar sein, der Zugang soll über komfortable Suchfunktionalitäten in einem Forschungsdatenportal erfolgen.

Zugangsmethoden

- Definition der Zugangsmöglichkeiten zu den Datensätzen
- Download einzelner Datensätze oder Batchmodus
- Visualisierung und Mapping Applikationen
- Zugang über andere Web Services

Konzeptstudie:

Das Archiv stellt eine Downloadmöglichkeit für den gesamten Datensatz einschließlich Metadaten zur Verfügung.

Weitere Spezifikationen sind Visualisierungen, andere Web Services usw.

Ein Harvesting der Metadaten erfolgt zwecks Vernetzung und besserer Auffindbarkeit.

Verwendung und Nachnutzung von Datensätzen

- Beschränkungen für die Nachnutzung von Datensätzen:
- keine Nachnutzung erlaubt
- Beschränkungen hinsichtlich Umformatieren oder Weitergabe
- sonstige Auflagen

Konzeptstudie:

Die Daten sind in einer Public Domain und die Nachnutzung kann uneingeschränkt erfolgen bzw. ist ausdrücklich erwünscht.

Das Copyright ist einzuhalten.

Copyright

- Definition eines Lizenzmodells für Datensätze

Zum Schutz der intellektuellen Leistung der Wissenschaftler sollten die Daten im Langzeitarchiv mit Lizenzen versehen sein, die die Bedingungen einer Nachnutzung regeln. Entsprechende Vorarbeiten sind bereits in den Projekten Creative Commons[53] (CC) und Science Commons [128] (SC) geleistet worden. Science Commons ist eine Initiative von Creative Commons mit dem speziellen Ziel, Strategien und Werkzeuge für eine schnellere web-gestützte Wissenschaft zu entwickeln. Die Diskussion, welche Lizenzen für Daten empfohlen werden, ist noch offen.

Weitere klärungsbedürftige Aspekte

- Zitierungen
- Kopien
- User Tracking und Statistik

Datenerhalt

- Aufbewahrungsdauer
- Festlegung der Aufbewahrungsfrist:
- dauerhafte Archivierung
- definierte Jahreszahl
- Lebenszeit des Repositories
- individuelle Fristen je nach Datensatz

Konzeptstudie: Im Institutionellen Repository werden die Daten bis zu zehn Jahre aufbewahrt, im Langzeitarchiv erfolgt eine dauerhafte Archivierung. Optionen für individuelle Fristen sind zu prüfen.

Preservation Planning für archivierte Formate

- Definition von Supportlevels für verschiedene Dateiformate
- Migration in neue Dateiformate

Konzeptstudie: Für die Dateiformate der chemischen Forschungsdaten müssen geeignete Erhaltungsstrategien entwickelt werden (s. Kap. 3).

Integrität und Authentizität

Sicherheitsmaßnahmen sind erforderlich, um die Integrität und Authentizität der Datensätze zu gewährleisten. Es muss belegt werden, dass die Originaldaten des benannten Datenproduzenten vorliegen und diese nicht im Nachhinein manipuliert wurden.

- Definition von Sicherheitsmaßnahmen: Fixity Checks über Prüfsummen
- Festlegung von Zeitpunkten für Fixity Checks
- Protokollierung des Zugriffs auf einzelne Datensätze durch Archivbetreiber

Konzeptstudie:

Für jeden Datensatz werden Fixity Checks über Prüfsummen durchgeführt, sinnvolle Zeitpunkte sind zu definieren.

Jedes Datenpaket wird mit einer digitalen Signatur versehen.

Die Nutzerverwaltung erfolgt über einen anerkannten Authentifizierungsdienst wie DFN-AAI (Shibboleth[111])

Sperrung von Datensätzen

- Analyse der Bedingungen, die zur Sperrung eines Datensatzes führen
- Definition der Vorgehensweise für gesperrte Datensätze

Konzeptstudie: Datensätze sollten in bestimmten Fällen vom Archivbetreiber gesperrt werden können, denkbare Szenarien für eine Sperrung sind:

Verletzung des Copyrights

verfälschte Daten

vertrauliche Daten

Aus Gründen der Beweissicherung sollte von einer Löschung des Datensatzes abgesehen werden.

6.2.3 Qualitätssicherung von Forschungsdaten

Ein offener Zugang zu Forschungsdaten fördert sowohl die Transparenz der Forschung als auch die Forschungseffektivität, da für den Wissenschaftler die Möglichkeit einer Nachnutzung bereits erhobener Daten besteht. Die wesentliche Voraussetzung für Transparenz und Nachnutzbarkeit von Forschungsdaten ist deren hohe Qualität. Um diese sicher zu stellen, müssen Prozesse für eine Qualitätskontrolle definiert und realisiert werden. Dabei wird zwischen einer inhaltlichen Qualitätskontrolle der Daten und einer technischen Kontrolle unterschieden.

Die inhaltliche Qualitätskontrolle erfolgt durch den Datenproduzenten und kann nicht durch den Archivbetreiber geleistet werden, da dieser Prozess kaum automatisiert ablaufen kann. Die Hauptaufgabe des Archivbetreibers im Bereich der Qualitätssicherung besteht darin, einen zuverlässigen Speicher zur Verfügung zu stellen und einen sicheren Zugriff auf die Daten zu gewährleisten. Der Archivbetreiber ist also für die Integrität und Vertraulichkeit von Infrastruktur und Daten verantwortlich.

Entsprechend führt der Archivbetreiber technische Qualitätskontrollen der Daten durch, die sich allerdings vorrangig auf die Überprüfung der formalen Qualität beschränken. Dazu bietet er z. B. für den Datentransfer in das Langzeitarchiv Services wie eine automatische Validitäts- und Vollständigkeitsprüfung von Daten oder Metadaten an.

Die Methoden der Qualitätssicherung sind disziplinspezifisch und abhängig von der Art der Forschungsdaten. Ein universeller Lösungsansatz existiert nicht.

Inhaltliche Qualitätskontrolle

Die inhaltliche Qualität kann ausschließlich durch die Wissenschaftler selbst gesichert werden. Diese Art der Qualitätssicherung muss damit in der jeweiligen Arbeitsgruppe, die die Daten produziert hat, geleistet werden. Eine Community-basierte Qualitätsprüfung kann unterstützend eingesetzt werden.

Gerade die zu definierenden Maßnahmen einer inhaltlichen Qualitätssicherung sind stark disziplinabhängig. An die jeweilige Problematik angepasste Richtlinien für eine inhaltliche Qualitätsprüfung sind zu entwickeln. Derartige Prüfverfahren dürfen nicht starr sein, sondern müssen flexibel und anpassungsfähig sein.

In einer Art Review-Prozess könnten chemische Forschungsdaten beispielsweise auf ihre Qualität, Plausibilität, Vollständigkeit, Genauigkeit und ihr Format geprüft werden. Die Infrastruktureinrichtung kann unterstützend tätig sein, indem sie den Wissenschaftlern einen Workflow für eine inhaltliche Prüfung zur Verfügung stellt.

Technische Qualitätskontrolle

Der Umfang der technischen Qualitätskontrolle ist abhängig von der Art der Messung und der Quelle der Daten. Eine weitgehend automatisierte Prüfung kann nur für Standardverfahren, wie z. B. NMR-Messungen, und für Standard-Dateiformate erreicht werden (s. Kap. 6.1).

Im ersten Schritt ist eine Prüfung der erfassten Metadaten auf Vollständigkeit (z. B. sind alle Pflicht-Datenfelder ausgefüllt?) und Plausibilität (z. B. liegen die angegebenen Temperaturen in einem realistischen Bereich?) erforderlich. Ist die Prüfung positiv, kann in einem zweiten Schritt die Integrität der Daten (z. B. stimmen Prüfsummen und Signaturen? Erfolgte die Datenübertragung fehlerfrei?) geprüft werden (s. Kap. 5.5.2).

Liegen keine Fehler vor, erfolgt eine Überprüfung der Messdaten. Eine solche Überprüfung kann nur bei Daten durchgeführt werden, deren Dateiformat bekannt und dokumentiert ist. Verfügbare Tools decken sehr gut den Bereich der Spektroskopie ab. Sie können sowohl die Validität der Datei wie auch in Teilen die der darin enthaltenen Messdaten prüfen und sie können verwendet werden, um ältere oder ausgefallenere Dateiformate in aktuelle Standardformate zu konvertieren und für eine Langzeitverfügbarkeit zugänglich zu machen.

6.2.4 Informationssicherheit und Vertrauenswürdigkeit

Wie in den vorhergehenden Kapiteln ausführlich beschrieben, erfordert die Langzeitarchivierung Maßnahmen, die sicherstellen, dass die zu archivierenden Daten sorgfältig ausgewählt, erschlossen und nach einer Qualitätsprüfung gespeichert werden. Weiterhin müssen die logische und physische Integrität einschließlich der Authentizität langfristig erhalten bleiben und die archivierten Daten dauerhaft verfügbar und nachnutzbar sein.

- **Integrität:** unverändertes Vorliegen der digitalen Objekte
- **Authentizität:** Echtheit der digitalen Objekte, insbesondere der Aspekt der Nachweisbarkeit der Identität des Urhebers bzw. Autors
- **Vertraulichkeit:** kein Zugang zu den digitalen Objekten durch unberechtigte Dritte
- **Verfügbarkeit:** Zugänglichkeit zum digitalen Objekt
- **Nutzbarkeit:** Interpretierbarkeit der Daten sichergestellt

Digitale Informationen in einem Archiv sind bedroht durch Einbußen in ihrer Integrität, Authentizität und Vertraulichkeit sowie den gänzlichen Verlust ihrer Verfügbarkeit und Nutzbarkeit.

Um diesen Gefahren entgegen zu wirken, muss das Archiv organisatorische und technische Maßnahmen ergreifen. Es gibt festgelegte Ziele und Spezifikationen, innerhalb derer das Archiv operieren muss. Die Prüfung und Bewertung der eingesetzten Maßnahmen erbringt den Nachweis der Vertrauenswürdigkeit eines Langzeitarchivs.

Integrität und Authentizität

Um die Integrität und Authentizität von digitalen Daten gewährleisten zu können, sind Sicherheitsmaßnahmen erforderlich. So muss belegt werden, dass die Originaldaten des benannten Datenproduzenten vorliegen und dass diese nicht im Nachhinein manipuliert worden sind. Der Archivbetreiber muss entsprechende technische Kontrollen durchführen.

So genannte Fixity Checks über Prüfsummen geben Aufschluss darüber, ob ein digitales Objekt zwischen zwei Zeitpunkten verändert wurde. Bei der Ermittlung der Prüfsumme eines Objekts wird ein Hashcode berechnet, also eine Zahl, die ein Dokument eindeutig identifiziert. Jegliche Änderungen des digitalen Objekts haben eine Veränderung der Prüfsumme zur Folge. Beim Übertragen eines Dokuments muss der Hashcode verschlüsselt sein.

Eine Lösung zur Sicherstellung der Authentizität ist die digitale Signatur, die rechtlich der handschriftlichen Unterschrift gleichgestellt ist und den Sender eindeutig identifiziert. Eine digitale Signatur wird mittels asymmetrischer Kryptographie hergestellt.

Die Herkunft eines digitalen Objekts kann ebenso durch Implementierung eines anerkannten Dienstes zur verteilten Authentifizierung sichergestellt werden.

Kriterienkataloge und Zertifizierungsverfahren

Die Vertrauenswürdigkeit als Eigenschaft des digitalen Langzeitarchivs kann anhand von so genannten Kriterienkatalogen geprüft und bewertet werden. Auf nationaler Ebene hat nestor mit dem Dokument „Kriterienkatalog vertrauenswürdige digitale Langzeitarchive“[129] Kriterien zur Feststellung der Vertrauenswürdigkeit publiziert. Ein weiterer bekannter Kriterienkatalog wurde 2007 unter dem Titel „Trustworthy Repositories Audit & Certification“ (TRAC)[130] von der OCLC/RLG-NARA Digital Repository Certification Task Force herausgegeben.

Wesentliche internationale Vertreter des Themas Vertrauenswürdigkeit – das Center for Research Libraries (CRL), das Digital Curation Center (DCC) und das Projekt DigitalPreservationEurope (DPE) – haben zusammen mit nestor zehn gemeinsame Prinzipien herausgearbeitet:

Das digitale Langzeitarchiv

- übernimmt die Verantwortung für die dauerhafte Erhaltung und kontinuierliche Pflege der digitalen Objekte für die identifizierten Zielgruppen,
- belegt die organisatorische Beständigkeit (Finanzierung, Personal, Prozesse),
- verfügt über die erforderlichen Rechte (per Vertrag oder Gesetz),
- besitzt ein effektives und effizientes Geflecht von Grundsätzen (Policies),
- erwirbt und übernimmt digitale Objekte auf der Grundlage definierter Kriterien gemäß seiner Verpflichtungen und Fähigkeiten,
- stellt die Integrität, Authentizität und Nutzbarkeit der dauerhaft aufbewahrten Objekte sicher,
- dokumentiert alle Maßnahmen, die während des gesamten Lebenszyklus auf die digitalen Objekte angewendet werden, durch angemessene Metadaten,
- übernimmt die Bereitstellung der digitalen Objekte,
- verfolgt eine Strategie zur Planung und Durchführung von Langzeiterhaltungsmaßnahmen und
- besitzt eine angemessene technische Infrastruktur zur dauerhaften Erhaltung und Sicherung digitaler Objekte.

Das digitale Langzeitarchiv sollte diese wichtigen Kriterien zur Vertrauenswürdigkeit nicht nur erfüllen, sondern dieses auch nach außen darstellen. Das festigt zum einen die Position des Archivs in der wissenschaftlichen Community und stellt die Basis für einen erfolgreichen Betrieb dar, zum anderen wächst mit der Vertrauenswürdigkeit die Akzeptanz des Archivs bei den Datenproduzenten und -nutzern und führt zu einer Motivationssteigerung bei der Datenablage.

Vertrauenswürdigkeit lässt sich in Form einer Zertifizierung nachweisen. Ein Prüfsiegel ist beispielsweise das Data Seal of Approval[131], welches über eine Art Assessment-Verfahren erlangt werden kann.

C 7. Perspektiven und Realisierung

Das in dieser Studie entworfene Konzept eines vernetzten Langzeitarchivs zur Speicherung chemischer Forschungsdaten ist aufgrund der Komplexität der Aufgabe nur in mehreren, aufeinanderfolgenden Entwicklungsphasen zu realisieren.

Phase 1	Implementierung	Aufbau Repository-Prototyp für Bitstream-Preservation
Phase 2	Erprobung	Inbetriebnahme des Prototyps und Optimierung, Kooperation mit ausgewählten Arbeitsgruppen
Phase 3	Etablierung	Erweiterung auf alle Universitäten, Skalierung und Weiterentwicklung
Phase 4	Umwandlung	Überführung des Repositories in Langzeitarchiv, Entwicklung einer Langzeitarchivierungsstrategie
Phase 5	Dauerbetrieb	Zertifizierung, Übertragung auf andere Disziplinen
Flexibel	Vernetzung	Harvester, Zentrales Forschungsdatenportal, Schnittstellen zu Institutionellen Daten-Repositories

Tabelle 4: Phasen der Realisierung einer vernetzten Forschungsdaten-Infrastruktur

Die Realisierung des vernetzten Langzeitarchivs startet mit einer *Implementierungsphase*, in der zunächst ein Repository-Prototyp entwickelt wird. Der Prototyp basiert auf einer bereits voll funktionsfähigen Archiv-Basisarchitektur und kann Daten aufnehmen und wieder bereitstellen. So erfolgt in der Implementierungsphase neben dem Aufbau des Hardware-Systems eine intensive Softwareentwicklung. Aufbauend auf einer Repository-Installation sind leicht zu bedienende Werkzeuge für die Wissenschaftler mit innovativen Funktionalitäten für die Datenablage und -bereitstellung zu entwickeln. Am Ende dieser ersten Entwicklungsphase steht ein Repository zur Verfügung, welches den Erhalt des Bitstreams garantiert, also eine dauerhafte und sichere Speicherung von Daten ermöglicht. Aspekte der Langzeitarchivierung müssen bei der Entwicklung des Prototyps zwar bereits berücksichtigt werden, aber es existiert zu diesem Zeitpunkt noch keine konkrete Langzeitarchivierungsstrategie.

In einer sich der Implementierungsphase anschließenden *Erprobungsphase* wird der Repository-Prototyp anhand von ausgewählten Arbeitskreisen aus dem Bereich der Synthesechemie evaluiert. Die Wissenschaftler dieser Arbeitskreise erhalten Zugang zum Prototyp und füllen das Repository mit ersten Datensätzen. Die aus einer engen Kooperation

mit den Wissenschaftlern resultierenden Verbesserungswünsche fließen in die Weiterentwicklung und Optimierung des Repositories ein. Am Ende dieser Phase ist der Prototyp so weit entwickelt, dass ein realer Einsatz im wissenschaftlichen Betrieb bereits möglich ist.

Deshalb sollte in der nachfolgenden *Etablierungsphase* der Zugang zum Repository auf alle deutschen Universitäten ausgedehnt werden. An diesem Punkt existiert ein System zur Datenspeicherung mit garantierter Bitstream Preservation, aber eben noch kein Langzeitarchiv, welches die Nachnutzbarkeit, Qualität und Pflege der Daten garantiert. Das System sollte dennoch bereits zu diesem Zeitpunkt mit Daten gefüllt werden, weil die Dauer der Entwicklung von Langzeitarchivierungsstrategien die Lebenszeit der Datenformate unterschreitet. In dieser Phase ist nicht nur die technische Weiterentwicklung des Repositories wichtig, sondern zum einen muss mit Unterstützung von Fachgesellschaften, wie der GDCh, und weiteren Organisationen das Bewusstsein der Wissenschaftler für die Relevanz dieser Problematik geschärft werden und zum anderen müssen Anreize für den Datenproduzenten zur Datenablage geschaffen werden. Ein erfolgreicher Betrieb eines Archivs ist nicht nur abhängig vom Archivbetreiber selbst, sondern viel mehr noch von den Datenproduzenten und Datennutzern. Das Repository für chemische Forschungsdaten muss in die Wissenschaftslandschaft und den Forschungsalltag eingebettet werden.

An die Erprobungsphase schließt sich die *Umwandlungsphase* an, in der das Repository in ein dauerhaft betriebenes, skalierbares Langzeitarchiv überführt wird. Es müssen Langzeitstrategien entwickelt werden, die über den reinen Bitstream-Erhalt weit hinausgehen. Auf Software-Seite sind Verfahren zu implementieren, mit denen die rasante technologische Weiterentwicklung bewältigt werden kann. Des Weiteren bedarf es einer dokumentierten Organisation des Archivs als Zusammenspiel aus Mensch und Technik. Die Langzeitarchivierung stellt eine langfristige, große, administrative und auch verantwortungsvolle Aufgabe für den Betreiber des Archivs dar.

Am Ende dieser Kette steht der *Dauerbetrieb* eines gut organisierten, dokumentierten Langzeitarchivs, welches dauerhaft die Verantwortung für die Daten – deren Pflege, Integrität, Authentizität, Verfügbarkeit und Nachnutzbarkeit – übernommen hat. Aber nicht nur die Langzeitarchivierung selbst ist eine Aufgabe mit neuen Herausforderungen, sondern auch für das Datenarchiv und seine Funktionalitäten wird es zukünftig einen hohen Anpassungs- und Weiterentwicklungsbedarf geben.

So müssen z. B. spätestens zu dem Zeitpunkt, wenn die Universitäten Institutionelle Daten-Repositories für ihre Wissenschaftler zur Verfügung stellen, im Langzeitarchiv die zur *Vernetzung* mit den Repositories notwendigen Schnittstellen implementiert sein. Harvester, zentrales Forschungsdatenportal und eine Repository-übergreifende Suche müssen flexibel

und nach Bedarf entweder parallel zum Aufbau des Archivs oder auch erst im Anschluss daran entwickelt und bereitgestellt werden.

Auf langfristige Sicht existieren für das Archiv folgende Entwicklungsperspektiven:

- Erweiterung auf andere Teilgebiete der Chemie:
  Der Archiv-Prototyp fokussiert sich vorrangig auf die im Bereich der Synthesechemie anfallenden Daten. Langfristig sollte das Archiv auf andere Teilgebiete der Chemie erweitert werden.
- Übertragung auf andere Disziplinen:
  Nach erfolgreichem Aufbau und Etablierung des Archivs ist die Übertragbarkeit der entworfenen technischen Konzepte und Lösungen auf andere Disziplinen zu prüfen. Alternativ könnte sich das Archiv für chemische Forschungsdaten auch für andere Fächer öffnen.
- Aufbau einer vernetzten Forschungsdaten-Infrastruktur:
  Im Rahmen der Realisierung einer vernetzten Forschungsdaten-Infrastruktur werden zukünftig immer mehr Universitäten Institutionelle Repositories für die Datenablage zur Verfügung stellen. Deshalb wird es von Relevanz sein, derartige Daten-Repositories als Input-Quelle mit dem Langzeitarchiv zu verknüpfen und einen automatischen Datentransfer zu ermöglichen.
- Zertifizierung:
  Langfristig sollte ein vertrauenswürdiges Langzeitarchiv aufgebaut werden, welches aufgrund der nachgewiesenen Einhaltung definierter Qualitätskriterien eine Zertifizierung erhält. Die Bewertungskriterien für eine Zertifizierung des Archivs zielen nicht auf die Beurteilung eines technischen Produkts, sondern eines digitalen Archivs als Gesamtsystem aus Strategien, Methoden, Organisation, Funktionen, Prozessen, Management und technischer Infrastruktur. Eine Zertifizierung, die die Vertrauenswürdigkeit des Archivs nachweist und das Risiko für einen Wertverlust der Archivinhalte minimiert, ist eine wichtige Voraussetzung für die Akzeptanz des Datenarchivs in der chemischen Community und für seine Nutzung.

Der Aufbau eines digitalen und zudem vernetzten Langzeitarchivs, zugeschnitten auf die fachspezifischen Bedürfnisse der Wissenschaftler und versehen mit innovativen Funktionalitäten, stellt aufgrund seiner Komplexität eine disziplinübergreifende Aufgabe dar, die nur durch eine enge Zusammenarbeit von Fachwissenschaftlern verschiedener Disziplinen gelöst werden kann.

Abbildungsverzeichnis

Abkürzungsverzeichnis

Abb.	Abbildung
AIP	Archival Information Package
AKS	Altman & King Schema
ANDS	Australian National Data Service
AWI	Alfred-Wegener-Institut für Polar- und Meeresforschung
BABS	Bibliothekarisches Archivierungs- und Bereitstellungssystem
BMBF	Bundesministerium für Bildung und Forschung
bzw.	beziehungsweise
CC	Creative Commons
CISTI	Canada Institute for Scientific and Technical Information
CRL	Center for Research Libraries
CSD	Cambridge Structural Database
DANS	Data Archiving and Networked Services
DC	Dublin Core
DCMI	Dublin Core Metadata Initiative
DCC	Digital Curation Center
DFG	Deutsche Forschungsgemeinschaft
DINI	Deutsche Initiative für Netzwerkinformation
DIP	Dissemination Information Package
DMP	Datenmanagementplan
DOI	Digital Object Identifier
DPE	DigitalPreservationEurope
FID	Flammenionisationsdetektor
FIZ	Fachinformationszentrum
GDCh	Gesellschaft Deutscher Chemiker
GFZ	Geoforschungszentrum Potsdam
GWDG	Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen
HSM	Hierarchisches Speichermanagement
ICDP	International Continental Scientific Drilling Program
ICSU WDCs	International Council for Science World Data Centers
IDC	International Data Cooperation
IDF	International DOI Foundation
InChI	International Chemical Identifier
INIST	L'Institut de l'Information Scientifique et Technique
IR	Infrarot
JCAMP-DX	Joint Committee on Atomic and Molecular Physical Data Exchange
Kap.	Kapitel
KoLaWiss	Kooperative Langzeitarchivierung für Wissenschaftsstandorte
koLibRI	kopal Library for Retrieval and Ingest
kopal	Kooperativer Aufbau eines Langzeitarchivs digitaler Information
LDR	Labelled Data Record
MARC	Machine-Readable Cataloging
MARUM	Zentrum für Marine Umweltwissenschaften
MDZ	Münchener Digitalisierungszentrum

METS	Metadata Encoding & Transmission Standard
MIT	Massachusetts Institute of Technology
NASA	National Aeronautics and Space Administration
nestor	Network of Expertise in Long-term Storage of Digital Resources
netCDF	Network Common Data Format
NFS	Network Files System
NMR	nuclear magnetic resonance
OAI	Open Archives Initiative
OAI-MPH	Open Archives Initiative Protocol for Metadata Harvesting
OAIS	Open Archival Information System
OCLC	Online Computer Library Center
OECD	Organisation for Economic Co-operation and Development
OSG	Operational Support Group ICDP
PANGAEA	Publishing Network for Geoscientific & Environmental Data
PI	Persistent Identifier
PICA	Project of Integrated Catalogue Automation
PLANETS	Preservation and Long-term Access through Networked Services
PLM	Preservation Layer Model
PREMIS	PREservation Metadate: Implementation Strategies
RDF	Resource Description Framework
RLG	Research Libraries Group
SAM-FS/QFS	Storage- and Archive-Manager / Quick File System
SC	Science Commons
SDDB	Scientific Drilling Database
SIP	Submission Information Package
SPECTRa	Submission, Preservation and Exposure of Chemistry Teaching and Research Data
SUB Göttingen	Staats- und Universitätsbibliothek Göttingen
TByte	Terabyte
TIB	Technische Informationsbibliothek
TRAC	Trustworthy Repositories Audit & Certification
TSM	Tivoli Storage Manager
u. a.	unter anderem
URN	Uniform Resource Name
USA	United States of America
usw.	und so weiter
UV	Ultraviolett
UVC	Universal Virtual Computer
VPN	Virtual Private Network
WDC MARE	World Data Center for Marine Environmental Sciences
www	World Wide Web
XML	Extensible Markup Language
z. B.	zum Beispiel
z. T.	zum Teil
ZBMed	Deutsche Zentralbibliothek für Medizin
ZEND	Zentrale Erfassungs- und Nachweisdatenbank

Literatur

[1] Vincent S. Smith, Data publication: towards a database of everything, BMC Research Notes 2009, 2, 113

[2] http://www.ccdc.cam.ac.uk/

[3] http://www.fiz-karlsruhe.de/icsd.html

[4] Christoph Steinbeck, Stefan Kuhn , NMRShiftDB – compound identification and structure elucidation support through a free community-built web database, *Phytochemistry*, *Volume 65, Issue 19*, *October 2004*, *Pages 2711-2717*

[5] *http://www.tib-hannover.de/de/dietib/aktuelles/aktuelles/id/120/*

[6] H. T. Tran-Thien, *Diplomarbeit,* Paderborn **2006**

[7] K. Krohn, H. T. Tran-Thien, J. Vitz, A. Vidal, *Eur.J.Org.Chem*. **2007**, 1905–191. **DOI:10.1002/ejoc.200700015**

[8] Hoang Trang Tran-Thien, Dissertation „Synthesen auf dem Gebiet der Anthrapyran-Antibiotika“, Paderborn **2010**

[9] A. Treloar "The Curation Continuum and its application within the Australian National Data Service (ANDS)", http://oa.helmholtz.de/fileadmin/user_upload/Data_Continuum/treloar.pdf

[10] *http://www.tib-hannover.de/de/dietib/aktuelles/aktuelles/id/120/*

[11] *The PICA webpage,* http://oclcpica.org/

[12] *The Marc 21 webpage*, http://www.loc.gov/marc/

[13] *Dublin Core Metadata Element Set*, Version 1.1 http://dublincore.org/documents/dces/

[14] *DCMI Metadata* Terms, http://dublincore.org/documents/dcmi-terms/

[15] http://www.iso.org/iso/catalogue_detail.htm?csnumber=25921

[16] Brase, J. (2004) Using Digital Library Techniques - Registration of Scientific Primary Data. Lecture Notes in Computer Science 3232, 488-494.

[17] Altman M., King G., *A Proposed Standard for the Scholarly Citation of Quantitative Data*, D-lib Magazine, March/April 2007, Vol 13 No.3/4

[18] Green, T (2009), *We Need Publishing Standards for Datasets and Data Tables*, OECD Publishing White Paper, OECD Publishing. DOI: 10.1787/603233448430

[19] JCAMP-DX web site, http://www.jcamp-dx.org

[20] R. S. McDonald, P. A. Wilks, *Appl Spectrosc* **1988** *42,* 151-162.

[21] Draft Reports, http://www.jcamp-dx.org/drafts/JCAMP6_2b%20Draft.pdf

[22] JCAMP-DX V.6.00 Draft Report, http://www.jcamp-dx.org/drafts/Chromatography%20&%20MS%20JCAMP-DX%206%20-%20Draft%2031%20May%202005.pdf

[23] Freie Spektroskopiesoftware, http://www.effemm2.de/spekwin/spekwin_links_en.html

[24] Datenvisualisierung und Formatkonversion für MS, http://www.ms-utils.org/wiki/pmwiki.php/Main/SoftwareList

[25] JSpecView, http://jspecview.sourceforge.net

[26] R. J. Lancashire, Chemistry Central Journal **2007**, 1

[27] JSpec View, JSpecView 1.0.20060627-2100

[28] JCAMP-DX bei Wareseeker, http://wareseeker.com/free-jcamp/

[29] FID to JCAMP-DX, FIDtoJCAMP 0.6a

[30] Spekwin, http://wareseeker.com/Home-Education/spekwin32-1.71.3.zip/7de238f32

[31] Produktinformation Grams/Al, http://www.chemie.de/products/d/82952/

[32] Produktinformation ACD-Labs, http://www.acdlabs.com/products/adh/

[33] International Union of Crystallography, http://www.iucr.org/resources/cif

[34] S. R. Hall, F. H Allen, I. D. Brown, *Acta Cryst.* **1991** A47, 655-685

[35] S. Dohmeier-Fischer, G. Fels, *Nachrichten aus der Chemie* **2010**, 58, 650

[36] DFG Denkschrift, Sicherung guter wissenschaftlicher Praxis, Empfehlung 7: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf, Zugriff am 10.06.2010

[37] http://archives.arte.tv/hebdo/archimed/19990504/dtext/sujet1.html, Zugriff am 10.06.2010

[38] Fels, G., Dohmeier-Fischer, S., Umfragebericht: Fragebogen zum Umgang mit Primärdaten in der Chemie, **2009**

[39] InChI: http://www.iupac.org/inchi/, Zugriff am 10.06.2010

[40] nestor: http://www.langzeitarchivierung.de/, Zugriff am 10.06.2010

[41] nestor-Handbuch: Eine kleine Enzyklopädie der Langzeitarchivierung, **2009**: http://nestor.sub.uni-goettingen.de/handbuch/index.php, Zugriff am 10.06.2010

[42] PLANETS: http://www.planets-project.eu/, Zugriff am 10.06.2010

[43] Open Planets Foundation: http://www.openplanetsfoundation.org/, Zugriff am 10.06.2010

[44] World Data Center System der ICSU: http://www.ngdc.noaa.gov/wdc/, Zugriff am 10.06.2010

[45] AWI: http://www.awi.de/de/, Zugriff am 10.06.2010

[46] marum: http://www.marum.de/, Zugriff am 10.06.2010

[47] PANGAEA: http://www.pangaea.de/, Zugriff am 10.06.2010

[48] WDC-MARE: http://www.wdc-mare.org/, Zugriff am 10.06.2010

[49] GFZ: http://www.gfz-potsdam.de/portal/gfz/home, Zugriff am 10.06.2010

[50] Scientific Drilling Database: http://dc110dmz.gfz-potsdam.de/contenido/lakedb/front_content.php, Zugriff am 10.06.2010

[51] Klump, J., Conze R., The Scientific Drilling Database (SDDB) – Data from Deep Earth Monitoring and Sounding, *Scientific Drilling*, (4), 30-31, **2007**, doi:10.2204/iodp.sd.4.06.2007

[52] OSG ICDP: http://www.icdp-online.org/front_content.php?idcat=341, Zugriff am 10.06.2010

[53] Creative Commons (CC): http://de.creativecommons.org/, Zugriff am 17.06.2010

[54] DataCite: http://www.datacite.org/, Zugriff am 15.06.2010

[55] kopal: http://kopal.langzeitarchivierung.de/index.php.de, Zugriff am 10.06.2010

[56] Deutsche Nationalbibliothek: http://www.d-nb.de/, Zugriff am 10.06.2010

[57] GWDG: http://www.gwdg.de/, Zugriff am 10.06.2010

[58] DIAS: http://www-935.ibm.com/services/nl/dias/is/implementation_services.html, Zugriff am 10.06.2010

[59] IBM: http://www.ibm.com/, Zugriff am 10.06.2010

[60] koLibRI: http://kopal.langzeitarchivierung.de/index_koLibRI.php.de, Zugriff am 10.06.2010

[61] SUB: http://www.sub.uni-goettingen.de/, Zugriff am 10.06.2010

[62] KoLaWiss: http://kolawiss.uni-goettingen.de/, Zugriff am 10.06.2010

[63] BABS: http://www.babs-muenchen.de/, Zugriff am 10.06.2010

[64] Bayerische Staatsbibliothek: http://www.bsb-muenchen.de/, Zugriff am 10.06.2010

[65] Leibniz-Rechenzentrum: http://www.lrz-muenchen.de/, Zugriff am 10.06.2010

[66] ZEND: http://www.digitale-sammlungen.de/index.html?c=digitalisierung-zend&l=de, Zugriff am 15.06.2010

[67] DigiTool: http://www.exlibrisgroup.com/category/DigiToolOverview, Zugriff am 10.06.2010

[68] TSM von IBM: http://www-01.ibm.com/software/tivoli/products/storage-mgr/, Zugriff am 10.06.2010

[69] BABS2: http://www.lrz-muenchen.de/projekte/langzeitarchivierung/babs21.html, Zugriff am 10.06.2010

[70] Digitales Archiv: http://www.bundesarchiv.de/fachinformationen/00895/index.html.de, Zugriff am 10.06.2010

[71] Treloar, A., Data management and the curation continuum: how the Monash experience is informing repository relationships, VALA2008 Conference, **2008**: http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf, Zugriff am 10.06.2010

[72] OAI-PMH: http://www.openarchives.org/pmh/, Zugriff am 10.06.2010

[73] Informationsplattform Open Access: http://www.open-access.net/, Zugriff am 10.06.2010

[74] DRIVER: http://search.driver.research-infrastructures.eu/, Zugriff am 10.06.2010

[75] OAIster: http://oaister.worldcat.org/, Zugriff am 10.06.2010

[76] BASE: http://www.base-search.net/, Zugriff am 10.06.2010

[77] DINI: http://www.dini.de/, Zugriff am 10.06.2010

[78] DINI-Zertifikat Dokumenten- und Publikationsservice 2007: http://edoc.hu-berlin.de/series/dini-schriften/2007-3/PDF/3.pdf, Zugriff am 15.06.2010

[79] Ullrich, D., nestor-Handbuch, Kapitel 8.2, Bitstream Preservation, **2009**: http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_346.pdf, Zugriff am 15.06.2010

[80] VMWare: http://www.vmware.com/resources/compatibility/pdf/VMware_OS_Compatibility_Guide.pdf , Stand 17.06.2010

[81] Stromasys: http://www.stromasys.ch/products/charon-axp/, Zugriff am 17.06.2010

[82] IDC: http://www.idc.de/, Zugriff am 14.06.2010

[83] Hartwich, R., Sun SAM-FS / QFS intelligentes Data Management: http://www.init-bs.com/download/SunStorageDay_SAM-FS-Roadshow.pdf, Zugriff am 15.06.2010

[84] Borghoff, U. M. et al., nestor-materialien 3: Vergleich bestehender Archivierungssysteme, **2005**, http://files.d-nb.de/nestor/materialien/nestor_mat_03.pdf, Zugriff am 15.06.2010

[85] Niederländische Nationalbibliothek: http://www.kb.nl/index-en.html, Zugriff am 10.06.2010

[86] Rosetta: http://www.exlibrisgroup.com/category/RosettaOverview, Zugriff am 10.06.2010

[87] British Library: http://www.bl.uk/, Zugriff am 10.06.2010

[88] DSpace: http://www.dspace.org/, Zugriff am 10.06.2010

[89] DuraSpace: http://duraspace.org/index.php, Zugriff am 10.06.2010

[90] Smithsonian Institution: http://si-pddr.si.edu/dspace/, Zugriff am 10.06.2010

[91] Pandektis at National Research Foundation of Greece: http://pandektis.ekt.gr/dspace/?locale=en, Zugriff am 10.06.2010

[92] Texas Digital Library: http://repositories.tdl.org/, Zugriff am 10.06.2010

[93] EPrints: http://www.eprints.org/, Zugriff am 10.06.2010

[94] Organic EPrints: http://orgprints.org, Zugriff am 10.06.2010

[95] Fedora: http://www.fedora-commons.org/, Zugriff am 10.06.2010

[96] eSciDoc: http://www.escidoc.org/, Zugriff am 10.06.2010

[97] RODA: http://portal.roda.dgarq.gov.pt/en, Zugriff am 10.06.2010

[98] MyCoRe: http://www.mycore.de/, Zugriff am 10.06.2010

[99] Papyrus-Projekt Halle-Jena-Leipzig: http://papyri.uni-leipzig.de/, Zugriff am 10.06.2010

[100] VMWare View: http://www.vmware.com/de/products/view/, Zugriff am 17.06.2010

[101] Klump, J., Managing the Data Continuum, **2009**: http://oa.helmholtz.de/fileadmin/user_upload/Data_Continuum/klump.pdf, Zugriff am 10.06.2010

[102] ACD/Labs: http://www.acdlabs.com/products/auto_int/auto/auto_serv/, Zugriff am 17.06.2010

[103] U.S. National Library of Medicine: http://www.nlm.nih.gov/mesh/meshhome.html, Zugriff am 14.06.2010

[104] RSC Ontologies: http://www.rsc.org/ontologies/, Zugriff am 14.06.2010

[105] JChemPaint: http://sourceforge.net/apps/mediawiki/cdk/index.php?title=JChemPaint, Zugriff am 10.06.2010

[106] OrChem: http://orchem.sourceforge.net/, Zugriff am 10.06.2010

[107] checkmol: http://merian.pch.univie.ac.at/~nhaider/fga.php, Zugriff am 10.06.2010

[108] JSpecView: http://sourceforge.net/projects/jspecview/, Zugriff am 10.06.2010

[109] ImageJ: http://rsb.info.nih.gov/ij/, Zugriff am 10.06.2010

[110] DFN-AAI: https://www.aai.dfn.de/, Zugriff am 10.06.2010

[111] Shibboleth: http://shibboleth.internet2.edu/, Zugriff am 10.06.2010

[112] DOI: http://www.icdp-online.org/contenido/std-doi/front_content.php, Zugriff am 10.06.2010

[113] CCSDS: Reference Model for an Open Archival Information System (OAIS), **2002**: http://public.ccsds.org/publications/archive/650x0b1.pdf, Zugriff am 15.06.2010

[114] CCSDS: Producer-Archive Interface Methodology Abstract Standard (PAIMAS), **2004**: http://public.ccsds.org/publications/archive/651x0b1.pdf, Zugriff am 15.06.2010

[115] nestor-Materialien 10, Wege ins Archiv – Ein Leitfaden für die Informationsübernahme in das digitale Langzeitarchiv, **2008**: http://files.d-nb.de/nestor/materialien/nestor_mat_10.pdf, Zugriff am 15.06.2010

[116] Van der Werf, T., The Deposit System for Electronic Publications: A Process Model, **2000**: http://nedlib.kb.nl/results/DSEPprocessmodel.pdf, Zugriff am 14.06.2010

[117] KoLaWiss: AP 3 – Technik, **2009**: http://kolawiss.uni-goettingen.de/projektergebnisse/AP3_Report.pdf, Zugriff am 14.06.2010

[118] Knight, G., Anderson, S., SHERPA DB: Final report of the SHERPA DP project, **2007**: http://www.sherpadp.org.uk/documents/sherpadp_finalreport.pdf, Zugriff am 15.06.2010

[119] Handle-System: http://www.handle.net/, Zugriff am 10.06.2010

[120] IDF: http://www.doi.org/, Zugriff am 10.06.2010

[121] DOI Info & Guidelines: http://www.crossref.org/01company/15doi_info.html, Zugriff am 15.06.2010

[122] METS: http://www.loc.gov/standards/mets/, Zugriff am 10.06.2010

[123] PREMIS: http://www.loc.gov/standards/premis/, Zugriff am 10.06.2010

[124] LMER http://www.d-nb.de/standards/lmer/, Zugriff am 10.06.2010

[125] Caplan, P., Understanding PREMIS, Library of Congress Network Development and MARC Standards Office, **2009**: http://www.loc.gov/standards/premis/understanding-premis.pdf, Zugriff am 14.06.2010

[126] PLATO: http://www.ifs.tuwien.ac.at/dp/plato/intro.html, Zugriff am 10.06.2010

[127] Green, A., Macdonald, S., Rice, R., Policy-making for Research Data in Repositories: A Guide, **2009**: http://www.disc-uk.org/docs/guide.pdf, Zugriff am 14.06.2010

[128] Science Commons (SC): http://sciencecommons.org/, Zugriff am 17.06.2010

[129] nestor-Materialien 8, Kriterienkatalog vertrauenswürdige digitale Langzeitarchive - Version 2, **2008**: http://www.langzeitarchivierung.de /publikationen/expertisen/expertisen.htm#nestor-materialien8, Zugriff am 15.06.2010

[130] Trustworthy Repositories Audit & Certification: Criteria and Checklist, **2007**: http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf, Zugriff am 10.06.2010

[131] Data Seal of Approval: http://www.datasealofapproval.org/?q=frontpage, Zugriff am 10.06.2010