Handbuch Forschungsdatenmanagement

Herausgegeben von
Stephan Büttner, Hans-Christoph Hobohm, Lars Müller

BOCK + HERCHEN Verlag
Bad Honnef
2011

ISBN 978-3-88347-283-6

BOCK+HERCHEN Verlag, Bad Honnef
Printed in Germany

3.1 Institutionalisierte „*Data Curation Services*“

Michael Lautenschlager
ICSU World Data Center Climate
Deutsches Klimarechenzentrum GmbH

Der Lebenszyklus wissenschaftlicher Daten (*Data Life Cycle*) durchläuft die Stadien Erzeugung, Bearbeitung, Archivierung und Wiederverwendung. Die Wiederverwendung von Daten führt zur Erzeugung neuer Daten und leitet den nächsten Zyklus ein.

Die Archivierung von Daten ist hier nicht zu verstehen als Zwischenspeicherung von Ergebnissen im Rahmen eines wissenschaftlichen Bearbeitungsprozesses, sondern als Langzeitarchivierung mit dem Ziel, die archivierten Daten auch nach vielen Jahren für wissenschaftliche, interdisziplinäre Nachnutzung bereitzustellen (Wiederverwendung). Im Rahmen der Regeln zur Sicherung guter wissenschaftlicher Praxis der Forschungsgesellschaften werden hier Zeiträume von 10 Jahren und mehr gefordert. Zentrale Forderung in der Langzeitarchivierung ist die Sicherstellung der Datenintegrität. Im Zeitalter der elektronischen Speicherung von Daten ist das ein aktiver Prozess, der kontinuierliche Pflege der archivierten Daten erfordert.

Wesentliche Bestandteile der Integritätssicherung elektronischer Daten sind:

- Sicherstellung der Unversehrtheit (*Bit-stream Preservation*)
- Sicherstellung der Lesbarkeit
- Sicherstellung der Interpretierbarkeit

Für eine Wiederverwendung wissenschaftlicher Daten müssen diese identisch zum Original sein, aus dem *Bit-stream* müssen Informationen (z. B. Ziffern und Buchstaben) auslesbar sein und diese Ziffern und Buchstaben müssen interpretiert werden können, um die ursprünglichen Informationen wieder zu gewinnen. „*Data Curation Services*“ unterstützen die Integritätssicherung elektronischer Daten. Sicherstellung der Les- und Interpretierbarkeit elektronischer Daten ist eine Voraussetzung für eine zukünftige Überprüfung wissenschaftlicher Ergebnisse. Damit unterstützen „*Data Curation Services*“ direkt die Regeln zur guten wissenschaftlichen Praxis, wie sie von den Wissenschaftsgesellschaften formuliert wurden.

3.1.1 Sicherstellung der Unversehrtheit

Sicherstellung der Unversehrtheit oder auch „*Bit-stream Preservation*“ garantiert die Bit-genaue Erhaltung der archivierten Daten ohne Ansehen ihres Inhalts. Die Umsetzung dieser Anforderung wird durch technische und organi-

satorische Maßnahmen im Arbeitsablauf der Archivierung und der Archivbetreuung erreicht.

Einem möglichen Datenverlust wird vorgebeugt durch Sicherheitskopien der archivierten Datenentitäten, die mit unterschiedlichen Technologien erstellt und an unterschiedlichen Orten gespeichert werden. Die Anzahl der Sicherheitskopien hängt ab, vom Grad der Sicherheit, der erreicht werden soll, dem Aufwand, den man treiben kann, und zum Teil auch von gesetzlichen Vorschriften die erfüllt werden müssen. Wird an dieser Stelle noch Dokumentensicherheit gefordert, muss zusätzlich noch sichergestellt werden, dass die Datenentitäten nachträglich nicht mehr verändert werden können.

Einem möglichen Datenverlust bei Speicherung auf Festplatten wird häufig durch Speicherung in der *Raid* (*Redundant Array of Independent Disks*) Architektur vorgebeugt. Dabei wird in der Praxis unterschieden zwischen *Raid* 0, keine Sicherung, *Raid* 1, komplette Spiegelung der Daten, und *Raid* 5 bzw. *Raid* 6, bei denen redundante Informationen zur Rekonstruktion der Daten bei Verlust auf einer oder zwei zusätzlichen Festplatten abgespeichert werden.

Die Unversehrtheit von Datenentitäten kann geprüft werden durch Prüfsummen bzw. *Check Sums*. Dabei wird jeder Datenentität während ihrer Erzeugung eine Bit-genaue Prüfsumme mitgegeben, deren Unveränderlichkeit die Unversehrtheit einer Datenentität nach Kopiervorgängen oder Transport über das Netzwerk garantiert. Ist die Prüfsumme vor und nach der Operation identisch, ist auch die zugehörige Datenentität identisch. Ein einfaches Beispiel für eine Prüfsumme ist die Quersumme der Ziffern einer Zahl. Allerdings werden mit diesem Verfahren beispielsweise „Zahlendreher“, also ein häufig vorkommender Fehler in der Eingabe von numerischen Informationen durch Menschen, nicht erkannt.

Eine weitere Maßnahme zur Sicherung der Unversehrtheit von Datenentitäten ist die regelmäßige Erneuerung von elektronischen Speichermedien. Alle Speichermedien unterliegen einem Alterungsprozess, der gemessen wird in Standzeit oder Nutzungsintensität. Wird das „Haltbarkeitsdatum“ überschritten, werden in Langzeitarchiven die betroffenen Datenentitäten automatisch und weitgehend transparent für den Nutzer auf ein frisches Medium kopiert. Bei diesem Prozess findet auch häufig ein Generationswechsel im Speichermedium mit z. B. höherer Kapazität statt. Die Unversehrtheit der kopierten Datenentitäten kann wieder mit Hilfe ihrer Prüfsummen sichergestellt werden. Im Fehlerfall wird der automatische Prozess unterbrochen und manuelles Eingreifen durch einen Betreuer des Datenarchivs (Daten-Kurator) wird erforderlich.

3.1.2 Sicherstellung der Lesbarkeit

Abhängig vom archivierten Datenvolumen und der wissenschaftlichen *Community* werden zwei Strategien zur Sicherstellung der Lesbarkeit von Datenentitäten verfolgt: Formatkonvertierung oder Migration der Lese-Werkzeuge.

Kleinere Datenvolumina werden häufig mit Werkzeugen erzeugt, deren Quelle für die Wissenschaft nicht zugreifbar ist und damit auch deren Ausgabeformate nicht im Detail bekannt oder beeinflussbar sind. Neue Generationen von Werkzeugen bieten häufig neue Ausgabeformate mit Lesbarkeit der oder des alten Formats. Diese Lesbarkeit alter Formate ist aber nicht auf Dauer garantiert. Beispiele solcher Werkzeuge sind die die typischen *Office*-Anwendungen oder Geographische Informationssysteme (GIS), bei denen nicht das Programm selbst gekauft wird, sondern Nutzungsrechte. Zur Sicherstellung der Lesbarkeit dieser Formate über lange Zeiträume müssen *Data Curation Services* existieren, die Datenentitäten mit abgekündigten Formaten aufspüren und in neue Formate wandeln. An dieser Stelle ist die *Bit-stream* Sicherung verletzt (Prüfsummen sind nicht identisch) und zur Sicherung der Datenintegrität sind Qualitätssicherungen erforderlich, die Unversehrtheit des Inhalts sicherstellen. Die Entwicklung solcher Dienste steht noch ganz am Anfang und speziell die wissenschaftlichen Bibliotheken bemühen sich im Rahmen der elektronischen Informationsversorgung um Systematisierung bestehender Ansätze. Formate dieser Kategorie und deren Eignung für die Langzeitarchivierung werden im Detail im NESTOR Handbuch diskutiert. Ein Beispiel für einen aktuellen Standard für elektronische Dokumente ist das PDF (*Portable Document Format*).

Die zweite Kategorie sind große bis sehr große Datenmengen aus den Bereichen numerische Modellierung, Satellitendaten, *Monitoring*-Systeme, Daten aus Biodiversitätsuntersuchungen oder Hochenergiephysik. Allen diesen Daten ist ihre maschinelle Erzeugung gemeinsam, bei der die Datenproduktion nur durch die Leistungsfähigkeit der Maschinen beschränkt wird. Die diskutierten Archivvolumina wachsen alle 3 bis 5 Jahre um eine Größenordnung von Terabyte über aktuell Petabyte zu Exabyte Datenarchiven. Die Archive verwalten zwar riesige Datenmengen, aber sie sind weitgehend homogen und in einer überschaubaren Anzahl von Formaten gespeichert. Diese Formate sind in der jeweiligen Wissenschafts-*Community* definiert, sind Speicherplatz sparend und enthalten häufig Angaben zur weiteren, maschinellen Verarbeitung der Daten (z. B. Variable, Einheit, Raum-Zeit-Bezug). Es sind selbstbeschreibende Binärformate, die nicht wie ASCII Daten direkt lesbar sind, sondern für deren Erzeugung und weiteren Verarbeitung spezielle Computerprogramme (*Libraries* oder Bibliotheken) benötigt werden. Diese Format-Bibliotheken sind in der Verantwortung der jeweiligen Wissenschaftsdisziplin. Aufgrund der Datenmenge verbietet sich hier der oben skizzierte Weg der Formatkonvertierung zur Sicherstellung der Lesbarkeit. Für diesen Typ Daten wird der Weg der Migration der Format-Bibliotheken

von einer Rechnergeneration auf die nächste gewählt, wobei darauf geachtet wird, dass bei Weiterentwicklung des Formatstandards auch alte Daten noch gelesen werden können (Abwärtskompatibilität). Beispiele für diese Formate sind die den Naturwissenschaften verwendeten Formate NetCDF (*Network Common Data Form*) und HDF (*Hierarchical Data Format*). Neben der Unterstützung der *Data Curation Services* erleichtern derartige Standardformate den Datenaustausch und die maschinelle Verarbeitung.

3.1.3 Sicherstellung der Interpretierbarkeit

Dieser Teil der Integritätssicherung ist der am schwierigsten zu standardisierende Bereich, da es hier um (wissenschaftliche) Inhalte von Datenentitäten geht. Während *„Data Curation Services“* die Unversehrtheit und Lesbarkeit elektronischer Daten auch über Disziplingrenzen hinweg sicherstellen, ist eine Disziplin übergreifende Interpretierbarkeit konzeptionell schwierig und erst in Ansätzen realisiert (Beispiel: Verwendung elektronischer Daten in der Klimafolgenforschung). Zu unterschiedlich erscheinen die Begriffswelten und Beschreibungstraditionen, die über die Wissenschaftsdisziplinen hinweg aufeinander abgebildet werden müssten, um eine breitere Umsetzung zu etablieren.

Interpretierbarkeit und Wiederverwendung von Daten ist eng verknüpft mit Kontextinformation bzw. Metadaten („Daten über Daten“). Hier sind Informationen enthalten zum Inhalt der Datenentität, zur technischen Verarbeitung, zur Qualität und zur Datenhistorie (*data provenance*). Die Suche nach Datenentitäten zu bestimmten Begriffen kann ebenfalls durch Metadaten unterstützt werden. Dafür werden die Metadaten in Katalogen zusammengefasst und in Repositorien abgelegt. Standardisierung von Metadaten in Schemata erleichtert die Suche nach Schlüsselbegriffen und steigert die Effizienz. Häufig werden für die Suche in standardisierten Datenkatalogen relationale Datenbanksysteme verwendet. Sowohl Metadaten Schema (auch Datenmodell genannt) als auch der Inhalt der Metadaten sind abhängig von der Wissenschaftsdisziplin (also dem Dateninhalt selbst) und dem Anwendungsgebiet bzw. der Nutzergruppe dieser Metadaten.

Ein Ziel der Langzeitarchivierung muss sein, die Nachnutzbarkeit von Daten sicher zu stellen. Dafür müssen die Metadaten so vollständig sein, dass Datenentitäten unabhängig vom Erzeuger auf ihre Verwendbarkeit hin beurteilt und letztlich auch verwendet werden können. Die so definierte Vollständigkeit ist dabei nicht unabhängig von der Nutzergruppe des Datenarchivs. Grob unterschieden wird

- Fachnutzung durch Wissenschaftler aus der entsprechenden Fachdisziplin,
- interdisziplinäre Nutzung durch Wissenschaftler aus unterschiedlichen Fachdisziplinen und
- Datennutzung im Rahmen Öffentlichkeit und Politik.

Neben dem Dateninhalt selbst stellen diese Nutzergruppen unterschiedliche Anforderungen an die Komplexität von Metadaten.

Metadaten werden häufig in strukturierten Datenmodellen abgelegt zur Unterstützung der Vergleichbarkeit der beschriebenen Datenentitäten. Feste Wertelisten für Schlüsselbegriffe (wie z. B. für Variable oder die Raum-Zeit Zuordnung) unterstützen Datensuche, Datenaustausch und Vergleichbarkeit. Strukturierte Datenmodelle formalisieren und standardisieren die Datendokumentation und erleichtern die Sicherstellung der Interpretierbarkeit im Rahmen der Datenpflege (*Data Curatio*n) im Langzeitarchiv. Im Zuge von Datenpflege und Qualitätssicherung müssen auch Verarbeitungsschritte und Historie der Daten (*data provenanc*e) in den Metadaten aktualisiert werden.

Die Definition strukturierter Datenmodelle und die Festlegung von Wertelisten für Schlüsselbegriffe erscheinen durchführbar innerhalb einer Wissenschafts-*Community*, die mit gleichen Datentypen in einer definierten Begriffswelt arbeitet. Eine interdisziplinäre Verwendung stellt Anforderungen an die Abbildung von Begrifflichkeiten verschiedener Wissenschaftsdisziplinen aufeinander (Ontologien). Namen für bestimmte Größen und Sachverhalte sind durchaus unterschiedlich in verschiedenen Wissenschaftsdisziplinen. Werden Datenarchive für den interdisziplinären Zugriff geöffnet, sollte der Fachspezifik der jeweiligen Ontologien und Begriffswelten Rechnung getragen werden, damit bei einer Datensuche auch die erwarteten Größen geliefert werden.

Eine weitere Komplexitätsstufe in der Sicherstellung der Interpretierbarkeit ergibt sich im Übergang von Datenmanagement zum Informationsmanagement. Im Informationsmanagement werden strukturierte Information (Metadaten und Daten) verknüpft mit unstrukturierter Information wie Zeitschriftenveröffentlichungen, Texten und Grafiken. Während die strukturierten Informationen als Tabellen und Hierarchien in relationalen Datenbanken effizient organisiert werden können, bieten sich für unstrukturierte Informationen XML-Datenbanken und das *Ressource Description Framework* (RDF) an, die erlauben, eine Netzwerktopologie der Information frei zu definieren. Elektronische Datenentitäten können so flexibel verknüpft werden mit Zusatzinformationen wie z. B. graphischen Darstellungen, Literatur oder verwandten Datenquellen. Das so entstehende Informationsnetzwerk unterscheidet sich deutlich von der Matrix-Struktur für Informationen in relationalen Datenbanksystemen.

3.1.4 Data Curation Services

Das Referenzmodell für ein *Open Archival Information System* (OAIS) definiert Informationen, Services und Prozesse, die in einem Langzeitdatenarchiv implementiert werden sollten. Die hier gewählte und diskutierte Schichteneinteilung beginnend mit der Maschinen bezogenen Ebene (Sicherstellung der Unversehrtheit), fortschreitend mit der syntaktischen Ebene (Sicherstellung der Lesbarkeit)

und endend mit der semantischen Ebene (Sicherstellung der Interpretierbarkeit) ist verbunden mit einer Abnahme von Gemeinsamkeiten über die Wissenschafts-Communities hinweg. Während die Sicherstellung der Unversehrtheit als *Curation Service* für alle Daten auf der Ebene der Rechenzentren organisiert wird, wird für die ‚Sicherstellung der Lesbarkeit' bereits in den *Curation Services* zwischen heterogenen Daten geringen Volumens und homogenen Daten großen Volumens unterschieden. In der Ebene ‚Sicherstellung der Interpretierbarkeit' sind bisher erst in einzelnen Wissenschaftsdisziplinen *Curation Services* für Metadaten eingerichtet. Definition und Entwicklung von Ontologien sind in Forschungsgebieten mit inter-disziplinärer Ausrichtung zu beobachten. Die Vereinheitlichung von Strukturen und Metdatenmodellen zur Beschreibung von Daten ist parallel zur Entwicklung von Ontologien in einzelnen Forschungsgebieten zu erkennen. Ein Beispiel ist die EU-Direktive INSPIRE zur Beschreibung und Organisation von Raum bezogenen Daten nicht nur im Forschungsbereich, sondern gerade auch für Behörden und Ämter in den europäischen Mitgliedsstaaten.

Das ICSU *World Data Center Climate* (WDCC) hat als Arbeitsschwerpunkt die Langzeitarchivierung von Klimamodelldaten und verwandter Beobachtungsdaten. Das WDCC unterstützt mit seinen „*Data Curation Services*" alle drei Ebenen der Integritätssicherung elektronischer, wissenschaftlicher Daten. „*Bitstream Preservation*" ist im Rahmen des Massenspeichersystems am Deutschen Klimarechenzentrum (DKRZ) realisiert. Es werden zwei Magnetbandkopien gespeichert und die Anzahl der Bandzugriffe wird protokolliert. Erreicht der Zugriffszähler seinen Maximalwert, wird der Bandinhalt automatisch auf ein frisches Medium kopiert und das alte Magnetband wird im Silo ausgetauscht. Lesbarkeit der Daten im WDCC wird im Rahmen der Qualitätssicherung kontrolliert. Es wird geprüft, ob die gespeicherten Daten lesbar sind und die ausgelesenen Werte den Erwartungswerten der gespeicherten Variablen entsprechen. Die Interpretierbarkeit wird mit dem verwendeten, umfangreichen Metadatenmodell umgesetzt. Ziel ist, dass im WDCC gespeicherte wissenschaftliche Daten auch nach 10 Jahren und mehr direkt und ohne Nachfrage beim Datenautor in wissenschaftlichen Arbeiten verwendet werden können. Besondere Anforderungen an die Datenbeschreibung stellt die Klimafolgenforschung mit ihrem interdisziplinären Ansatz.

Literaturhinweise

CCSDS (Consultive Committee for Space Data Systems), 2002. *OAIS Reference Model for an Open Archival Information System (OAIS).* Blue Book. (Jan. 2002) Online: http://public.ccsds.org/publications/archive/650x0b1.pdf [Zugriff am 09.08.2010].

DFG (Deutsche Forschungsgemeinschaft), 1998. *Vorschläge zur Sicherung guter wissenschaftlicher Praxis.* Denkschrift. Weinheim: Wiley-VCH.

European Commission, o.J. *INSPIRE Directive*. Online: http://inspire.jrc.ec.europa.eu/index.cfm [Zugriff am 14.08.2011].

Neuroth, H. et al. Hrsg., 2010. *NESTOR-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*. (Version 2.3.) Online: http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-2010071949.

Wikipedia, 2011a. *RAID*. (Version vom 12.08.2011, 18:34 h) Online: http://de.wikipedia.org/w/index.php?title=RAID&oldid=92366907 [Zugriff am 09.08.2011].
(Originaldokument: Patterson, D. A. Gibson, G. & Katz, R. H., 1988. *A Case for Redundant Arrays of Inexpensive Disks*. Online: http://www-2.cs.cmu.edu/~garth/RAIDpaper/Patterson88.pdf [Zugriff am 09.08.2011].)

Wikipedia, 2011b. *Prüfsumme*. (Version vom 19.04.2011, 21:45 h) Online: http://de.wikipedia.org/w/index.php?title=Pr%C3%BCfsumme&oldid=87905903 [Zugriff am 14.08.2011].
(Rivest, R., 1992. *The MD5 Message-Digest Algorithm*. (Network Working Group – Request for Comments: 1321). Online: http://people.csail.mit.edu/rivest/Rivest-MD5.txt [Zugriff am 09.08.2011].)

Wikipedia, 2011c. *PDF*. (Version vom 08.08.2011, 15:31 h) Online: http://de.wikipedia.org/w/index.php?title=Portable_Document_Format&oldid=92200868 [Zugriff am 09.08.2011]
(Adobe Systems, 2011. *PDF Reference and Adobe Extensions to the PDF Specification*. Online: http://www.adobe.com/devnet/pdf/pdf_reference.html [Zugriff am 09.08.2011].)

Wikipedia, 2011d. *NetCDF*. (Version vom 07.04.2011, 10:09 h) Online: http://de.wikipedia.org/w/index.php?title=NetCDF&oldid=87398911 [Zugriff am 14.08.2011]
(Unidata Program Center, o.J. *NetCDF (Network Common Data Form)*. Online: http://www.unidata.ucar.edu/software/netcdf/ [Zugriff am 14.08.2011].)

Wikipedia, 2011e. *HDF.* (Version vom 18.06.2011, 9:08 h) Online: http://de.wikipedia.org/w/index.php?title=Hierarchical_Data_Format&oldid=90174104
(HDF Group, o.J. *Welcome*. Online: http://www.hdfgroup.org/ [Zugriff am 14.08.2011].)

Wikipedia, 2011f. *Resource Description Framework.* (Version vom 16.06.2011, 10:14 h) Online: http://de.wikipedia.org/w/index.php?title=Resource_Description_Framework&oldid=90099766 [Zugriff am 09.08.2011]. (W3C, 2010. Resource Description Framework (RDF). (Stand: 07.03.2010, 7:34 h) Online: http://www.w3.org/RDF/ [Zugriff am 14.08.2011].)

Wissenschaftlicher Rat der Max-Planck-Gesellschaft, 2001. *Verantwortliches Handeln in der Wissenschaft – Analysen und Empfehlungen.* (Max-Planck-Forum Bd. 3). München: Max-Planck-Gesellschaft.