Datenpublikation im Rahmen von Digital Classics Online

Hinweise an die Datengeber

Verantwortlich für den Lebenszyklus der Forschungsdaten sind primär die Autoren der Beiträgen des elektronischen Journals Digital Classics Online (DCO)[1].

Teil eines jeden Forschungsprojektes sollte ein Plan für das Datenmanagement sein, von der Erzeugung bis zur endgültigen Löschung der Forschungsdaten. Dieser Plan stellt den Zugriff und die Nutzung unter Einhaltung von ethischen und Open Access-Prinzipien unter geeigneten Sicherheitsmaßnahmen sicher.

DCO kann im Rahmen der Veröffentlichung von Forschungsergebnissen über Open Journals Systems (OJS)[2] bei der Archivierung und Veröffentlichung von 'Forschungsdaten gemäß der Grundsätze von Open Access, wie sie in der „Berliner Erklärung über offenen Zugang zu wissenschaftlichem Wissen“ von 2003 beschrieben sind, in Zusammenarbeit mit dem Kompetenzzentrum Forschungsdaten (KFD) behilflich sein.

Welche Daten sind Forschungsdaten im Rahmen von DCO?

Die in den Altertumswissenschaften erzeugten Forschungsdaten sind so mannigfaltig, wie die Fachdisziplinen selbst. Sie umfassen unter anderem

- digitale Fotografien, Luft- und Satellitenbilder;
- Textkorpora;
- Datenbanken und Statistiken;
- Vermessungsdaten, Punktwolken und Fotogrammetriedaten;
- 3-D-Rekonstruktionen und -modelle;
- Audio- und Videodateien;
- Vektorzeichnungen;
- klein- und großformatige Scans.

1 URL: https://journals.ub.uni-heidelberg.de/index.php/dco/index.

2 Opern-Source-Software für die Verwaltung und Veröffentlichung von wissenschaftlichen Zeitschriften. URL: https://pkp.sfu.ca/ojs/.

Zu einem zentralen Problem der institutionellen und fachspezifischen Vielfalt gehört eine gewachsene heterogene Landschaft mit verschiedenen Systemen und Insellösungen sowie dem noch fehlenden Bewusstsein für die Orientierung an Mindeststandards.[3]
Zum mittlerweile akzeptierten, aber bei weitem nicht realisierten Quasistandard gehöre die Akzeptanz, dass die verwendeten Softwarelösungen und Dateiformate nicht proprietär sein dürfen. Bei den verwendeten Formaten für Einzeldateien habe sich die Situation klar zugunsten von TIFF und XML entwickelt, anders als bei Vektor- oder 3-D-Daten, wo aufgrund komplexer Architekturen die Dateninteroperabilität zwischen verschiedenen GIS-Systemen sich noch außerordentlich schwierig gestalte.[4]

Bei der Auswahl eines Dateiformates und damit implizit auch die Auswahl der verwendeten Software sollte auf folgende Eigenschaften besonders geachtet werden. Das Format sollte sein:

- nicht proprietär, also unabhängig von der Software eines bestimmten Herstellers;
- standardisiert und weit verbreitet;
- offen dokumentiert mit frei verfügbaren technischen Spezifikationen;
- unmittelbar lesbar oder einfach dekodierbar;
- unkomprimiert oder verlustfrei komprimiert.

Für die Langzeitarchivierung von Forschungsdaten kann DCO im Rahmen der Vermittlung von Datenkuration nur für bestimmte Typen von Daten Empfehlungen aussprechen. Für einzelne Kategorien, wie beispielsweise Software selbst, werden keine Empfehlungen ausgesprochen. Sie sollten in dafür spezialisierte, öffentlich zugängliche Repositorien eingespeist werden und können dann im Verlauf des Veröffentlichungsprozesses im KFD registriert werden.

Welche Dateiformate für Forschungsdaten werden empfohlen?

Bei der Wahl von Dateiformaten bei der eigenen Forschung sollten neben der fachlichen Eignung zwei weitere Kriterien maßgebend sein: **Nachnutzbarkeit** und **Archivierbarkeit**. Nachfolgende Übersicht soll eine Entscheidungshilfe bei der Auswahl der Formate sein, mit denen erste Erfahrungswerte im Umgang vorliegen und deren Empfehlungen in den zugrundeliegenden Quellen nachvollzogen werden können.

Literatur zu Dateiformaten:

- https://wiki.de.dariah.eu/pages/viewpage.action?pageId=38080370
- http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

---

3 Vgl. Kapital 8 in Heike Neuroth (Hrsg.), Langzeitarchivierung von Forschungsdaten: eine Bestandsaufnahme, Boizenburg, 2012. URL: https://nbn-resolving.org/urn:nbn:de:0008-2012031401

4 Ebenda.

- https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-formats
- http://www.digitalpreservation.gov/formats/
- http://guides.library.oregonstate.edu/research-data-services/data-management-types-formats
- http://www.forschungsdaten-bildung.de/formate
- http://www.ianus-fdz.de/it-empfehlungen/dateiformate

Kategorie	empfohlene Formate	Bemerkungen
Bild	TIFF .tif, .tiff Baseline TIFF Version 6, unkomprimiert	Tagged Image File Format • keine Patenteinschränkungen oder technische Schutzmechanismen • gute Akzeptanz • Baseline ist eine Untermenge von formatspezifischen Eigenschaften, die zwingend unterstützt werden müssen • unterstützt das CMYK-Farbmodell (Vierfarbdruck) • Quasistandard für Bilder mit hoher Qualität
Bild	PNG .png	Portable Network Graphics • keine Patenteinschränkung • gute Akzeptanz • weite Verbreitung • besonders zur Webanzeige geeignet
Vektorgrafik	SVG .svg, .svgz	Scalable Vector Graphics • keine Patenteinschränkungen oder technische Schutzmechanismen • breite Akzeptanz • hohe Transparenz • basiert auf XML
statischer Text	PDF .pdf in der ISO-Standardform PDF/A	Portable Document Format • gute Akzeptanz zum Zweck der Langzeitarchivierung • nicht zur Nachnutzung geeignet • wenn möglich, sollte die zur Erzeugung des PDF vorliegende Quelldatei (.docx, .odt etc.) mit archiviert werden
strukturierter Text	XML .xml auf XML basierende	Extensible Markup Language • vollständig strukturiert • von Menschen lesbar

	Formate wie TEI	• gut in andere Formate (PDF, HTML) konvertierbar • es sollte auf Wohlgeformtheit und Validität geachtet werden
einfacher Text	TXT .txt UTF-8 kodiert	Text File • sehr gute Akzeptanz von allen Betriebssystemen und den meisten Textprogrammen • bietet keine Seitenbeschreibung oder Strukturauszeichnung • Kodierung sollte möglichst in UTF-8 sein
Text (Office)	ODF .odt, .fodt	Open Document Format for Office Applications • offener Standard • hohe Transparenz • basiert auf XML • für die Nachnutzung eigener Inhalte geeignete • das "f" steht für flat und ist die unkomprimierte Form der ansonsten mit ZIP komprimierten Standardversion • nur bedingt zur Langzeitarchivierung geeignet
Text mit Formeln	TeX .tex	TeX / LaTex • offener Standard • sehr gut dokumentiert • hoher Verbreitungsgrad in einzelnen Wissenschaftsdisziplinen • da es sich um ein Textsatzsystem handelt, sollten die Dateien immer nur zusätzlich zu einer fertig erzeugten Dateiversion (PDF/A) archiviert werden • lokale Spezifikationen, Makros, beteiligte LaTeX/TeX-Pakete sollten dokumentiert werden
Tabelle	.csv UTF-8 kodiert	Comma Separated Values • offener Standard • weit verbreitet • bietet keine Struktur- oder Gestaltungsauszeichnung • eignet sich besonders als Austauschformat • Kodierung sollte möglichst in UTF-8 sein • Texttrenner, Feldtrenner und Kodierung sollten dokumentiert werden
Tabelle	ODF .ods, .fods.	Open Document Format for Office Applications • offener Standard • hohe Transparenz

		• basiert auf XML • das "f" steht für flat und ist die unkomprimierte Form der ansonsten mit ZIP komprimierten Standardversion
Relationale Datenbank	SQL .sql	Structured Query Language • Tabelleninhalte und Strukturinformationen in textbasiertem Format • geeignet für die Langzeitarchivierung bei Verwendung eines offiziellen ISO/IEC 9075 Standards • neben den Daten müssen die Datenbankstrukturdefinitionen, wie Attributdatentypen, Schlüssel und Relationen zwingend mit archiviert werden
Audio	WAV .wav	Waveform Audio File Format • offen dokumentiert, jedoch proprietär • weit verbreitet • gilt noch als Quasistandard • Audiodaten sollten als lineares PCM gespeichert werden
Audio	FLAC .flac	Free Lossless Audio Codec • offen dokumentiert • verlustfrei komprimierend • frei verfügbar
Video	MKV .mkv	Matroska • offenes Containerformat • unterstützt eine Vielzahl von Codecs • für die Archivierung geignete Codes sind FFV1, FLAC, H.264/MPEG-4 AVC und MPEG-2
Video	MP4 .mp4	MPEG-4 • der unter ISO/IEC 14496 zertifizierte MPEG-4-Standard verwendet den Codec H.264/MPEG-4 AVC • kann bei verlustfreier Kompression zur Langzeitarchivierung genutzt werden

Welche Nutzungslizenzen sind gebräuchlich?

Eine Veröffentlichung von Forschungsdaten durch das KFD wird nur unter den nachfolgend skizzierten Voraussetzungen ermöglicht. Detaillierte Regelungen finden sich in der Archivierungs- und Veröffentlichungsvereinbarung, die mit dem KFD[5] abgeschlossen wird.

- Die Datengeber sind Eigentümer oder sonstige Rechteinhaber der abgelieferten Daten und deren erläuternde Dokumente.
- Die Datengeber sind alleiniger Inhaber von Nutzungsrechten an den abgelieferten Daten und deren erläuternden Dokumenten mit dem Recht, diese vervielfältigen, verbreiten und öffentlich wiedergeben zu können.
- Die Datengeber liefern dem KFD bzw. der redaktionellen Betreuung von DCO die Daten sowie die obligatorischen Metadaten.
- Alle beteiligten Datengeber unterzeichnen mit dem KFD eine Archivierungs- und Veröffentlichungsvereinbarung.
- Die Datengeber räumen dem KFD räumlich und zeitlich unbegrenzt das nicht-ausschließliche Recht zur Nutzung der Daten und deren erläuternde Dokumente im Rahmen des Repositoriums heiDATA ein. Davon umfasst sind insbesondere die Vervielfältigung, Bearbeitung und öffentliche Zugänglichmachung.
- Das KFD ist berechtigt, Daten und Dokumente nach der von den Datengebern im Anhang des Vertrags gewählten Open Content-Lizenz öffentlich zugänglich zu machen. Die Zugänglichkeit des gesamten Datenbestandes besteht ohne technische Zugangsbeschränkung ab dem Zeitpunkt der Freischaltung oder mit technischer Zugangsbeschränkung bis zu einem angegebenen Zeitpunkt.

Elementare Voraussetzung für die Zugänglichmachung der abgelieferten Datengegenstände unter einer Open Content-Lizenz ist, dass die Datengeber als Rechteinhaber alle relevanten Nutzungsrechte einräumen können. Urheberrechtlich nicht geschütztes Material (z. B. wissenschaftliche Rohdaten, Fakten, Erkenntnisse) ist von der Lizenz auszunehmen bzw. entsprechend zu kennzeichnen.
Für das Zugänglichmachen der Forschungsdaten sind folgende Standardlizenzen zu berücksichtigen:

- Creative Commons Lizenzen (CC Version 4.0)[6]: Die modular aufgebaute CC-Lizenzen decken neben Urheberrechten auch Leistungsschutzrechte ab.
- Open Data Commons (ODC)[7]: Diese Lizenz enthält im Unterschied zu CC-Lizenzen eine vertragsrechtliche Komponente zur Erfassung des Datenbankherstellerrechts (§§ 87a ff. UrhG)[8].

5 URL: https://data.uni-heidelberg.de/md/data/autorenvereinbarung.pdf.

6 URL: https://creativecommons.org/licenses/.

7 URL: https://opendatacommons.org/licenses/.

8 URL: http://www.gesetze-im-internet.de/urhg/__87a.html.

Aus der Sicht des KFD und der Herausgeber von DCO gewährleistet die Lizenzvariante **CC BY** am ehesten die freie Nutzbarkeit der abgelieferten Datengegenstände in Übereinstimmung mit den Open Access und Open Science-Forderungen. Diese Lizenzvariante wird den Autoren ausdrücklich empfohlen.

Welche Metadaten sollten vorhanden sein?

Neben den Forschungsdaten selbst sind zusätzliche Informationen für das Auffinden und die Einordnung der Datensätze notwendig. Einige wenige Pflichtangaben sind:

- ein Titel des Datensatzes,
- die Benennung der Datenautorenschaft, also Vor- und Zunamen der Beteiligten,
- eine Beschreibung des Datensatzes in wenigen Sätzen (wie Abstract möglichst in Englisch),
- und einige wenige Schlüsselwörter.

Weitere Metadaten sind optional und können umfassen:

- weitere Angaben zu den Datenautoren wie Organisationszugehörigkeit oder unabhängige Personen-Identifier für Wissenschaftler (z.B. ORCID),
- geografische, astronomische oder biowissenschaftliche Metadaten.

Die Datengeber versichern mit einer Archivierungs- und Veröffentlichungsvereinbarung, die mit dem Kompetenzzentrum Forschungsdaten abzuschließen ist, dass

- die Metadaten unter Einhaltung der datenschutzrechtlichen Bestimmungen erstellt worden sind,
- der abgelieferte Datengegenstand und eventuell erläuternde Dokumente keine personenbezogenen Daten im Sinne der geltenden datenschutzrechtlichen Vorschriften enthalten.

Was ist noch zu beachten?

Wegen der gewachsenen heterogene IT-Landschaft mit verschiedenen Systemen und Insellösungen und der fachspezifischen Vielfalt an Datentypen ist es dringend angeraten, zu dem Datensatz eine Beschreibungsdatei (z.B.: readme.txt) mitzuliefern. In dieser Datei sollten Informationen hinterlegt werden, wie mit den Daten umzugehen ist, was beachtet werden sollte, welche Datei welche Daten vorhält, gegebenenfalls mit welchem Programm / Softwarepaket wurden sie erstellt oder können sie nachgenutzt werden etc.
Diese Beschreibungsdatei sollte so angelegt sein, dass die Nachnutzung möglichst unkompliziert ohne großen Rechercheaufwand zu bewerkstelligen ist.