e-infrastructures austria · Deliverable

> *„Metadata ist the data warehouse‘s Gordian knot, and Alexander and his sword are nowhere in sight“*
> Ralph Kimball

Metadaten und Forschungsdaten

e-infrastructures austria · Deliverable

Metadaten und Forschungsdaten

Work-Package-Cluster:	**Cluster I: Metadatenkomplex aus nicht-technischer und technischer Sicht**	
Leitung des Clusters:	Susanne Blumesberger	Universitätsbibliothek Wien susanne.blumesberger@univie.ac.at
Datum:	25.05.2016	
Version	1.1	
AutorInnen:	Silvia Gstrein	Universitäts- und Landesbibliothek Tirol silvia.gstrein@uibk.ac.at
	Andreas Krexhammer	Österreichische Akademie der Wissenschaften andreas.krexhammer@oeaw.ac.at
	José Luis Preza	Universitätsbibliothek Wien jose.luis.preza@univie.ac.at
	Yukiko Sakabe	Österreichische Akademie der Wissenschaften yukiko.sakabe@oeaw.ac.at

Kurzbeschreibung (Deutsch):	Es geht hier um digitale Forschungsdaten, für die ein professioneller Umgang hinsichtlich der Langzeitarchivierung erforderlich ist. Um eine langfristige und zuverlässige Verfügbarkeit der Forschungsdaten zu gewährleisten, ist eine strukturierte Beschreibung der Daten notwendig.

	Unterschiedliche Anforderungen von verschiedenen wissenschaftlichen Fachbereichen bedürfen passender Metadatenformate, wobei eine klare Trennung zwischen den Fachdisziplinen nicht möglich ist.
Description (English):	This document discusses the area of digital research data and the importance of professional data management, in particular when data is to be long-term archived. Descriptive metadata is required for long-term and reliable availability of research data. Different requirements of various scientific disciplines ask for a suitable metadata format, but a clear separation of disciplines is not possible.

Schlagwörter (Deutsch):	Metadaten – Forschungsdaten – Repositorien – Österreich – Metadatenformate
Keywords (English):	Metadata – Research data – Repositories – Austria – Metadata format

Metadaten und Forschungsdaten

Daten, ob analog oder digital, sind eine zentrale Ressource jeder wissenschaftlichen Forschung. Durch die fortschreitende Technologisierung und vor allem Digitalisierung ist aber ein rasanter Anstieg des digitalen Datenvolumens in der Forschung zu verzeichnen. Damit steigen auch die Anforderungen an einen professionellen Umgang mit digitalen Informationsobjekten. Die Herausforderung ist dabei eine langfristige, referenzierte und verlässliche Verfügbarkeit der Forschungsdaten bzw. der Zugriff darauf. Die Notwendigkeit dafür ergibt sich aus verschiedenen Gründen:

- Zum einen sind strukturierte Forschungsdaten notwendig um darauf aufbauende Forschungsergebnisse verifizieren zu können.
- Zum anderen kann nur dadurch eine Nachnutzung der Daten gewährleistet werden. Hier spielt vor allem die Integration neuer Daten, in alte Datensets eine Rolle aber auch die Chance der Beantwortung alter Forschungsfragen durch neue Möglichkeiten.
- Durch die langzeitliche Verfügbarkeit und Interpretierbarkeit werden Langzeitstudien überhaupt erst möglich.
- Durch adäquates Datenmanagement, der daraus resultierenden Sichtbarkeit, Referenzierbarkeit und durch den freien Zugang wird eine interdisziplinäre Nutzung möglich.

Wenn auch die Notwendigkeit eines adäquaten Datenmanagements noch nicht in allen Bereichen der Forschung angekommen ist, haben Förderorganisationen bereits darauf reagiert und stellen Erwartungen an ein kontinuierliches und langfristiges Datenmanagement. In vielen Ausschreibungen ist ein DMP (Datamanagementplan) bereits ein notwendiger Antragsinhalt. Beispielhaft zeigt sich dies auch in Horizon 2020, unter gewissen Voraussetzungen, Teilnahme an „Open Research Data Pilot“ ist auch hier ein DMP bereits verpflichtend. Der DMP beschreibt den Lebenszyklus aller Datensets die im Zuge eines Forschungsprojekts erstellt, gesammelt, verarbeitet und publiziert werden. Dabei müssen folgende Aspekte berücksichtigt werden[1]:

- Der Umgang mit Forschungsdaten während und nach dem Projekt
- Welche Daten werden verwendet
- Welche Methoden und Standards finden Anwendung
- Werden die Daten zu Verfügung gestellt und wie
- Wie werden die Daten kuratiert und konserviert

[1] Guidelines on Data Management in Horizon 2020 : Version 2.1 / EUROPEAN COMMISSION Directorate-General for Research & Innovation [http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf] (abgerufen am 23.05.2016)

Vor allem Datenkuratierung und -konservierung ist ein Aspekt der unter Wissenschaftlern selbst, noch wenig Beachtung findet. Eine österreichweite Umfrage unter Forschenden und deren Umgang mit Daten zeigt, dass ein Großteil der Forschungsdaten immer noch auf dienstlichen oder privaten Rechnern und externen Festplatten gespeichert werden. Auch die Beschreibung der Daten erfolgt in überwiegender Mehrheit individuell und nicht standardisiert[2]. Unter solchen Voraussetzungen ist es allerdings schwer die oben erwähnten Aspekte zu berücksichtigen und die gestellten Anforderungen zu erfüllen, vor allem in Bezug auf eine Langzeitarchivierung. Die Langzeitarchivierung ist eine unter dem jeweiligen Kontext definierte Zeitspanne der Bewahrung der digitalen Objekte auch über technologische und soziokulturelle Wandlungsprozesse hinaus. So unterschiedlich dabei die Objekte und deren Formate sein können, so unterschiedlich sind dabei auch die Definitionen von Zeiträumen der Langzeitarchivierung.
Die Herausforderung der Verfügbarkeit der Daten, über unterschiedlich definierte Zeiträume, liegt dabei auf der Hand. Da zukünftige technologische Entwicklungen und Nutzungsumgebungen schwer prognostizierbar sind, ist es umso wichtiger die Objekte adäquate und standardisiert zu beschreiben. Dabei sind aber nicht nur deskriptive Metadaten notwendig auch administrative und technische Metadaten werden benötigt um Emulation Migration und Zugang für die Zukunft gewährleisten zu können.
Was aber sind Forschungsdaten tatsächlich? Welche Informationen sind erhaltenswert um eine gute wissenschaftliche Praxis zu gewährleisten? Dazu gibt es verschiedene Auffassungen und unterschiedlichste Definitionen.[3]
Im Allgemeinen aber, bezeichnen Forschungsdaten alle Ressourcen bzw. Informationen,

- die Gegenstand eines Forschungsprozesses sind
- die während des Prozesses entstehen
- oder dessen Ergebnisse sind

Die Daten werden dabei unter Anwendung verschiedener Methoden und Standards erzeugt, z.B. durch Quellenforschungen, Experimente, Messungen, Beschreibungen, Erhebungen, Befragungen etc..
Die Anwendung der verschiedenen Methoden und Standards bei der Datengenerierung zeigt die Heterogenität von Forschungsdaten. In der Praxis bedeutet das ein projekt- vor allem aber disziplinspezifisches Verständnis von Forschungsdaten mit divergierenden Anforderungen für Aufbereitung und Verwaltung. Diese unterschiedlichen fachspezifischen Anforderungen haben zur Herausbildung verschiedener Metadatenformate geführt.
So unterschiedlich die Daten und Formate auch sind, durch die Anwendung von standardisierten und normierten Metadatenformaten, können nicht nur innerhalb der verschiedenen Wissenschaftsdisziplinen die Ziele der Interoperabilität und langfristigen Verfügbarkeit gewährleistet werden.
Nachfolgend werden Metadatenformate verschiedener Fachdisziplinen vorgestellt.

---

[2] Forschende und ihre Daten: Ergebnisse einer österreichweiten Befragung : Report 2015 / e-infrastructures austria ; Hauptautorinnen und - autoren: Bruno Bauer. – Wien : e-infrastructures austria, 2015, Seite 28ff als e-book unter http://phaidra.univie.ac.at/o:407736 abrufbar.
[3] www.forschungsdaten.org (abgerufen am 23.06.2016)

Metadatenformate

Hier werden ein paar ausgewählte Beispiele für Metadatenformaten nach Wissenschaftsdisziplinen kurz erwähnt und aufgelistet. Jeder Fachbereich besteht jedoch oft aus einer Vielzahl von Einzeldisziplinen, und selbst durch eine interdisziplinäre Zusammenarbeit ist eine klare Eingrenzung vielleicht nicht unbedingt zielführend. Die hier erwähnten Metadatenstandards treten auch gleich bei mehreren Forschungsbereichen auf.

Sozialwissenschaften

Die Sozialwissenschaften umfassen verschiedene Wissenschaftsdisziplinen wie Anthropologie, Demografie, Soziologie oder auch Wirtschaftswissenschaften. Die Data Documentation Initiative (DDI) hat einen gleichnamigen Metadaten-Standard entwickelt, der sozial- und wirtschaftswissenschaftliche Daten beschreibt und weltweit von zahlreichen Organisationen genutzt wird. Der vollständige Datenlebenszyklus, der sogenannte Data Life Cycle, wird mittels XML erfasst. Forschungsdaten können somit bezüglich ihrer Konzepte, Erhebung, Verarbeitung, Verteilung, Exploration, Analyse, Wiederverwendung und Archivierung beschrieben werden. Die DDI ist auch als XML-Schema mit anderen Metadatenstandards, etwa mit ISO/IEC 11179 oder Dublin Core, kompatibel.

DDI (Data Documentation Initiative)	Standard für die Beschreibung von sozialwissenschaftlichen Daten	XML-Schema
SDMX (Statistical Data and Metadata eXchange)		SDMX-ML (using XML syntax), SDMX-EDI
TEI (Text Encoding Initiative)	Tiefere Erschließung von Texten der Geisteswissenschaften, Linguistik	XML

Geisteswissenschaften

Durch die Vielfältigkeit des Forschungsbereichs ist es hier besonders schwierig ihn klar einzugrenzen. Auch eine Trennung zwischen Geistes- und Naturwissenschaften verschwimmt bereits in einigen Teilbereichen. Immer häufiger wird auch durch „digitale“ Methoden eine interdisziplinäre Zusammenarbeit gefördert. [4] Im Bereich der eHumanities

[4] nestor Handbuch: eine kleine Enzyklopädie der digitalen Langzeitarchivierung / hg. V. H. Neuroth. http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-2010071949 (abgerufen am 25.05.2016)

bilden sich sogar nicht selten institutionsübergreifende Projekte der Geisteswissenschaften.[5]
Für den geisteswissenschaftlichen Bereich werden beispielsweise Digitalisate von Texten, wie Scans von Drucken, Manuskripten oder Fotografien von Forschungsobjekten, wie Inschriften etc. verwendet. Aber auch Textdaten, die während der wissenschaftlichen Arbeit entstanden sind, können als typische Daten der Geisteswissenschaft bezeichnet werden.

Dublin Core	Beschreibung von Dokumenten und anderen Objekten im Internet	15 Kernfelder (*core elements*) **RDF/XML**
MARC (Machine-Readable Cataloging)	Bibliothekarische Austauschformate	MARC 21
MODS (Metadata Object Description schema)	Bibliothekarische Austauschformate	XML-Format Unterstützt von Literaturverwaltungsprogram men (BibDesk, JabRef, Zotero)
METS (Metadata Encoding and Transmission Standard)	Bibliothekarische Austauschformate	
RDF (Resource Description Framework)	Beschreibung von Webressourcen	
TEI (Text Encoding Initiative)	Tiefere Erschließung von Texten	XML

Altertumswissenschaften

Die sehr komplexe archäologisch-altertumswissenschaftlichen Arbeitsfelder umfassen: „Ausgrabungen (Dokumentation von Architektur, Stratigrafie, Gräbern etc.), Prospektionen/
Surveys (Begehungen, Testgrabungen, Keramiksammlungen), Fundbearbeitung (z.B. Keramikanalyse), Fotogrammetrie (z.B. Aufnahme von 162 Langzeitarchivierung von Forschungsdaten Gebäuden), Chronometrie (unterschiedliche naturwissenschaftliche und
kunstgeschichtliche Methoden zur Datierung), Klima- und Landschaftsgeschichte (Geologie, Geomorphologie, Hydrologie), anthropologische Untersuchungen (Untersuchungen von Skeletten, Ernährungsgewohnheiten, Krankheiten und der genetischen Hinweise auf Verwandtschaftsbeziehungen) sowie epigrafische,

[5] Ebd.

philologische, linguistische Untersuchungen (Editions- und Corpusprojekte) und die Dokumentation und Restaurierung von Gebäuden.“[6]

Dublin Core	Beschreibung von Dokumenten und anderen Objekten im Internet	15 Kernfelder (*core elements*) RDF/XML
LIDO (Lightweight Information Describing Objects)	Harvesting-Format zur Weitergabe von Objekten aus Museen	
MODS (Metadata Object Description Schema)	Bibliothekarische Austauschformate	XML-Format Unterstützt von Literaturverwaltungsprogrammen (BibDesk, JabRef, Zotero)
RDF (Resource Description Framework)	Beschreibung von Webressourcen	

Geowissenschaften

Die Geowissenschaften beschäftigen sich im Allgemeinen mit der Erforschung des Systems Erde, dabei beschränken sich die Themen nicht mehr nur auf rein naturwissenschaftliche Aspekte. Durch das immer stärkere Eingreifen der Menschen in das System der Erdatmosphären, gewinnen auch „nicht-physische Disziplinen“ wie die Humangeographie an Bedeutung innerhalb der Geowissenschaften. Die etablierte interdisziplinäre Arbeit des Fachbereichs wird dadurch noch stärker forciert und folglich steigt auch die Heterogenität der Forschungsdaten. Diese reichen von großen, automatisch prozessierten Daten aus der Fernerkundung bis zu individuell erzeugten Datensätzen einer Quellenrecherche. So unterschiedlich die Forschungsdaten und Formate auch sind, die Daten der Geowissenschaften haben, bis auf wenige Ausnahmen, immer eine Gemeinsamkeit, den Raumbezug. Vor allem für den Aufbau von nationalen und internationalen Geodaten-Infrastrukturen (GDI) wurde die Notwendigkeit einer internationalen Standardisierung für Metadaten zu raumbezogenen Daten erkannt. Hierfür wurde 2003 der ISO-Standard 19115 veröffentlicht, der sich auch durch die Anwendung in INSPIRE (Infrastructure for Spatial Information in the European Community) etablieren konnte.

ISO 19115	Eine **Norm** der **Internationalen**	

[6] nestor Handbuch / hg. v. Heike Neuroth

(DIN EN ISO 19115)	**Organisation für Normung** (ISO) – Geoinformationen und Geodaten	
CSDGM (Content Standard for Digital Geospatial Metadata)		
CERA-2 (Climate and Environmental Retrieval and Archive)	Klimaforschung	Simulationsdaten zur Klimaentwicklung nach den Regeln des WDV-V (World Data Center for Climate)

Literatur:

- Data Cite (2015): DataCite Metadata Schema for the Publication and Citation of Research Data. Version 3.1. Abgerufen unter: https://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.1.pdf doi:10.5438/0010

- Data Documentation Alliance - Metadatenstandard für Forschungsdaten in den Sozialwissenschaften, Online abrufbar unter http://www.ddialliance.org/

- GESIS (2006): Klassifikation Sozialwissenschaften. Abgerufen unter: http://www.gesis.org/fileadmin/upload/dienstleistung/tools_standards/klass.pdf

- Gregory, Arofan; Heus, Pascal; Ryssevik, Jostein (2009): Metadata. Working Paper Series of the Council for Social and Economic Data. RatSWD Working Paper No. 57. Abgerufen unter: http://www.ratswd.de/download/RatSWD_WP_2009/RatSWD_WP_57.pdf

- Jensen, Uwe / Katsanidou, Alexia / Zenk-Möltgen, Wolfgang (2011): Metadaten und Standards. In: Büttner, Stefan et. al Handbuch Forschungsdatenmanagement. Bad Honnef. Online abrufbar unter www.forschungsdatenmanagement.de bzw. https://opus4.kobv.de/opus4-fhpotsdam/files/198/2.4_Metadaten_und_Standards.pdf

- Jensen, Uwe (2012): Metadaten für Forschungsdaten: Welche Standards gibt es? Online abrufbar unter: https://opus4.kobv.de/opus4-bib-info/files/1176/Metadatenstandards_Welche_gibt_es_Btag2012_Uwe_Jensen.pdf

- NISO (2004): Understanding Metadata Abgerufen unter: http://www.niso.org/publications/press/UnderstandingMetadata.pdf

- Toussaint, Frank (1999): Wissenschaftliches Datenmanagement: Das „CERA-2 Daten- und Metadatenmodell“. Abgerufen unter: http://www.mad.zmaw.de/uploads/media/9911_ptb_01.pdf

- Vardigan, Mary; Heus, Pascal; Thomas, Wendy (2008): Data Documentation Initiative. Toward a Standard for the Social Sciences. In: The International Journal of Digital Curation, 3(1). Abgerufen unter: http://www.ijdc.net/index.php/ijdc/article/view/66/45
- http://nestor.sub.uni-goettingen.de/bestandsaufnahme/nestor_lza_forschungsdaten_bestandsaufnahme.pdf - Eine Bestandsaufnahme. (2012). Neuroth et.al. (Hrsg)
- http://nestor.sub.uni-goettingen.de/handbuch/nestor-handbuch_23.pdf : Eine kleine Enzyklopädie der digitalen Langzeitarchivierung (2010). Neuroth et.al. (Hrsg.)

Link:

- Glossary of Metadata Standards (2010). Riley und Becker
- http://www.forschungsdaten.org/
- http://libreas.eu/ausgabe23/inhalt.htm
- https://wiki.dnb.de/display/DINIAGKIM/Diverses+zum+Thema+Metadaten+zu+Forschungsdaten
- https://wiki.de.dariah.eu/pages/viewpage.action?pageId=38080370
- http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2012/TechnicalReport_2012-07.pdf
- https://www.cms.hu-berlin.de/de/ueberblick/projekte/dataman/teilen/dokumentation/metadaten
- http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

e-Infrastructures Austria

Sustainable data storage and the provision of data for use by third parties are the central roles of science. e-Infrastructures Austria is a federally funded program for the coordinated expansion and continued development of data repositories across Austria, and is made possible by a grant from the Austrian Ministry of Science, Research and Commerce (BMWFW). This program enables the safe archival and lasting availability of electronic publications, multimedia objects and other digital data from the research and teaching fields. Concurrently, topics relating to research data management and digital archiving workflows will be addressed.

The working area is organized in twelve Work-Package-Clusters:

Cluster A	Monitoring of Document Repositories within the Partner Network *Patrick Danowski (IST Austria)*
Cluster B	Planning and Implementation of a „National Survey" for Research Data *Christian Gumpenberger (University of Vienna)*
Cluster C	Designing a Knowledge Network: Development of a reference structure for the construction of Repositories *Paolo Budroni (University of Vienna)*
Cluster D	Infrastructure *Raman Ganguly (Vienna University Computer Center)*
Cluster E	Legal and Ethical Issues *Seyavash Amini (Counsellor-at-law, University of Vienna)*
Cluster F	Open Access *Andreas Ferus (academy of fine arts vienna)*
Cluster G	Visual Data modeling *Martin Gasteiner (University of Vienna)*
Cluster H	Life Cycle Management *Andreas Rauber (Technical University Vienna)*
Cluster I	Metadata *Susanne Blumesberger (University of Vienna)*
Cluster J	Permanent backup of the data *Adelheid Mayer (University of Vienna)*
Cluster K	Data from scientific and artistic-scientific research processes *Bernhard Haslhofer (Austrian Institute of Technology)*
Cluster L	Cross-project issues (technical and non-technical) *Andreas Jeitler (University of Klagenfurt)*

University of Vienna Library and Archive Services, 1010 Vienna,
Universitätsring 1
T: +43-1-4277-15172, office@e-infrastructures.at