s Austria führte An- fang
eine umfassende Untersu-
g mit Forschungsdaten in Öster-
000 WissenschafterInnen von 20
täten sowie drei
Forschungseinrich-

E-INFRASTRUCTURES AUSTRIA DELIVERABLE Cluster D

Im Rahmen des
Projekts LEARN3
raum von Juli
rsitätsbibliothek Wien unterschiedliche europäi-
olicies sowohl formell als auch inhaltlich analy-
dingungen für ein kompetentes Forschungsda-
zu Beginn 2016 die ExpertInnengruppe - Strate-

t e r -
s c h u n g s -
licy an österreichi-
ıtungen zu erstellen. Das vorliegende Doku-
orschungseinrichtung lokalisiert und an die
eigenen Institution angepasst werden
n der Sitzung der ExpertInnengrup-

e-infrastructures austria · Deliverable

Bewertung von Archivsystemen

Work-Package-Cluster:	Cluster D: Aufbau Infrastruktur
Leitung des Clusters:	Raman Ganguly, Universität Wien, raman.ganguly@univie.ac.at
Datum:	17.08.2016
Version:	1.0

AutorInnen/ Sonstige Beteiligte:	Michael Birkner	AK Bibliothek Wien, michael.birkner@akwien.at
	David Mitterhuber	Akademie der bildenden Künste Wien, d.mitterhuber@akbild.ac.at
	Raman Ganguly	Universität Wien, raman.ganguly@univie.ac.at
	Gerhard Gonter	Universität Wien, gerhard.gonter@univie.ac.at
	Christoph Ladurner	TU Graz, christoph.ladurner@tugraz.at

Kurzbeschreibung (Deutsch):	Das Modell, das in diesem Dokument beschreiben wird, soll eine Hilfestellung geben, um zu entscheiden ob gewissen Daten in allgemein gehaltenen Datenrepository passen oder nicht. Dabei werden die zu archivierenden Daten und die Repositorien mit einem Bewertungssystem beurteilt.
Description (English):	The model, which is described in the document, should give help for in order to decide whether certain data fit into a general data repository. The archival data and repositories are evaluated with a assessment system.

Schlagwörter (Deutsch):	Dokumentenserver, Repositories, Schnittstellen, API, OAI-PMH
Keywords (English):	document server, repositories, interfaces, API, OAI-PMH

Inhalt

1. Einleitung

Die Vielzahl an unterschiedlichen Daten stellt eine große Herausforderung im Research Data Management da. Man ist mit unterschiedlichen Dateiformaten und Dateigrößen konfrontiert sowie auch mit unterschiedlichsten Anforderungen, wie die Daten bereitgestellt werden sollen. Es sollte nicht versucht werden, ein System für alles zu finden, da dies auf Grund der heterogenen Anforderungen nicht möglich ist. Es würde nur ein Kompromiss entstehen, der vielen Aufgaben nicht lösen kann.

Sinnvoller ist es, ein Ökosystem von generellen und spezialisierten Systemen aufzubauen, in dem es möglich ist, die Daten aus den verschiedenen Systemen in Beziehung zu setzen. Bei einem spezialisierten System ist die Frage einfacher zu beantworten, welche Daten dort archiviert werden sollen, als bei generellen Datenarchiven. So ist es klar, dass Software in Software-Repositorien archiviert werden können. Die Beurteilung, in welches Datenrepositorium, das eine institutionelle Datensammlung darstellt, eine Datei mit statistischen Daten passt, ist dagegen nicht eindeutig.

Das Modell, das in diesem Dokument beschrieben wird, soll eine Hilfestellung geben, um zu entscheiden, ob gewisse Daten in allgemein gehaltene Datenrepositorien passen oder nicht. Dabei werden die zu archivierenden Daten und die Repositorien mit einem Bewertungssystem beurteilt.

2. Grundlegende Idee des Modells

Wie schon in der Einleitung erwähnt, geht es hier um allgemeine Datenrepositorien, die eine Vielzahl unterschiedlicher Daten sammeln sollen.

Bewertet werden zunächst einmal drei wesentliche Faktoren die für die Archivierung wichtig sind und mit folgenden Merkmalen beschrieben werden:

- Dauer der Archivierung
- Komplexität der Formate
- Datenmenge

2.1 Dauer der Archivierung

Langzeitarchivierung ist sehr kostspielig und sollte daher bewusst durchgeführt werden. Sämtliche Daten aufzuheben macht wenig Sinn, da nicht alles im Laufe der Zeit nachgenutzt werden kann. Um Daten auch wirklich langfristig nutzbar zu machen, müssen sie auch dementsprechend aufbereitet werden. Beispielsweise müssen entsprechende Metadaten vergeben werden.

Das bedeutet aber nicht, dass man nur jene Daten archivieren sollte, die für die Langezeitarchivierung notwendig sind. Auch andere Daten, die während des Projekts oder für einen gewissen Zeitraum nach den Projekt zugänglich gemacht werden sollen, müssen vom Research Data Management berücksichtig werden.

2.2 Komplexität der Formate

Die Formate, in denen die Daten gespeichert sind, eignen sich unterschiedlich gut für die Archivierung. Es gibt aber auch einen Unterschied in der Komplexität der Daten. So sind verschiedene Formate leichter als andere zu bedienen. Dokumentformate, auch jene die für die Langzeitarchivierung nur bedingt geeignet sind, sind in der Regel einfacher zu archivieren als Videodaten. Die Komplexität äußert sich sowohl durch den Aufwand beim Ingest, als auch durch den Aufwand die Formate über einen längeren Zeitraum lesbar zu halten. Datenbanken sind hier eines der komplexesten Formate, da auch die genaue Beschreibung der Datenstruktur und der Software, die die Daten interpretiert, dokumentiert und archiviert werden muss.

2.3 Datenmenge

Die Datenmenge ist am leichtesten zu beurteilen, da die Datenmenge einfach zu messen ist. Die Schwierigkeit hier ist, dass oft die Datenmenge schon zu Beginn eines Projekts beurteilt werden muss, um entsprechende Speicher zur Verfügung stellen zu können. Diese Einschätzung wird oft getroffen, ohne alle Faktoren der Datenerhebung zu kennen. Einfache Einteilungen sind hier daher auch hilfreich.

3. Bewertungsmodell

Die drei oben beschrieben Merkmale werden als Achsen dargestellt und so eine einfache Einteilung vorgenommen. Die Einteilung sollt nicht zu feingliedrig sein, da dieses Modell einen schnellen Überblick geben soll, um Entscheidungen treffen zu können. Stufen von eins bis vier sind ein gutes Raster. Hier ein Beispiel für mögliche Stufen.

Dauer der Archivierung

1. Strukturierte Aufbewahrung über die Laufzeit des Projekts
2. Kurzzeitige Archivierung, bis zu 5 Jahre, z.B. für die forschungsgestützte Lehre
3. Mittelfristige Archivierung, bis zu 10 Jahre, z.B. für die Dokumentation von Publikationen
4. Langzeitarchivierung, über 10 Jahre, Daten die einen wesentlichen Wert für zukünftige Forschung haben.

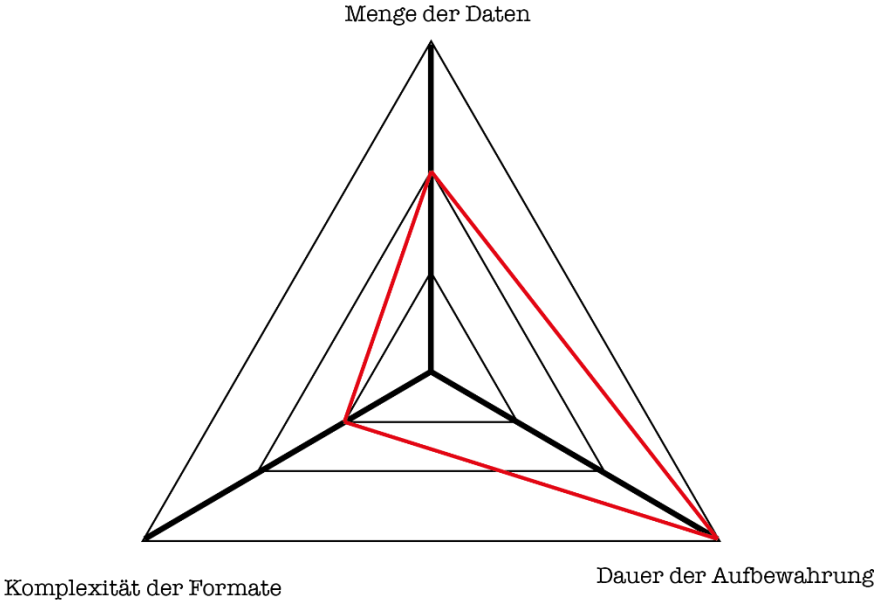
Komplexität der Formate

1. Text-Files (z.B. CSV), Dokument und Bildformate
2. Audio/Video-Formate
3. Datenbanken, 3D-Formate
4. Datenbanken, die laufend noch geändert werden müssen

Menge der Daten

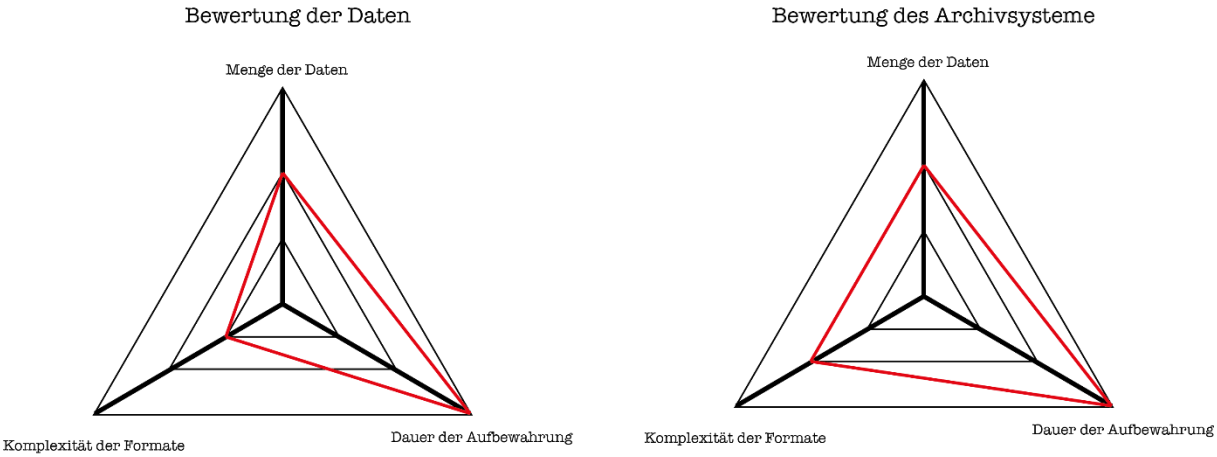
1. Mehrere MB pro Projekt möglich
2. Mehrere GB pro Projekt möglich
3. Bis zu einem TB pro Projekt möglich
4. Mehrere TB pro Projekt möglich

Die Skala der Merkmale sollte sorgfältig dokumentiert werden. Anschließend kann man folgendes Diagramm erstellen.



Das rote Dreieck markiert dabei die Ausprägung der Daten. Nun kann das Archivsystem, in dem die Daten archiviert werden sollen, auf die gleiche Art und Weise bewertet werden und anschließend den Daten gegenüber gestellt werden.

Die nachfolgende Grafik zeigt so ein Beispiel und lässt erkennen, dass die Daten in den Archivsystemen aufbewahrt werden können.



4. Erweitertes Bewertungsmodell

Es kann aus institutionellen, technischen oder projektbezogen Gründen notwendig sein, weitere Merkmale zu dem Modell hinzuzufügen. Wichtig ist dabei, dass immer die Daten und die Archivsysteme mit den gleichen Merkmalen verglichen werden.

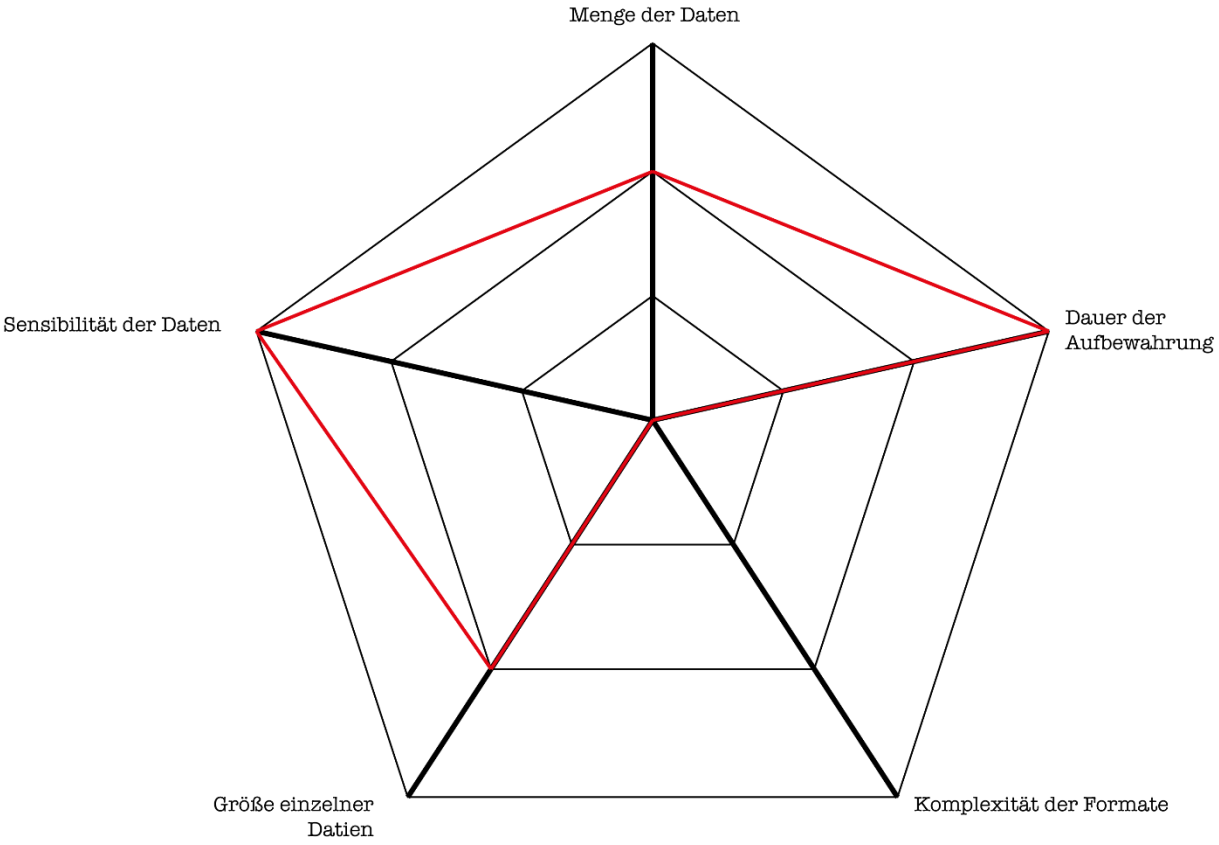
In dem nachfolgenden Beispiel wurden folgende Merkmale hinzugefügt:

Größe einzelner Datei

1. Bis zu 4 GB
2. Bis zu 100 GB
3. Bis zu 500 GB
4. Über 500 GB

Sensibilität der Daten

1. Open Data
2. Zugangsbeschränkung notwendig
3. Sensible Daten
4. Personenbezogene Daten



Das rote Vieleck markiert wieder die Ausprägungen der Merkmale der Daten. Hierbei könnte es sich z.B. um statistische Erhebungen handeln, deren Rohdaten mit Personenbezug in CSV Files gespeichert sind.

e-Infrastructures Austria

Nachhaltige Datensicherung und das Bereitstellen von Daten für Dritte ist eine zentrale Aufgabe der Wissenschaft. e-Infrastructures Austria ist ein vom Bundesministerium für Wissenschaft, Forschung und Wirtschaft (MBWFW) gefördertes Hochschulraumstrukturmittel-Projekt für den koordinierten Ausbau und die Weiterentwicklung von Repositorien in ganz Österreich. Dadurch wird die sichere Archivierung und dauerhafte Bereitstellung von elektronischen Publikationen, Multimedia-Objekten und anderen digitalen Daten aus Forschung und Lehre gewährleistet. Eng damit zusammenhängend werden Themen im Bereich Forschungsdatenmanagement und Workflows von digitaler Archivierung bearbeitet.

Cluster A	Monitoring und Austausch zum Aufbau von Dokumentservern in den lokalen Einrichtungen *Patrick Danowski (IST Austria)*
Cluster B	Planung und Durchführung einer österreichweiten Umfrage zu Forschungsdaten *Christian Gumpenberger (Universität Wien)*
Cluster C	Aufbau eines Wissensnetzwerks: Erarbeitung eines Referenzmodells für den Aufbau von Repositorien *Paolo Budroni (Universität Wien)*
Cluster D	Aufbau Infrastruktur *Raman Ganguly (Zentraler Informatikdienst Universität Wien)*
Cluster E	Legal and Ethical Issues *Seyavash Amini (Rechtsberater Universität Wien)*
Cluster F	Open Access *Andreas Ferus (Akademie der bildenden Künste Wien)*
Cluster G	Visuelle Datenmodellierung – Generierung von Wissenschaftsräumen *Martin Gasteiner (Universität Wien)*
Cluster H	Life Cycle Management *Andreas Rauber (Technische Universität Wien)*
Cluster I	Metadatenkomplex *Susanne Blumesberger (Universität Wien)*
Cluster J	Dauerhafte Sicherung der Daten (aus nicht-technischer & technischer Sicht) *Adelheid Mayer (Universität Wien)*
Cluster K	Daten aus wissenschaftlichen und künstlerisch-wissenschaftlichen Forschungsprozessen (Entwicklung und Erschließung der Künste) *Bernhard Haslhofer (Austrian Institute of Technology)*
Cluster L	Projektübergreifende Fragen (aus nicht-technischer & technischer Sicht) *Andreas Jeitler (Universität Klagenfurt)*