E-INFRASTRUCTURES AUSTRIA DELIVERABLE Cluster H

Lifecycle Management

Work-Package-Cluster:	Cluster H: Lifecycle Management
Leitung des Clusters:	Andreas Rauber, TU Wien, rauber@ifs.tuwien.ac.at
Datum:	30.07.2016
Version:	1.0

AutorInnen/ Sonstige Beteiligte:		
	Michael Birkner	AK Bibliothek Wien, Michael.BIRKNER@akwien.at
	Raman Ganguly	ZID der Universität Wien, raman.ganguly@univie.ac.at
	Gerhard Gonter	ZID der Universität Wien, gerhard.gonter@univie.ac.at
	Silvia Gstrein	ULB Tirol, silvia.gstrein@uibk.ac.at
	Ralf Pausz	UB Wien, ralf.pausz@univie.ac.at
	José Luis Preza	ZID der Universität Wien, jose.luis.preza@univie.ac.at
	Wolfram Seidler	UB Wien, wolfram.seidler@univie.ac.at
	Elisabeth Trinkl	Universität Graz, elisabeth.trinkl@uni-graz.at

Kurzbeschreibung (Deutsch):	Cluster H beschäftigte sich mit vorwiegend technischen Aspekten des Data Lifecycle Management. Um Forschungsdaten als vertrauenswürdige Archive langfristig verfügbar und nutzbar zu halten, sind in naher Zukunft eine Reihe von Services zu erwarten. Bedeutsame Bereiche sind hier Repositorienlösungen, Unterstützung (Tools, APIs, Services) im Bereich Data Ingest, Data Citation Services, Data/Process Management Plans, Cost Estimation Tools, Bitstream-Preservation und Infrastrukturen für Software und

	Forschungsprozesse.

Schlagwörter (Deutsch):	Data Lifecycle Management, Repositorien, Schnittstellen, Data Ingest, Data Citation
Keywords (English):	Data Lifecycle Management, repositories, API, Data Ingest, Data Citation

Cluster H – Lifecycle Management

Cluster H beschäftigte sich mit vorwiegend technischen Aspekten des Data Lifecycle Management. Dabei wurden vor allem Services und Schnittstellen erörtert, welche Datenrepositorien anbieten sollten, um Forschungsdaten als vertrauenswürdige Archive langfristig verfügbar und nutzbar zu halten. Dies umfasst eine Reihe von Prozessen in den Bereichen Ingest (z.B. Datenübernahme, Metadaten, Qualitätssicherung), Verwaltung der Daten und Metadaten (z.B. Versionierung bei Ergänzungen, Korrekturen, Neuhinzunahme von Daten, Signatur, Provenance Trails, Migrationen), sowie Zugriff (z.B. Suche, Zitierbarkeit von arbiträren Subsets von Daten, Darstellung, Einzel- vs. Massenzugriff auf Daten).

Zu Beginn der Clusteraktivitäten wurden eine Reihe von Services bzw. Infrastrukturkomponenten identifiziert, die im Rahmen einer Forschungsdateninfrastruktur angeboten werden sollten, um Forschende, Fördergebern, und Institutionen einen effizienten und nachhaltigen Umgang mit Forschungsdaten zu ermöglichen. Diese wurden in der Folge diskutiert, um zu konkreten Empfehlungen bezüglich ihrer Umsetzungen zu kommen.

Hier wurde insbesondere offensichtlich, dass die einzelnen Partnerinstitutionen sehr heterogen sind, sowohl was die anfallenden Datentypen, -mengen, und Problemstellungen betrifft, als auch den Reifegrad der Institutionen und ihre Lage, entsprechende Services umsetzen und anbieten zu können. Weiters zeigte sich in den Diskussionen eine große Kluft hinsichtlich des Bewusstseins, welche Services benötigt werden.

Da oftmals auch auf Seite der Forschenden die Fragen des Forschungsdatenmanagements, Open Access und Nachnutzung nicht sehr präsent sind, gibt es von Seiten der Nutzer nur geringe bzw. keine Anforderungen an die Institutionen, entsprechende Services anzubieten. Dies ändert sich jeweils schlagartig, wenn derartige Services zunehmend verpflichtend für Forschungsprojekte erforderlich werden. Als Beispiel sei hier die Erstellung von Data Management Plänen (DMPs) genannt, welche von Fördergebern in zunehmenden Detailierungsgraden vorgeschrieben wurde und so innerhalb relativ kurzer Zeit ein großer Bedarf an Unterstützung in diesem Bereich spürbar wurde.

Dies lässt in sehr naher Zukunft eine massive Zunahme an den Services erwarten, welche zur Umsetzung der in den DMPs angegebenen „Versprechen“ erforderlich sind (Datenhaltung, Zertifikation von Trustworthy Repositories welche entsprechende Services anbieten, etc.).

Die als am bedeutsamsten eingestuften Bereiche umfassen:

1. Repositorienlösungen

Die Repositorienlösungen müssen für unterschiedliche Datenarten optimiert werden. Weiters sind Empfehlungen notwendig, wie geeignete Speicherinfrastrukturen für die jeweiligen Datentypen aussehen können, um z.B. die unterschiedlichen Anforderungen auf Storage-Ebene an *high-volume* RDBMS Systeme im Unterschied zu file-basierten Lösungen für große Videodatenbestände abbilden zu können.

In diesem Zusammenhang wären vorkonfigurierte Sets von Repository Tools wünschenswert, welche das rasche Aufsetzen einer Reihe von Repositorien, optimiert für

unterschiedliche Datentypen inklusive verbindender Interfaces, Querverknüpfungen zwischen Repositorien mit nach außen einheitlichen Schnittstellen, erlauben. Diese sollen auf technischer Ebene (in Kombination mit entsprechenden organisatorischen Maßnahmen) die Anforderungen an Trustworthy Repositories erfüllen, sollten im Idealfall mandantenfähig sein, um kooperative Strukturen zu erlauben und dabei hohe Security und Privacy-Anforderungen für sensible Daten erfüllen.

Prinzipielle Typen umfassen dabei

- File-basierte Repositorien (weiter unterschieden nach Anzahl und durchschnittlicher Dateigröße) für Dokumente, Bilder, Videos, NETCDF, uvm.
- SQL Datenbanken/RDBMS
- XML Datenbanken (bzw. evtl. auch dateibasiert)
- komplexe Objekte (z.B. Kombination von Bild plus positionsbezogene Annotationen; 3D-Objekte plus Metainformation wie z.B. Gebäudemodelle für Facility-Management)

2. Unterstützung (Tools, APIs, Services) im Bereich Data Ingest

Für verschiedene Datentypen, deren Qualitätssicherung sowie metadaten-Extraktion, Collection Profiling, bzw. Standardisierungs-/Migrationstools. Diese sollen flexibel über Plug-ins erweiterbar sein und sowohl vollautomatische Prozesse, als auch interaktive/teilautomatisierte Prozesse zulassen. Datentypen umfassen dabei:

- File-basierte Daten (Sammlung von Objekten/Dokumenten in unterschiedlichen Dateiformaten)
- SQL Daten: Datenbank-Engine mit Abfrage/Suchmöglichkeiten
- CSV Dateien, (evl. intern mit Umwandlung in SQL)
- Datenstreams (file-basiert, SQL-basiert, Anforderungsunterschiede in Organisation bzw. Ingest)
- Binary Data Files

3. Data Citation Services

Die Dara Citation Services sollen es Forschenden ermöglichen, beliebige Subsets von potentiell dynamischen Daten präzise und eindeutig zu identifizieren und zu referenzieren. Referenzen müssen korrekt aufgelöst werden, auch wenn die zugrundeliegenden Daten sich mittlerweile geändert haben (neue Daten hinzugekommen, alte gelöscht, Werte korrigiert wurden). In diesem Bereich sollte auf die Empfehlungen der Arbeitsgruppe zu Dynamic Data Citation (WGDC) der Research Data Alliance (RDA) zurückgegriffen werden.

4. Data / Process Management Plans

Es sollten Software Tools zur Unterstützung im Data / Process Management geschaffen werden. Diese können konzeptionell ähnlich wie entsprechende existierende Lösungen wie z.B. das Data Management Plan (DMP) Toolkit der DCC (https://dmponline.dcc.ac.uk/) aufgebaut sein, müssen jedoch über die dort gebotene Funktionalität der textuellen Erstellung eines Plans hinausgehen. Insbesondere sollen diese Tools maschinen-verarbeitbare Darstellungen eines Data Management Plans liefern und somit sowohl Forschenden, als auch Datenzentren und Fördergebern in den weiteren Phasen des Data Lifecycles unterstützen.

In diesem Zusammenhang ist eine engere Integration mit den jeweiligen Repositorien anzustreben um z.B. eine automatische Verifikation der im DMP angegebenen Elemente (wie z.B. die redundante Speicherung, Vergabe von PIDs, Zugriffsregelungen, etc.) zu ermöglichen. Weiters sollten die Data Management Pläne massiv um Komponenten zur Erfassung von Prozessmetadaten erweitert werden.

5. Cost Estimation Tools

Cost Estimation Tools sollen bei der Abschätzung der Kosten für die LZA und Verfügbarhaltung von Daten unterstützen. Diese sollen sowohl für Datenzentren in der Planung ihrer Ressourcen bzw. zur Berechnung der Kosten der angebotenen Services, als auch Forschenden in der Budgetplanung bei Projektanträgen helfen, entsprechende Kostenschätzungen für die langfristige Sicherstellung der Nachnutzbarkeit der Daten abzugeben. Als Vorlagen können hier die Arbeiten des 4C Projekts (collaboration to Clarify the Cost of Curation, http://4cproject.eu/) herangezogen werden.

6. Bitstream-Preservation

Es sollten kooperative Speicherlösungen angedacht werden, welche eine geographisch redundante Sicherung der Forschungsdaten erlauben. Services umfassen dabei insbesondere

- redundante Speicherung / Replikation
- automatische Verifikation, Abgleich, Signatur
- Sicherstellung von Security und Privacy Anforderungen, evtl. mit Unterstützung hinsichtlich K-Anonymity, l-Diversity, *watermarking*, *fingerprinting* bzw. Verschlüsselung soweit erforderlich.

7. Infrastrukturen für Software und Forschungsprozesse

als Teil von Forschungsdaten: Es wurde die Notwendigkeit ersichtlich, das Konzept von Forschungsdaten auf die im Rahmen der Forschungsarbeiten erstellte bzw. zum Einsatz kommende Software zu erweitern, da Letztere wesentlicher Bestandteil der Forschungsergebnisse ist und für die Nachvollziehbarkeit der Daten (Provenance) als auch Nachnutzbarkeit bzw. Reproduzierbarkeit von Forschungsergebnissen essentiell ist. Das umfasst unter anderem Services zur

- Archivierung von Software
- Version Control Systems (SVN, GIT, ...) unmittelbar im Umfeld der Erstellung von Forschungssoftware bzw. bei deren Einsatz (z.B. YesWorkflow, NoWorklflow, Workflow Engines wie Taverna, Kepler, etc.)

- Unterstützung von Continuous Integration wie z.B. Automatischer build, durch committ ausgelöst, welcher kompilieren, testen, *packaging* etc. beinhaltet und es so erlaubt, auch einfach ein *nightly release package* zu erstellen.
- Software Package Distribution wie z.B. Launchpad PPAs (https://launchpad.net/ubuntu/+ppas), z.B. : Git Package, https://launchpad.net/~git-core/+archive/ubuntu/ppa
- Execution platforms, e.g. Taverna Server (http://www.taverna.org.uk/documentation/taverna-2-x/server/). Diese würden es vor allem erleichtern, einen Workflow mit anderen zu teilen, da kein lokales software setup nötig wäre.
- Monitoring und Log-File Analyse

Weiters wurde in Kooperation mit Cluster K in der Folge ein Pilot zur Umsetzung eines ausgewählten Services (Data Citation) gestartet. Hier wurde im Rahmen von Workshops mit dem Climate Control Centre Austria (CCCA) die Anforderungen erhoben sowie ein Implementierungsplan zur Umsetzung der Empfehlungen der RDA WGDC ausgearbeitet. (siehe Deliverable Cluster K).

Zusammenfassend kann gesagt werden, dass es einerseits eine große Heterogenität der beteiligten Institutionen hinsichtlich der Anforderungen, besonders aber hinsichtlich des Reifegrads in der Umsetzung derartiger Services gibt, jedoch andererseits ein massiver Bedarf sowie großes Potential in der Kooperation und Spezialisierung der einzelnen Institutionen gesehen wird. Da de-fakto alle Institutionen die oben angeführten Services (tlw. in unterschiedlicher Ausprägung) benötigen, die Institutionen jedoch unterschiedliche Schwerpunktexpertisen besitzen, erscheint eine kooperative Entwicklung bzw. weiterführender kooperativer Betrieb einzelner Services sinnvoll.

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)*