ne umfassende Untersu-
mit Forschungsdaten in Öster-
0 WissenschafterInnen von 20
täten sowie drei
Forschungseinrich-
reich beteiligten
frage2. In Bezug

# E-INFRASTRUCTURES AUSTRIA DELIVERABLE Cluster K

raum von Juli
tätsbibliothek Wien unterschiedliche europäi-
cies sowohl formell als auch inhaltlich analy-
ngungen für ein kompetentes Forschungsda-
Beginn 2016 die ExpertInnengruppe - Strate-

ter-
schungs-
licy an österreichi-
ngen zu erstellen. Das vorliegende Doku-
schungseinrichtung lokalisiert und an die
enen Institution angepasst werden
er Sitzung der ExpertInnengrup-
Juni 2016 verabschiedet.

e-infrastructures austria · Deliverable

# Bericht Data Citation Pilot

## Zwischenergebnisse des Data Citation Piloten für Klimadaten am CCCA Datenzentrum

| | |
|---|---|
| Work-Package-Cluster: | Cluster K: Daten aus wissenschaftlichen und künstlerisch-wissenschaftlichen Forschungsprozessen (Entwicklung und Erschließung der Künste) |
| Leitung des Clusters: | Bernhard Haslhofer AIT bernhard.haslhofer@ait.ac.at<br>Michela Vignoli AIT michela.vignoli@ait.ac.at |
| Datum: | 17.06.2016 |
| Version: | 1.1 |
| AutorInnen/ Sonstige Beteiligte: | Chris Schubert CCCA chris. schubert@ccca.ac.at<br>Michela Vignoli AIT michela.vignoli@ait.ac.at<br>Andreas Rauber TU Wien rauber@ifs.tuwien.ac.at<br>Raman Ganguly Uni Wien raman.ganguly@univie.ac.at<br>Georg Seyerl CCCA georg.seyerl@ccca.ac.at<br>Armin Leuprecht Wegener Center armin.leuprecht@uni-graz.at<br>Tomasz Miksa TU Wien tmiksa@sba-research.org |
| Kurzbeschreibung (Deutsch): | Im Rahmen des e-Infrastructures Austria Projektes wurde ein Pilot für die Zitierung von Klimadaten initiiert, der vorsieht eine entsprechende technische Lösung am CCCA |

| | |
|---|---|
| | Datenzentrum zu implementieren. Der Pilot wurde in Zusammenarbeit mit Cluster D (R. Ganguly), Cluster H (A. Rauber), der RDA Data Citation Arbeitsgruppe und dem CCCA Datenzentrum umgesetzt. Dieser Bericht stellt die Zwischenergebnisse mit Stand Juni 2016 vor. |
| Description (English): | In context of the e-Infrastructures Austria project a pilot aiming at implementing a technical solution for data citation at the CCCA Data Center was initiated. The pilot implementation was done in close collaboration with clusters D (R. Ganguly) and H (A. Rauber), the RDA Data Citation Working Group, and the CCCA Data Center. This report presents the interim results achieved by June 2016. |


| | |
|---|---|
| Schlagwörter (Deutsch): | Data Citation, Pilot |
| Keywords (English): | data citation, pilot |
| Verwandte Dokumente/ Related Documents | Rauber et al, Data Citation of Evolving Data. Recommendations of the Working Group on Data Citation (WGDC), Revision of October 20th 2015. |

# Inhalt

# 1. Einleitung

Die Veröffentlichung von Daten, insbesondere von raumbezogenen Daten in einer Web basierten Dateninfrastruktur erfordert ein Maß an strukturierter Beschreibung von Meta-Informationen, den Metadaten. Anhand der Metadaten ist es technisch möglich die Daten selbst aufzufinden, zu katalogisieren, zu klassifizieren und automatisiert abzufragen. Die Struktur und der Inhalt von Metadaten werden in Standards, Profilen und individuellen Adaptionen festgelegt. Anhand dieser Vorgaben können für Datensätze Referenzen, Quellverweisen zu Rohdaten, Daten Aggregate, angewandten Methoden, Algorithmen, Publikationen, etc. beschrieben werden. Jedoch kann die technische Repräsentation dieser Zitate auf sehr heterogene Art und Wiese erfolgen, so dass das explizite Wissen für andere Anwender nicht zugänglich ist.

Erst langsam, mit der immer größeren Etablierung von Daten und interoperablen Daten Diensten in einer SDI (Spatial Data Infrastructure), der Open Data Initiative, dem Daten Management und der Daten Archivierung findet eine Sensibilisierung in der administrativen und wissenschaftlichen Praxis in der gleichen Weise statt, wie auf Publikationen oder Artikel in Forschung und Wissenschaft verwiesen wird.

Mit dem Climate Change Center Austria[1] (CCCA), das sich gerade im Aufbau befindet, fand das e-Infrastructures Austria Projekt einen engagierten Partner, um die Umsetzung eines Data Citation Piloten in die Wege zu leiten und zu begleiten. Der Pilot ermöglicht dem CCCA Datenzentrum die eigene Infrastruktur dahingehend zu verbessern, dass sowohl Qualität als auch Erreichbarkeit der darauf verfügbaren Daten deutlich erhöht werden, was der gesamten CCCA Community zu Gute kommt.

Das e-Infrastructures Austria Projekt profitiert vom Piloten dahingehend, dass im Data Citation Bereich bestehende Synergien genutzt und international erhöhte Sichtbarkeit erlangt werden kann. Dies wird insbesondere durch die enge Kooperation mit der Research Data Alliance[2] Arbeitsgruppe zu Data Citation gewährleistet.

Der Zweck des Data Citation Piloten ist die vornehmlich textuellen Projekt-Outputs mit einer Praxisanwendung zu bereichern. Die Zielsetzung im Rahmen des Projektes ist die Pilotumsetzung bis zum Ende der technischen Implementierung Ende 2016 fachlich und beratend zu begleiten. Die erste Implementierungsphase läuft gerade und wird im August 2016 abgeschlossen werden. Um die Funktionalitäten, die in der zweiten Implementierungsphase vorgesehen sind umzusetzen, besteht das Interesse den Piloten über das e-Infrastructures Projekt hinaus weiterzuführen (siehe **Fehler! Verweisquelle konnte nicht gefunden werden.**).

[1] http://www.ccca.ac.at
[2] https://rd-alliance.org/

## 2. Arbeitsbeschreibung

Im Rahmen des e-Infrastructures Austria Projektes wurde ein Pilot für die Zitierung von Klimadaten initiiert, der vorsieht eine entsprechende technische Lösung am CCCA zu implementieren. Mit der Einführung einer konsequenten und konsistenten Data Citation mit klar definierter Angabe, maschinenlesbare Repräsentation zur Versionierung, Zeitangaben, verfügbarerer Ort, Data Policies und die Vergabe eindeutiger Identifier (PID), die auf eine echte „lesbare" Web Ressource verweisen, steht damit das implizite Wissen von Daten und der dahinter stehenden Informationen sowie Prozesse der wissenschaftlichen Gemeinschaft und der Allgemeinheit zur Verfügung.

Der Pilot wurde in Zusammenarbeit mit Cluster D (R. Ganguly), Cluster H (A. Rauber), der RDA Data Citation Arbeitsgruppe und dem CCCA Datenzentrum umgesetzt, wobei das CCCA für die technische Implementierung zuständig ist. e-Infrastructures Austria und die RDA Arbeitsgruppe fungieren als Consultants. Somit nützt die UAG vorhandene Synergien optimal aus und erweitert die Sichtbarkeit über e-Infrastructures hinaus.

Die Implementierung der technischen Komponente, die durch das CCCA (Ch. Schubert) durchgeführt wird, ist noch nicht abgeschlossen. Mit Ende des Projektberichtzeitraumes wurde das technische Konzept für die Implementierung durch die UAG Mitarbeiter ausgearbeitet.

Das CCCA Datenzentrum stellt den Zugang von verteilten Informationen der CCCA Mitglieder und anderen Institutionen, als Daten Provider, zur Klimaforschung in Österreich sicher. Unter Zugang der verteilten Informationen ist die Integration, Auffinden und Verfügbarmachen verschiedener Datentypen (Primärdaten, Metadaten, Projektberichte etc.), zur Klimaforschung relevanter Daten, Modelle als auch Modellergebnisse vorgesehen.

## 3. Ziel des Piloten

Auf technischer Ebene bezweckt das Vorhaben einerseits die Entwicklung von Datenzitierungs-methoden im Allgemeinen voranzutreiben, andererseits die Funktionalitäten der CCCA Daten Plattform zu erweitern und die technische Implementierung (wissenschaftlich) zu begleiten. Aktuelle Standards und Empfehlungen werden auf ausgewählten Testdatensets angewandt und in der CCCA Daten Plattform implementiert. Auf praktischer Ebene soll den CCCA Partnern und Nutzern ermöglicht werden auf bestimmte Datensätze, Sub-Datensätze und deren Versionen mittels eines Persistent Identifiers eindeutig zu verweisen (z.B. auf einen bestimmten Datensatz, der einem Paper zugrunde liegt).

Der Data Citation Pilot für Klimadaten und dessen Implementierung im CCCA Datenzentrum erfolgt in enger Abstimmung mit den Anforderungen und Ergebnissen aus dem Projekt e-Infrastructures Austria und in Kooperation mit der RDA Data Citation Arbeitsgruppe, die umfangreiche Empfehlungen zu Data Citation[3] ausgearbeitet haben.

[3] https://rd-alliance.org/group/data-citation-wg/outcomes/data-citation-recommendation.html

Mit dem Data Citation Ansatz wird jede individuelle Datenbank-Abfrage und deren Ausgabe zwischengespeichert und kann nach Belieben erweitert werden. Diese Abfrage generiert einen sogenannten Subdatensatz. Wird dieser für den Anwender und seine (wissenschaftliche) Fragestellungen als hinreichend angesehen wird ein Prozess gestartet, der eine Persistierung mit einem wiederaufrufbaren Unique Identifier (PID) vornimmt. Dieser wird permanent in einer Datenbank, einem Register, gespeichert. Dadurch wird dem Nutzer ermöglicht in seiner Publikation, Präsentation bzw. wissenschaftlichen Untersuchung den Subdatensatz auf Dauer zu zitieren. User können auf den referenzierten Datensatz zugreifen ohne den Gesamtdatensatz herunterzuladen oder auf dessen Aktualität, Version achten zu müssen.

Für den Data Citation Piloten werden die Daten aus dem Projekt ÖKS15[4] (Österreichische Klimaszenarien bis 2100) für eine Best Practice Umsetzung verwendet. Mit Hilfe von Empirisch-Statistischen Downscaling (ESD)-Methoden werden wichtige klimatologische Parameter aus den EURO-CORDEX RCMs (regional climate models) auf einen hochaufgelösten 1x1 km² Raster gebracht und daraus Klimaänderungssignale über verschiedenste Klimaindizes von 1970 bis 2100 berechnet. Die ausgegebenen Klimaparameter sind Temperatur, Niederschlag und Globale Sonnenscheindauer auf ein Raster und Stationen bezogen, Die abgeleiteten Klimaindizes (insgesamt 33) sind u.a. das Temperatur Mittel, Sommer-, Hitzetage, tropische Nächte, Dauer der Hitze-, Kältewelle, etc.

Diese Daten sind im NetCDF (Network Common Data Format) abgelegt. NetCDF ist ein offener Standard, als maschinenunabhängiges Datenformat entwickelt wurde und hauptsächlich in der Wissenschaft für die strukturierte Ablage mehrdimensionaler Daten, in einer Art Container gehalten. NetCDF-Daten beinhalten Attribute, Dimensionen und Variablen. Ein Attribut hat einen Namen und einen Wert und kann mit einer Variablen assoziiert werden. Dimensionen werden benutzt um die Größe der Variablenfelder zu definieren. Die Variable ist der Datencontainer für einen Einzelwert oder einer komplexen Daten-Matrix. Der Datentyp, die Anzahl an Dimensionen und notwendige Attribute werden deklariert. Als Beispiel: Die Temperaturdaten werden als gerasterte, georeferenzierte Daten in mehreren Einzeldateien getrennt nach Region für jeweils einen Tag, eine Woche und einen Monat bereitgestellt. Die Werte eines Rasterpixels entsprechen den gemittelten Temperaturwerten für den jeweiligen Zeitraum am entsprechenden Ort.

Der gesamte Datensatz (ca. 3,5 TB) wird über das CCCA Datenzentrum frei verfügbar gemacht. Die Zusammenhänge zwischen den Datensätzen, den verwendeten Methoden und den Resultaten werden als Verlinkungen abgebildet. Damit wird erstmals der gesamte Prozess von der meteorologischen Messung bis zum Factsheet für Entscheidungsträger völlig transparent für jedermann nachvollziehbar.

[4] Barbara Chimani et al., ÖKS 15: Hochaufgelöste, biaskorrigierte Klimaszenarien für Österreich. In Tagungsband 17. Österreichischer Klimatag, 6.–8. April 2016, Graz, S. 34-35

# 4. Status und weitere Schritte über e-Infrastructures Austria hinaus

Dieser Bericht stellt die Zwischenergebnisse mit Stand Juni 2016 vor, der auch am Fortbildungsseminar für Forschungsdaten und e-Infrastrukturen als Use Case aus der Klimaforschung vorgestellt wurde[5]. Es ist vorgesehen im letzten Projekthalbjahr weitere Dokumentation im **RDA-Wiki**[6] zur Verfügung zu stellen. Die Dokumentation im Wiki wird im Dezember 2016 abgeschlossen werden.

Die Arbeitsschritte sind in zwei Phasen untergliedert:

1. Phase: Implementierung von Data Citation auf File basierter Abfrage (ca. 1700 Files);
2. Phase: Implementierung einer parameterfreien Abfrage zu filtern, um Subsets File-unabhängig zu generieren.

Das CCCA Datenzentrum führt derzeit die technische Implementierung der 1. Phase des Data Citation Piloten durch. Mit Abschluss dieses Piloten wird eine abgestimmte Handlungsempfehlung entstehen, die wissenschaftlich fundiert ist, und sowohl Anforderungen internationaler Standards als auch denen des österreichischen Projektes e-Infrastructures Austria entspricht.

Ende Mai stellte das CCCA Datenzentrum der Community einen Prototypen [7] zu Testzwecken bereit und integrierte den ÖKS Datensatz die Datenbank. Dieser Datensatz steht als Originalfile über sftp den Anwendern als Download zur Verfügung. Phase 1 des Prototypen sieht vor gesamte NetCDF Files herunterzuladen, die im ersten Schritt mit einem PID versehen werden können.

In **Phase 1** wird die Versionierung für die NETCDF Dateien auf File-Ebene (z.B. durch Git) sowie der Metadaten-Tabelle, in welcher die NETCDF Dateien beschrieben werden umgesetzt (siehe Abbildung 1). User wählen über ein Interface ein Set aus solchen Files aus. Diese Query auf die Metadatenbank wird versioniert und gespeichert. Für diesen Ansatz können Standard-Lösungen, d.h. primär die Versionierung der Metadaten-Table (plus Git oder ähnliches für die Files), eingesetzt werden.

[5] http://e-seminar.univie.ac.at/ueber-die-bedeutung/
[6] https://rd-alliance.org/group/data-citation-wg/wiki/wgdc-pilot-ccca
[7] https://sandboxdc.ccca.ac.at/

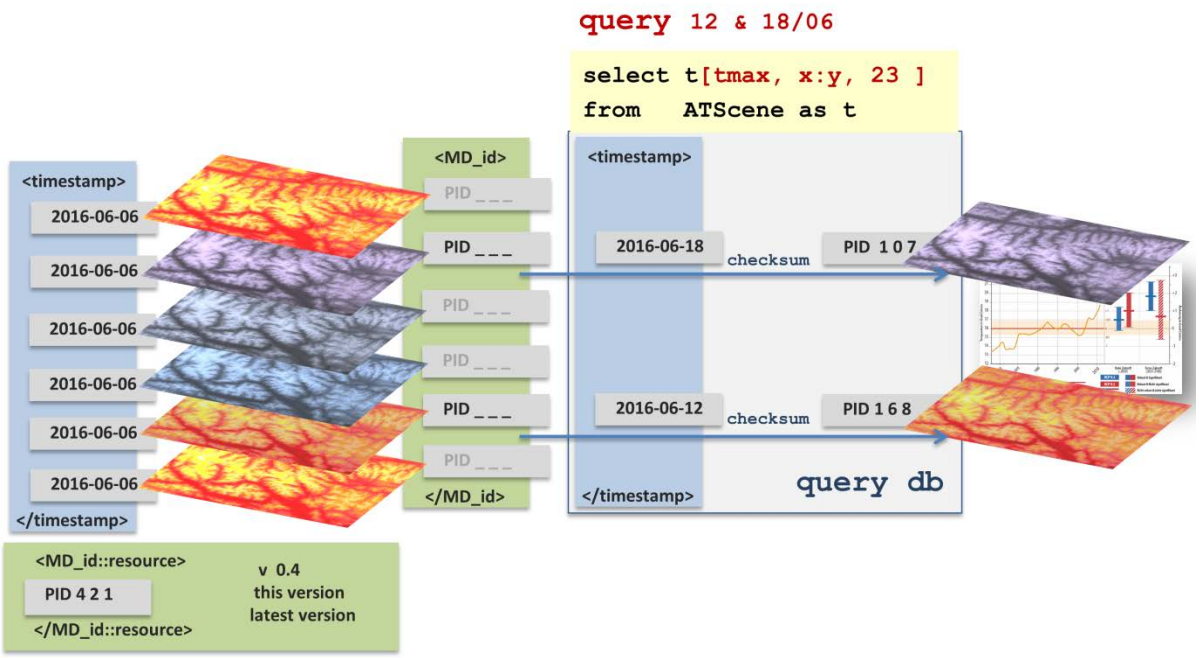


**Abbildung 1 - Phase 1: File basierte Abfrage**

Es wird aber davon ausgegangen, dass sich die potentiellen Nutzer, insbesondere bei so einem hochaufgelösten Datensatz, auf spezielle Regionen konzentrieren, bestimmte Zeitintervalle oder bestimmte Variablen abfragen. Deswegen sieht eine weitere Ausbaustufe für die Zukunft Abfrage auf Sub-Set Ebene vor. Damit User nicht hunderte Files herunterladen müssen, wenn sie jeweils nur einen Slice aus jedem File haben wollen, sollen in **Phase 2** Queries innerhalb von NETCDF Dateien gespeichert werden können (siehe Abbildung 2).

Um diese Funktionalität umzusetzen besteht das Interesse diese Aktivität als Piloten unter Schirmherrschaft des CCCA Datenzentrums und in Kooperation mit der RDA Arbeitsgruppe Data Citation längerfristig laufen zu lassen. Für die Fortführung des Piloten über das e-Infrastructures Projekt hinaus gibt es zurzeit ein Budget von 5000 Euro von Seiten der CCCA.

Im Oktober/November 2016 ist angedacht einen **Workshop** abzuhalten, sobald die erste Implementierung des Data Citation Prototypen abgeschlossen ist und der weiteren Community vorgestellt werden kann. Ziele des Workshops sind die Community auf den Prototypen und die erzielten Ergebnisse aufmerksam zu machen, sowie weiteren Input für die finale Implementierungsphase über das e-Infrastructure Projekt hinaus zu sammeln.

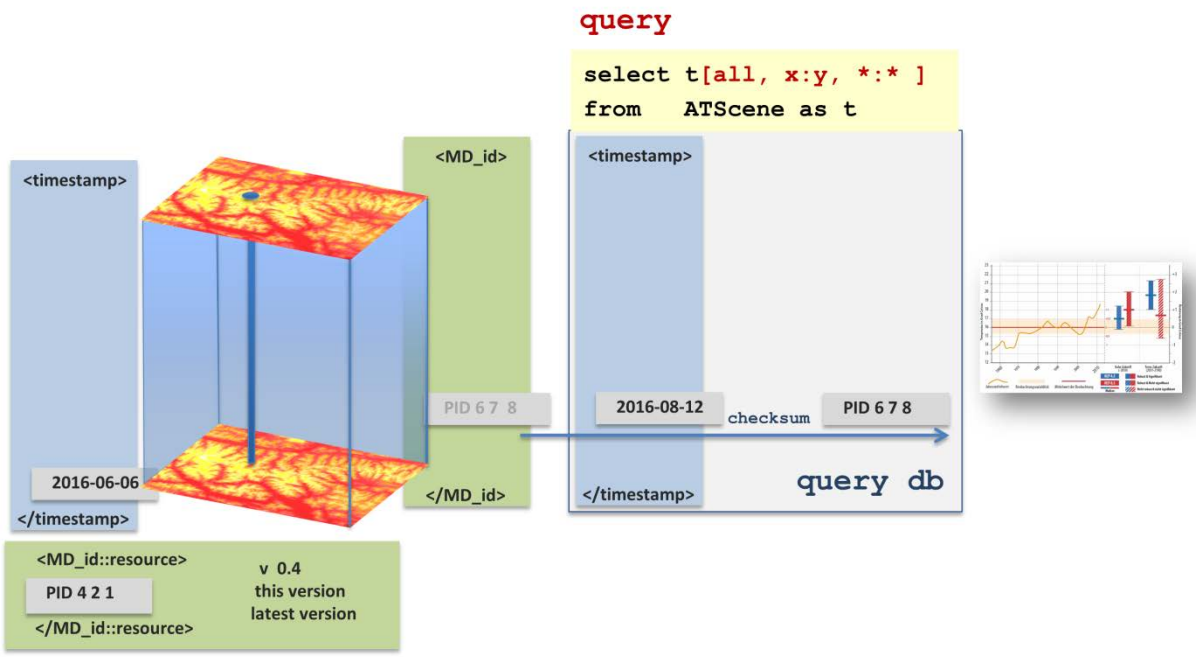


**Abbildung 2 - Phase 2: Sub-Set Abfrage**

## 5. Schlussfolgerungen

Durch die Implementierung des Data Citation Piloten am CCCA Datenzentrum können wertvolle Praxiserfahrungen und Best Practices gesammelt werden, um ein verlässliches und maßgeschneidertes Zitiersystem für dynamische Daten in bestehenden oder entstehenden Datenrepositorien umzusetzen. Dies kommt potenziell allen e-Infrastructures Austria Partnerinstitutionen zu Gute, die gedenken eine Data Citation Lösung für ihr institutionelles Datenrepositorium zu implementieren. Die aus dem Piloten hervorgehenden Handlungsempfehlungen werden Anforderungen internationaler Standards entsprechen. Obwohl der Pilot auf Klimadaten und ein entsprechendes Repositorium angewendet wird ist die angewandte technische Lösung höchst versatil und kann auf verschiedene Datenformate und Dateninfrastrukturen angewendet und angepasst werden.

## 6. Zeitplan

| Task Beschreibung | | Mär | Apr | Mai | Jun | Jul | Aug | Sep | Okt | Nov | Dez |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Elaboration state of the art data citation | | | | | | | | | | | |
| Technical recommendations, data format, data representat. | | | | | | | | | | | |
| Beschreibung der Daten und Prozesskette für ÖKS15 | | | | | | | | | | | |
| Technische Implementierung im CCCA Datenzentrum | Phase I: File based appr. | | | | | | | | | | |
| | Phase II: Sub-Set b. appr. | | | | | | | | | | |

Abgeschlossene Arbeiten

Offene Arbeiten

## 7. Kontakt

**Cluster Leitung:** Michela Vignoli michela.vignoli@ait.ac.at

**Technische Implementierung:** Chris Schubert chris.schubert@ccca.ac.at

**Data Citation:** Andreas Rauber rauber@ifs.tuwien.ac.at

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