

e-Infrastructures Austria

Life Cycle Management: Das digitale Objekt, Content Modelle

Fortbildungsseminar für Forschungsdaten und e-Infrastrukturen

Universität Wien, 6.-9. Juni 2016

Chris Schubert, Andreas Rauber

e-infrastructures
austria

Diese Unterlagen wurden im Rahmen der e-Infrastructures Austria Veranstaltung
Fortbildungsseminar für Forschungsdaten und e-Infrastrukturen
erstellt und stehen im Web unter folgender CC-by Lizenz zur Verfügung:



Dieses Werk bzw. dieser Inhalt steht unter einer Creative Commons Namensnennung 4.0 International Lizenz. <http://creativecommons.org/licenses/by/4.0/>

Seminar-Website: <http://e-seminar.univie.ac.at/>

Projekt-Website: <http://e-infrastructures.at/>

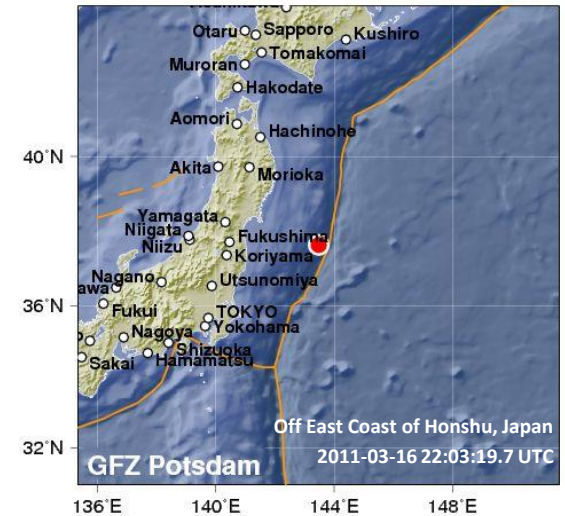
Agenda

- Motivation
- Data Life Cycle
 - Definition, Grundlagen
 - Life Cycle Modelle
 - das DCC Model
- Aktionen zum Data Life Cycle
 - Erfahrungen zur Datenkonzipierung
 - Fallbeispiel zum ‚create or receive‘
 - Daten Transformation
- Access: Daten- vs. Dokumentrepositorien?
- Provenance: Was & Wie?
- Zusammenfassung und Ausblick

Motivation

Datenverlust

- Naturkatastrophen (Feuer, Wasser, Erdbeben, ...)
- Cyber-attacks, Human-attacks,
- ...



<http://iwgcr.org>

Motivation

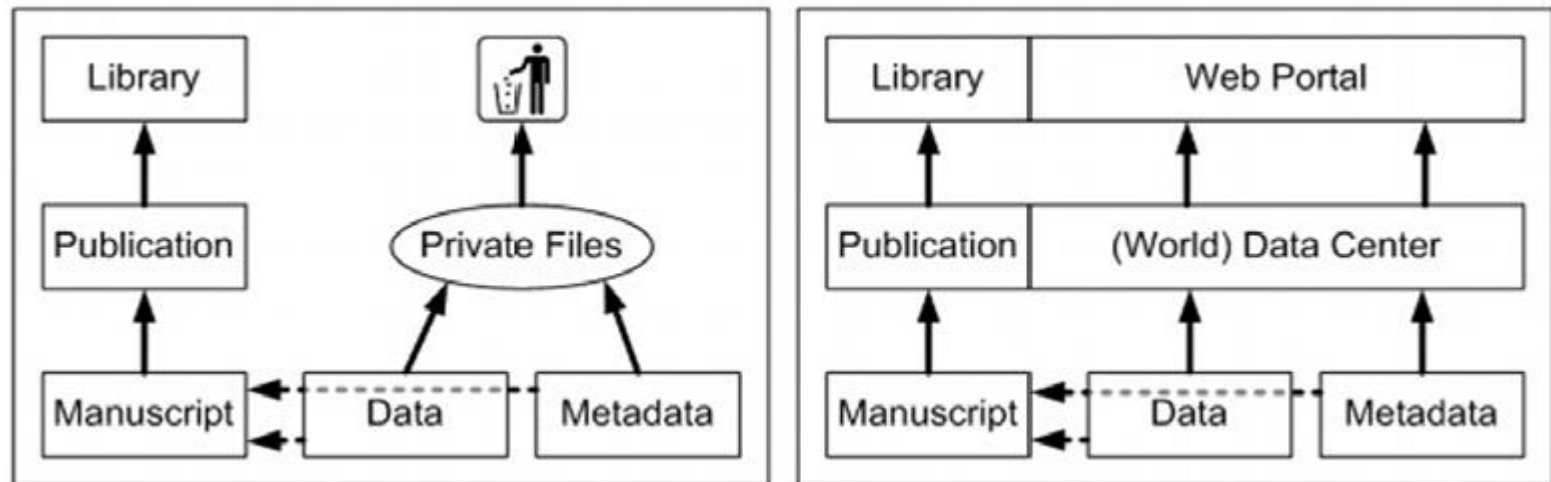
Datenverlust - alltäglich?

- Verlust von Personal, Wissen- & Prozess Kompetenzen
- Verlust des institutionellen Engagements, Finanzierung, etc.

Anforderungen

- Nachhaltigkeit & Nachvollziehbarkeit von Forschungsergebnissen
- Re-usability , Reproducibility, Data Mining, Metastudien rücken in den Vordergrund
- Verwertungskonzept abgelöst durch Data Management Plan

Motivation



aus Handbuch zum Forschungsdatenmanagement, 2011, ISBN 978-3-88347-283-6, ch. 1.2, S. Rümpel

Motivation

- Was sollte getan werden, Aktionen:
 - Daten Management (Planung)
 - welche Daten, Forschungsdomäne, Institution, etc.,
 - Metadaten & Dokumentation (MD Profile & Standards),
 - wer sollte involviert werden in der Planung,
 - Schaffung einer kollaborativen Struktur,
 - Investition in die Infrastruktur, Datenerfassung,
 - Unterstützung der Überprüfung zur Herkunft,
 - Unterstützung durch kontinuierliche, transparente Services

Verständnis erreichen, was passiert mit Daten

Definition des Lebenszyklus von Daten

Data Life Cycle

CEOS Data Life Cycle Models and Concepts
CEOS.WGISS.DSIS.TN.01 Issue 1.0 September 2011

CEOS
Working Group on Information Systems and Services
Data Stewardship Interest Group

Data Life Cycle Models and Concepts Version 1.0

Doc. Ref.: CEOS.WGISS.DSIS.TN.01
Date: 26 September 2011
Issue: 1.0

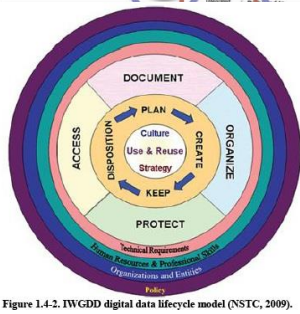
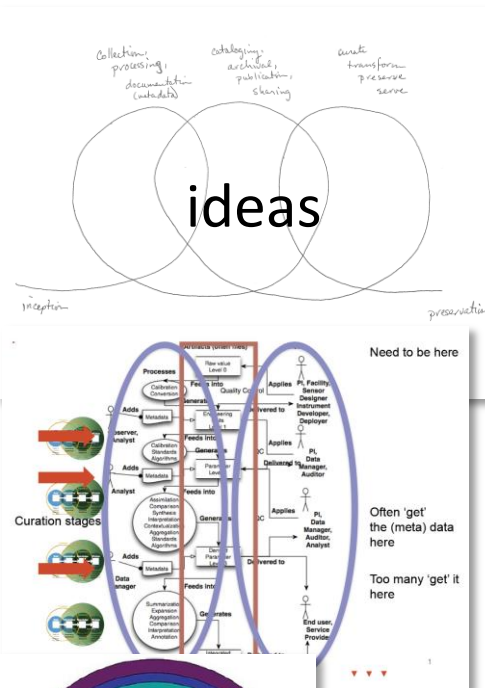
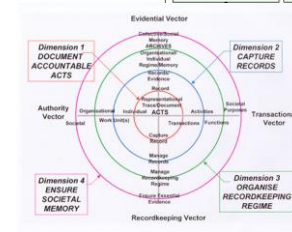
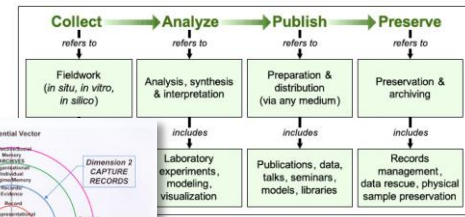
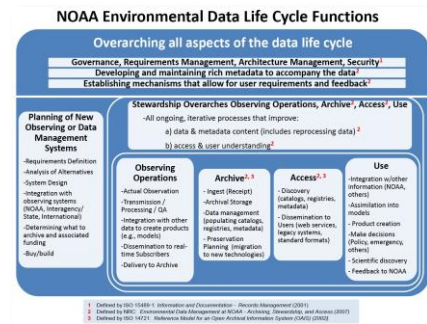
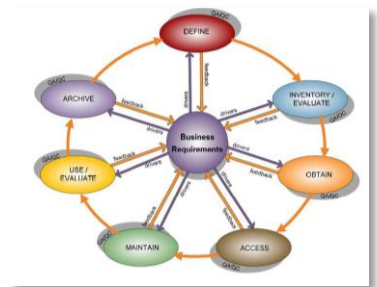


Figure 1.4-2. IWGDD digital data lifecycle model (NSTC, 2009).



Data Life Cycle

„Curation Lifecycle Model“

Digital Curation Centre, UK

“alle” Aktionen als Ideal des Datenmanagements werden abgebildet

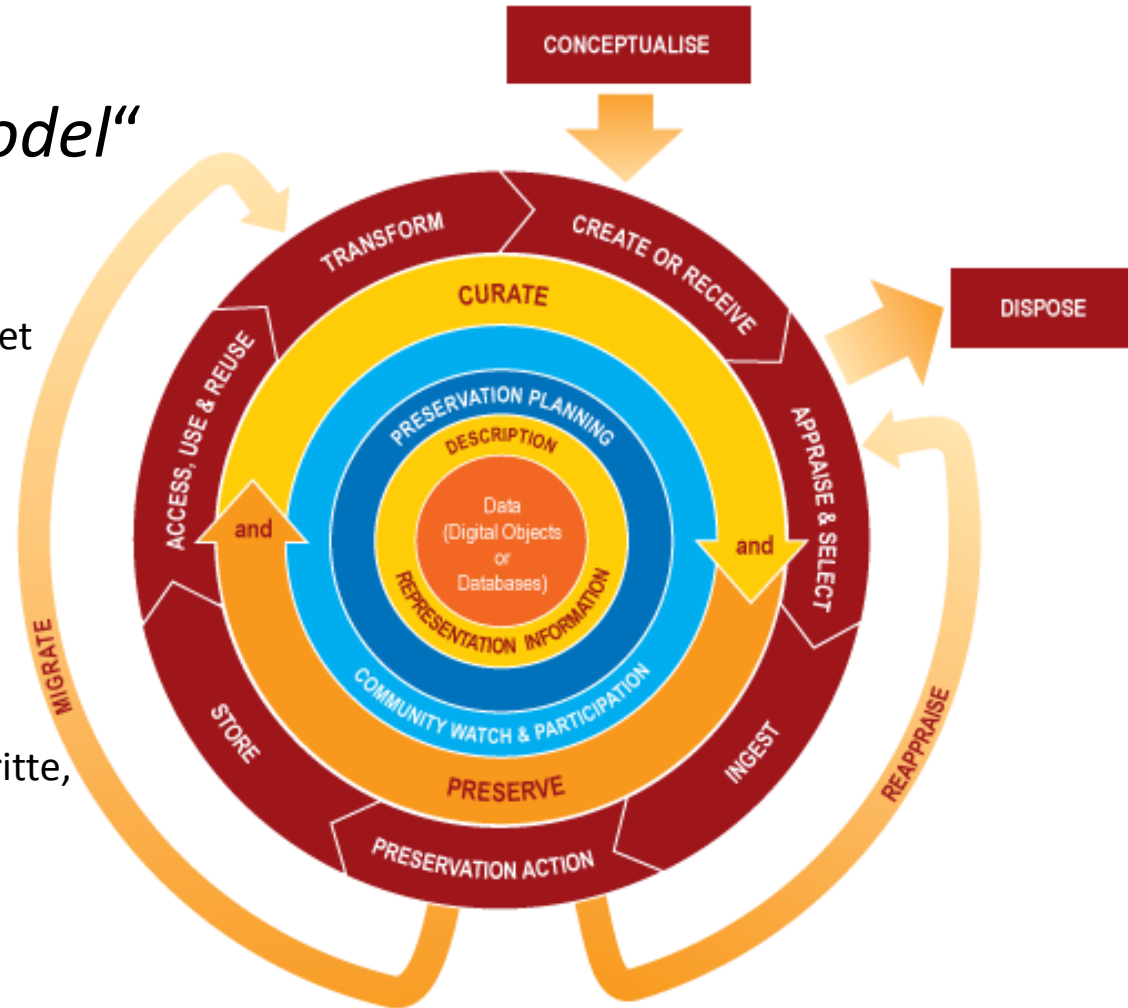
zentriert (Full Lifecycle Actions):

„Data Preservation“ (Datenerhaltung)

„Data Curation“ (Datenpflege)

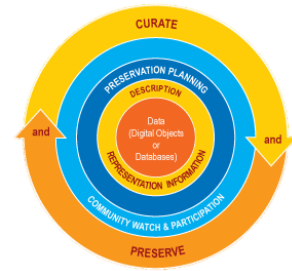
außen (Sequential Lifecycle Actions):

Sequenzielle Aktionen, sind Arbeitsschritte, die definiert werden müssen und über Services, transparent gestaltet.

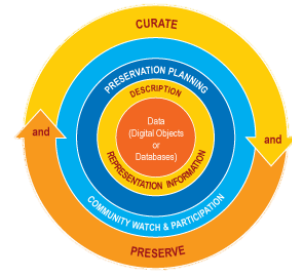


<http://www.dcc.ac.uk/resources/curation-lifecycle-model>

Daten



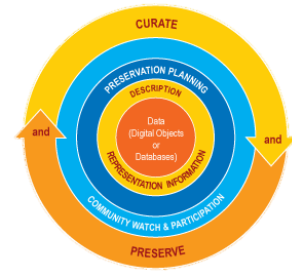
- Daten, Dokumenten, Informationen in binärer digitaler Form
 - einfache digitale Objekte
 - komplexe digitale Objekte
 - Datenbanken, ...



Services Full Lifecycle

Beschreibung und Repräsentation

- Beschreibung von Metainformationen,
- Strukturiert, standardisiert, langfristig verfügbar
 - Metadaten, welches Profil, insbesondere Aussagen (welche Elemente)
 - zum Datenformat, als syntaktische Informationen,
 - Verwendung von kontrolliertem Vokabular, Resource der Thesauri als Semantische Informationen
 - Zusatzinformationen: Software, Algorithmen, Prozesskette inkl. Versionen



Services Full Lifecycle

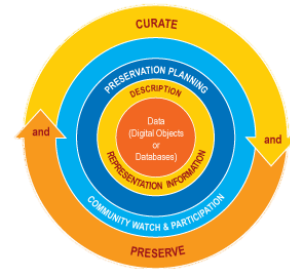
Observation (Community Watch and Participation)

- Transparenz, Erhöhung der Akzeptanz, des Wissens und Know How
- Ermöglichen von Systemspiegelungen, kollaborativen Systemen

Data Curation

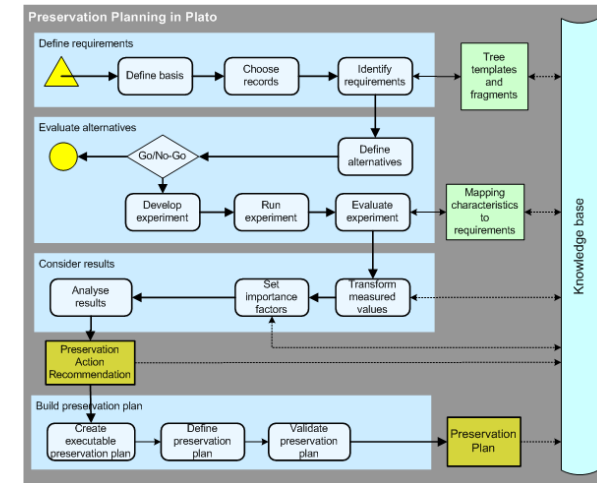
- Organisation der Sicherstellung für die Unversehrtheit von Daten organisiert
- Open Archival Information System (OAIS)

Services Full Lifecycle



Preservation

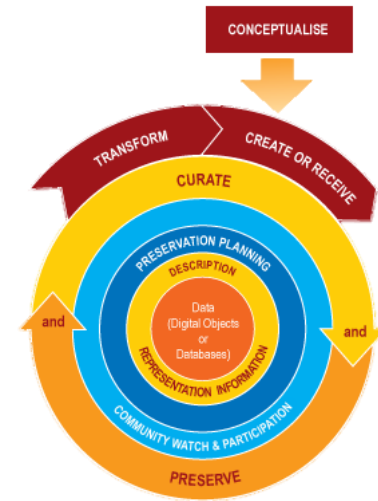
- Sicherstellung des Zuganges
- Hochgradig spezifisch für die jeweilige Anwendungsdomäne und Datenstruktur
- Preservation Planning
 - Verringerung der Komplexität
 - Unabhängig von proprietärer Technologie
 - Stabilitäts-optimiert statt Performance-optimiert
- Langzeitarchivierungsmaßnahmen
 - Konvertierung in XML, CSV
 - Migration in simple Standard SQL-Technologie
 - Robuste Dateiformate (Videocodecs, PDF/A, ...)



Services Sequential Lifecycle

Conceptualise

- Welche Daten?
- Welche Datenmodelle / formate werden unterstützt?
- Speicherbedarf ?
- Bereitstellung eines workflows ?
 - QA Prozess, e.g. ATS (Abstract Test Suite)

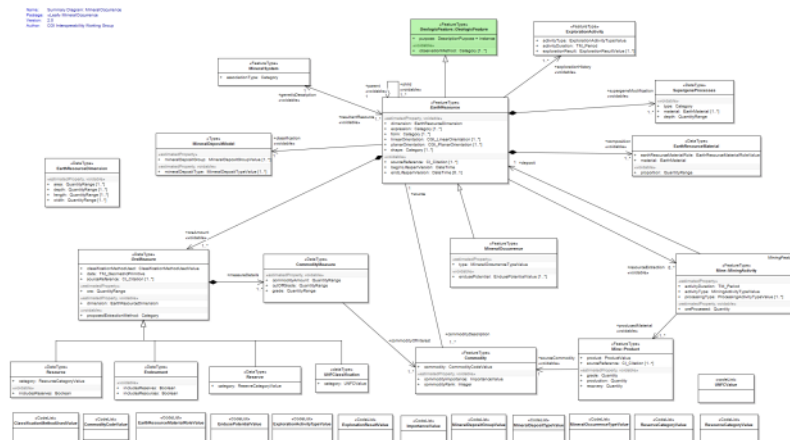


Conceptualise

-



- konzeptionelles Datenmodell

[illegible]

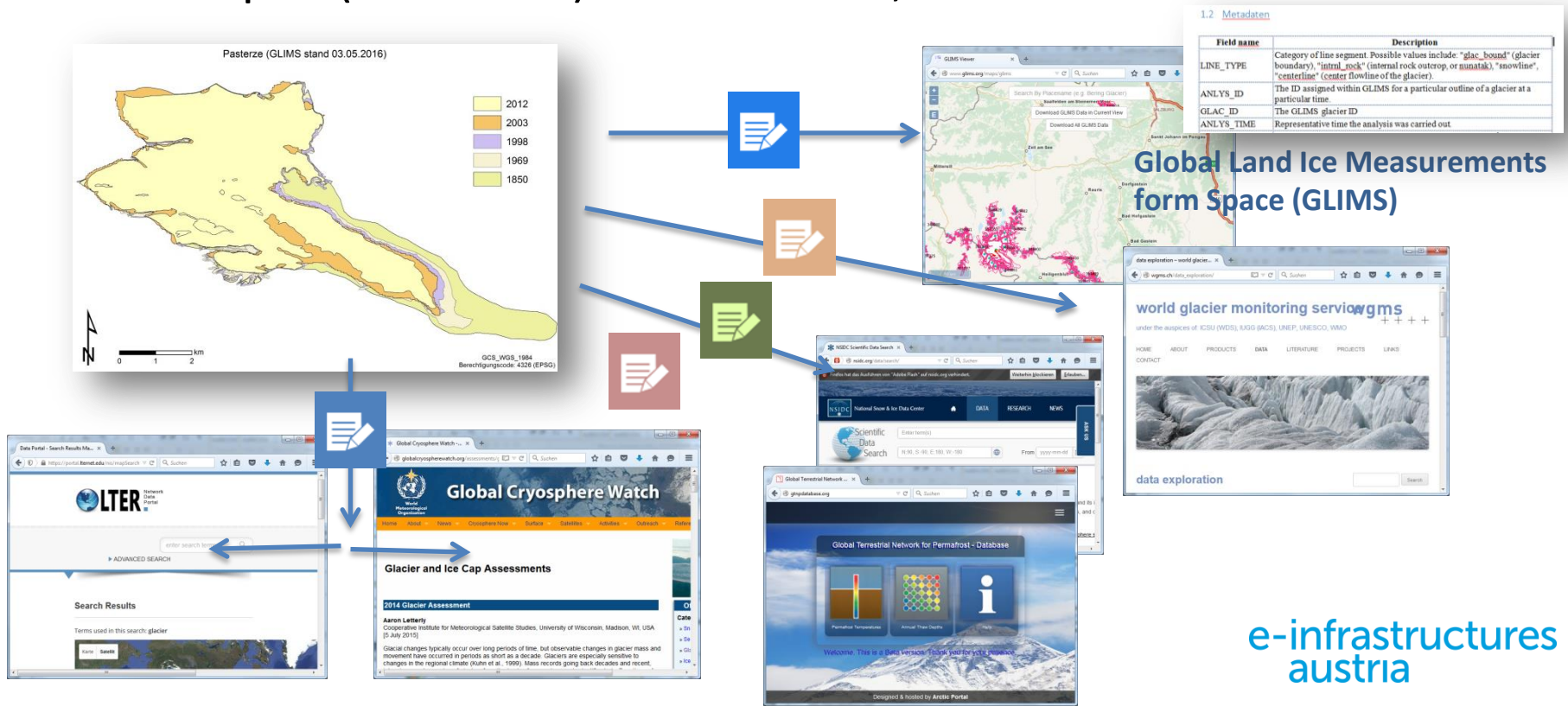
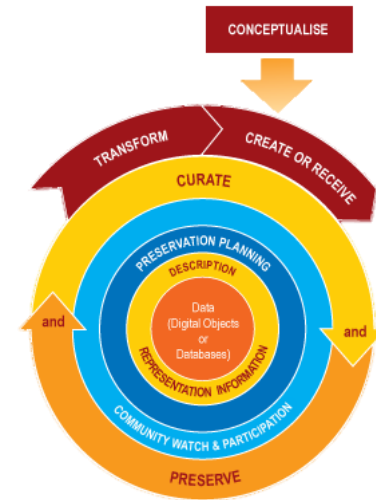
Conceptual data models	Encoding	Harmonised vocabularies	Registers
- objects types, properties & relationships - cross-domain harmonization - based on a common modelling framework - managed in a common UML repository	- conceptual models independent of concrete encodings - standard encoding: GML, but also possible to derive other encodings (e.g. based on RDF)	- to overcome interoperability issues caused by free-text and/or multi-lingual content - allow additional terms from local vocabularies	- provide unique and persistent identifiers for reference to resources - allow their consistent management and versioning

INSPIRE Mineral Resources, EU MS abgestimmtes Model

Services Sequential Lifecycle

Create or receive + (appraise & select)

Fallbeispiel (Metadata): Permafrost -, Gletscher Daten in Österreich

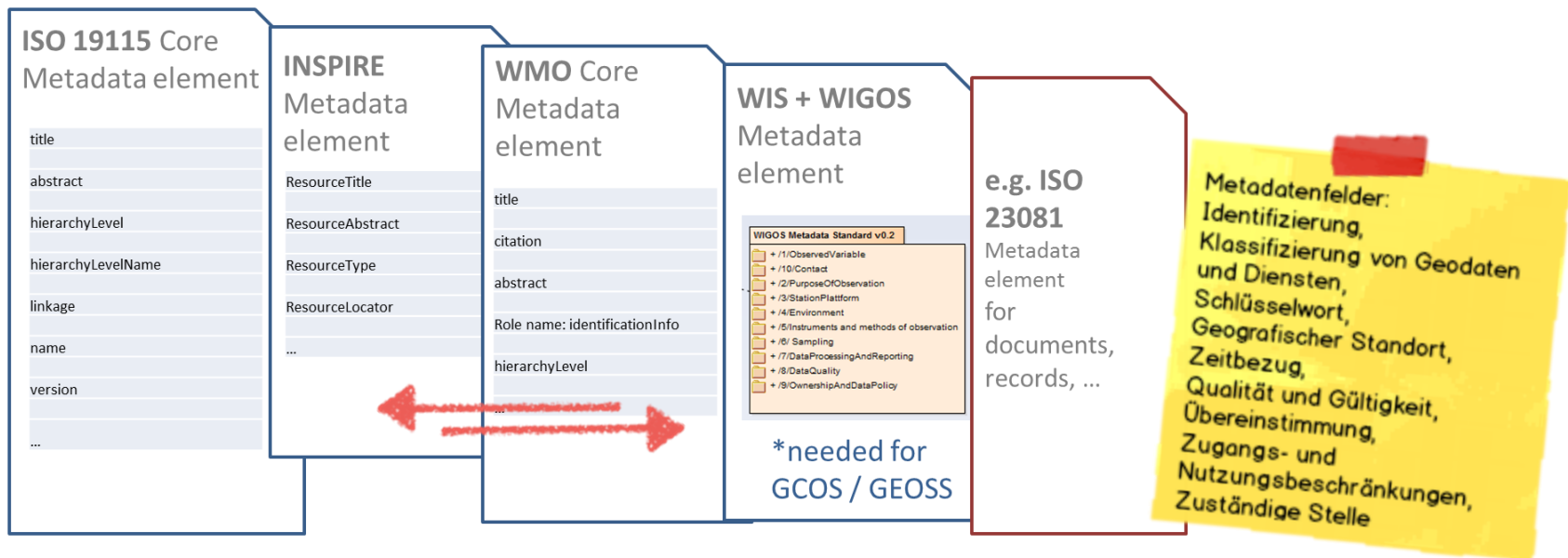


Services Sequential Lifecycle

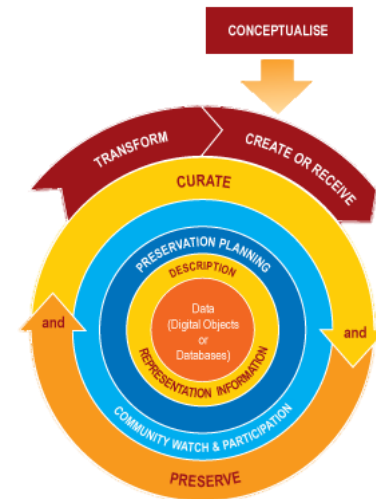
Create or receive + (appraise & select)

- z.B. Metadaten Editor im CCCA Data Centre

Metadatenüberführung (Mapping):



- das Harvesten (bi - directional), CSW als Standard ?



Services Sequential Lifecycle

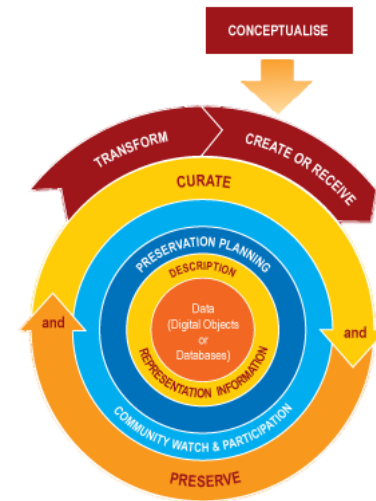
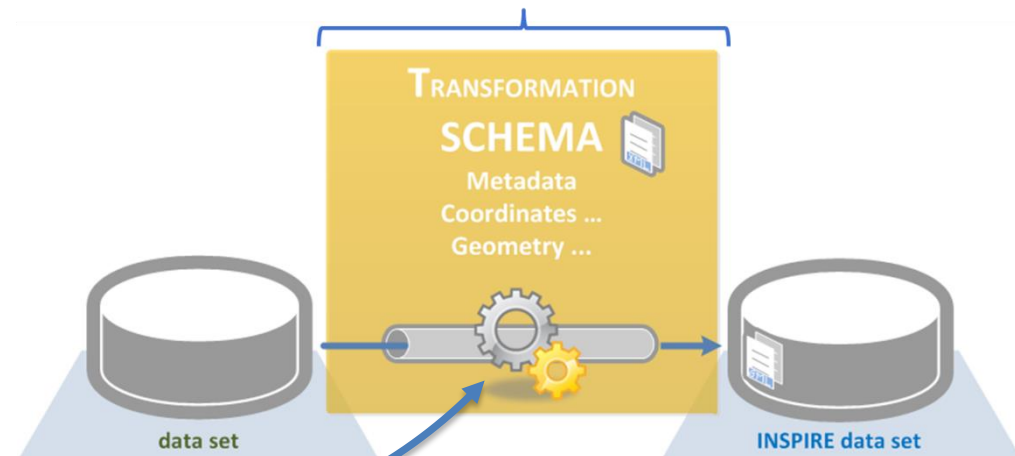
Transform

- erzeugt einen „neuen“ Datensatz aus dem Original
- betrifft auch die Erzeugung von subsets



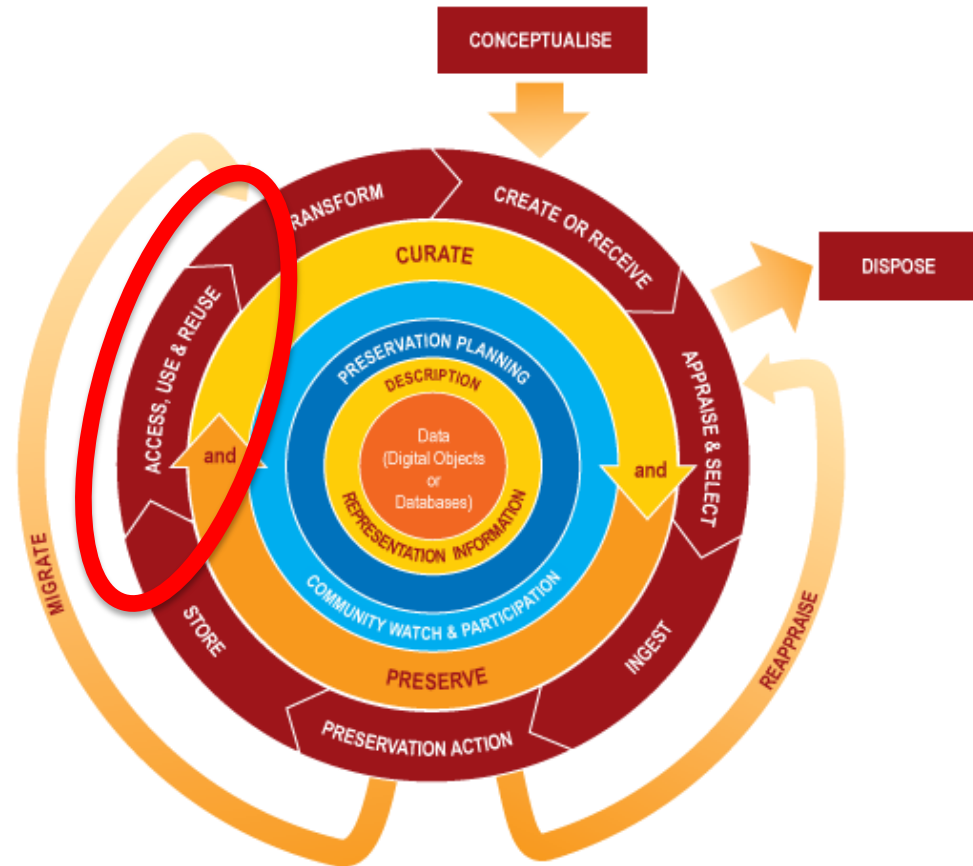
Transformation

... is an ETL repeatable process



Data Life Cycle

- Access & Reuse
 - Zugriff auf die Daten
 - Nutzung / Nachnutzung
- Lösung:
 - Daten bekommen einen PID
 - Datenrepositorium für Datenfiles
 - PID führt zu Datenfile



<http://www.dcc.ac.uk/resources/curation-lifecycle-model>

- Unterschiede zu Standard Repositorien?

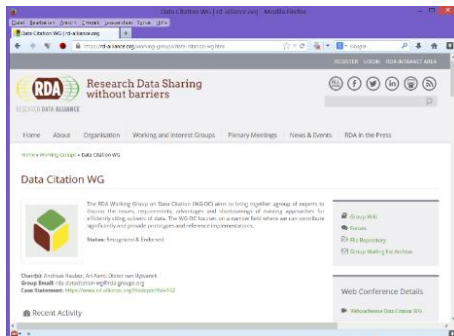
Services - Access für Daten

- Unterschiede zu Standard Repositorien?
 - Selten Abfrage einzelner Elemente, üblicherweise Subsets der gesamten Datenmenge
 - Datenbasis ändert sich kontinuierlich (Fehlerkorrekturen, neue Daten)
 - Forscher holen sich unterschiedliche Subsets zu unterschiedlichen Zeitpunkten
 - wollen diese später wieder abrufbar haben
- Lösung
 - Jedes Datentupel ein AIP, aggregierte AIPs?
 - Jedes Subset ein einzelnes File, re-ingest?



Services - Access für Daten

- Ziel: Service, der Zugriff auf Daten erlaubt
 - wie sie zum jeweiligen Zeitpunkt existiert haben
 - für jedes beliebige Subset
 - ohne redundante Kopien der Daten anzulegen
 - für diese Subsets jeweils eigene PIDs liefert (Identifizieren, Zitieren) Parameter in Analyseprozessen)
- Research Data Alliance (RDA) –
Working Group on Dynamic Data Citation (WGDC)



<https://rd-alliance.org/groups/data-citation-wg.html>

Services - Access für Daten

Datenidentifikation: Daten + Zugriffsmechanismus

- Daten → Zeitstempel & Versionierung (aka history)

Forscher erstellen Subsets über Interfaces:

- Abfrage → speichere **PID für die QUERY**, plus
 - **Timestamp** der Query zur Wiederausführung gegen versionierte DB
 - **Anpassen** der Query (Normalisierung, Versionierungsadaption)
 - **Prüfsumme** des Ergebnisses zur Verifikation
- PID führt zur Landing Page des Datensets
- Prinzip gültig für alle Arten von Daten:
SQL, XML, CSV, Datei-basierte Datensammlungen, ...

S. Pröll, A. Rauber. **Scalable Data Citation in Dynamic Large Databases: Model and Reference Implementation**. In IEEE Intl. Conf. on Big Data 2013 (IEEE BigData2013), 2013

http://www.ifs.tuwien.ac.at/~andi/publications/pdf/pro_ieeebigdata13.pdf

Prototype for CSV: <http://datacitation.eu/>

Services - Access für Daten

Nutzung des Interfaces

- Forscher wählt Subset der Daten aus (Web-UI, API), erhält
 - Daten (ZIP-Datei, CSV, Datenstrom über API, Input in Analyseprozess, ...)
 - PID (z.B. DOI) (Query wird mit Zeitstempel gespeichert)
 - Prüfsumme wird berechnet und gespeichert
 - Empfohlener Zitationstext wird angezeigt (z.B. BibTeX, Endnote, Text)
- PID führt zu Landing Page
 - Liefert detaillierte Metadaten, Link zur Datenbasis, ...
 - Erlaubt Zugriff auf Originaldaten ODER aktuelle Version ODER Differenz
- Bei Anforderung der Daten
 - Query wird gegen die Datenbank ausgeführt
 - Ergebnisse werden mit Prüfsumme verifiziert und geliefert
- Query Store erlaubt Nutzungsanalysen der Daten

Services - Access für Daten

- N Query liefert automatisch präzise Metainformationen (Provenance) für Subset!
- Query (z.B. über API), erhält
 - Daten (ZIP-Datei, CSV, Datenstrom über API, Input in Analyseprozess, ...)
 - PID (z.B. DOI) (Query wird mit Zeitstempel gespeichert)
 - Prüfsumme wird berechnet und gespeichert
 - Empfohlener Zitationstext wird angezeigt (z.B. BibTeX, Endnote, Text)
 - PID führt zu Landing Page
 - Liefert detaillierte Metadaten, Link zur Datenbasis, ...
 - Erlaubt Zugriff auf Originaldaten ODER aktuelle Version ODER Differenz
 - Bei Anforderung der Daten
 - Query wird gegen die Datenbank ausgeführt
 - Ergebnisse werden mit Prüfsumme verifiziert und geliefert
 - Query Store erlaubt Nutzungsanalysen der Daten

Services - Access für Daten

- Query liefert automatisch präzise Metainformationen (Provenance) für Subset! (API), erhält
 - Daten (ZIF)
 - PID (z.B. D)
 - Prüfsumme
 - Empfohlener Zitationstext wird angezeigt (z.B. BibTeX, Endnote, Text)
- PID führt zu Landing Page
 - Liefert detaillierte Metadaten, Link zur Datenbasis, ...
 - Erlaubt Zugriff auf Originaldaten ODER aktuelle Version ODER Differenz
- Bei Anforderung der Daten
 - Query wird gegen die Datenbank ausgeführt
 - Ergebnisse werden mit Prüfsumme verifiziert und geliefert
- Query Store erlaubt Nutzungsanalysen der Daten

Großer Vorteil gegenüber Standardansätzen wie Speicherung von redundanten Datendumps!

Services - Access für Daten

- Query liefert automatisch präzise Metainformationen (Provenance) für Subset!
 - Daten (ZIF, ...), API, erhält
 - PID (z.B. D) **Großer Vorteil gegenüber Standardansätzen wie Speicherung von redundanten Datendumps!**
 - Prüfsumme
 - Empfohlener Zitationstext wird angezeigt (z.B. BibTeX, Endnote, Text)
- PID führt zu Längstzeitarchivierung
 - Liefert detaillierte Informationen
 - Erlaubt Zugriffsstatistiken
- Bei Anforderung der Daten
 - Query wird gegen die Datenbank ausgeführt
 - Ergebnisse werden mit Prüfsumme verifiziert und geliefert
- Query Store erlaubt Nutzungsanalysen der Daten

Welche Daten werden wann von wem wie oft genutzt? Welche Studien sind von Änderungen in den Daten betroffen?

Services - Access für Daten

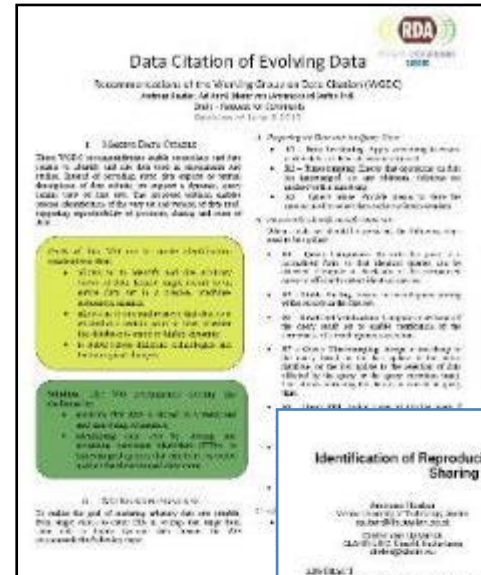
- 14 Punkte Empfehlung gruppiert in 4 Phasen:
 - Vorbereitung der Daten und des Query Store
 - Persistente Identifikation von Daten(sub)sets
 - Auflösen von PIDs
 - Änderung der Dateninfrastruktur

- 2-seitiger Flyer

- Detaillierter Technical Report:

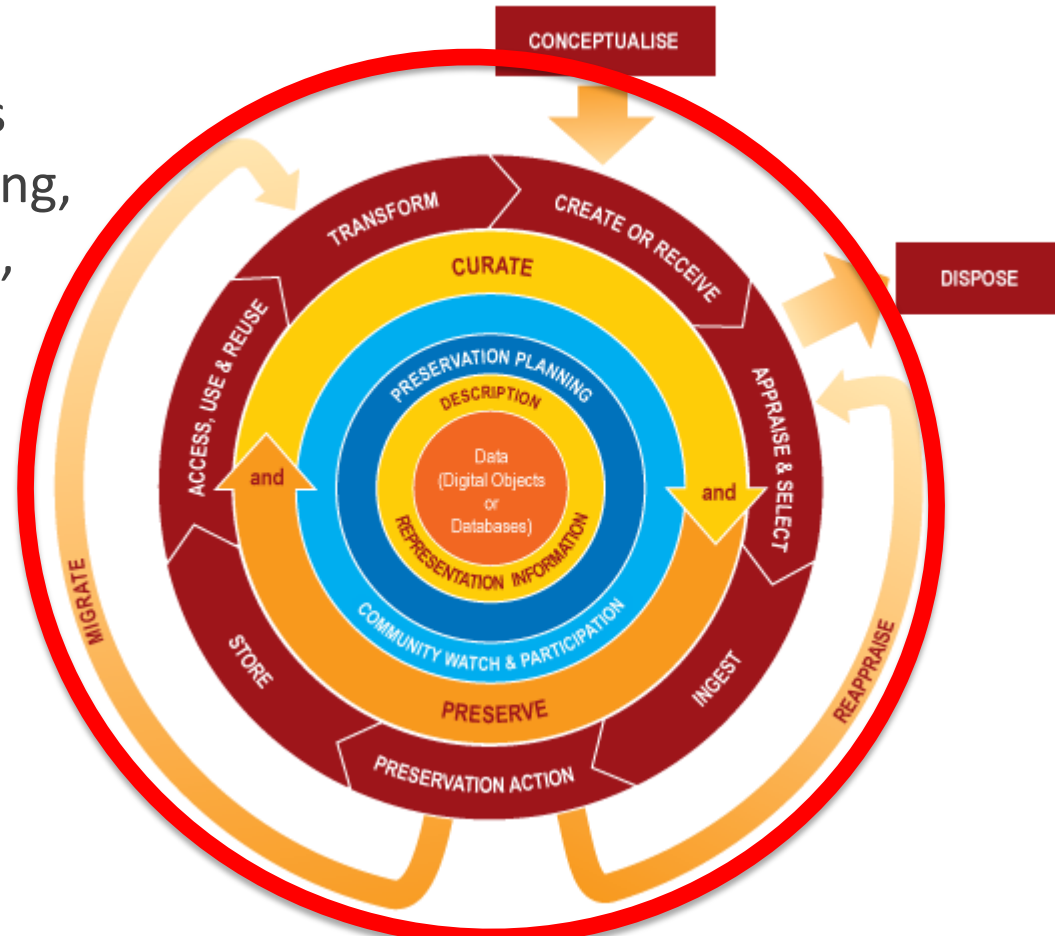
https://rd-alliance.org/system/files/documents/RDA-Guidelines_TCDL_draft.pdf

- Referenzimplementierungen und Pilotanwendungen für SQL, CSV, XML



Data Life Cycle

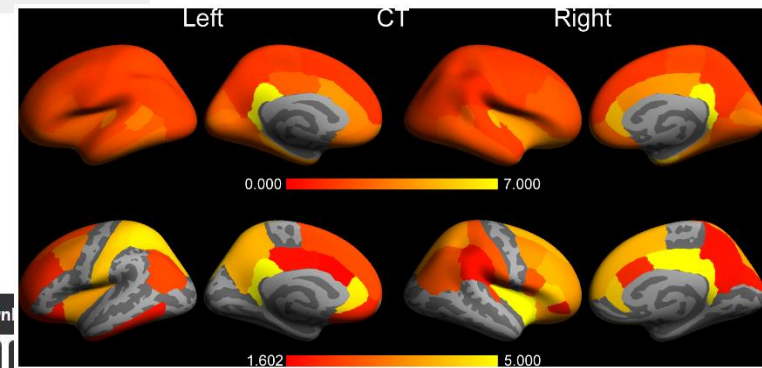
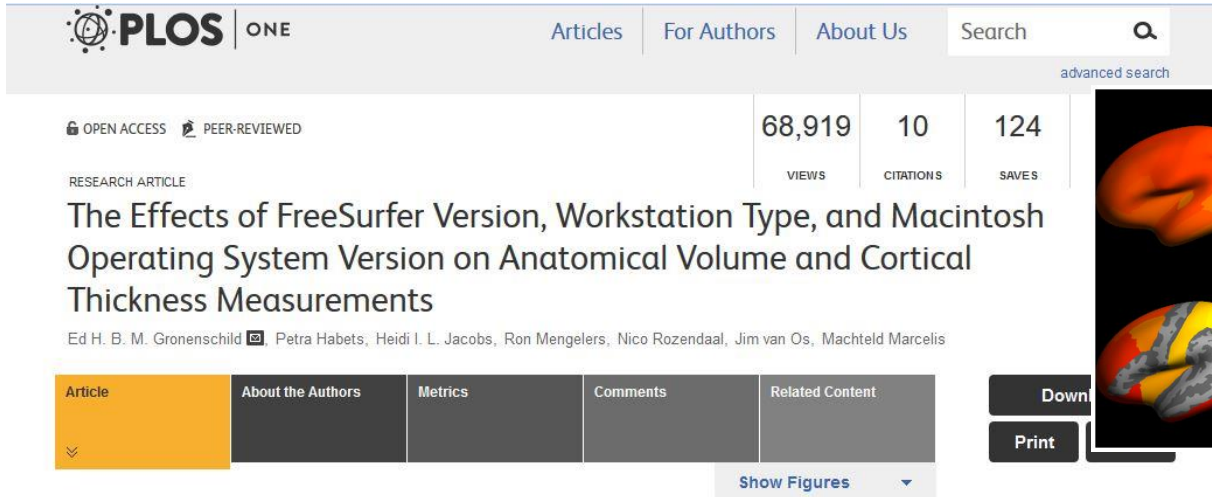
- Daten und Provenance
 - Daten durchlaufen Prozess (Erzeugung, Vorverarbeitung, Analyse, Nachbearbeitung, ...)
 - Verwendung zahlreicher SW Tools, diverse HW, ...
 - All das solle (muss?) dokumentiert werden.



<http://www.dcc.ac.uk/resources/curation-lifecycle-model>

Daten Provenance

<http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0038234>



► Abstract

Introduction

Materials and Methods

Results

Discussion

Supporting Information

Acknowledgments

Author Contributions

References

Reader Comments (5)

Figures

Abstract

FreeSurfer is a popular software package to measure cortical thickness and volume of neuroanatomical structures. However, little if any is known about measurement reliability across various data processing conditions. Using a set of 30 anatomical T1-weighted 3T MRI scans, we investigated the effects of data processing variables such as FreeSurfer version (v4.3.1, v4.5.0, and v5.0.0), workstation (Macintosh and Hewlett-Packard), and Macintosh operating system version (OSX 10.5 and OSX 10.6). Significant differences were revealed between FreeSurfer version v5.0.0 and the two earlier versions. These differences were on average 8.8±6.6% (range 1.3–64.0%) (volume) and 2.8±1.3% (1.1–7.7%) (cortical thickness). About a factor two smaller differences were detected between Macintosh and Hewlett-Packard workstations and between OSX 10.5 and OSX 10.6. The observed differences are similar in magnitude as effect sizes reported in accuracy evaluations and neurodegenerative studies.

The main conclusion is that in the context of an ongoing study, users are discouraged to update to a new major release of either FreeSurfer or operating system or to switch to a different type of workstation without repeating the analysis; results thus give a quantitative support to successive recommendations stated by FreeSurfer developers over the years. Moreover, in view of the large and significant cross-version differences, it is concluded that formal assessment of the accuracy of FreeSurfer is desirable.

Comments

In praise of prog

Posted by GedR

Media Coverage

Article

Posted by

PLoS ONE Grc

Comments made

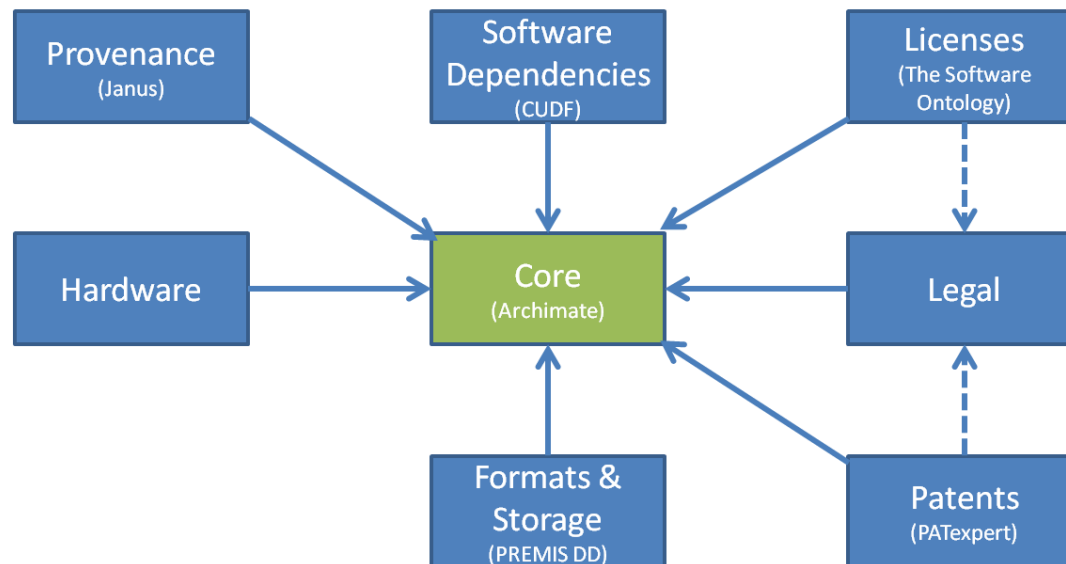
authors
Revised by EdCs

Posted by EdGro

[illegible]

Prozess Kontextmodell

- Definiert, welche information wie abgelegt wird
- Metamodell für Prozess & Kontext
 - Kernmodell mit Erweiterungsontologien
 - Basierend auf ArchiMate, umgesetzt in OWL
 - Nutzung von existierenden Modellen (PREMIS, CUDF, ...)



Kontextmodell – Statische Analyse

- Analysiert Prozessschritte, Plattform, Services, SW Tools
- Abhängigkeiten (SW Pakete, Bibliotheken, Treiber)
- HW, SW Lizenzen, ...

```
#!/bin/bash

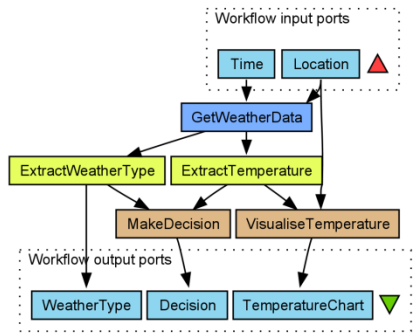
# fetch data
java -jar GestBarragensWSClientIQData.jar
unzip -o IQData.zip

# fix encoding
#iconv -f LATIN1 -t UTF-8 iq.r > iq_utf8.r

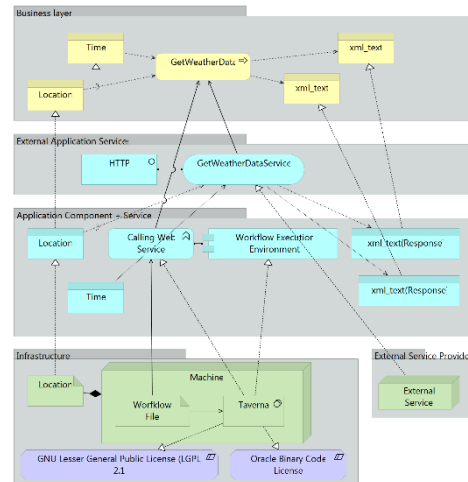
# generate references
R --vanilla < iq_utf8.r > IQout.txt

# create pdf
pdflatex iq.tex
pdflatex iq.tex
```

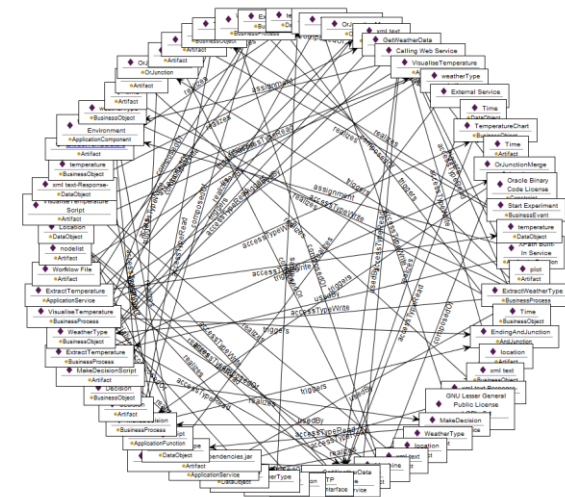
Skript



Taverna Workflow



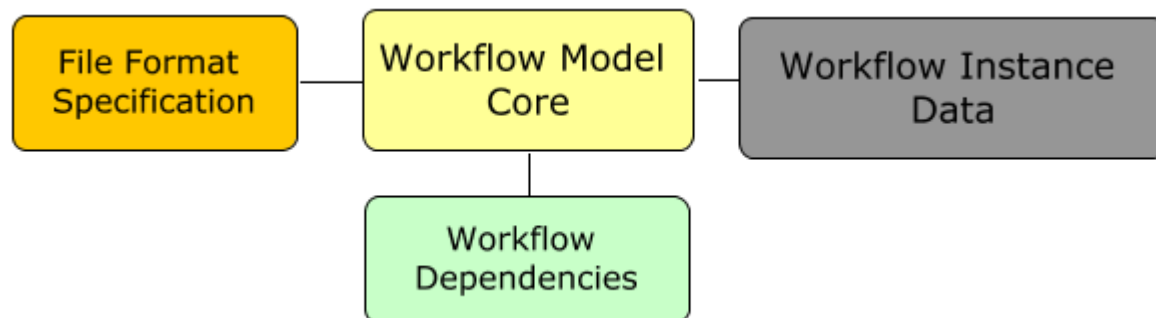
ArchiMate Modell



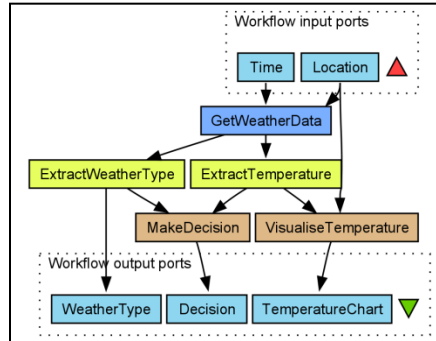
Kontextmodell
(OWL Ontology)

Kontextmodell – Dynamische Analyse

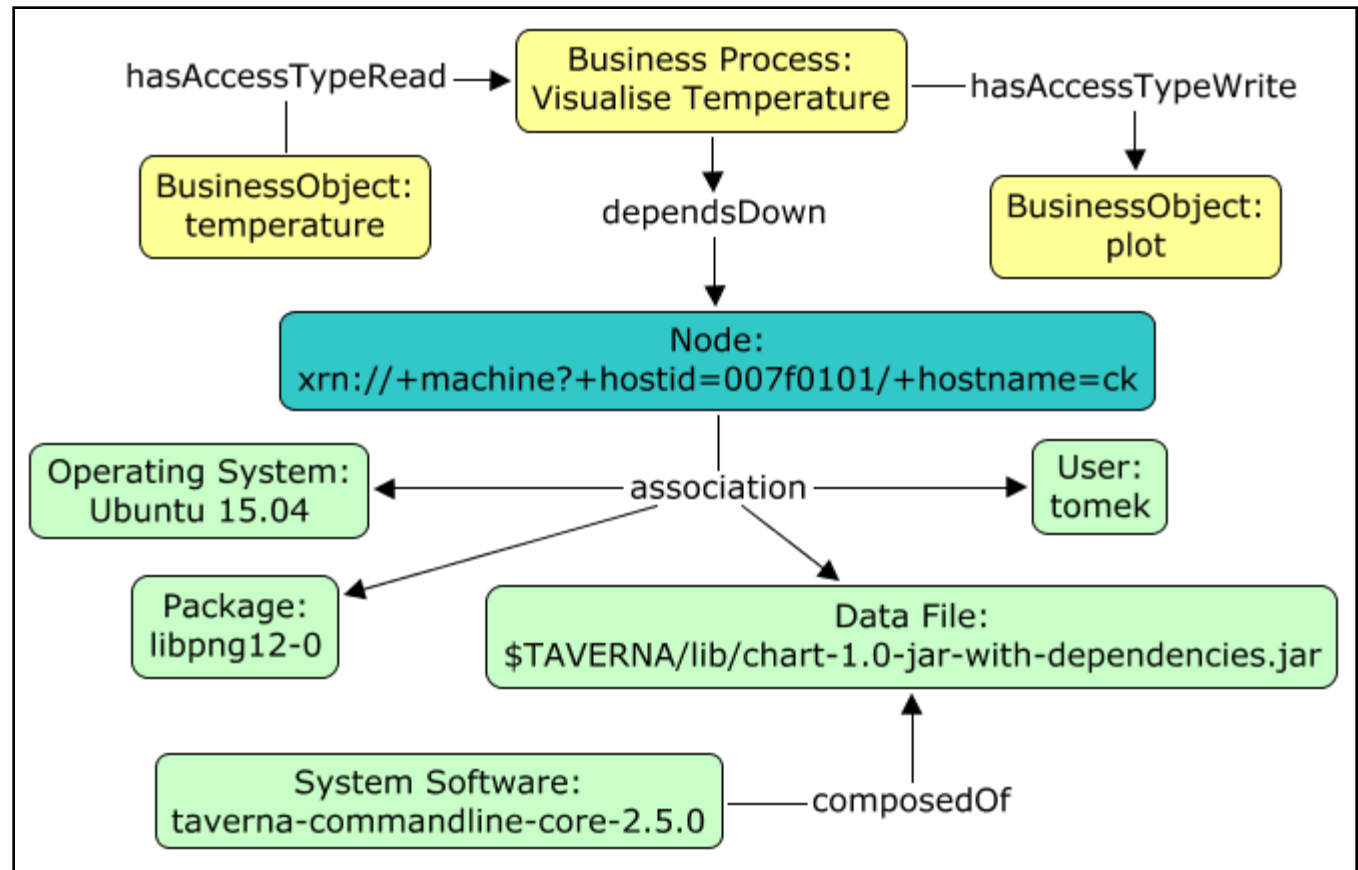
- Process Migration Framework (PMF)
 - Entwickelt um automatisch virtuelle Maschinen für Prozesse zu erstellen
 - Verwendet *strace* um system calls zu überwachen
 - komplettes log aller betroffenen Ressourcen (Dateien, Ports)
 - Speichert sämtliche Daten einer Prozessinstanz
 - Analysiert betroffene Dateien (Dateiformate via PRONOM, PREMIS)



Kontextmodell – Dynamische Analyse

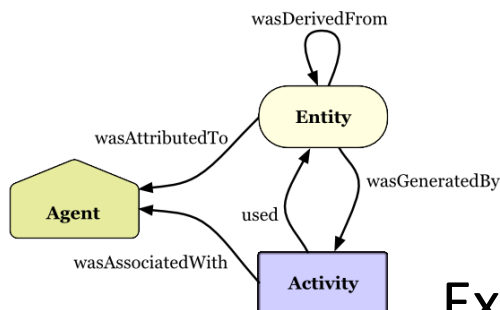
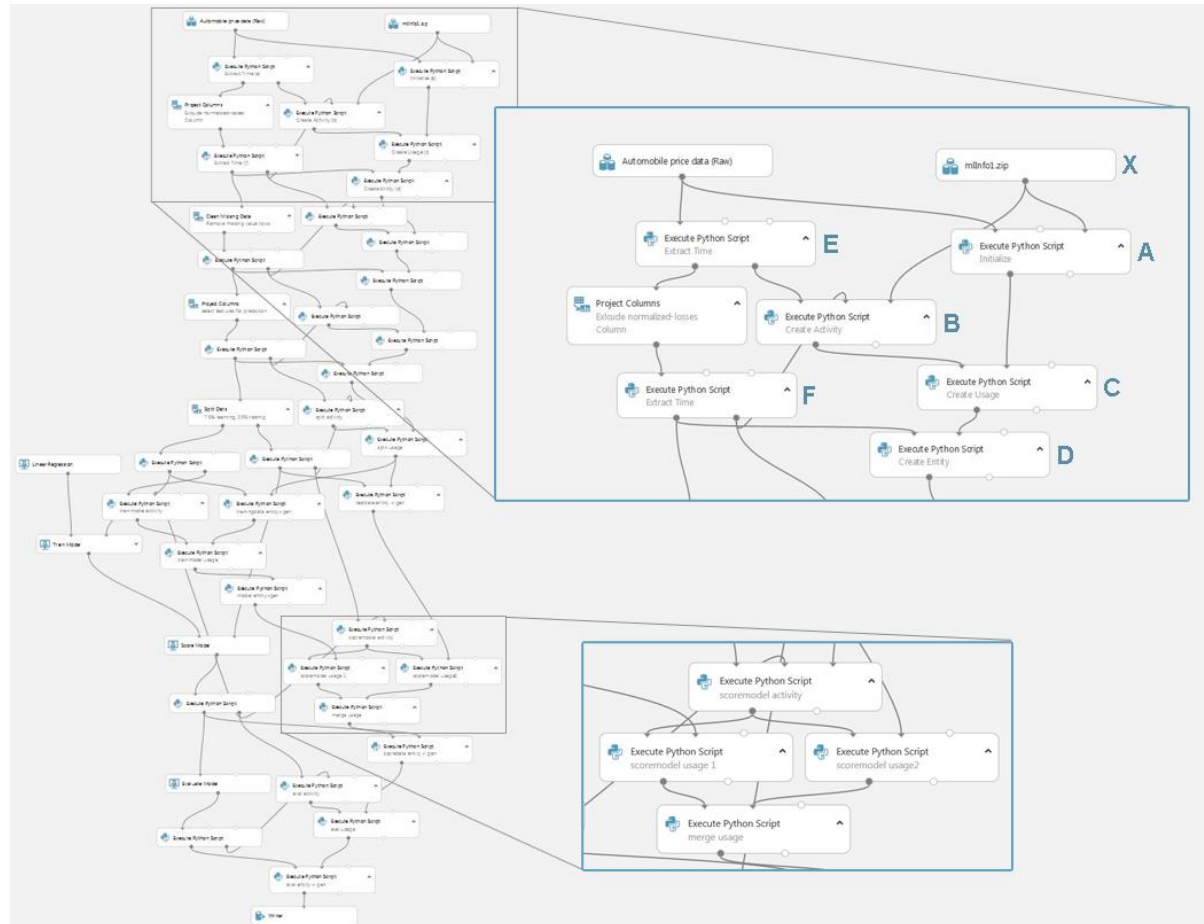


Taverna Workflow



Kontextmodell – Integriert in WF-Engine

MS Azure ML:



Export als W3C Prov-O

Source:
www.w3.org

Prozess Kontextmodell

Preservation und Re-deployment

- „Kapselung“ als Research Object (RO)
- DP: Re-Deployment in anderer Umgebung
 - Formatmigration der Elemente des ROs
 - Cross-compilation von Quellcode
 - Emulation-as-a-Service
- Verifikation nach Re-Deployment



Evaluation result: PASS
All Significant Properties are OK. All metrics were fulfilled.
Comparison performed using following workflow execution traces

Original Workflow
ID: 70264734-cdda-4630-8ecd-27ba30f11d8f
Timestamp: 2015-04-21 13:39:03.499

Compared Workflow
ID: 70264734-cdda-4630-8ecd-27ba30f11d8f
Timestamp: 2015-04-21 13:39:03.499

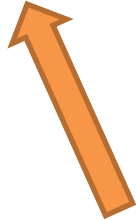
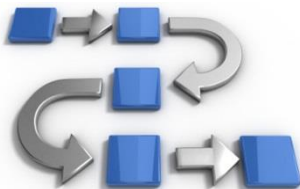
Table 1: Overview of significant properties

Significant Property	Description	Is Fulfilled
SP1_hjove2_input	The workflow step 'hjove2_input' has identical outputs.	True
SP2_WorkflowCorrectInputs	The inputs to the workflow are the same.	True
SP3_BeanShellCopy	The workflow step 'BeanShellCopy' has identical outputs.	True
SP4_WorkflowCorrectOutputs	The outputs of the workflow are the same.	True
SP5_hjove2_output	The workflow step 'hjove2_output' provides the	True



VFramework

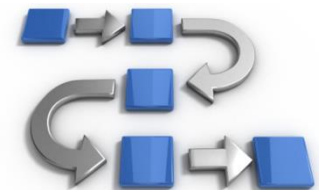
Originalumgebung



Repository

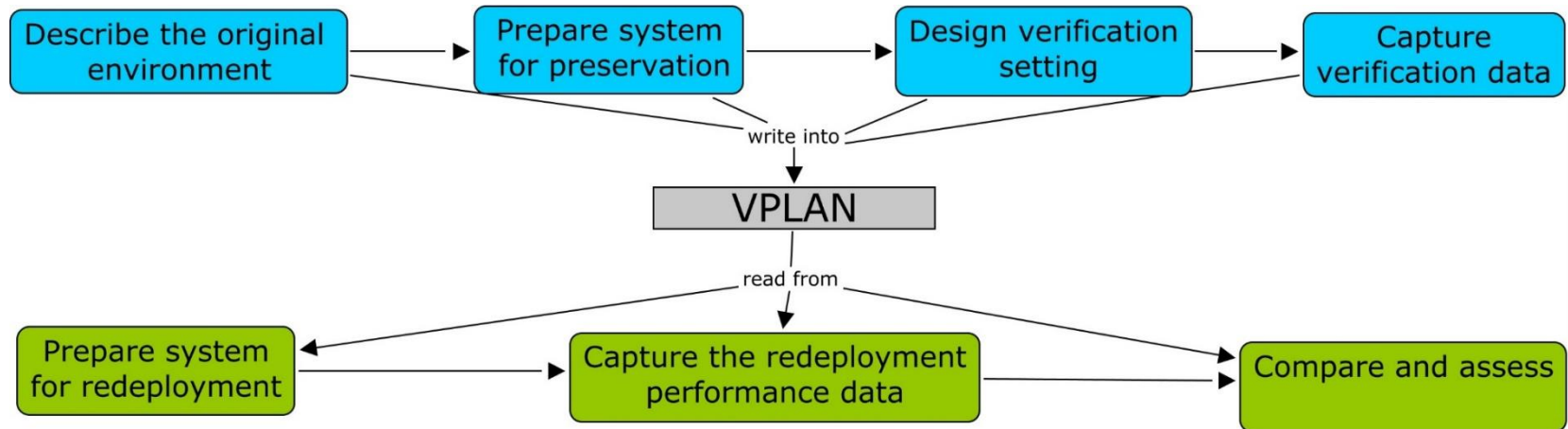


Redeployment Umgebung



Verhalten sich diese Prozesse gleich?

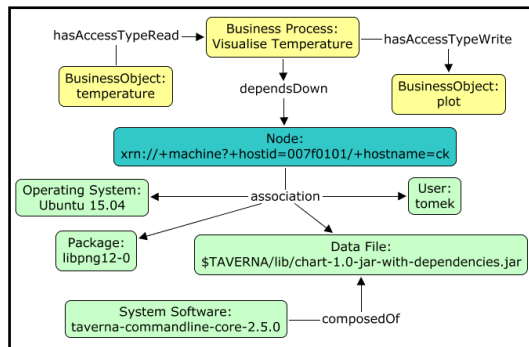
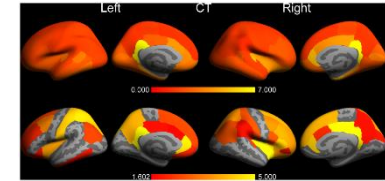
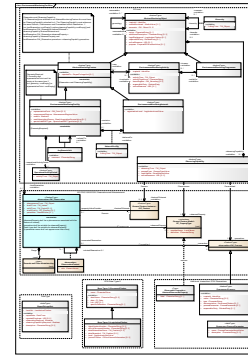
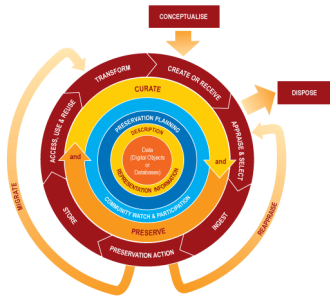
VFramework



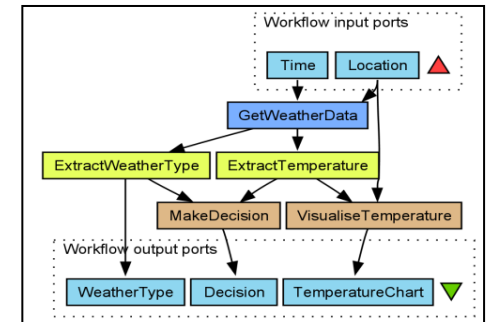
- Prozesskontextmodell plus Instanz-Daten erlauben
 - Verifikation der Umgebung sowie
 - Validierung der Prozessergebnisse
- “Ideallösung” – vollständige Integration
- Teillösungen/Teilautomatisierungen als erster Schritt

Zusammenfassung

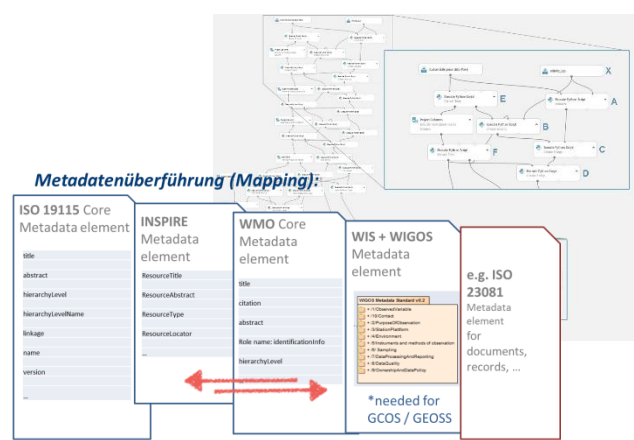
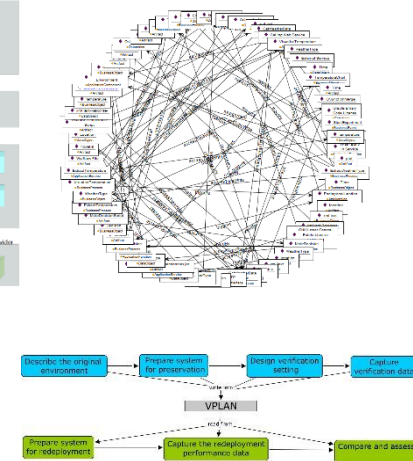
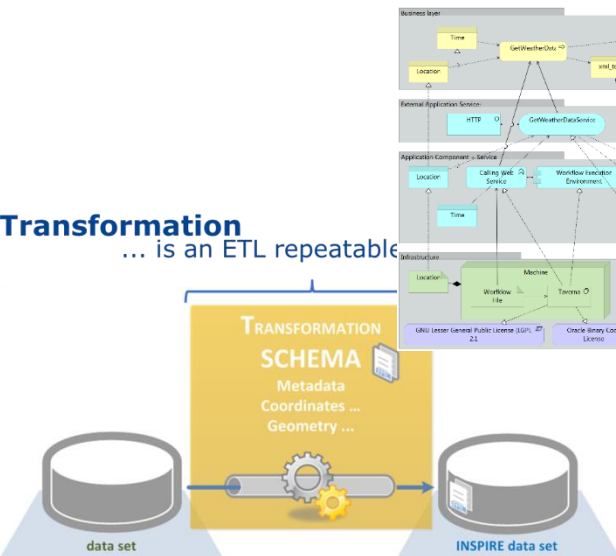
- Data Management Lifecycle essentiell um
 - Investitionen in Daten zu sichern
 - Re-Use, Verifikation, Metastudien etc. zu ermöglichen
- Ähnlich zu klassischen Aktivitäten im Bereich
 - Rechenzentren: Datenmanagement
 - Bibliotheken: Repositorien
- Neue Anforderungen durch eScience, “Big Data”
 - Geringere Trennung Live-Daten und Archivdaten
 - Integration in Forschungsumgebungen, SW & HW
 - Vielfältigere Nutzungsszenarien
 - Höhere Dynamik
- Erfordert zusätzliche Services und Automatisierung
 - Access, Provenance, ...
- Schrittweise Einführung von modularen Teillösungen



Vielen Dank



Transformation ... is an ETL repeatable



Kontakt

Chris Schubert

CCCA Datenzentrum

chris.schubert@ccca.ac.at

Andreas Rauber

Institute of Software Technology and Interactive Systems

Technische Universität Wien

rauber@ifs.tuwien.ac.at