

Schriftliche Zusammenfassung des Vortrages zum Thema

# **FORSCHUNGS- DATENMANAGEMENT**

Scientific Data Management

---

## **Oberseminar I „Datenbanksysteme – Aktuelle Trends“**

Master Medieninformatik (MIM 18)

Sommersemester 2019

**Prof. Dr. Thomas Kudraß**

### Verfasserin:

Elizabeth Orta

Matrikelnr.: 73343

# Inhaltsverzeichnis

|                                                          |          |
|----------------------------------------------------------|----------|
| <b>1. Einführung.....</b>                                | <b>1</b> |
| 1.1 Problemstellung.....                                 | 1        |
| 1.2 Beispiel.....                                        | 1        |
| <b>2. Forschungsdaten .....</b>                          | <b>1</b> |
| 2.1 Definition.....                                      | 1        |
| 2.2 Forschungsdaten vs. geschäftsorientierte Daten ..... | 1        |
| 2.3 Scientific Data Lifecycle .....                      | 2        |
| <b>3. Forschungsdatenmanagement.....</b>                 | <b>2</b> |
| 3.1 Definition.....                                      | 2        |
| 3.2 Ziele.....                                           | 2        |
| 3.3 Status Quo.....                                      | 2        |
| 3.4 Relationale Datenbanksysteme.....                    | 3        |
| 3.4.1 Vorteile .....                                     | 3        |
| 3.4.2 Nachteile .....                                    | 3        |
| 3.5 Anforderung an das Forschungsdatenmanagement .....   | 4        |
| <b>4. Metadaten.....</b>                                 | <b>4</b> |
| 4.1 Definition.....                                      | 4        |
| 4.2 Relevanz .....                                       | 4        |
| 4.3 Metadatenstandards.....                              | 4        |
| 4.4 Wissenschaftliche Datenformate.....                  | 4        |
| <b>5. Ontologien .....</b>                               | <b>5</b> |
| 5.1 Definition.....                                      | 5        |
| 5.2 Relevanz .....                                       | 5        |
| <b>6. Data Provenance .....</b>                          | <b>5</b> |
| 6.1 Definition.....                                      | 5        |
| 6.2 Relevanz .....                                       | 5        |
| 6.3 Datenverarbeitungsarchitekturen.....                 | 6        |
| 6.4 Taxonomie .....                                      | 6        |
| 6.5 Techniken.....                                       | 6        |
| 6.6 Problematik.....                                     | 6        |
| <b>7. Scientific Workflow .....</b>                      | <b>7</b> |
| 7.1 Definition.....                                      | 7        |
| 7.2 Ziele.....                                           | 7        |
| 7.3 Phasen .....                                         | 8        |
| <b>8. Anwendungsbeispiele.....</b>                       | <b>8</b> |
| 8.1 Problematik.....                                     | 8        |
| 8.2 Auswahl .....                                        | 8        |

|     |                           |    |
|-----|---------------------------|----|
| 8.3 | Vorführung.....           | 8  |
| 9.  | Fazit .....               | 8  |
| 10. | Diskussion .....          | 9  |
| 11. | Literaturverzeichnis..... | 10 |
| 12. | Anhang.....               | 14 |

# 1. Einführung

## 1.1 Problemstellung

Die Verfügbarkeit von Daten ist für Forschende in den Wissenschaften eine notwendige Voraussetzung ihres wissenschaftlichen Arbeitens. In den letzten Jahren wuchsen die Verfügbarkeit und Komplexität von Daten kontinuierlich. Dieser Zuwachs beruht u. a. auf methodische und technische Neuerungen (z. B. modernere wissenschaftliche Instrumente, Analysemethoden oder Computersimulationen mit hoher Präzision und Datenqualität), wachsende verfügbare Speicherkapazitäten, das Internet und auf die interdisziplinäre Verknüpfung von Datentypen (z. B. Verbindung von Geo- oder Gesundheitsdaten mit traditionellen Umfragedaten der empirischen Sozialforschung).<sup>1</sup>

Es wird erwartet, dass die Verfügbarkeit und Komplexität von Daten in Zukunft weiterhin schnell wächst. Dies bietet einzigartige Möglichkeiten für neue wissenschaftliche Entdeckungen sowie erhebliche Herausforderungen für das Management, das Kuratieren, die Analyse, die Pflege, die Provenienz, der Zugriff und das Computing von Daten. Es werden daher vor allem Maßnahmen zum Erhalt der Daten, der Qualität, der Verständlichkeit und der Bereitstellung der Daten (auch über die Forschung hinaus) benötigt.<sup>2</sup>

## 1.2 Beispiel

Die Nasa speichert derzeit über 100 Petabyte (PB) Beobachtungs- und Modelldaten. Innerhalb von fünf Jahren werden alle vier NASA-Wissenschaftsbereiche voraussichtlich über 100 PB Daten pro Jahr generieren.<sup>3</sup>

# 2. Forschungsdaten

## 2.1 Definition

„Als Forschungsdaten gelten (unabhängig von ihrer Form oder Darstellung) alle Informationen, einschließlich Kontextinformationen, die benötigt werden, um die Entwicklung, Ergebnisse, Beobachtungen oder Erkenntnisse einer Forschungstätigkeit zu stützen oder zu validieren.“<sup>4</sup>

Beispiele dafür sind Laborwerte, Surveydaten, Statistiken, Interviews, Simulationen, experimentelle Messdaten, methodische Testverfahren, 3D-Scans, audiovisuelle Informationen, Bilder, Code etc.<sup>5</sup>

## 2.2 Forschungsdaten vs. geschäftsorientierte Daten

Forschungsdaten sind im Vergleich zu geschäftsorientierten Daten hauptsächlich:<sup>6</sup>

### heterogen:

Forschungsdaten bestehen nicht nur aus Bytes, sondern auch aus Workflow-Definitionen, Berechnungsparametern, Umgebungseinstellungen usw.

### verteilt:

Aufgrund von domänenübergreifende wissenschaftliche Kollaborationen.

---

<sup>1</sup> Vgl. Gray et al. 2005, S. 1; Nasa 2019a

<sup>2</sup> Vgl. Nasa 2019b

<sup>3</sup> Vgl. Nasa 2019a

<sup>4</sup> TU Wien 2018

<sup>5</sup> Vgl. TU Wien 2018; HU Berlin 2015

<sup>6</sup> Vgl. Kindling et al. 2019; Valle 2019; Roy 1998; Gray et al. 2005, S. 1; Van den Eynden, Veerle 2013, S. 1 ff.

### komplex:

Die Bedeutung der Forschungsdaten ist nur durch Datenanalysewerkzeuge möglich.

### entdeckungsorientiert:

Durch das Analysieren von Forschungsdaten, können weitere subtile Informationen/Effekte gefunden werden, die von früheren Studien übersehen wurden.

### validierbar:

Bei Forschungsdaten geht es meistens um Sinn und Inhalt.

### wiederverwendbar:

Daten können für zukünftige Forschungen wiederverwendet werden, wenn diese geteilt, gut verwaltet, ordnungsgemäß aufbewahrt und verfügbar gemacht werden.

### langlebig:

Ein Forschungsdatensatz hat normalerweise eine längere Lebensdauer als das eigentliche Forschungsprojekt, welches es produziert.

## 2.3 Scientific Data Lifecycle

Der Forschungsdatenlebenszyklus beschreibt die verschiedenen Phasen die Daten durchlaufen, um von der Datengenerierung bis zur Wissensgenerierung zu gelangen. (vgl. Abb. 1 im Anhang).<sup>7</sup>

Die Universität von Ottawa unterscheidet zwischen sieben Phasen: *Plan, Create, Process, Analyze, Preserve, Share, Reuse* (vgl. für detailliertere Erläuterungen Abb. 2 im Anhang).<sup>8</sup>

## 3. Forschungsdatenmanagement

### 3.1 Definition

Das Forschungsdatenmanagement (FDM) hat die Aufgabe Strategien, Verfahren, Konventionen und Arbeitsabläufe zu entwickeln, um die Qualität, die Validierung, die Dokumentation, die dauerhafte Auffindbarkeit und Nutzbarkeit von Forschungsdaten (inkl. Schutz vor Verlust oder Manipulation) zu gewährleisten.<sup>9</sup>

### 3.2 Ziele

Die Ziele des FDM werden wie folgt zusammengefasst:<sup>10</sup>



### 3.3 Status Quo

In der Aktualität ist eine zuverlässige Datenqualitätsüberprüfung oft nicht möglich, da Wissenschaftler/innen nur über „konventionellen“ Verfahren/Tools und beschränkte Mitteln

<sup>7</sup> Vgl. University of Ottawa o. J.; Strasser et al (o. J.), S. 4 ff.

<sup>8</sup> ebd.

<sup>9</sup> Vgl. Valle 2019; GESIS o. J.

<sup>10</sup> Vgl. Valle 2019

verfügen. Zudem ist ein dateiorientierter Ansatz sehr verbreitet. Dateien werden in Verzeichnissen organisiert. Je weiter die Forschung voranschreitet, desto mehr Verzeichnisse entstehen und die Verzeichnishierarchie wird größer. Die Datenanalyse bedeutet hier das Durchsuchen aller relevanten Dateien, das Öffnen jeder einzelnen Datei, das Extrahieren der relevanten Daten und das Bewegen auf die nächste Datei. Wenn alle relevanten Daten gesammelt und im Speicher hinterlegt wurden (oder in Zwischendateien) kann das Programm seine Analyse beginnen.<sup>11</sup> Einige Projektdatenrepositorys sind einfach große flache Verzeichnisse die über begrenzte Benennungsschemata verfügen. Eine strikte Metadatenkultur ist nicht vorhanden. Viele Wissenschaftler/innen beschränken die Bedeutung des Datenmanagements normalerweise auf die reine physische Datenspeicherungs- und Zugriffsebene und arbeiten mit nur einer CPU und einer Festplatte.<sup>12</sup> Das Verarbeiten und das Analysieren großer Datensätze mittels konventionellen Verfahren/Tools, wird mit zunehmendem Datenvolumen immer langsamer ausgeführt. Aus diesen Grund ziehen viele Wissenschaftler/innen es vor, einen Großteil ihrer Analysen bei Rechenzentren durchzuführen, weil sie bspw. keine lokale Daten- und Computerfarmen verwalten müssen und allgemein möglicherweise keine passende Analysetools besitzen oder keine Zeit darin investieren möchten neue Verfahrensmethoden zu erlernen.<sup>13</sup>

Nicht zuletzt ist oft kein zuverlässiger Sicherheitsmechanismus vorhanden welches die Probleme (Löschen, Viren etc.) des Zugriffs auf wissenschaftliche Daten verhindern bzw. beseitigen könnte.<sup>14</sup>

## 3.4 Relationale Datenbanksysteme

### 3.4.1 Vorteile

Relationale Datenbanksysteme verfügen über eine leistungsstarke assoziative Suche (bspw. Suche nach Wert und nicht nach Ort), eine nicht-prozedurale Abfrageanalyse und automatische parallele Zugriffe, die die Datensuche, die Aufführung, das schnelle Auffinden von Datenteilmengen und die Petascale-Datenanalyse erleichtern. Sie besitzen anspruchsvolle Werkzeuge für die Automatisierung von Datenentwürfe und -verwaltung, für assoziative, zeitliche und räumliche Suche und für die Definition von Daten(formate). Sie ermöglichen das routinemäßiges Replizieren und Kombinieren von Daten, sodass diese auf verschiedene Arten (nach Zeit, nach Raum, nach anderen Attributen) organisiert werden können (Ansichten). Ebenfalls läuft der OLAP-Würfel in Richtung maschinelles Lernen, durch die Integration von Algorithmen die Trends ableiten, Daten clustern und Anomalien erkennen.<sup>15</sup>

### 3.4.2 Nachteile

Herkömmliche tabellarische Datenbanksysteme sind ausreichend für die Analyse von Objekten (bspw. Galaxien, Spektren, Proteinen, Ereignissen usw.). Diese unterstützen aber traditionell nicht das Kerndatentyp der Wissenschaft: das N-dimensionale Array. Arrays mussten seit jeher in den meisten Systemen als *Blobs* (binäre große Objekte) ausgegeben werden. Auch die Unterstützung für die Zeitreihenfolge, räumliche, Text- und andere Datentypen ist oft umständlich.<sup>16</sup>

Neben dies, macht die Impedanzfehlانpassung (Impedance-Mismatch) es schwer, wissenschaftliche Anwendungen auf herkömmliche tabellarische Datenbanksysteme abzubilden. Datenbanktabellen werden zum Teil meistens als Dateien gespeichert. Dateisysteme erlauben keine parallele,

---

<sup>11</sup> Vgl. Gray et al. 2005, S. 1 ff.

<sup>12</sup> Vgl. Valle 2019; Gray et al. 2005, S. 1 ff.

<sup>13</sup> Vgl. Gray et al. 2005, S. 1 ff.

<sup>14</sup> Vgl. Valle 2019

<sup>15</sup> Vgl. Gray et al. 2005, S. 1 ff.

<sup>16</sup> Vgl. Gray et al. 2005, S. 1 ff.

assoziative, zeitliche oder räumliche Suche. Ihnen fehlt auch eine allgemeine Abfragesprache. Ebenfalls haben Dateisysteme keine Metadaten jenseits einer hierarchischen Verzeichnisstruktur und Dateinamen.<sup>17</sup>

Wissenschaftler/innen meinen u. a. auch, dass die Kosten zum Erlernen neuer Werkzeuge (Datendefinition, Laden und Abfrage von Daten etc.) sich nicht lohnen würden. Sie meinen vorhandene Systeme hätten keine gute Visualisierungs-/Plotterwerkzeuge und Manipulationsmöglichkeiten und seien zudem oft zu langsam.<sup>18</sup>

### 3.5 Anforderung an das Forschungsdatenmanagement

Um die Ziele des FDM (vgl. Punkt 3.2) erreichen zu können müssen Daten selbstbeschreibend sein. Eine Selbstbeschreibung der Daten kann durch die Definition von Metadaten erfolgen.<sup>19</sup>

## 4. Metadaten

### 4.1 Definition

Metadaten sind Daten über Daten. Es sind zusätzliche Informationen, die erforderlich sind, damit Daten nützlich sind.<sup>20</sup>

### 4.2 Relevanz

Metadaten erlauben es den Anwendungen Daten zu verstehen, weil sie beschreibende Information von Daten enthalten. Diese Information detailliert, wie die Daten gemessen, erworben oder berechnet wurden. Bspw. werden damit die gemessenen Attribute, ihre Namen, Einheiten, Präzision, Genauigkeit, Datenformat erklärt. Somit ermöglichen Metadaten *Data Access* (Datenzugriff), *Data Independence* (physische und logische Datenunabhängigkeit) und *Data Provenance* (Datenherkunft).<sup>21</sup>

### 4.3 Metadatenstandards

Metadatenstandards sind für die Wiederverwendung wissenschaftlicher Daten essentiell. Jeder Standard ermöglicht einen einfachen Datenaustausch unter Wissenschaftler/innen.<sup>22</sup> Ein Metadatenstandard muss implementierungsneutral und mit explizit eingeschränkter Erweiterbarkeit sein. Je nach Disziplin gibt es fachspezifische Metadaten (vgl. Abb. 3 und 4 im Anhang).<sup>23</sup>

### 4.4 Wissenschaftliche Datenformate

Datenformate werden für die Beschreibung von Metadaten verwendet. Einige Datenformate eignen sich jedoch besser zur Unterstützung von Metadaten und sind beim Erreichen der verschiedenen Datenverwaltungsziele hilfreicher. Zu den bekanntesten Forschungsdatenformate zählen XML, XSIL (Extensible Scientific Interchange Language), BFD (Binary Format Description Language), XDF (Extensible Data Format), FITS (Flexible Image Transport System), XDMF (eXtensible

---

<sup>17</sup> Vgl. Gray et al. 2005, S. 1 ff.

<sup>18</sup> Vgl. Gray et al. 2005, S. 1 ff.

<sup>19</sup> Vgl. Gray et al. 2005, S. 1

<sup>20</sup> Vgl. Bretherton 1994; Vgl. Gray et al. 2005, S. 1

<sup>21</sup> Vgl. Gray et al. 2005, S. 1

<sup>22</sup> Vgl. Gray et al. 2005, S. 1

<sup>23</sup> Vgl. Hoffmann 2018; Wandt o. J.; RADAR 2016

Data Model and Format), ICE (Interdisciplinary Computing Environment), DFDL (Data Format Description Language), HDF5 (Hierarchical Data Format), HDF-EOS (Earth Observing System), CGNS (General Notation System).<sup>24</sup>

## 5. Ontologien

### 5.1 Definition

Die Struktur der Metadaten stellt Regeln dar, durch die die Information organisiert wird (z. B. alphabetische, zeitliche, hierarchische Ordnungstypen). Aus diesen Regeln können wiederum Beziehungen (Relationstypen) zwischen Metadaten und Information (Daten) hergestellt werden.<sup>25</sup>

Ontologien sind also gemeinsame Vokabeln und logische Aussagen zu der jeweils interessierenden Anwendungsdomäne, die Komponenten einer bestimmten Disziplin und die Beziehungen zwischen diesen Komponenten beschreiben (vgl. Abb. 5 im Anhang).<sup>26</sup> Sie stellen Modelle dar, die eine Kommunikation von Mensch zu Maschine oder von Maschine zu Maschine erlauben. Voraussetzungen für die Kommunikation i. w. S. mit der Maschine ist allerdings, dass das explizite Wissen gleichzeitig für lesende Menschen und Maschinen verständlich ist.<sup>27</sup>

Im Wissenschaftssektor verfügt fast jede Disziplin über eine gemeinsame Terminologie (kontrollierter Wortschatz wie Einheiten, Messungen etc.).<sup>28</sup>

### 5.2 Relevanz

Ontologien schränken die Menge der möglichen Interpretationen für Worte und ihren zugehörigen Beziehungen ein.<sup>29</sup> Sie erhöhen die Datenqualität, die Auffindbarkeit und die Provenienz indem sie die Daten strukturieren und semantisch in Verbindung setzen (Semantic Web) und ermöglichen es Daten „richtig“ einzuordnen und besser zu verstehen.<sup>30</sup>

## 6. Data Provenance

### 6.1 Definition

*Data Provenance* (auch: *Data-Lineage*, *Datenherkunft*, *Data Pedigree*, *Datenabstammung*) bezieht sich auf Informationen über die Herkunft, den Kontext, die Abstammung, das Eigentum oder die Ableitungshistorie von Daten.<sup>31</sup> Provenienzinformationen von Daten basieren grundsätzlich auf den (Stamm-)Daten und die Transformationen, die sie durchgemacht haben, um neue Daten zu produzieren.

### 6.2 Relevanz

Herkunftsangaben sind von entscheidender Bedeutung, um beispielsweise zu gewährleisten, dass eine festgelegte wissenschaftliche Schlussfolgerung reproduzierbar ist (vgl. Beispiel in der PowerPoint-Präsentation, Seiten 53-63).<sup>32</sup>

---

<sup>24</sup> Vgl. Valle 2019; HDF Group 2019; HEASARC 2017; Barnes o. J.

<sup>25</sup> Vgl. Payrhuber 2009

<sup>26</sup> Vgl. Staab 2002, S. 6; Yale University 2019; Coustaty 2011

<sup>27</sup> Vgl. Staab 2002, S. 6

<sup>28</sup> Vgl. Gray et al. 2005, S. 1

<sup>29</sup> Vgl. Yale University 2019

<sup>30</sup> Vgl. Winkelmann et al. 2008; Ontotext 2019; Gray et al. 2005, S. 1

<sup>31</sup> Vgl. Simmhan et al. 2005; Cheney et al. 2012, S. 1 ff.

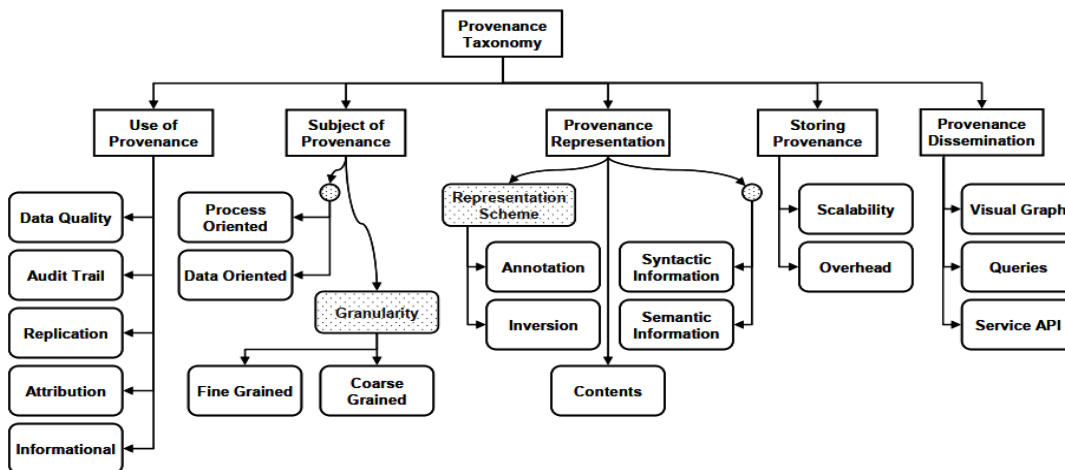
<sup>32</sup> Vgl. Cheney et al. 2012, S. 1 ff.; Cui/Widom 2001, S. 2 ff.

### 6.3 Datenverarbeitungsarchitekturen

Datenbankarchitekturen beschreiben die Art und Weise indem Provenienzdaten gesammelt werden. Es werden grundsätzlich zwischen vier Arten von Datenverarbeitungsarchitekturen unterschieden: serviceorientierte Architektur, Datenbankarchitektur, Befehlsverarbeitungsarchitekturen und skriptbasierte Architekturen (vgl. Abb. 6-9 im Anhang).<sup>33</sup>

### 6.4 Taxonomie

Simmhan et al. 2005 präsentieren folgende Taxonomie der Provenienz:<sup>34</sup>



Bildquelle: Simmhan et al. 2005

Für weiterführende Erläuterungen vgl. PowerPoint-Präsentation, Seiten 75-87.

### 6.5 Techniken

Bereits 2005 untersuchten Simmhan et al. acht unterschiedliche Datenprovenienztechniken nach dem im Punkt 6.4. genannten Kriterien. Eine Übersicht ist in Abb. 10 im Anhang zu finden.<sup>35</sup>

### 6.6 Problematik

Viele Projekte verfügen über eigene Protokolle für das Provenienzmanagement. Es fehlen offene Standards für das Sammeln, Darstellen, Speichern und Abfragen von Provenienz. Dies ist ein offensichtliches Hindernis für Förderung der Interoperabilität. Durch Standards und Middleware könnte die automatische Provenienz-Sammlung vorangetrieben werden. Die Vereinheitlichung semantischer Begriffe zur Beschreibung der Herkunft in Verbindung mit domänenspezifischen Ontologien würde reichere Abfragen, eine eindeutigere Provenienzinterpretation sowie eine automatisierte Überprüfung der Herkunft ermöglichen.<sup>36</sup>

Ebenfalls fallen oft die Kosten für das Sammeln und Darstellen von Provenienz umgekehrt zu ihrer Granularität an. In dieser Hinsicht müssten die Vorteile und Nachteile zwischen den

<sup>33</sup> Vgl. Simmhan et al. 2005, S. 1 ff.

<sup>34</sup> Vgl. Simmhan et al. 2005, S. 1 ff.

<sup>35</sup> Vgl. Simmhan et al. 2005, S. 1 ff.

<sup>36</sup> Vgl. Simmhan et al. 2005, S. 1 ff.

unterschiedlichen Provenienztechniken abgewogen werden, um bestmögliche Entscheidungen treffen zu können.<sup>37</sup>

Hinsichtlich des Datenschutzes, sind Herkunftsangaben zwar sehr nützlich, enthalten jedoch häufig vertrauliche Informationen, die mit Vorsicht verwendet werden sollten.<sup>38</sup>

Neben dies, erschwert die Heterogenität von Daten oft ihre Verteilung und ihre gemeinsame bzw. branchenübergreifende Nutzung.<sup>39</sup>

## 7. Scientific Workflow

### 7.1 Definition

Ein Scientific Workflow besteht aus einer Reihe von Aktivitäten bzw. eine Abfolge von Vorgängen, die durch die systematische Organisation von Ressourcen ermöglicht werden. Diese Aktivitäten transformieren Materialien, stellen Dienstleistungen bereit oder verarbeiten Informationen. Wissenschaftliche Arbeitsabläufe können einfache und lineare Aufgabenketten sein aber auch komplexere, grafisch strukturierte Abhängigkeiten repräsentieren.<sup>40</sup>

Übliche Verarbeitungsbeispiele sind z. B. Daten (neu) formatieren, Teilmengen bereinigen oder analysieren, Berechnungen oder computerwissenschaftliche Simulationen durchführen.<sup>41</sup>

### 7.2 Ziele

Die Ziele eines Scientific Workflows sind u. a. (1) den Wissenschaftler/innen bei sich wiederholende und fehleranfällige Aufgaben zu unterstützen bzw. Aufgaben zu automatisieren, (2) wissenschaftliche Entdeckungen zu beschleunigen, (3) spezielle Anwendungen und neue Datenanalysemethoden miteinander zu verknüpfen, (4) Maschinenzyklen einzusparen durch Optimierung von Workflow-Ausführungen auf verfügbaren Ressourcen, (5) kostspielige Ressourcen („menschliche Zyklen“) einzusparen, (6) die Dokumentation und Reproduzierbarkeit der zugrunde liegende experimenteller Prozesse effektiv zu unterstützen, (7) den Wissenschaftler/innen es ermöglichen, sich auf die domänenspezifische (naturwissenschaftliche) Aspekte ihrer Arbeit zu konzentrieren, anstatt sich mit komplexen Datenmanagement- und Softwareprobleme zu befassen.<sup>42</sup>

Darüber hinaus sind wissenschaftliche Workflows auch für die Modellierung, Design, Analyse und Wiederverwendung weiterer Workflows da.<sup>43</sup>

---

<sup>37</sup> Vgl. Simmhan et al. 2005, S. 1 ff.

<sup>38</sup> Vgl. Cuevas-Vicentín et al. 2012, S. 198 ff.

<sup>39</sup> Vgl. Simmhan et al. 2005

<sup>40</sup> Vgl. Cuevas-Vicentín et al. 2012, S. 193

<sup>41</sup> Vgl. Cuevas-Vicentín et al. 2012, S. 193

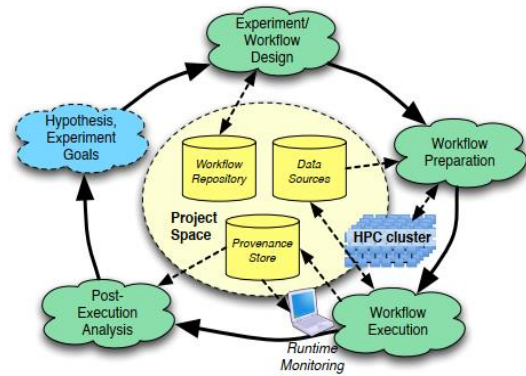
<sup>42</sup> Vgl. Cuevas-Vicentín et al. 2012, S. 193; Ludäscher, et al. 2009, S. 1 ff.

<sup>43</sup> Vgl. Cuevas-Vicentín et al. 2012, S. 197

## 7.3 Phasen

Laut Ludäscher 2009 besteht ein Scientific Workflow aus folgenden typischen Phasen (s. Bild rechts).<sup>44</sup>

Für detailliertere Erläuterungen vgl. Abb. 11 im Anhang.



## 8. Anwendungsbeispiele

### 8.1 Problematik

Heutzutage existiert noch kein einheitliches etabliertes wissenschaftliches Scientific Workflow Managementsystem, welches wissenschaftlichen Workflows über seinen gesamten Lebenszyklus verwaltet. Hauptursache dafür sind komplexe und oft schwer umsetzbare Anforderungen aus Sicht der Technik und der Benutzer/innen (vgl. Abb. 12 im Anhang für weiterführende Erläuterungen zu den Anforderungen).<sup>45</sup>

### 8.2 Auswahl

Je nach Phase des Scientific Workflows ist in der Literatur eine große Auswahl an wissenschaftlichen Werkzeugen zu finden.<sup>46</sup> Beispielsweise existieren bis heute 254 Workflow-Systeme<sup>47</sup> und zahlreiche kostenpflichtige und opensource Browser-, Middleware-, Softwareanwendungen (vgl. PowerPoint-Präsentation, Seiten 109-111).

### 8.3 Vorführung

- [myExperiment](#)
- [Kepler](#)

## 9. Fazit

In Bezug auf das Forschungsdatenmanagement kann Folgendes zusammengefasst werden:<sup>48</sup>

- Die Verfügbarkeit an komplexe und hochqualitative Forschungsdaten wächst stetig aufgrund u. a. moderner wissenschaftlicher Instrumente.
- Datenexploration, interaktive Datenanalyse und -integration von Forschungsdaten sind für den Fortschritt der Wissenschaft essentiell.
- Herkömmliche Datenbanksysteme haben lange wissenschaftliche Datentypen ignoriert.

<sup>44</sup> Vgl. Ludäscher, et al. 2009, S. 1 ff.

<sup>45</sup> Vgl. Ludäscher, et al. 2009, S. 1 ff.

<sup>46</sup> Vgl. Valle 2019; Mathews 1996; Liu et al. (2015), S. 2 ff.

<sup>47</sup> Vgl. Pellmann 2019

<sup>48</sup> Vgl. Gray et al. 2005, S. 1

- Traditionelle Datenbanksysteme haben die Vorteile der automatischen Parallelität, der Indizierung und des nicht-prozeduralen Zugriffs.
- Um die Bereitstellung (auch für Dritte), die (langzeit-)Archivierung, die Wiederverwendung, die Qualität und Verständlichkeit von Forschungsdaten in den unterschiedlichsten Disziplinen gewährleisten zu können, wird ein ganzheitliches, multifunktionales Forschungsdatenmanagementsystem benötigt.
- Eine ausgeprägte standardisierte Metadaten-, Ontologien-, Datenprovenienz-Kultur sind für die Zukunft des Forschungsdatenmanagements bzw. für die Entwicklung von Forschungsmanagement-Anwendungen unerlässlich.
- Allgemein besteht ein Bedarf nach benutzerfreundliche Software-Infrastrukturen (intelligente Systeme) zum Teilen, Auffinden, Erfassen, Organisieren, Analysieren, Visualisieren und Veröffentlichen von Forschungsdaten.
- Es wird viel über Ontologien (Semantic Web, RDF etc.) diskutiert. Das Risiko besteht darin, die Implementierung des Datenmanagements so zu erschweren, dass das System unübersichtlich wird.

## 10. Diskussion

Ist eine einheitliche gut funktionierende domänenübergreifende Lösung für das Forschungsdatenmanagement möglich?

## 11. Literaturverzeichnis

- Barnes, Nataly o. J.: XDF: The Extensible Data Format Based on XML Concepts. URL: [https://nssdc.gsfc.nasa.gov/nssdc\\_news/june01/xdx.html](https://nssdc.gsfc.nasa.gov/nssdc_news/june01/xdx.html) [Letzter Zugriff: 04.07.2019]
- Bretherton, Francis (1994): Reference Model for Metadata: A Strawman DRAFT 3 / 2 / 94. URL: <https://www.semanticscholar.org/paper/Reference-Model-for-Metadata-%3A-A-Strawman-DRAFT-3-%2F-Bretherton/f9414454ef0e25ef102831ed8c7a4b6e9c094b00> [Letzter Zugriff: 04.07.2019]
- Cheney, James; Finkelstein, Anthony; Ludäscher Bertram; Vansummeren, Stijn (2012): Principles of provenance (Dagstuhl Seminar 12091). DOI: 10.4230/DagRep.2.2.84 [Letzter Zugriff: 04.07.2019]
- Coustaty, Mickaël (2011): Complex analysis of historical documents. Application to lettrines. URL: [https://www.researchgate.net/publication/278639053\\_Complex\\_analysis\\_of\\_historical\\_documents\\_application\\_to\\_lettrines](https://www.researchgate.net/publication/278639053_Complex_analysis_of_historical_documents_application_to_lettrines) [Letzter Zugriff: 04.07.2019]
- Cuevas-Vicenttín, Víctor; Dey, Saumen; Köhler, Sven; Riddle, Sean; Ludäscher; Bertram (2012): Scientific Workflows and Provenance: Introduction and Research Opportunities. In: Datenbank Spektrum. Zeitschrift für Datenbanktechnologien und Information Retrieval. November 2012, Vol. 12, Issue 3. S. 193-203. Berlin, Heidelberg. URL: <https://doi.org/10.1007/s13222-012-0100-z> [Letzter Zugriff: 04.07.2019]
- Cui, Yingwei and Widom, Jennifer (2001): Lineage Tracing for General Data Warehouse Transformations. URL: <https://www.researchgate.net/publication/229911112-Lineage-Tracing-for-General-Data-Warehouse-Transformations> [Letzter Zugriff: 04.07.2019]
- Europäische Kommission (2005) – ARION: An advanced lightweight architecture for accessing scientific collections. Factsheet. URL: <https://cordis.europa.eu/project/rcn/55454/factsheet/de> [Letzter Zugriff: 02.07.2019]
- GESIS – Leibniz-Institut für Sozialwissenschaften in Mannheim o. J.: Management von Forschungsdaten. URL: <https://www.gesis.org/angebot/archivieren-und-registrieren/cessda-training/forschungsdatenmanagement/> [Letzter Zugriff: 02.07.2019]
- Gray, Jim; T. Liu, David; Nieto-Santisteban, Maria; Szalay, Alexander; J. DeWitt, David; Heber, Gerd (2005): Scientific Data Management in the Coming Decade. In: SIGMOD Record, Vol. 34(4), March 2005. New York.
- GISMentors (2019): GRASS GIS Workshop in Jena. Unit 13 - PyGRASS intro. URL: <http://training.gismentors.eu/grass-gis-workshop-jena-2018/units/13.html> [Letzter Zugriff: 02.07.2019]
- HDF Group (2019): HDF-EOS Tools and Information Center. URL: <https://hdfEOS.org> [Letzter Zugriff: 04.07.2019]
- HEASARC – HighEnergy Astrophysics Science Archive Research Center (2017): FITS Support Office at NASA/GSFC. FITS – Flexible Image Transport System. URL: <https://fits.gsfc.nasa.gov> [Letzter Zugriff: 04.07.2019]

- Hoffmann, Andre (2018): Metadatenstandards. URL: <https://wiki.dnb.de/x/IYAHc> [Letzter Zugriff: 02.07.2019]
- HU Berlin – Humboldt-Universität zu Berlin (2015): Computer- und Medienservice – Forschungsdatenmanagement – Informationen – Relevanz von Forschungsdaten. URL: <https://www.cms.hu-berlin.de/de/dl/dataman/infos/relevanz> [Letzter Zugriff: 02.07.2019]
- Houstis, Catherine E., Lalis, Spyros (2001): ARION: An Advanced Lightweight Software System Architecture for accessing Scientific Collections. URL: [https://www.researchgate.net/publication/243788163\\_ARION\\_An\\_Advanced\\_Lightweight\\_Software\\_System\\_Architecture\\_for\\_accessing\\_Scientific\\_Collections](https://www.researchgate.net/publication/243788163_ARION_An_Advanced_Lightweight_Software_System_Architecture_for_accessing_Scientific_Collections) [Letzter Zugriff: 02.07.2019]
- Houstis, Catherine E.; Lalis, Spyros; Pitikakis, Marios; Vasilakis, George V.; Kritikos, Kyriakos; Smardas, Antonis. (2003): A Grid Service-Based Infrastructure for Accessing Scientific Collections: The Case of the ARION System. In: International Journal of High Performance Computing Applications, 17 (Fall 2003). S. 269-280. California. URL: [https://www.researchgate.net/publication/230578866\\_A\\_Grid\\_Service-Based\\_Infrastructure\\_for\\_Accessing\\_Scientific\\_Collections\\_The\\_Case\\_of\\_the\\_ARION\\_System](https://www.researchgate.net/publication/230578866_A_Grid_Service-Based_Infrastructure_for_Accessing_Scientific_Collections_The_Case_of_the_ARION_System) [Letzter Zugriff: 03.07.2019]
- Houstis, Catherine E.; Lalis, Spyros; Vavalis, Manolis; Pitikakis, Marios; Vasilakis, George V. (2005): Advanced Middleware for e-Science Applied to Environmental Integrated Systems. In: Hilty, Lorenz M.; Eberhard K. Seifert; Treibert, Rene (2005): „Information Systems for Sustainable Development“, Chapter VII. S. 76-93. Hershey, London, Melbourne, Singapore. URL: [https://www.researchgate.net/publication/230578788\\_Advanced\\_Middleware\\_for\\_e-Science\\_Applied\\_to\\_Environmental\\_Integrated\\_Systems](https://www.researchgate.net/publication/230578788_Advanced_Middleware_for_e-Science_Applied_to_Environmental_Integrated_Systems) [Letzter Zugriff: 03.07.2019]
- IVOA – International Virtual Observatory Alliance (o. J.): Homepage. URL: <http://www.ivoa.net> [Letzter Zugriff: 02.07.2019]
- Jensen, Uwe (2019): Forschungsdaten und Forschungsdatenmanagement in den Sozialwissenschaften. In: Jensen U., Netscher, Sebastian; Weller, Katrin (Hrsg.): Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten: Grundlagen und praktische Lösungen für den Umgang mit quantitativen Forschungsdaten. S. 13-36. Opladen, Berlin, Toronto. URL: <http://www.jstor.org/stable/j.ctvbkk1p8.5> [Letzter Zugriff: 05.07.2019]
- Kindling, Maxi; Schirnbacher, Peter; Simukovic, Elena (2019): Forschungsdatenmanagement an Hochschulen: das Beispiel der Humboldt-Universität zu Berlin. URL: <https://libreas.eu/ausgabe23/07kindling/#2> [Letzter Zugriff: 04.07.2019]
- Liu, Ji; Pacitti, Esther; Valduriez, Patrick; Mattoso, Marta (2015): A Survey of Data-Intensive Scientific Workflow Management. In: Journal of Grid Computing. December 2015, Volume 13, Issue 4. S. 457-493. URL: <https://link.springer.com/article/10.1007/s10723-015-9329-8> [Letzter Zugriff: 05.07.2019]
- Ludäscher, Bertram; Weske, Mathias; McPhillips, Timothy; Bowers, Shawn (2009): Scientific Workflows: Business as Usual? Lecture Notes in Computer Science. S. 31-47. URL: [https://www.researchgate.net/publication/221586063\\_Scientific\\_Workflows\\_Business\\_as\\_Usual](https://www.researchgate.net/publication/221586063_Scientific_Workflows_Business_as_Usual) [Letzter Zugriff: 05.07.2019]

- Mathews, Jason (1996): The World Wide Web: Interactive Interface to Scientific Data. Volume 12, Number 1, March 1996. URL: [https://nssdc.gsfc.nasa.gov/nssdc\\_news/march96/08\\_j\\_mathews\\_0396.html](https://nssdc.gsfc.nasa.gov/nssdc_news/march96/08_j_mathews_0396.html) [Letzter Zugriff: 02.07.2019]
- Moore, R.W. (2001): Data Management Systems for Scientific Applications. In: Boisvert R.F., Tang P.T.P. (Hsg.): The Architecture of Scientific Software. IFIP — The International Federation for Information Processing, vol. 60. Springer, Boston, MA.
- NASA - National Aeronautics and Space Administration (2019a): Ocean Color Image Gallery. <https://www.nasa.gov> [Letzter Zugriff: 02.07.2019]
- NASA - National Aeronautics and Space Administration (2019b): Researchers - Overview. <https://www.nasa.gov/researchers> [Letzter Zugriff: 02.07.2019]
- Ontotext – Polygraphia Office Center (2019): What are Ontologies? URL: <https://www.ontotext.com/knowledgehub/fundamentals/what-are-ontologies> [Letzter Zugriff: 02.07.2019]
- Payrhuber, Andrea (2009): Von Metadaten zu Ontologien. URL: <http://semanticweb.myxwiki.org/xwiki/bin/view/Da/SwWwwOntologies> [Letzter Zugriff: 02.07.2019]
- Pellman, John (2019): Existing Workflow systems. URL: <https://github.com/common-workflow-language/common-workflow-language/wiki/Existing-Workflow-systems> [Letzter Zugriff: 10.07.2019]
- Quinn, Peter; Lawrence, Andy; Hanisch, Bob (2004): The Management, Storage and Utilization of Astronomical Data in the 21st Century. A Discussion Paper for the OECD Global Science Forum. URL: <http://www.ivoa.net/documents/Notes/OECDWhitePaper/index.html> [Letzter Zugriff: 02.07.2019]
- RADAR - Research Data Repositorium (2016): Formatempfehlungen. URL: <https://www.radar-projekt.org/display/RD/Datenformate#Datenformate-Format> [Letzter Zugriff: 02.07.2019]
- Strasser, Carly; Cook, Robert; Michener, William; Budden, Amber (o. J.): Primer on Data Management: What you always wanted to know. URL: [https://www.dataone.org/sites/all/documents/DataONE\\_BP\\_Primer\\_020212.pdf](https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf) [Letzter Zugriff: 02.07.2019]
- Simmhan, Yogesh L.; Plale, Beth; Gannon, Dennis (2005): A Survey of Data Provenance Techniques. Technical Report IUB-CS-TR618. URL: [https://www.researchgate.net/publication/taxonomien/228636320\\_A\\_survey\\_of\\_data\\_provenance\\_techniques](https://www.researchgate.net/publication/taxonomien/228636320_A_survey_of_data_provenance_techniques) [Letzter Zugriff: 04.07.2019]
- Staab, Steffen (2002): Wissensmanagement mit Ontologien und Metadaten. In: Informatik-Spektrum. Band 25, Heft 3. S. 194-209. URL: <https://userpages.uni-koblenz.de/~staab/Research/Publications/spektrum2002.pdf> [Letzter Zugriff: 04.07.2019]
- TU Wien (2018): Forschungsdatenmanagement. Policy für Forschungsdatenmanagement an der TU Wien, Annex. URL: <https://www.tuwien.at/forschung/fti-support/forschungsdaten/forschungsdatenmanagement/uebersicht/> [Letzter Zugriff: 04.07.2019]

- University of Ottawa (o. J.): What is research data management? URL: <https://biblio.uottawa.ca/en/services/faculty/research-data-management/what-research-data-management> [Letzter Zugriff: 04.07.2019]
- University of Western Australia (2019): Research Data Management Toolkit: Metadata Standards. URL: <https://guides.library.uwa.edu.au/RDMtoolkit> [Letzter Zugriff: 04.07.2019]
- Valle, Mario (2019): Scientific Data Management. URL: <http://mariovalle.name/sdm/scientific-data-management.html> [Letzter Zugriff: 02.07.2019]
- Van den Eynden, Veerle (2013): Data Life Cycle & Data Management Planning. Research Data Management Team. UK Data Archive. University of Essex. URL: <https://www.ukdataservice.ac.uk/media/187718/dmplanningdm24apr2013.pdf> [Letzter Zugriff: 02.07.2019]
- Williams, Roy (1998): NSF Workshop on Interfaces to Scientific Data Archives. March 25-27, 1998. California. URL: <http://deslab.mit.edu/DesignLab/dicpm/isda.html> [Letzter Zugriff: 02.07.2019]
- Winkelmann, Axel; Matzner, Martin; Müller, Oliver; Becker, Jörg (2008): Ein Plädoyer für die Berücksichtigung von Semantik beim Stammdaten-Alignment. Vorgehensmodell und prototypische Anwendung im Einzelhandel. Conference: DW 2008 - Synergien durch Integration und Informationslogistik, 27.-28.10. 2008 in St. Gallen, Switzerland. URL: [https://www.researchgate.net/publication/221138345\\_Ein\\_Pladoyer\\_fur\\_die\\_Beruecksichtigung\\_von\\_Semantik\\_beim\\_Stammdaten-Alignment\\_-\\_Vorgehensmodell\\_und\\_prototypische\\_Anwendung\\_im\\_Einzelhandel](https://www.researchgate.net/publication/221138345_Ein_Pladoyer_fur_die_Beruecksichtigung_von_Semantik_beim_Stammdaten-Alignment_-_Vorgehensmodell_und_prototypische_Anwendung_im_Einzelhandel) [Letzter Zugriff: 02.07.2019]
- Wandt, Julia o. J.: Metadaten und Metadatenstandards. URL: <https://www.forschungsdaten.info/themen/aufbereiten-und-veroeffentlichen/metadaten-und-metadatenstandards/> [Letzter Zugriff: 04.07.2019]
- Yale University (2019): Research Data Management for the Health Sciences: Metadata and Ontology Schemas. URL: [https://guides.library.yale.edu/rdm\\_healthsci/metadata\\_schema](https://guides.library.yale.edu/rdm_healthsci/metadata_schema) [Letzter Zugriff: 04.07.2019]

## 12. Anhang

|                                                                                                                                                  |    |
|--------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Abbildung 1: Phasen des Scientific „Data Lifecycle“   © University of Ottawa o. J. ....                                                          | 15 |
| Abbildung 2: Erläuterung zu den Phasen des „Scientific Data Lifecycle“   Vgl. University of Ottawa<br>o. J.; Strasser et al o. J., S. 4 ff. .... | 15 |
| Abbildung 3: RADAR-Metadaten-Schemas Pflichtfelder   © RADAR 2016.....                                                                           | 17 |
| Abbildung 4: Metadaten bei Forschungsdaten   © University of Western Australia 2019 .....                                                        | 17 |
| Abbildung 5: Beispiel einer Ontologie   © Coustaty 2011, S. 80 .....                                                                             | 18 |
| Abbildung 6: Serviceorientierte Architektur   © Simmhan et al. 2005, S. 19.....                                                                  | 18 |
| Abbildung 7: Datenbankarchitektur   © Simmhan et al. 2005, S. 19.....                                                                            | 19 |
| Abbildung 8: Skriptbasierte Architektur   © GISMentors 2019.....                                                                                 | 19 |
| Abbildung 9: Datenverarbeitungsarchitektur   © Matlab 2019.....                                                                                  | 20 |
| Abbildung 10: Zusammenfassung der Provenienztechniken nach unterschiedlichen Merkmalen   ©<br>Simmhan et al., S. 21.....                         | 21 |
| Abbildung 11: Phasen eines Scientific Workflows   Vgl. Ludäscher et al., S. 1 ff.....                                                            | 22 |
| Abbildung 12: Anforderungen an einem Forschungsdatenmanagementsystem   Vgl. Ludäscher et<br>al., S. 1 ff. ....                                   | 23 |

Abbildung 1: Phasen des Scientific „Data Lifecycle“ | © University of Ottawa o. J.

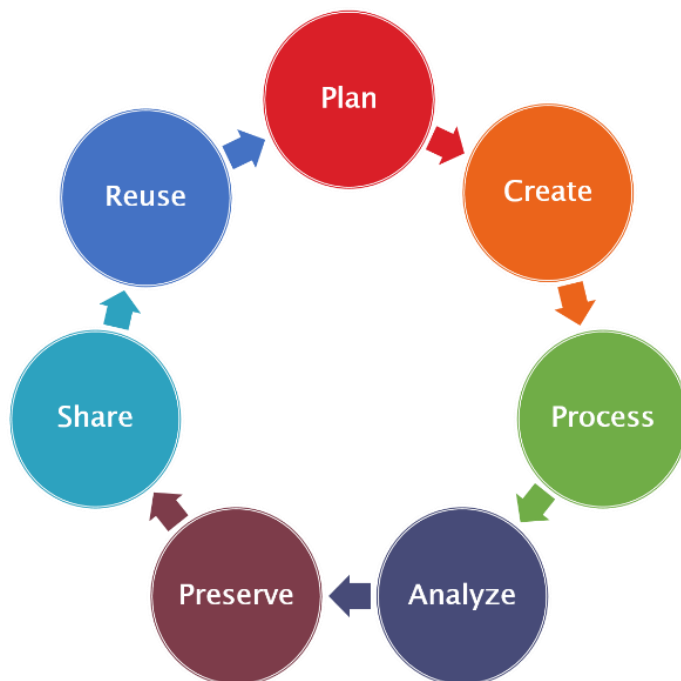


Abbildung 2: Erläuterung zu den Phasen des „Scientific Data Lifecycle“ | Vgl. University of Ottawa o. J.; Strasser et al o. J., S. 4 ff.

| Phase 1: Plan                                                                                                                                                                                                                                                                                                                                                                                                                               |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ul style="list-style-type: none"> <li>· Vorhandene Datenquellen werden durchsucht und analysiert</li> <li>· Daten werden gesammelt</li> <li>· Daten werden organisiert</li> <li>· Rollen werden zugewiesen: Wer ist für die Daten verantwortlich?</li> <li>· Berücksichtigung von Zeit, Kosten, (Cyber-)Infrastruktur (Rechnerarchitekturen, Repositories, Werkzeuge etc.)</li> <li>· Entscheidung über Metadata-Standards etc.</li> </ul> |
| Phase 2: Create                                                                                                                                                                                                                                                                                                                                                                                                                             |
| <ul style="list-style-type: none"> <li>· Organisieren von Daten und Materialien (auch von Drittanbietern)</li> <li>· Beschreibung von Daten, d. h. erfassen und erstellen von Metadaten und zusammenhängendes Material</li> <li>· Erstellen neuer Daten durch Experimente, Beobachtungen, Messungen, Simulationen</li> </ul>                                                                                                                |

|                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Phase 3: Process</b>                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| <ul style="list-style-type: none"> <li>· Daten werden gemäß Qualitätssicherungsstandards in ein digitales Format konvertiert (transkribiert, konvertiert, digitalisiert, kuratiert)</li> <li>· Daten werden überprüft, validiert, bereinigt, neu codiert, versioniert und bei Bedarf anonymisiert</li> <li>· Alle diese Prozesse werden dokumentiert und die Daten werden unter Verwendung des entsprechenden Metadatenstandards beschrieben</li> </ul> |
| <b>Phase 4: Analyze</b>                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| <ul style="list-style-type: none"> <li>· Daten werden durch Visualisierung, Überlagerung, Transformation, statistische Methoden oder Modellierung interpretiert und analysiert</li> <li>· Es entstehen neue Erkenntnisse, Forschungsergebnisse, Veröffentlichungen</li> </ul>                                                                                                                                                                           |
| <b>Phase 5: Preserve</b>                                                                                                                                                                                                                                                                                                                                                                                                                                |
| <ul style="list-style-type: none"> <li>· Daten werden in entsprechende Formaten gespeichert</li> <li>· Es werden Benutzerdokumente und weitere Such-Metadaten erstellt</li> <li>· Neue Daten bekommen eine digitale Kennung (DOI)</li> <li>· Neue Daten werden mit bereits veröffentlichten Daten(sätze) verknüpft</li> </ul>                                                                                                                           |
| <b>Phase 6: Share</b>                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| <ul style="list-style-type: none"> <li>· Aspekte des geistigen Eigentums werden überprüft</li> <li>· Zugriffsrechte werden definiert</li> <li>· Daten, Benutzerdokumentation und Metadaten werden zugänglich gemacht, z. B. auf einem Public Domain-Server oder in einem Repository</li> </ul>                                                                                                                                                          |
| <b>Phase 6: Reuse</b>                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| <ul style="list-style-type: none"> <li>· Anderweitige potenziell nützliche Daten, Benutzerdokumentation und Metadaten werden gefunden und abgerufen.</li> <li>· Erforderlichen Datentransformationen werden vollzogen</li> <li>· Eine Sekundäranalyse wird durchgeführt</li> <li>· Transformationen, Integrationen bzw. Veränderungen werden dokumentiert und Datenquellen zitiert</li> </ul>                                                           |

Abbildung 3: RADAR-Metadatenschemas Pflichtfelder | © RADAR 2016



**RADAR**

RADAR Metadata Kernel  
November 2014  
Version 0.2

1. Overview

9 Mandatory properties (Properties 1-4 are identical to DataCite mandatory properties):

| ID  | Element description                                                   | XML                  |
|-----|-----------------------------------------------------------------------|----------------------|
| 1.  | <b>identifier</b> (unique string which identifies a resource)         | <identifier>         |
| 1.1 | identifierType (Handle or DOI)                                        | <identifierType>     |
| 2.  | <b>creator</b> (persons involved in producing the data)               | <creator>            |
| 2.1 | creator name (name(s) of the creator(s))                              | <creatorName>        |
| 2.2 | creator affiliation (e.g. institution, company)                       | <creatorAffiliation> |
| 3.  | <b>title</b>                                                          | <title>              |
| 4.  | <b>publisher</b> (e.g. corporate/institutional or personal name)      | <publisher>          |
| 5.  | <b>production year</b> (year, in which data was created or refers to) | <productionYear>     |
| 6.  | <b>subject area</b> (scientific fields appropriate for the resource)  | <subjectArea>        |
| 6.1 | subject area detail (additional information on the subject area)      | <subjectAreaDetail>  |
| 7.  | <b>resource</b> (information on the resource's content)               | <resource>           |
| 7.1 | resource type (type of the resource to be archived/published)         | <resourceType>       |
| 8.  | <b>rights</b> (provides a rights management statement)                | <rights>             |
| 8.1 | rights type (opportunity to further specify selected rights)          | <rightsType>         |
| 9.  | <b>rightsholder</b> (institution/person owning property rights)       | <rightsholder>       |

Abbildung 4: Metadaten bei Forschungsdaten | © University of Western Australia 2019

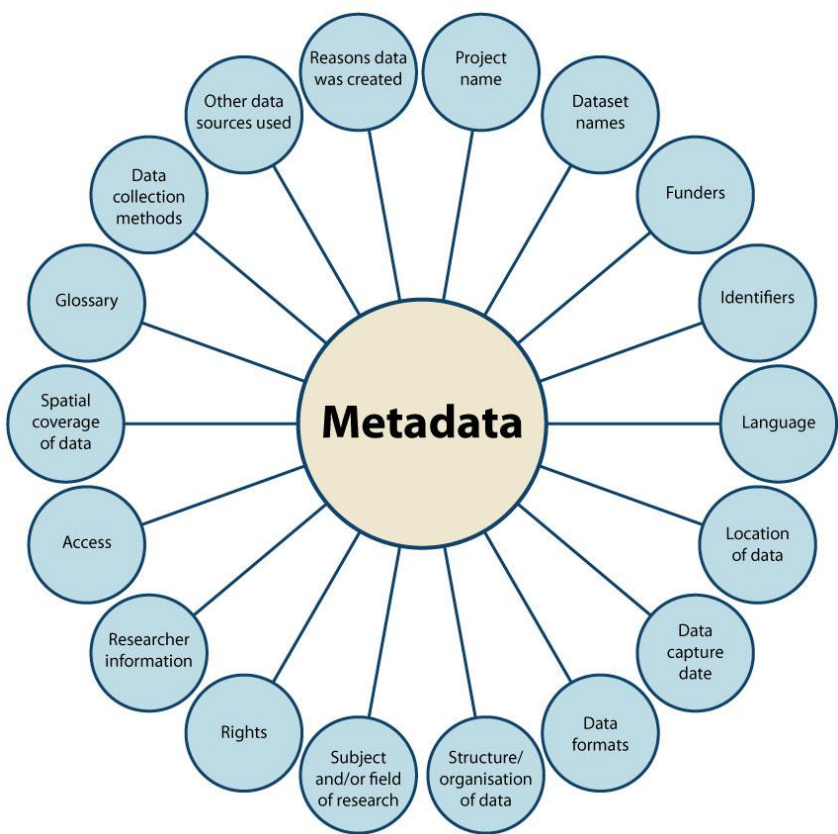


Abbildung 5: Beispiel einer Ontologie | © Coustaty 2011, S. 80

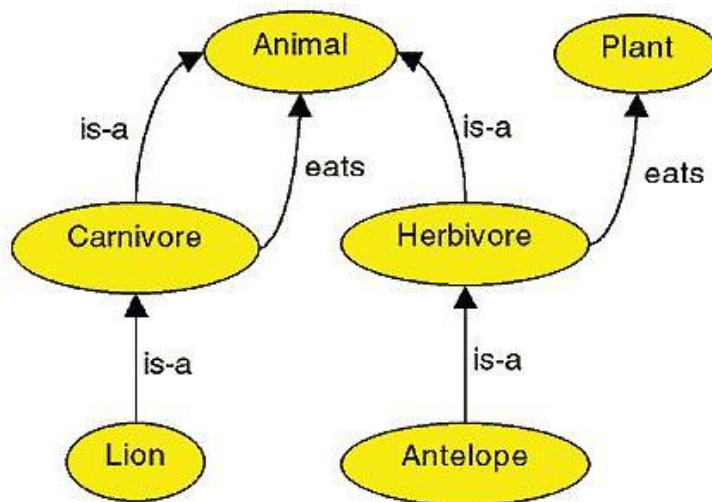


Abbildung 6: Serviceorientierte Architektur | © Simmhan et al. 2005, S. 19

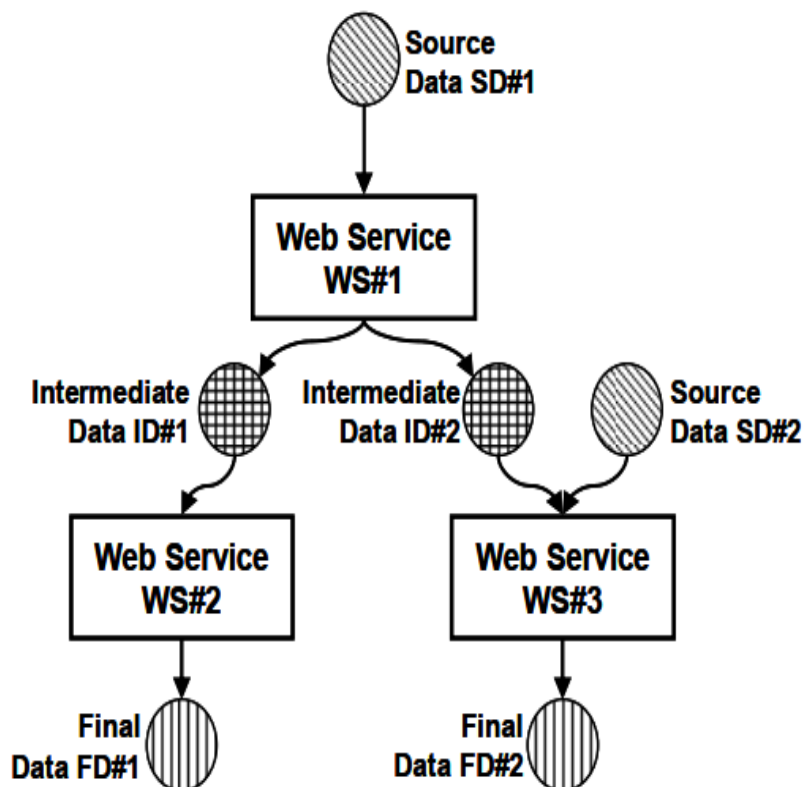


Abbildung 7: Datenbankarchitektur | © Simmhan et al. 2005, S. 19

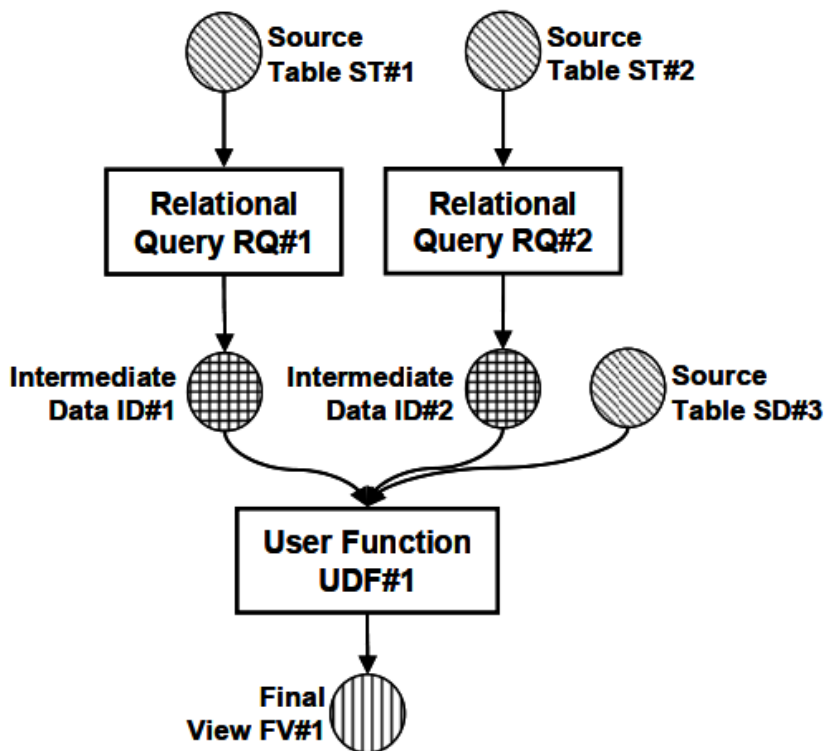


Abbildung 8: Skriptbasierte Architektur | © GISMentors 2019

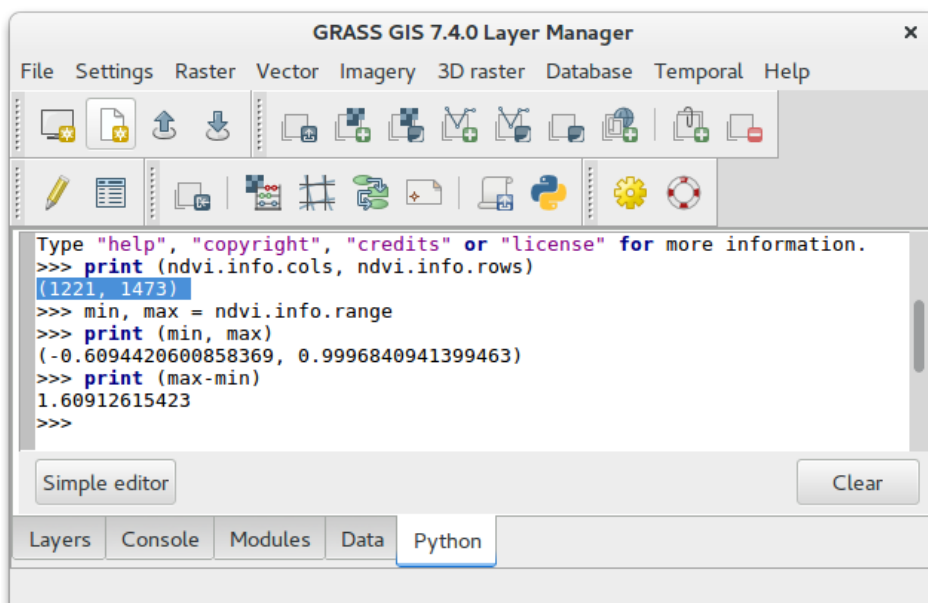


Abbildung 9: Datenverarbeitungsarchitektur | © Matlab 2019

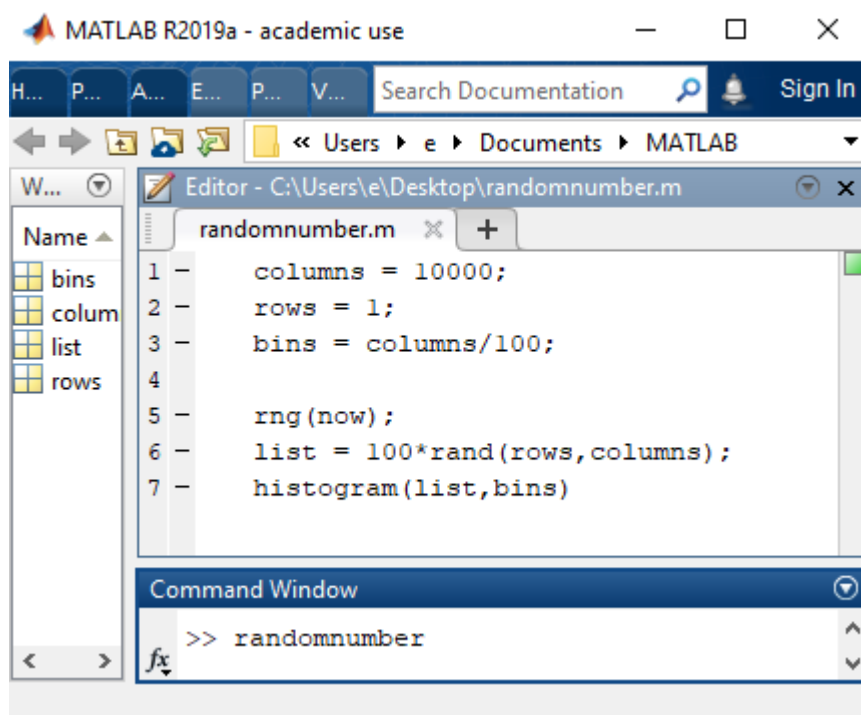


Abbildung 10: Zusammenfassung der Provenienztechniken nach unterschiedlichen Merkmalen | © Simmhan et al., S. 21

|                                | Lanter, D. P. (LIP)                                    | Chimera                                           | MyGRID                                             | CMCS                                        | PASOA                               | ESSW                                                | Tioga                                 | Buneman, P.                           | Cui, Y., Widom, J. (Trio)               |
|--------------------------------|--------------------------------------------------------|---------------------------------------------------|----------------------------------------------------|---------------------------------------------|-------------------------------------|-----------------------------------------------------|---------------------------------------|---------------------------------------|-----------------------------------------|
| Applied Domain                 | GIS                                                    | Physics, Astronomy                                | Biology                                            | Chemical Sciences                           | Biology                             | Earth Sciences                                      | Atmospheric Science                   | Generic (Scientific databases)        | Generic                                 |
| Data Processing Framework      | Command Processing                                     | Service Oriented                                  | Service Oriented                                   | Service Oriented                            | Service Oriented                    | Script Based                                        | Relational Database                   | Relational/Semi Structured Database   | Relational Database                     |
| Application of Provenance      | Informational; update stale, regenerate & compare data | Informational; Audit; Data Regeneration; Planning | Contextual Information; Re-enactment               | Informational; Data Update                  | Informational; Re-enactment         | Informational                                       | Informational; Track errors           | Annotation propagation; View Updation | Information; update propagation         |
| Data/Process Oriented          | Data                                                   | Process                                           | Process                                            | Data                                        | Process                             | Both                                                | Data                                  | Data                                  | Data                                    |
| Granularity                    | Spatial layers                                         | Abstract datasets (Currently files)               | Abstract resources having LSID                     | Files                                       | Abstract parameters to Workflow     | Files                                               | Attributes in Database                | Attributes & Tuples in Databases      | Tuples in Database                      |
| Representation Scheme          | Commands & Frames as Annotations                       | Virtual Data Language Annotations                 | XML/RDF Annotations                                | Dublin Core XML Annotations                 | Annotations                         | XML/RDF Annotations                                 | Inverse Functions                     | Inverse Queries                       | Inverse queries                         |
| Semantic Information           | No                                                     | No                                                | Yes                                                | Limited                                     | No                                  | No, Proposed                                        | No                                    | No                                    | No                                      |
| Storage Repository/ Backend    | MetaDatabase                                           | Virtual Data Catalog/ Relational DB               | mIR repository/ Relational DB                      | SAM over WebDAV/ Relational DB              | PReServ/ Relational DB, File System | Lineage Server/ Relational DB                       | Relational DB                         | N/A                                   | Relational DB                           |
| Provenance Collection Overhead | Store User commands; solicit metadata                  | User defines derivations; automated WF trace      | User defines service semantics; Automated WF Trace | Manual; Apps use DAV APIs, Users use portal | Manual; Actors use PReP API         | Libraries assist user to generate, store provenance | User registers inverse functions      | N/A                                   | Inverse queries automatically generated |
| Addressed Scalability          | No                                                     | Yes                                               | No                                                 | No                                          | No (Proposed)                       | No (Proposed)                                       | Yes                                   | N/A                                   | No                                      |
| Provenance Dissemination       | Queries                                                | Queries                                           | Semantic browser; Lineage graph                    | Browser; Queries; GXL/RDF                   | Queries                             | Browser                                             | Queries; box-and-arrows visualization | N/A                                   | SQL/TriQL Queries                       |

Abbildung 11: Phasen eines Scientific Workflows | Vgl. Ludäscher et al., S. 1 ff.

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Phase 1: Experiment/Workflow Design</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| <ul style="list-style-type: none"> <li>· Beschreibung des Experiments, welches ausgeführt werden soll</li> <li>· Entwurf eines Workflows, welches auf das beschriebene Experiment basiert</li> <li>· Suche nach bereits bestehenden Templates um diese eventuell zu übernehmen oder zu verändern</li> <li>· Neue oder veränderte Workflow-Entwürfe können über ein gemeinsam genutztes Repository zur Weiterverwendung öffentlich gemacht werden</li> </ul>                                                                                                                                                                                            |
| <b>Phase 2: Workflow Preparation</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| <ul style="list-style-type: none"> <li>· Ressourcen für die Ausführung werden ausgewählt und vorbereitet</li> <li>· Parameter werden von den Wissenschaftler/innen festgelegt</li> <li>· Für einige Workflows werden möglicherweise HPC-Ressourcen (High Performance Computing) wie lokale Cluster-Computer oder Remote-Ressourcen (Grid- oder Cloud-Computing) benötigt</li> <li>· Es kann auch sein, dass Rechenaufträge die auf dem HPC-Cluster laufen bestimmte Daten erwarten, dann müssen Daten entsprechend bereitgestellt werden</li> </ul>                                                                                                    |
| <b>Phase 3: Workflow Execution</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| <ul style="list-style-type: none"> <li>· Während der Ausführung des Workflows werden Eingabedaten verbraucht und neue Datenprodukte erstellt</li> <li>· Für groß angelegte rechnergestützte wissenschaftliche Simulationen ist die Laufzeitüberwachung von entscheidender Bedeutung (laufen auf Hunderten oder Tausenden von Knoten; über viele Stunden, Tage oder Wochen)</li> <li>· Datenzwischenprodukte und spezielle Provenienz-Informationen werden häufig auf ein webbasiertes Überwachungs-Dashboard angezeigt</li> <li>· So bleiben Wissenschaftler/innen über den Fortschritt und mögliche Probleme bei der Ausführung informiert</li> </ul> |
| <b>Phase 4: Post-Execution Analysis</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| <ul style="list-style-type: none"> <li>· Wissenschaftler/innen müssen häufig die Workflow-Ergebnisse nach der Ausführung überprüfen und interpretieren: Ist dieses Ergebnis sinnvoll?</li> <li>· Evaluation der Datenprodukte: Welche Ergebnisse wurden von diesem Eingabedatensatz „verdorben“?</li> <li>· Untersuchung von Ausführungsspuren und Datenabhängigkeiten: Warum ist Schritt X fehlgeschlagen?</li> <li>· Debug-Analyse: Welche Schritte wurden ausgeführt? Welcher hat am meisten gedauert?</li> <li>· Analyse der Leistung</li> </ul>                                                                                                   |
| <b>Phase 5: Hypothesis, Experiment Goals</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| <ul style="list-style-type: none"> <li>· Ursprünglichen Hypothesen oder experimentellen Ziele können überarbeitet oder verfeinert werden</li> <li>· Dies kann zu weiteren Workflow-(Re-)Designs führen oder zu einer neuen Iteration des Zyklus</li> <li>· Abhängig von den Ergebnissen des Workflows und den Analyseergebnissen können die ursprünglichen Hypothesen oder experimentellen Ziele überarbeitet oder verfeinert werden, was zu weiteren Workflow-(Re-)Designs führt und möglicherweise eine neue Iteration des Zyklus beginnen lässt.</li> </ul>                                                                                         |

Abbildung 12: Anforderungen an einem Forschungsdatenmanagementsystem | Vgl. Ludäscher et al., S. 1 ff.

| Anforderungen aus Sicht der Benutzer/innen                                                                                                                                                                                                                                                        | Anforderungen aus Sicht der Technik                                                                                                                                                                                                                                                                                                                                                                                          |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><u>Designtools</u></p> <ul style="list-style-type: none"> <li>Workflow-Design speziell für nicht erfahrene Benutzer, basierend auf Prozesse die von Wissenschaftler/innen definiert werden</li> </ul>                                                                                          | <p><u>Fehlererkennung und Wiederherstellung nach einem Fehler</u></p>                                                                                                                                                                                                                                                                                                                                                        |
| <p><u>Nahtloser Zugriff auf Ressourcen und Services</u></p>                                                                                                                                                                                                                                       | <p><u>Protokollinformationen für jeden Workflow</u></p>                                                                                                                                                                                                                                                                                                                                                                      |
| <p><u>Wiederverwendung allgemeiner Funktionen</u></p> <ul style="list-style-type: none"> <li>Servicezusammensetzung und -wiederverwendung</li> <li>Wie werden einfache, leicht wiederverwendbare und nicht anwendungsspezifische Dienste zur Ausführung komplexer Aufgaben komponiert?</li> </ul> | <p><u>Möglichkeit zur daten- und rechenintensive Aufgaben (im besten Fall auch gleichzeitig)</u></p>                                                                                                                                                                                                                                                                                                                         |
| <p><u>Benutzerinteraktion</u></p> <ul style="list-style-type: none"> <li>Beispielsweise sollten Benutzer Zwischenergebnisse überprüfen können</li> </ul>                                                                                                                                          | <p><u>Zugriff auf semantik- und metadatenbasierte Datasets</u></p>                                                                                                                                                                                                                                                                                                                                                           |
| <p><u>Registrierung und Veröffentlichung von Datenprodukten</u></p> <ul style="list-style-type: none"> <li>Registrierung und Veröffentlichung von Datenprodukten und „Prozessprodukten“ (Workflows)</li> </ul>                                                                                    | <p><u>Datenmanagement/-integration</u></p>                                                                                                                                                                                                                                                                                                                                                                                   |
| <p><u>Unterstützung von Provenienzdaten</u></p> <ul style="list-style-type: none"> <li>Herkunft der Ergebnisse sollten aufgezeichnet und die Laufzeitausführung überwachen werden.</li> </ul>                                                                                                     | <p><u>Visualisierungsmöglichkeiten</u></p>                                                                                                                                                                                                                                                                                                                                                                                   |
|                                                                                                                                                                                                                                                                                                   | <p><u>Zertifizierung, Vertrauen, Sicherheit</u></p>                                                                                                                                                                                                                                                                                                                                                                          |
|                                                                                                                                                                                                                                                                                                   | <p><u>Skalierbarkeit</u></p> <ul style="list-style-type: none"> <li>Einige Workflows erfordern große Datenmengen und/oder High-End-Rechenressourcen und benötigen Schnittstellen zu Grid-Middleware-Komponenten</li> <li>Workflows müssen skalierbar und fehlertolerant sein, damit Aufgaben bei rechenintensiven Anwendungen relativ schnell ausgeführt werden können.</li> <li>Bspw. kann es für ein Experiment</li> </ul> |

- Provenienzdaten sollten dann abgefragt, analysiert und gespeichert, visualisiert werden, um ein tieferes Verständnis der Ergebnisse bereitgestellt zu haben oder einfach, um einen Workflow oder Datensatz „debuggen“ zu können, indem seine Herkunft durch die Ausführung des Workflows zurückverfolgt wird: Welche Datenprodukte und Tools haben ein abgeleitetes Datenprodukt erstellt?
- Protokollsequenz von Schritten, Parametereinstellungen usw. soll möglich sein

### Generisch vs. spezifisch

- Es sollte generisch genug sein, um verschiedene *Communities* zu bedienen, aber spezifisch genug, um eine *Domain* zu bedienen

### „Intelligente“ semantische Links

- Diese sollten das Workflow-Design unterstützen, indem sie vorschlagen, welche Komponenten möglicherweise zusammenpassen

### „Smart“ re-runs → Ändern von Parametern

- Ändern eines Parameters nach Zwischenergebnissen sollte möglich sein, ohne den Workflow komplett neu auszuführen

### Leistungsstärke

- Freistehende Ausführung → Ausführung von Workflows mit langer Laufzeit im Hintergrund ermöglichen, bspw. auf einem Remoteserver

### Erweiterbar

- Erweiterbarkeitmöglichkeiten für

erforderlich sein, ein Programm tausendmal auszuführen mit leicht veränderten Eingabeparametern, wodurch viele Ressourcen (Rechenzyklen) verbraucht werden.

- Distributed Grid oder Cloud-Computing-Technologien und andere parallele Frameworks wie *MapReduce* können verwendet werden, um die Workflow-Ausführung durch Ausnutzung paralleler Ressourcen zu skalieren.

### Zuverlässigkeit und Fehlertoleranz

- Workflows könnten fehlschlagen, wenn der Webdienst ausfällt

Experten/innen

- Es sollte fast eine visuelle Programmierschnittstelle sein

### **Benutzerfreundlichkeit**

- Es ist eine ziemlich einfache Benutzeroberfläche mit komplexeren Funktionen im Hintergrund erwünscht