WissGrid

Deliverable 3.4.3

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

AP 3: Langzeitarchivierung von Forschungsdaten

Metadaten und Provenienz: Eine Übersicht[1]

Version 1.0:
Entwurfsversion zur öffentlichen Kommentierung

[1] Diese Übersicht wurde im Rahmen des Projektes „WissGrid“ erstellt. Dieses Projekt wird vom Bundesministerium für Bildung und Forschung (BMBF) gefördert.

Herausgegeben von

WissGrid – Grid für die Wissenschaft
Teil des D-Grid Verbundes und der deutschen e-Science Initiative
www.wissgrid.de

BMBF Förderkennzeichen: 01|G09005A-G (Verbundprojekt)
Laufzeit: Mai 2009 - April 2012
Dokumentstatus: Deliverable

Kontakt

Niedersächsische Staats- und Universitätsbibliothek Göttingen
Jens Ludwig
Abteilung Forschung & Entwicklung
Papendiek 14
37073 Göttingen

Leibniz-Institut für Astrophysik Potsdam
Harry Enke
Abteilung e-Science
An der Sternwarte 16
14482 Potsdam

Autor

Thomas Fischer, Niedersächsische Staats- und Universitätsbibliothek Göttingen

Inhaltsverzeichnis

1 Herkunftsmetadaten

Herkunftsinformation ist für elektronische Dokumente wie für digitale Daten aus mindestens zwei Gründen von Bedeutung:

- Zur Feststellung ihrer Vertrauenswürdigkeit: Wer hat sie erstellt, wurden bestehende Standards berücksichtigt, von wem sind sie verändert worden, wie wird garantiert dass niemand ihren Inhalt unbemerkt verändern konnte?
- Zur Bewahrung ihrer Nutzbarkeit: In welchem (digitalen) Format liegt die Dokumente bzw. die Daten vor, mit welchen Programmen in welcher digitalen Umgebung wurden sie erstellt bzw. bearbeitet, mit welchem Programm können sie präsentiert werden?

Während Herkunftsinformation oder Provenienz in der Kunst eine lange Tradition und offensichtliche Bedeutung hat (z.B. für die Frage des Wertes eines Kunstwerks), entwickelt sich das Verständnis von Provenienz digitaler Daten erst allmählich. Einige Anstrengungen auf nationaler und internationaler Ebene wurden unternommen, um den Begriff und den Umgang damit auf eine solide Basis zu stellen, zu nennen sind z.B. das britische PASOA-Projekt (http://www.pasoa.org/), das von der EU geförderte „Provenance Project“ (http://www.gridprovenance.org/) oder das internationale PREMIS-Projekt (http://www.loc.gov/standards/premis/), außerdem die vier „Provenance Challenges“ (http://twiki.ipaw.info/bin/view/Challenge/).

Will man Provenienz in einem System einsetzen, so müssen zumindest drei miteinander verbundene Aufgaben bewältigt werden:

- Es muss eine Vorstellung davon entwickelt werden, was die entsprechenden Provenienzinformationen sind, wie sie miteinander zusammenhängen und wie sie in einem Datenmodell dargestellt werden können.
- Für diese Daten muss eine verlässliche Methode gefunden werden, sie zu speichern, zu recherchieren, zu präsentieren und auszutauschen.
- Schließlich muss ein effektives System gefunden werden, wie diese Daten erhoben werden.

Alles dies muss unter Berücksichtigung gegebener Rahmenbedingungen gelöst werden: weder darf die Performanz des Basissystem ernsthaft leiden, noch können beliebig große zusätzliche Datenmengen verkraftet werden oder menschliche Arbeitskraft in großem Maße eingesetzt werden.

Innerhalb virtueller Umgebungen (z.B. Grid oder Cloud), die einerseits auf (viele) verschiedene Nutzer und andererseits eventuell auf eine lange Verwendungszeit angelegt sind, erhöht sich die Bedeutung der Provenienzinformationen, wird aber auch gleichzeitig komplizierter, weil vielfältige Interaktionen mit verschiedenen Teilsystemen der Umgebung zur Erzeugung, Speicherung, Authentifizierung, Weiterverarbeitung berücksichtigt werden müssen.
Zusätzlich werden in solchen Umgebungen nicht nur (mehr oder weniger) einfache Dokumente bearbeitet, sondern komplexe und z.T. sehr umfangreiche Forschungsdaten erzeugt,

verarbeitet, weitergeleitet und abgelegt. Damit verschärft sich die Frage der Herkunftsinformationen in ihrer Dringlichkeit einerseits und ihrer Schwierigkeit andererseits.

Zur Bearbeitung dieser Frage wird eine Sichtung relevanter Konzepte vorgelegt, die entweder allgemein Anforderungen an ein System zur Erzeugung und Pflege von Herkunftsmetadaten beschreiben oder konkret schon Schemata vorlegen, die festlegen, welche Informationen wie erfasste werden sollten.

1.1 Quellen

Beschrieben werden die folgenden Quellen:

1. Das „Metadata Framework to Support the Preservation of Digital Objects“ (siehe http://www.oclc.org/research/projects/pmwg/pm_framework.pdf) der *OCLC/RLG Working Group on Preservation Metadata* (2002)
2. Die „Open Provenance Specification“ (siehe http://www.gridprovenance.org/openSpecification/) des *EU Grid Provenance Project* (2005)
3. Das „Usage Record – Format Recommendation“ (http://www.ogf.org/documents/GFD.98.pdf) des *Open Grid Forum* (2006)
4. „The Origin of Data“, Dissertation von Paul T. Groth (http://eprints.ecs.soton.ac.uk/14649/) (2007)
5. Das „PREMIS Data Dictionary for Preservation Metadata (version 2.0)“ (http://www.loc.gov/standards/premis/v2/premis-2-0.pdf) (2005, 2008)
6. Das „Open Provenance Model“ (http://eprints.ecs.soton.ac.uk/18332/1/opm.pdf) als Ergebnis des *Third Provenance Challenge* (2008)
7. Das „DELOS Digital Library Reference Model“ (http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf) des *DELOS* Projektes (2007, 2010)
8. Das „Core Scientific Metadata Model“ (http://www.ijdc.net/index.php/ijdc/article/viewFile/149/211) (2010)
9. Das „Provenance Vocabulary“ (http://sourceforge.net/apps/mediawiki/trdf/index.php?title=Provenance_Vocabulary) (2010)
10. „The Foundations for Provenance on the Web“ Ein Übersichtsartikel von Luc Mureau (http://eprints.ecs.soton.ac.uk/21691/1/survey.pdf) (2010)

11. Die „DCMI Metadata Provenance Task Group“ (http://www.dublincore.org/groups/provenance/) (2010-2011)
12. Die „W3C Provenance (Incubator) Group“ (http://www.w3.org/2005/Incubator/prov/XGR-prov/, http://www.w3.org/2011/prov/) (2009-2012)
13. Projekte zur Verbindung wissenschaftlicher Workflow-Systeme mit der Erhebung und Verwaltung von Provenienzinformationen (Verweise in dem entsprechenden Abschnitt) (2010-12)

Die Darstellung erfolgt teilweise zweisprachig, da es mir nicht notwendig und unnötig aufwendig erscheint, alle englischen Zitate ins Deutsche zu übersetzen. Sie würden dadurch auch nicht unbedingt klarer, da sich die Begrifflichkeit dieses Arbeitsfeldes erst allmählich herausbildet und derzeit noch stark auf englischen Veröffentlichungen beruht. Wo englische Textteile ins Deutsche übersetzt wurden, stammt die Übersetzung von mir.

Natürlich kann ein solcher Überblick das Studium der entsprechenden Dokumente nicht ersetzen, Ziel ist vielmehr, Hinweise darauf zu geben, wo welche Fragen bearbeitet werden und welche Herangehensweise für das je eigene Projekt am erfolgversprechendsten erscheint.

1.2 Überblick

Die beschriebenen Papiere zeigen verschiedene Herangehensweisen an die Frage der Herkunft digitaler Dokumente. Während für DELOS die Herkunft ein einfacher Qualitätsparameter ist, bietet das „Usage Record“ des Open Grid Forum ein relativ reichhaltiges Format zur Beschreibung von Dokumenten im allgemeinen und ihrer Herkunft im Besonderen.
Die von Arbeitsgruppen der Library of Congress und der Research Library Group vorgelegten Dokumente zum „Metadata Framework“ und „PREMIS Data Dictionary“ bieten umfangreiches Material zur Beschreibung von Herkunftsmetadaten im Kontext allgemeiner Fragen der Implementierung eines „Open Archive Information System“, insbesondere zu Fragen der Langzeitarchivierung.
Das EU-Provenance Project legt mit der „Open Provenance Specification“ Regeln vor, die ein System zur Erfassung von Herkunftsinformationen zu erfüllen hätte, und das „Open Provenance Model“ bietet ein komplexes Model in Form eines gerichteten Graphen, in dem Herkunftsinformationen gefasst werden können.
Eine umfassende Beschreibung eines mögliches Systems zur Erfassung, Verwaltung und Nutzung von Herkunftsinformationen liefert die Dissertation von Paul Groth, die im Umfeld des EU Provenance-Projektes entstanden ist. Erkenntnisse daraus fließen in die umfassenden Arbeiten der W3C Provenance Group ein.
Aus diesem Umfeld stammt auch der Übersichtsartikel von Luc Moreau (Groths Doktorvater), der einen umfassenden Überblick über die Literatur zur „Provenance“ bis zum Herbst 2009 bietet.
Das „Provenance Vocabulary“ ist auf das „Web of Data“ ausgerichtet und stellt für diesen Kontext ein Vokabular und dazu passende Methoden der Erzeugung und Verarbeitung vor.
Die „DCMI Metadata Provenance Task Group“ untersucht in einer neuen Initiative die Möglichkeit, Herkunftsmetadaten für DC-Metadaten auszudrücken.

Eine sehr umfangreiche und sehr aktuelle Dokumentation liefert die „W3C Provenance Group“, die sich aus der gleichnamigen „Incubator Group“ entwickelt hat und sehr aktiv an einem webbasierten Rahmen für die Provenienzinformationen arbeitet und dazu vielfältige zusätzliche und weiterführende Dokumente bereitstellt.
In den letzten Jahren hat sich eine gewissen Konsolidierung der Vorstellungen von Provenienz ergeben (weitgehend auf der Basis des Open Provenance Model), so dass jetzt stärker Fragen der Implementierung in den Vordergrund getreten sind. Im Fokus des Interesses stehen dabei die „Scientific Workflow Engines“, an die geeignete Provenienzsystem „angeflanscht“ werden sollen. Der letzte Abschnitt gibt einen Überblick über diese Entwicklungen.

Gewisse Zusammenhänge und Ähnlichkeiten ergeben sich aus der Autorenschaft für die verschiedenen Beiträge, insbesondere die um das EU-Provenance-Projekt gebildete Gruppe ist sehr aktiv und auch in die aktuelle W3C-Provenance Group eingebunden.

Provenance Project: Paul Groth, Sheng Jiang, Simon Miles, Steve Munroe, Victor Tan, Sofia Tsasakou, Luc Moreau (siehe dazu auch die „Provenance Aware Service Oriented Architecture“ (PASOA), http://www.pasoa.org/)

OGF Usage Record: R. Mach, R. Lepro-Metz, S. Jackson, L. McGinnis

Delos: L. Candela, D. Castelli, N. Ferro, Y. Ioannidis, G. Koutrika, C. Meghini, P. Pagano, S. Ross, D. Soergel, M. Agosti, M. Dobreva, V. Katifori, H. Schuldt

PREMIS: Rebecca Guenther, Steve Bordwell, Olaf Brandt, Priscilla Caplan, Gerard Clifton, Angela Dappert, Markus Enders, Brian Lavoie, Bill Leonard, Zhiwu Xie

OPM: Luc Moreau, Ben Clifford, Juliana Freire, Yolanda Gil, Paul Groth, Joe Futrelle, Natalia Kwasnikowska, Simon Miles, Paolo Missier, Jim Myers, Yogesh Simmhan, Eric Stephan, Jan Van den Busscheh

Core Scientific Metadata Model: Brian Matthews, Shoaib Sufi, Damian Flannery, Laurent Lerusse, Tom Griffin, Michael Gleaves, Kerstin Kleese

Provenance Vocabulary Core Ontology: Olaf Hartig, Jun Zhao

DCMI Metadata Provenance Task Group: Kai Eckert, Magnus Pfeffer, Johanna Völker

W3C Provenance: Yolanda Gil, James Cheney, Paul Groth, Olaf Hartig, Simon Miles, Luc Moreau, Paulo Pinheiro da Silva

2 Metadata Framework

2002 veröffentlichtete die „OCLC/RLG Working Group on Preservation Metadata“ einen Report, in dem versucht wurde, die Prinzipien des „Open Archive Information System“ (OAIS)[2] mittels Metadaten zu beschreiben;
„A Metadata Framework to Support the Preservation of Digital Objects“, Dublin, Ohio: OCLC Online Computer Library Center, 2002
(http://www.oclc.org/research/projects/pmwg/pm_framework.pdf) oder
([OCLC/RLG Working Group on Preservation Metadata, 2002]: http://www.oclc.org/research/activities/past/orprojects/pmwg/pm_framework.pdf) Seitenangaben in diesem Abschnitt beziehen sich auf dieses Dokument, wenn keine andere Quelle vermerkt ist.

Schon für die Archivierung sind Herkunfts- und Authentizitätsinformationen von Bedeutung, daher gibt es hier schon Verweise auf solche Informationen. Sie treten insbesondere in zwei Teilabschnitten im Bereich „Preservation Description Information“ (PDI) auf: Unter „Provenance Information“ und „Fixity Information“, wobei sich der erste Bereich stark auf die Entstehungsgeschichte der Archivobjekte und der zweite eher auf deren Authentizität bezieht. Die entsprechenden Elemente werden hier aufgezählt, die anderen Bereiche nur kurz gestreift.

Das in einem Archiv zu speichernde Objekt wird im OAIS-Kontext als „Archive Information Package“ (AIP) bezeichnet und umfasst vier verschiedene Informationsobjekte (S. 8):

Content Information (CI): der tatsächliche Inhalt des Objektes

Preservation Description Information (PDI): Informationen zur Erhaltung der Nutzbarkeit der CI, es werden davon vier Typen unterschieden:

- „Reference Information“: Liste der für das CI benutzten internen und externen Identifikatoren (z.B. ISBN, URN)
- „Context Information“: Informationen über den Zusammenhang der CI mit ihrer Umgebung (z.B. mit anderen CI oder warum sie überhaupt erstellt wurde)
- „Provenance Information“: Geschichte der Information, ihre Herkunft, in wessen Verantwortung sie lag und welche Bearbeitungsstufen sie durchlaufen hat
- „Fixity Information“: Authentizitätsinformationen (z.B. Prüfsummen, digitale Signaturen)

Packaging Information (PI) verbinden Inhalt und Zusatzinformationen zu einem identifizierbaren Objekt oder Paket (einem „Archival Information Package“)

Descriptive Information (DI) unterstützen den Zugriff auf Inhalte durch Suchfunktionen.

[2] Reference Model for an Open Archival Information System (OAIS) (Washington, DC: Consultative Committee for Space Data Systems, 2002), http://public.ccsds.org/publications/archive/650x0b1.pdf

2.1 Preservation Description Information

Das OAIS Referenzmodell definiert „Preservation Description Information“ als „die Information, die notwendig ist, um die damit verbundene Inhaltsinformation angemessen zu bewahren. Sie legt besonderen Wert auf die Beschreibung früherer und gegenwärtiger Zustände der Inhaltsinformation und stellt sicher, das es eindeutig identifizierbar ist und noch unwissentlich verändert wurde.“ (S.27)

Das OAIS Informationsmodel teilt die „Preservation Description Information“ in vier Kategorien ein (S. 27):

Referenz: beschreibt Identifikationssysteme und die Mechanismen, mit denen Identifikatoren zugewiesen werden, um die Inhaltsinformation sowohl innerhalb als auch außerhalb des gegebenen Archivs eindeutig zu identifizieren.

Kontext: dokumentiert die Beziehung der Inhaltsinformation zu ihrer Umgebung, einschließlich der Gründe ihrer Erzeugung und der Beziehungen zu anderen Inhaltsinformationsobjekten.

Provenienz: dokumentiert die Geschichte der Inhaltsinformation, ihren Ursprung, Änderungen des Objektes oder seines Inhalts sowie der jeweiligen Zuständigkeiten.

Beständigkeit: liefert die Integritätsprüfungen oder Validierungen, die garantieren, dass das Inhaltsinformationsobjekt nicht unkontrolliert geändert wurde.

Zusammengefasst dokumentiert die „Preservation Description Information“ Identität, Beziehungen, Geschichte und die Integrität des archivierten Datenobjektes.

Für uns ist hier die Provenienzinformation von spezifischem Interesse.

Zusätzlich zu der „Chronologie“ des archivierten Datenobjektes können Provenienzinformationen auch als „ereignisbasierte“ Metadaten gefasst werden: der Entwicklungsprozess des Objektes wird durch das Auftreten von „Ereignissen“ hervorgerufen, wie die Erzeugung des Objektes, seine Eigentumsübertragung, die Aufnahme in das Archiv oder die Migration von einem Format in eine anderes. Diese Details dieser Ereignisse und ihre Auswirkungen auf das Datenobjekt zu dokumentieren ist dann eine weitere Schlüsselfunktion der Provenienzinformation. (S. 37)

Die folgenden Metadatenelemente beschreiben die für die Provenienzinformation benötigten Informationen (S. 37f):

Ursprung: beschreibt den Prozess, durch den das Objekt erzeugt wurde

Vorbereitung: beschreibt die Geschichte des Datenojektes, seine Pflege, Änderungen des Inhalts oder der Verantwortung usw. von seiner Erzeugung bis zur Anlieferung an das Archiv.

Aufnahme: beschreibt den Prozess der Aufnahme des Datenojektes in das Archiv.

Archivierung: beschreibt die Pflege, Änderungen des Inhalts oder der Verwaltung usw. während der Aufbewahrung im Archiv.

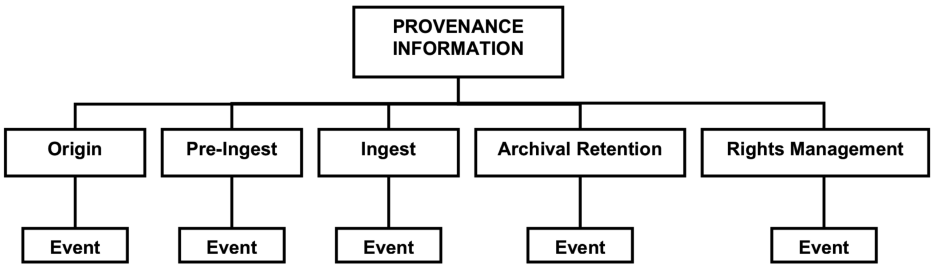

Abb. 1: Provenance Information

Rechte: Angabe der mit dem Datenobjekt verbundenen rechtlichen Beschränkungen.

Diese Elemente beschreiben die groben Phasen oder Aspekte der Chronologie des Datenobjektes bzw. seines Lebenszyklus. Innerhalb jeder dieser Kategorien nimmt Provenienzinformation die Form einer Samlung oder Abfolge von Ereignissen an [...] Daher müssen die die Provenienzinformation darstellenden Metadaten die Einzelheiten dieser Ereignisse aufzeichnen, um ihr Auftreten und ihre Ergebnisse zu dokumentieren, die Verwaltung der Erhaltungsprozesse unterstützen und die verlässliche Aufsicht über das Material dokumentieren, um die Integrität und Authentizität des Datenobjektes zu gewährleisten. (S. 38f)

Das folgenden Metadatum beschreibt ein allgemeines Ereignis, das mit Provenienzinformation zusammenhängt:

Ereignis: Ein Ereignis, das Auswirkungen auf einen oder mehrere Aspekte eines Datenobjektes hat: Inhalt, Format, Rechteverwaltung usw.

Jedes solche Ereignis wird durch Teilfelder genauer spezifiziert (S. 39f):

Bezeichnung: Name der Ereignissen (Beispiele: Änderung der Verantwortung; Migration; Auffrischung des Datenträgers)

Prozedur: Details der Prozedur, die das Ereignis ausmacht

Datum: Datum des Ereignisses

Verantwortliches Organ: die für das Auftreten des Ereignisses verantwortliche Einheit

Ergebnis: Beschreibung des Ergebnisses der letzten Ausführung des Ereignisses

Vermerk: Zusatzinformationen, die für das Ereignis von Bedeutung sind

2.2 Beständigkeit

Nach dem OAIS Referenzmodell liefert die Beständigkeitsinformation „die Integritätsprüfungen oder Validierungen, die garantieren, dass das Inhaltsinformationsobjekt nicht unkontrolliert geändert wurde. Beständigkeitsinformation schließt spezielle Verfahren zur Erkennung von Kodierung und Fehlern ein, die auf das gegebene Inhaltsobjekt ausgerichtet sind [...]

Die Beständigkeitsinformation kann für diese Mechanismen Minimalstandards für die Dienste festlegen.“ (S. 42f)

Für die „Object Authentication“ stehen die folgenden Elemente zur Verfügung:

Authentifizierungstyp: Die zur Authentifizierung des Fatenobjektes benutzte Technik (Beispiel: digitale Signatur) The technique used to authenticate the Content Data Object (Example: Digital signature)

Authentifizierungsverfahren: Die Schritte, mit denen der Authentifizierungstyp implementiert wird, einschließlich von Verweisen auf Dokumentation, Software etc.

Authentifizierungsdatum: Datum der letzten Durchführung des Authentifizierungstyp in dem Archiv

Authentifizierungsergebnis: Ergebnis der letzten Durchführung des Authentifizierungstyp

3 Open Provenance Architecture

Zwischen 2004 und 2006 wurde von der EU in ihrem sechsten Rahmenprogramm das „*EU Provenance Project*“ gefördert. ([EU Grid Provenance Project, 2005]: http://www.gridprovenance.org/openSpecification/) Ziel des Provenance Project was es, im Grid-Kontext die Dokumentation von Provenienz zu ermöglichen und zu unterstützen: „Das übergreifende Ziel des Provenance Project ist es, eine offene Provenienz-Architektur für Grid-Systeme zu entwerfen, zu entwickeln und zu implementieren, die höchsten Ansprüchen gewachsen ist. . . “ Im Rahmen des EU Provenance Project wurde eine Serie von Dokumenten entwickelt, die zusammen einen Entwurf für einen standardisierten Umgang mit Provenienz bilden.

Einen Überblick darüber erhält man auf http://www.gridprovenance.org/biblioPages/index.html#openspec. Die technischen Reports sind:

- The Provenance Standardisation Vision (http://eprints.ecs.soton.ac.uk/13201/).
- Data Model for Process Documentation (http://eprints.ecs.soton.ac.uk/13200/).
- Process Documentation Recording Protocol (http://eprints.ecs.soton.ac.uk/13053/).
- Provenance Query Protocol (http://eprints.ecs.soton.ac.uk/13204/).
- Process Documentation Query Protocol (http://eprints.ecs.soton.ac.uk/13052/).
- A Profile for Non-Repudiable Process Documentation (http://eprints.ecs.soton.ac.uk/13054/).
- A WS-Addressing Profile for Distributed Process Documentation (http://eprints.ecs.soton.ac.uk/13057/).
- Basic Transformation Profile for Documentation Style (http://eprints.ecs.soton.ac.uk/13202/).
- XPath Profile for the Provenance Query Protocol (http://eprints.ecs.soton.ac.uk/13203/).
- A SOAP Binding For Process Documentation (http://eprints.ecs.soton.ac.uk/13056/).
- WS Provenance Glossary (http://eprints.ecs.soton.ac.uk/13048/).

In einem eigenen Beitrag „An Architecture for Provenance Systems“ ([Groth et al., 2006]) werden die Anforderungen an ein Provenance-System noch einmal zusammengefasst, hier laufen viele Vorarbeiten aus dem Provenance- und dem PASOA[3]-Projekt zusammen. Auf diesen Artikel beziehen sich die bei den Zitaten gegebenen Seitenzahlen.

[3] siehe http://www.pasoa.org/

Das Projekt entwickelt an prototypischen Anwendungsbeispielen einen Rahmen für die Erstellung und Verwaltung von Herkunftsinformationen. Wesentliche Erkenntnisse daraus sind in das „Open Provenance Model“ (siehe 7) eingeflossen.

3.1 User Requirements Document

([Árpád Andics (editor), 2005a])
(Hierbei wurden die Empfehlungen der ESA für die Phase der Erhebung der Nutzungsanforderungen „ESA Software Engineering Standard PSS-05-02“ berücksicht.)

Grundlegende Definitionen:

- Actor: Eine Person oder Organisation die mit der Bearbeitung der Daten verbunden ist.
- Provenance: Die Provenienz von Daten besteht in der Dokumentation der Prozesse, die diese Daten erzeugt haben.
- Workflow: Der Prozess, der die Ausführung einer Reihe von Aufgaben in einer spezifischen Reihenfolge veranlasst, inklusive der Beschreibung wie die Ergebnisse einzelner Aufgabe an andere Aufgaben weitergeleitet werden, falls solche Aktionen erforderlich sind.
- Workflow enactment engine: Ein Programm, das die Ausführung eines Arbeitsablaufs (Workflow) in Übereinstimmung mit der Spezifikation des Workflow gewährleistet. In verteilten Umgebungen wird dazu normalerweise ein Dienst eingesetzt, der andere Dienste benutzt bzw. koordiniert, um den von einem Client an das System eingereichten Arbeitsablauf auszuführen.

Das Dokument bezieht sich im wesentlichen auf die folgenden Referenzen:

1. Benutztes Dokument: „ESA Software Engineering Standards“ by C. Mazza, J. Fairclough, B. Melton, D. de Pablo, A. Scheffer, R. Stevens. Published by Prentice Hall 1994.
2. Referenzen: [FKT01] Ian Foster, Carl Kesselman, and Steve Tuecke. The Anatomy of the Grid. Enabling Scalable Virtual Organizations. International Journal of Supercomputer Applications, 2001. [FKNT02] Ian Foster, Carl Kesselman, Jeffrey M. Nick, and Steven Tuecke. The Physiology of the Grid -An Open Grid Services Architecture for Distributed Systems Integration. Technical report, Argonne National Laboratory, 2002. [PASOA] Miles, S., Groth, P., Branco, M. and Moreau, L. The requirements of recording and using provenance in e-Science experiments. Technical Report, Electronics and Computer Science, University of Southampton, 2005. http://eprints.ecs.soton.ac.uk/10269/

Im Projekt werden zwei „Demo applications“ bearbeitet:

1. Luftfahrttechnik (das *TENT*-System): Großangelegte Flugsimulation
2. Verwaltung von Organtransplantationen: Abgleich von Organspendern und -empfängern

Darüber hinaus werden die folgenden Anwendungsszenarien beachtet:

1. eDiamond: ein Forschungsprojekt zu Aufbau einer nationalen Datenbank von Mammografie-Bildern (http://www.ediamond.ox.ac.uk)

2. Healthcare and Life Sciences Framework: System zur Kontrolle der Anwendung der angemessenen Verfahren und zur Entwicklung von „best practises“
3. Combechem: Grid-basierte kombinatorische Chemie, Entwicklung eines e-Lab (http://www.combechem.org)
4. myGrid: ein umfassender Satz von lose gekoppelten Middleware-Komponenten für datenintensive virtuelle biologische Experimente (http://www.mygrid.org.uk)
5. GENSS (Grid-Enabled Numerical and Symbolic Services): Kombination von Grid Computing mit mathematischen Webdiensten (http://genss.cs.bath.ac.uk)
6. Traffic management: Nutzung einer Grid-basierten Infrastruktur zur Lösung von Verkehrsproblemen (http://www.kwfgrid.net)
7. DataMiningGrid: Entwicklung allgemeiner fachunabhängiger Werkzeuge und Dienste für Data-Mining im Grid (http://www.datamininggrid.org)
8. DILIGENT (A DIgital Library Infrastructure on Grid ENabled Technology): sichere, koordinierte, dynamische und kostengünstige Zusammenarbeit und Nutzung gemeinsamen Wissens (http://diligentproject.org)

Es ergeben sich wesentliche allgemeine Anforderungen, von denen eine – etwas willkürliche, weil im Dokument nicht gegebene – Auswahl hier aufgelistet wird:

3.1.1 Abstrakte Anforderungen

- verfolgen wo Daten herkamen und welche Prozesse sie durchlaufen haben
- Auszug einer gesamten Fallspur: sammele alle Datensätze, die sich auf ein Ereignis beziehen, in einer einzelnen Falldatei
- Befolgung vorgegebener Regeln
- Verlässliche Aufzeichnungen über legitimen oder unzulässigen Zugriff
- Beweis, dass ein durchgeführtes Experiment den geforderten Standards entsprach, korrekt durchgeführt wurde und die benutzten Diensten richtig benutzt wurden und sie so funktionierten wie sie sollten

3.1.2 Technische Anforderungen

Automatische Log-Erstellung:

- Aufzeichnung aller Dienstaufrufe und deren Ergebnisse mit Zeit und Nutzerdaten
- Aufzeichnung von Versionsinformationen und der benutzen Eingabedateien
- Aufzeichnung zurückgewiesener Jobs

Weitere Anforderungen zum möglichen Umgang mit Provenienzdaten wurden formuliert:

- Export- und API-Format für Provenienzdaten: XML-basiert mit der Möglichkeit zur Weiterverarbeitung

- Speicherung und Export von Provenienzdaten: mehrfache Kopien und verschiedene Sichten, Austausch und Migration, Langzeitarchivierung
- Nutzung von Provenienzdaten: Abfrage, Abruf und Speicherung der Analysen
- Arbeitsweise der Provenienzarchitektur: automatisch und regelbasiert
- Schnittstellen: umfangreicher Satz an allgemeinen APIs, Merhsprachigkeit, Programmschnittstelle zur Systemverwaltung (die Forderung nach einer graphischen Nutzerschnittstelle wurde abgelehnt)
- Systemdokumentation: Verschiedene Niveaus der Dokumentation: API, Verwaltungsschnittstelle, Mensch-Computer-Schnittstellen

3.1.3 Beschränkungen

Außerdem wurden zu beachtende Rahmenbedingungen für den Betrieb des Provenienzsystems formuliert:

- Performanz: keine Behinderung bei der Bedienung, minimale Einbußen bei der Performanz, mäßige Speicheransprüche
- Qualität der Dienste: generelle Zuverlässigkeit
- Rechtliche und ethische Fragen: Befolgung von Gesetzen und Regeln, Schutz des geistigen Eigentums
- Sicherheitsaspekte: Zugangskontrolle, einige un veränderbare Daten

Weitere Bedingungen: Kosteneffizienz und Stabilität, Fähigkeit zur Verarbeitung großer Mengen von Provenienzdaten, keine Engpässe, Anwendungsunabhängigkeit

3.2 Software-Anforderungen

Hier wird im wesentlichen eine Auswertung der Nutzeranforderungen im Hinblick auf die notwendige Softwareentwicklung gegeben und eine Abbildung von Nutzer zu Software-Anforderungen und umgekehrt erstellt. Einige Nutzeranforderungen werden zurückgewiesen. ([Árpád Andics (editor), 2005b])
Außerdem gibt es hier einige Definitionen und Bemerkungen von allgemeiner Bedeutung.

Provenance record: Provenienzdaten, die an die Schnittstelle des Provenienzsystems zur Aufzeichnung weitergereicht werden

„Das Provenienzsystem dient der Dokumentation, Analyse und dem Nachweis, wie Daten in einer Anwendungen produziert wurden. Das Provenienzsystem unterscheidet sich von einem Logsystem, da es weitere und andere Funktionen bietet. Es muss nicht alle Informationen, nicht die direkt die Anwendung oder nicht einmal die den Betrieb der Anwendung betreffenden Informationen aufzeichnen. Im Allgemeinen unterscheidet sich die Information, die in

ein Provenienzsystem eingegeben oder von ihm zurückgegeben wird, von einer Logdatei. Das Provenienzsystem ist auch kein Überwachungssystem, da es keine Echtzeitinformationen bereitstellt, stattdessen unterstützt es die Nachuntersuchung. Performanzbezogene Information betrifft das Provenienzsystem nur, wenn sie für die Erzeugung von Daten Bedeutung hat. Das Provenienzsystem liefert Informationen, die aus der Dokumentation der Erzeugung von Daten abgeleitet werden." ([GridProv], S. 9)

„Das Provenienzsystem unterstützt die Aufzeichnung von Informationen, die sich auf die Erstellung von Daten innerhalb einer Anwendung beziehen. [. . .] Das Provenienzsystem bietet einige Funktionen für diese Aufzeichnung, aber die aufzuzeichnenden Informationen müssen von der Anwendung selbst bereitgestellt werden." ([GridProv], S. 12)

„Das Provenienzsystem unterstützt die die Basisfunktionen zur Verwaltung von Provenienzinformationen, aber komplexe und anwendungsspezifische Funktionen zur Verwaltung der Provenienzinformationen müsse für jede Anwendung auf dieser Basis aufgebaut werden." ([GridProv], S. 12)

Aus der Analyse etlicher Nutzungsszenarien ergeben sich zwei wichtige Typen von Provenienzinformationen:

Interaction provenance: eine Aufzeichnung der Wechselwirkung zwischen den Diensten und den Daten, die zwischen ihnen ausgetauscht wurden

Actor provenance: Zusatzinformation eines Dienstes, der an einem Workflow beteiligt war als der Dienst benutzt wurde (Nutzer- und Versionsinformation, Konfigurationsparameter?)

Das Dokument liefert in „3 Specific Requirements" eine Liste der Konsequenzen der allgemeinen Anforderungen und in einer umfangreichen Tabelle „4 User Requirements vs Software Requirements Traceability matrix" eine Übersicht über das Verhältnis von Nutzerzu Software-Anforderungen.

4 Usage Record – Format Recommendation

Das *Open Grid Forum* ([Open Grid Forum], http://www.ogf.org/) ist eine Gruppe von Grid-Nutzern und Entwicklern, die sich um die Standardisierung von Grid-Anwendungen bemühen. In Bezug auf die Erzeugung und Verwaltung von Provenienzinformationen sind ihre Empfehlungen „Usage Record – Format Recommendation“ ([Open Grid Forum: Usage Group, 2006]: http://www.ogf.org/documents/GFD.98.pdf) von Interesse.

4.1 Übersicht

Das in diesem recht umfangreichen Artikel beschriebene Datenmodell des „Open Grid Forum“ (OGF) kann als Teil eines Datenmodells zur Beschreibung von Herkunftsinformationen nützlich sein, da sich ein Großteil der Herkunftsinformationen auf „accounting and usage data“ bezieht.
Es wird ein umfassendes Scheman vorgeschlagen, mit dem die wesentlichen Informationen von Grid-Prozessen beschrieben und ausgetauscht werden können.

Aus dem „Abstract“ („accounting“ kann Buchhaltung, Abrechnung, Kontrolle, Rechnungsführung, Rechnungslegung bedeuten und ist somit ein sehr umfassender Begriff) (S. 1): „For resources to be shared, sites must be able to exchange basic accounting and usage data in a common format. This format must encompass both job level accounting and aggregate accounting. This document describes a common format with which to exchange basic accounting and usage data over a grid instantiation. This record format is intended to facilitate the sharing of usage information among grid sites, particularly in the area of job accounting. This document describes the requirements in natural language form for a Usage Record standard. The usage record is then represented in an XML format.
This document does not address how these records should be used, nor does it attempt to dictate the format in which the accounting records are stored at a local site; instead, it defines a common exchange format.“

Die zentralen Grundannahmen des Modell werden wie folgt beschrieben (S. 4):

- The fundamental component of a grid is a resource;
- The fundamental consumer of a resource is a job on that resource;
- Jobs may be batch (i.e. queued) or interactive.

Die Formatbeschreibung unterscheidet vier Eigenschaften des Datensatzes: Base Properties, Differentiated Properties, Extensions, Aggregation. Für unseren Zusammenhang sind die „Base Properties“ von Interesse: Basiseigenschaften („Base properties“) sind jene Punkte, die alle oder die meisten Sites für entscheidend halten, um die Nutzung ihrer Ressourcen genau aufzuzeichnen. Diese umfassen die Identifikation von Job und Nutzer wie auch die meisten der üblichen Ressourcenarten, die von den Sites gemessen werden müssen. (S.5)

Zur Beschreibung wird eine Liste dieser Basiseigenschaften vorgelegt (S. 7):

1. RecordIdentity: ein eindeutiger Identifikator des Datensatzes

2. GlobalJobId: der vom Scheduler vergebene eindeutige globale Jobidentifikator
3. LocalJobId: der lokale Jobidentifikator, wie er von der Stapelverabeitung oder dem Skript vergeben wird
4. ProcessId: der Prozess-ID des Jobs
5. LocalUserId: der lokale Identifikator des Auftraggebers
6. GlobalUsername: der globale Identifikator des Nutzers
7. JobName: der name des Jobs oder des Programms
8. Charge: die Gesamtgebühr des Jobs in den Zuweisungseinheiten des Systems
9. Status: der Gesamtstatus des Jobs
10. WallDuration: die Zeit, die die Ausführung des Jobs gedauert hat („Wall clock time“)
11. CpuDuration: die von dem Jobs insgesamt beanspruchte CPU-Zeit
12. EndTime: der Zeitpunkt, zu dem der Job beendet wurde
13. StartTime: der Zeitpunkt, zu dem der Job gestartet wurde
14. MachineName: eine beschreibender Name der Maschine, auf der der Job lief
15. Host: der Name des Systems auf dem der Job lief
16. SubmitHost: der Name des Systems von dem aus dem der Job in Auftrag gegeben wurde
17. Queue: der name der Prozessreihe („queue“), von der der Jobs ausgeführt oder aufgegeben wurde
18. ProjectName: das Projekt, für das die Ressourcen aufgewandt wurden

(Es sieht aus als sei $WallDuration = EndTime - StartTime$ und damit redundant, auch scheint der Unterschied zwischen „time“ und „Wall clock time“ nur im Jargon zu bestehen.)

Außerdem gibt es zusätzlich die Möglichkeit, weitere spezielle Information aufzunehmen, den Inhalt von Datenfeldern genauer zu spezifizieren, das Schema zu erweitern und die Daten zusammenzufassen.

Differentiated Properties: Zusätzliche Eigenschaften, deren Angabe von der Grid-Organisation nicht verlangt wird und die in einem validen Schema nicht notwendigerweise angegeben werden müssen. Sie können nach verschiedenen Ressourcentypen getrennt werden (Als Beispiele werden erwähnt: Network, Disk, Memory, Swap, NodeCount, Processors, TimeDuration, TimeInstant, ServiceLevel, Extension)

Meta Properties: Metaeigenschaften sind mit den einzelnen Basiseigenschaften verbunden und liefern zusätzliche Informationen oder Bedeutungszusammenhänge. Die folgenden Metaeigenschaften sind häufig anzutreffen und sollten für die entsprechenden Basiseigenschaften unterstützt werden: Beschreibung, Einheitendefinition, Maßeinheit, Phaseneinheit, Maßstab, Zeitstempel.

Extensions: Das Schema erlaubt Site- oder Grid-spezifische Erweiterungen, diese müssen nur nach Form und Inhalt mit allen Grid-Partnern abgestimmt werden.

Aggregation: Die meisten Sites haben angegeben, dass für ihre aktuellen Bedürfnisse die Zusammenfassung auf Job-Niveau ausreicht. Weitergehende Ansprüche können innerhalb des Schemas realisiert werden und erfordern wie die Erweiterungen die Absprache mit den entsprechenden Partnern.

Für das Datenmodell wird ein XML-Schema mit Hinweisen für die Implementierung der verschiedenen Felder und Eigenschaften vorgelegt, das die letzten sieben Seiten des Dokumentes füllt.

5 The Origin of Data

Diese Dissertation von Paul Groth ([Groth, 2007]: The Origin of Data, Verweise beziehen sich auf dieses Dokument) liefert eine umfassende Bearbeitung der Provenance-Problematik und sollte zur Bearbeitung dieser Aufgabe im Grid mit herangezogen werden. Allerdings sind nicht alle Elemente des Systems vollständig beschrieben. Eine Definition des zur Beschreibung der Interaktionen benutzten Vokabulars, inklusive eines XML-Schemas, findet sich in [Munroe et al., 2006]: Data model for Process Documentation. Die Erzeugung der Prozessdokumentation durch die einzelnen Programme wird nicht erläutert.

Diese Dissertation entstand im Zusammenhang mit dem EU-„Provenance Project“ (siehe Abschnitt 3, Seite 12) und dem PASOA-Projekt (http://www.pasoa.org/), der Autor ist auch im Bereich des „Open Provenance Model“ (siehe Abschnitt 7, Seite 34) und in der W3C Provenance Group (siehe Abschnitt 13, Seite 53) aktiv.

Groth drückt seine Herangehensweise so aus (S. i):

> „Daher tritt diese Arbeit für eine neue Herangehensweise ein, nämlich dass durch die unabhängige Erzeugung, skalierbare Aufzeichnung und regeltreue Organisation der Dokumentation der Prozesse eines Systems die Bestimmung der Provenienz von Erzeugnissen verteilter komplexer wissenschaftlicher Systeme ermöglicht wird.“

und verspricht vier Neuerungen:

1. Provenienz ist eine Anfrage an die Dokumentation der vergangenen Prozesse eines Systems
2. Ein offenes, allgemeines, verteiltes regelbasiertes Datenmodell für die Prozessdokumentation
3. Spezielle Speicher – „provenance stores“ – die ein formal festgelegtes Protokoll verwenden
4. Beschreibung der Nutzung der Prozessdokumentation für die Beantwortung von Fragen zur Provenienz digitaler Objekte und der Auswirkungen, die die Aufzeichnung auf die Performanz der Anwendung hat

Wesentlich sind die Trennung von Prozessdokumentation und Herkunftsanalyse, die Entwicklung eines dafür geeigneten allgemeinen Datenmodells („P-Structure“), eines dazu passenden verteilten Speichersystems („Provenance Store“) und eines Protokolls zum Transport und Austausch dieser Daten („P-assertion Recording Protocol“, PReP).

5.1 Provenance

Provenienz wird in dieser Arbeit auf wissenschaftliche Systeme bezogen, die über verschiedene Institutionen verteilt sind und dabei von hunderten von Personen genutzt werden. Dadurch wird es für Wissenschaftler, Rezensenten und das allgemeine Publikum schwierig, die darin

erzeugten Ergebnisse so nachzuvollziehen, dass sie sich von ihrer Korrektheit überzeugen können.

Formal wird einfach gesetzt:

Definition 5.1. (Provenance) Die Provenienz eines Ergebnisses ist der Prozess, der zu diesem Ergebnis geführt hat.

„Um die Provenienz eines Ergebnisses zu verstehen, muss man die verschiedenen Dokumentationen sichten bis man diejenige findet, die am besten den Prozess repräsentiert, der zu dem gegebenen Ergebnis geführt hat.
Daher ist Provenienz eine **Abfrage**, die durch Suche in der Dokumentation beantwortet wird. Das Problem in verteilten wissenschaftlichen Systemen ist, dass die Dokumentation über verschiedenen Orte verteilt ist, in verschiedenen Formaten vorliegt und nicht einfach befragt werden kann. Darüber hinaus geht Dokumentation oft verloren, wird gelöscht oder verändert, so dass sie nicht mehr als zutreffender Beleg der Provenienz gefunden oder benutzt werden kann. Daher ist die Bestimmung der Provenienz eines durch solche Systeme erzeugten Ergebnis schwierig. Dieses Problem bezeichnen wir als Proveninenzproblem.
Um es zu lösen, schlagen wir vor dass die Dokumentation aller Prozesse innerhalb eines Systems einem gemeinsamen Modell folgen sollte.“ (S. 6)

Groth identifiziert sechs Charakteristika, die diese Dokumentation besitzen sollte: sachlich, zurechenbar, unabhängig zu erzeugen, prozessorientiert, unveränderlich, finalisierbar.

„Wenn die Dokumentation erst einmal nach dem gegebenen Modell erstellt wurde, dann kann sie in spezielle Repositorien, „provenance stores“ genannt, gespeichert werden, die die Verantwortung für ihre Aufrechterhaltung und Bewahrung über das Bestehen der sie erzeugenden Komponenten hinaus übernehmen. Wir bezeichnen eine Anwendung, die Prozessdokumentation erzeugt und aufzeichnet als provenienzbewusste („provenance-aware“) Anwendung. Nachdem die Prozessdokumentation in diesen Repositorien gesammelt wurde, kann sie befragt werden um die Dokumentation zu finden, die die Provenienz eines speziellen digitalen Objektes repräsentiert. Damit kann die Provenienz eines digitalen Objektes durch eine Abfrage der Prozessdokumentation bestimmt werden.“ (S. 7)

Das Hauptresultat der Dissertation wird wie folgt formuliert:
„Die unabhängige Erzeugung, skalierbare Aufzeichnung und regeltreue Organisation der Dokumentation der Prozesse verteilter wissenschaftlicher Systeme ermöglicht die Lösung des Problems der Bestimmung der Provenienz von Ergebnissen dieser Systeme.“ (S.7)

Insgesamt verspricht der Autor eine vollständige, umfassende und realistische Lösung des „provenance problem“, die auch schon unter realen Bedingungen getestet wurde.

Der Autor führt „virtual organisation“ (VO) als Modell für die Kooperation dieser Systeme ein, wobei VOs gegründet werden, arbeiten und nach Beendigung der Arbeit aufgelöste werden.

Verteilte wissenschaftliche Systeme bieten derzeit aber noch keine Lösung des Provenienzproblems. Sie nutzen typischerweise Grid-Technologie und kommunizieren über Web Services. Damit stehen auch die Vielzahl Internet-gestützter Ressourcen innerhalb des Grids

zur Verfügung, insbesondere im Kontext der zunehmend genutzten dienstorientierten Architektur („Service Oriented Architecture“, SOA) für verteilte Systeme. Verschiedene Dienste werden dann gerne in einem Arbeitsablauf („Workflow“) zusammengebunden, der wiederum von geeigneter Software („workflow enactment engine“) abgearbeitet wird. Derzeit ergibt sich daraus aber noch adäquate Beschreibung der Provenienz von deren Resultaten, damit ergeben sich Probleme in der Reproduzierbarkeit wissenschaftlicher Analysen oder Ergebnisse.

In der Dissertation werden zwei Aspekte von Prozessen unterschieden, die sich auf die Vergangenheit („past“) bzw. die Zukunft („prospective“) beziehen:

> In the context of multi-institutional scientific systems, prospective processes are denoted by workflows, programs for services, and policy statements. Fundamentally, these define what could happen in the system. In contrast, a past process is the execution of a workflow, program, or policy. This execution is what has happened in the system. Past processes have information that a workflow or plan cannot contain.
> Therefore, when answering questions related to past processes, workflows provide inadequate information and may result in possibly inaccurate answers. For example, a workflow may show that a particular service was called when, in fact, an error occurred and the service was never contacted. (S.18)

5.2 Analyse von Provenance-Systemen

Der Autor betrachtet diverse Provenance-Systeme:

- Version Control Systems,
- Application-Specific Systems,
- Operating System Level Provenance Systems,
- Provenance in Database Systems,
- Distributed Debugging, Monitoring and Recovery,
- Workflow-centric Systems,

die sich jedoch auf die eine oder andere Weise als nicht hinreichend für eine allgemeine Lösung des „provenance problem“ erweisen. Zentral ist dabei die Frage der verteilten Bearbeitung verteilter Ressourcen: Weder können identische Workflow-Systeme vorausgesetzt werden noch ist klar, wie die verschiedenen Abläufe in einem Provenienz-System zusammengefasst werden können. Die Konsequenz ist die Forderung nach einem einheitlichen Datenmodell.

In einer weiteren Analyse betrachtet er verschiedene Aspekte von Provenance-Systemen:

1. **Abstraktionsniveau:** Er lehnt den Begriff der „Granularität“ als unpräzise ab und betrachtet stattdessen die Abstraktionsniveaus von Software (Schachtelung von Objekten oder Funktionen), Daten und Begriffen. Die Vorstellung ist dann, dass eine Herkunftsrecherche auf einem hohen Abstraktionsniveau ansetzt und dann gegebenenfalls tiefer gebohrt („Drill down“) wird.

2. **Herkunftsrecherche:** Der Autor gibt die folgenden Fragen als Beispiele
   - What were the inputs to this experiment?
   - When did the experiment run?
   - Where did the experimental data come from?
   - How fast did the experiment execute?
   - Which data sources were accessed while running the experiment?
   - Why did this part of the experiment fail?

   [. . .] A common thread to all [considered] systems is that, to answer questions related to the provenance of data, they rely on the equivalent of a dependency graph between data or events, which is then traversed to obtain an answer. [. . .]
   Hence, the framework through which we view provenance queries is three steps.
   (a) Identify the data to find the provenance of.
   (b) Extract the causality graph representing the provenance from documentation of process.
   (c) Traverse the causality graph to answer a specific question.

   (S. 33f)
3. **Kausalität:** Kausalbeziehungen erlauben die Erstellung des Kausalitätsgraphen. In einem verteilten System wird als „Beobachtung durch Teilnahme“ bezeichnet, wenn eine Komponente Daten oder Ereignisse aufzeichnen kann, während sie die Daten verarbeitet oder Ereignisse erzeugt.

Damit wird der Begriff der Kausalität zentral, diese muss vom System selbst in der Verarbeitung von Daten beobachtet werden, die Beziehungen werden dann als Kausalitätsgraph gespeichert.

Der Autor fasst dies in sechs Konsequenzen zusammen (S. 11):

1. The Service Oriented Architecture style is the primary software engineering approach to designing multi-institutional applications.
2. Provenance systems should take into account the important distinction between past processes and prospective processes.
3. A data model for provenance should be well-defined and independent from any one execution environment to cater for multiple platforms, programs and domains.
4. Multiple levels of abstraction must be supported by a provenance system to satisfy a range of queries.
5. The storage of provenance information should be separated from its collection point to ease management and query processing.
6. Causal dependency tracking is critical for understanding the provenance of data.

5.3 Prozess-Dokumentation

Der Autor beschreibt ein komplexes allgemeines „Model of Process Documentation“, das den von ihm aufgestellten Ansprüchen entspricht, hier aber nicht umfassend dargestellt werden kann.

> [...] the provenance of a particular digital object may include processes that occurred at different sites, at different institutions, and at different times. Because these processes may be different in terms of domain focus, underlying assumptions, and implementation technology, it is helpful to have a generic data model for their documentation so that the provenance of results can be traced back through these various interconnected processes. (S. 42)
>
> The perspective we take is to view applications as composed of entities, called actors, each of which represents a set of functionality within the application. Actors interact with other actors by the sending and receiving of messages through well-defined points of communication. (S. 46)
>
> Actors can also play different roles. An actor may have the role of a message sender in one interaction and may play the role of a message receiver in another. An interaction can have metadata associated with it. (S. 49)

Das Model wird mittels „concept maps“ dargestellt, als Beispiel kann die Concept Map angegeben werden, die den einzelnen Prozess beschreibt (siehe Abb. 2) (S. 47).

Die Prozessdokumentation wird aus Aussagen über die einzelnen Wechselwirkungen („interactions“) aufgebaut, die von den Prozessen aufgezeichnet werden (S. 50):

> Interactions are represented by **interaction p-assertions**, which contain four parts:
>
> 1. An asserter identity.
> 2. An event identifier.
> 3. A representation of the message exchanged in the interaction.
> 4. A documentation style describing how the representation was generated.

Darüber hinaus gibt es „relationship p-assertions“, die interne Zusammenhänge zwischen Vorkommnissen beschreiben, die als Ereignisse oder mit Ereignissen verbundene Daten gefasst werden, diese können *strukturelle Beziehungen* (z.B. Kompositionen) oder *Transformationen* beschreiben.

Zusammen liefern die „P-Aussagen“ über Wechselwirkungen und Beziehungen die Informationen, die zur Dokumentation der Prozesse benötigt werden; sie ermöglichen diese Dokumentation auch auf verschiedenen Abstraktionsniveaus.

Zur Vereinfachung werden schließlich interne Informations-P-Aussagen („Internal Information P-assertions“) eingeführt:

> It is often the case that a piece of data plays an important role in a process but the manner of its generation is not of interest. Examples of this include the time, the memory usage of an actor, and the configuration of an actor. All of these data items can be represented using relationship p-assertions and interaction p-assertions. However, using **internal information p-assertions**, the detail of how these data items were obtained can be abstracted away and one is left with just the data item and its basic causal connection to the process. (S. 55)

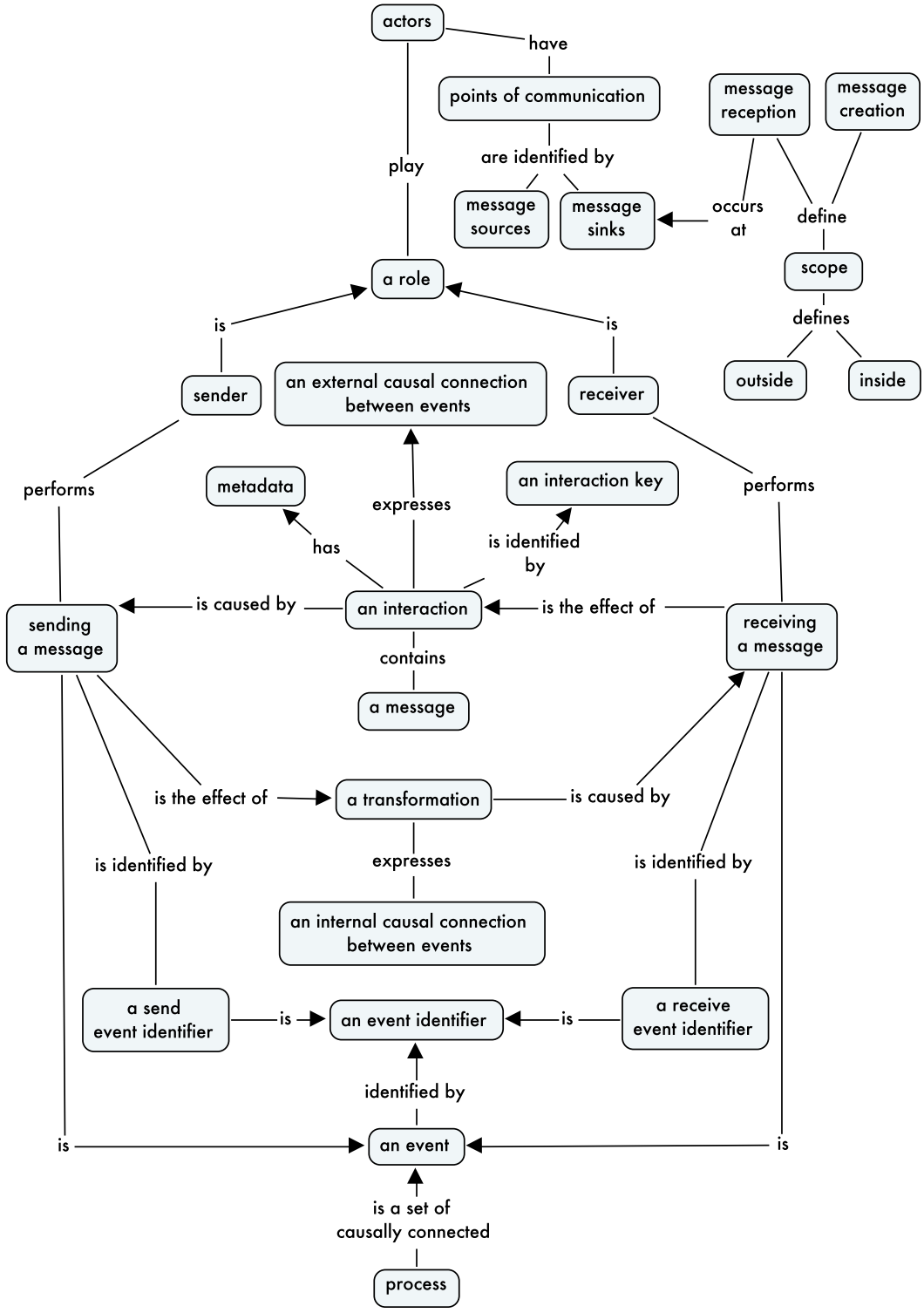

Abb. 2: Concept map describing process

Diese Informationen werden in einer P-Struktur („p-structure“) zusammengefasst:
Die P-Struktur ist eine Organisation von P-Aussagen, die es ermöglicht, Sätze von P-Aussagen zu isolieren, zu finden und zu verstehen. P-Aussagen werden entlang zugehöriger Ereignisse in einzelnen Sichten zusammengefasst.

Damit dies funktionieren kann, müssen die Akteure, die die P-Aussagen aufzeichnen, gewissen Regeln gehorchen, die sicherstellen, dass die Ereignisse korrekt identifiziert und zugeordnet werden können, dazu werden eindeutige „interaction keys“ vergeben und benutzt.

Aus der P-Struktur lassen sich dann die Herkunftsinformationen extrahieren. Schematisch wird das in Abb. 3 dargestellt.

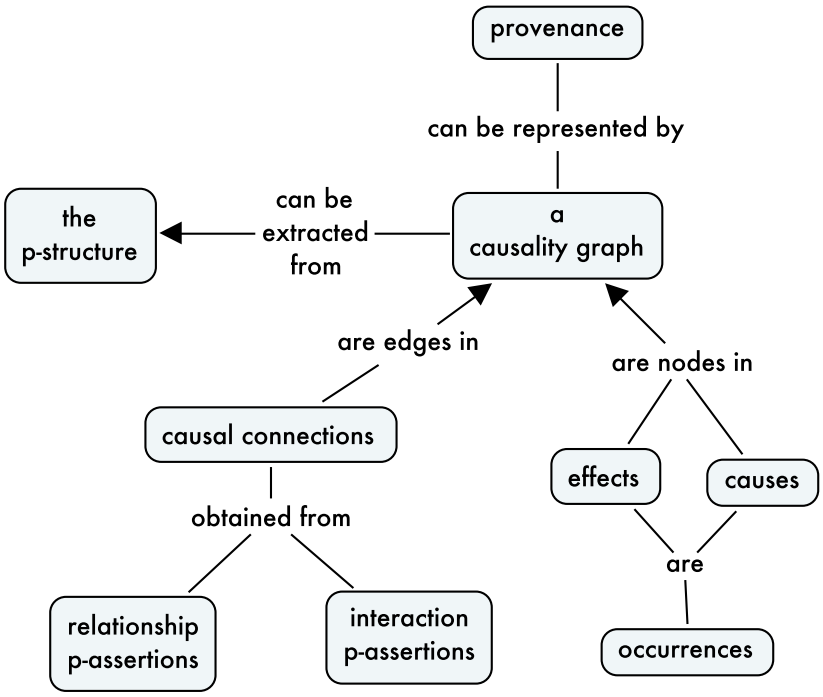

Abb. 3: Concept map describing provenance

5.4 Recording Process Documentation

Die vorherigen Abschnitte beschreiben die unabhängige Erzeugung von Prozessdokumentation durch einzelne Akteure. Hier wird jetzt eine Lösung vorgeschlagen, wie diese Informationen zusammengeführt und dauerhaft gespeichert werden können (S. 68):

> The solution presented consists of four main contributions:
>
> 1. An architectural element for the storage of process documentation known as the provenance store.
> 2. Patterns for the appropriate deployment of provenance stores within applications.

3. A technique for connecting process documentation stored in a distributed manner.
4. A formally defined protocol that actors use in order to record process documentation into provenance stores. The protocol is asynchronous, stateless, and helps ensure that high quality process documentation is recorded.

Der „Provenance Store“ wird als verteiltes System von Graphen-Datenbanken angelegt, für die einerseits Regeln für die Verteilung von Daten und Speichern und andererseits eine geeignetes Protokoll („P-assertion Recording Protocol“, PReP) eingeführt werden, die zusammen die Kommunikation zwischen den Informationsknoten und die sicher Speicherung gewährleisten. Das Protokoll wird vorgestellt, und etliche seiner Eigenschaften werden formal bewiesen. Insbesondere wird gezeigt, dass das Protokoll die von der Prozessdokumentation geforderten Eigenschaften besitzt. (Die Argumentation wurde allerdings nicht überprüft.) Implementierungen des Systems sind in Java und Python realisiert worden.

5.5 Evaluation

Als Fallbeispiel wird ein System zur Analyse von biologischen Daten (Amino Acid Compressibility Experiment, ACE) betrachtet, das auf einem Cluster (Iridis Computing Cluster, University of Southampton) mit Globus-Oberfläche und Condor-G läuft. Sechs Fragen werden als „Use Cases“ für die Beantwortung von Herkunftsfragen formuliert.

Das Gesamtkonzept wird in dieser realen Umgebung getestet. Dazu wird ein entsprechendes Client-Server-System für das P-assertion Recording Protocol aufgebaut und mit den die P-Aussagen erzeugenden Systemen verbunden.
Die genaue Erzeugung und Verarbeitung der Herkunftsdaten wird nicht beschrieben.

Das Ergebnis wird wie folgt zusammengefast:

> First, we analyse the scalability of PReServ in a controlled environment and demonstrate that it can handle up to 560 simultaneous connections recording p-assertions without reaching a plateau in throughput. Second, we analyse the performance impact on the execution time of ACE and find that their is a 13% overhead for p-assertion recording. Third, we demonstrate that the six provenance use cases can be answered practically by queries over process documentation. (S.131)
>
> PReServ achieved a throughput of 234,025 10k p-assertions in 10 minutes or, from a different perspective, 2.2 GB of p-assertions were created and recorded in a 10 minute period. Thus, every 2.3 milliseconds a p-assertion was recorded. (S. 162)
>
> It is still an open question as to the appropriate mechanisms necessary to *automatically integrate* with applications. (S. 170)
>
> This dissertation has shown that the problem of determining the provenance of results produced by complex multi-institutional scientific systems can be solved through the autonomous creation, scalable recording, and principled organisation

> of documentation of these systems' processes. This dissertation is a small building block towards the grand vision of being able to know the provenance of everything.[4] (S. 171f)

[4] Das bezieht sich auf Herkunftsfragen von Gebrauchsgegenständen (Wo wurde dieser Computer unter welchen Bedingungen produziert?), Lebensmitteln (Womit wurden die Hühner gefüttert, die diese Eier gelegt haben?) oder anderen Waren (Unter welchen ökologischen und sozialen Bedingungen wurden diese Rosen produziert?). Es gibt schon verschiedene Systeme, die mit der Kodierung einzelner Waren zumindest teilweise Antworten auf solche Fragen erlauben.

6 PREMIS

Vom Projekt PREMIS (Preservation Metadata: Implementation Strategies) von OCLC und RLG wurde 2005 ein „Data Dictionary for Preservation Metadata“ veröffentlicht. Seitdem wird das System von der „PREMIS Maintenance Activity“ bei der *Library of Congress* verwaltet und gepflegt, was in Zusammenarbeit mit internationalen Partnern zu einer Überarbeitung des Verzeichnisses geführt hat: „PREMIS Data Dictionary for Preservation Metadata version 2.0“ ([PREMIS Maintenance Activity (Library of Congress), 2008], Verweise beziehen sich auf dieses Dokument).

Offenbar geht es bei Erhaltungsmetadaten auch um einige Fragen, die für die Provenienz von Bedeutung sind. Außerdem wurde ein XML-Schema zur Umsetzung dieses Datenbeschreibungsverzeichnisses in digitalen Archivierungssystemen bereitgestellt, das auch dem Austausch von Informationen dienen kann.

> The Data Dictionary defines preservation metadata that:
>
> – Supports the viability, renderability, understandability, authenticity, and identity of digital objects in a preservation context;
> – Represents the information most preservation repositories need to know to preserve digital materials over the long-term;
> – Emphasizes „implementable metadata“: rigorously defined, supported by guidelines for creation, management, and use, and oriented toward automated workflows; and
> – Embodies technical neutrality: no assumptions made about preservation technologies, strategies, metadata storage and management, etc.
>
> (S. 1)

Das PREMIS Datenbeschreibungsverzeichnisses bezieht sich auf das OAIS-Modell und baut auf dem „Metadata Framework“ (Siehe Abschnitt 2 , S. 8) auf. Während jenes als Ausarbeitung des OAIS-Informationsmodell gesehene werden kann, in der Erhaltungsmetadaten in das OAIS-Konzept abgebildet werden, liefert das PREMIS Datenbeschreibungsverzeichnis eine Übersetzung davon in einen Satz praktisch umsetzbarer Begriffe.

Das PREMIS-Verzeichnis definiert Erhaltungsmetadaten als die von Repositorien zur Bewahrung digitaler Daten genutzten Informationen. Besonderen Wert wird dabei auf die Provenienz von und die Beziehungen zwischen solchen Objekten gelegt.

Das Verzeichnis soll von konkreten Implementierungen unabhängig sein, dazu wird ein Kern von Metadaten definiert, die jedes Repositorium kennen muss, und zwar unabhängig davon, wie oder gar ob, diese Informationen gespeichert werden. Deswegen wird im Verzeichnis nicht von „Metadaten“, sonden von „Begriffen“ (*semantic units* gesprochen.

Von der PREMIS-Gruppe wurde ein einfaches Modell entwickelt, um dieses Begriffe zu gliedern. Das Modell definiert fünf Einheiten, die für die Bewahrungsaktivitäten als besonders

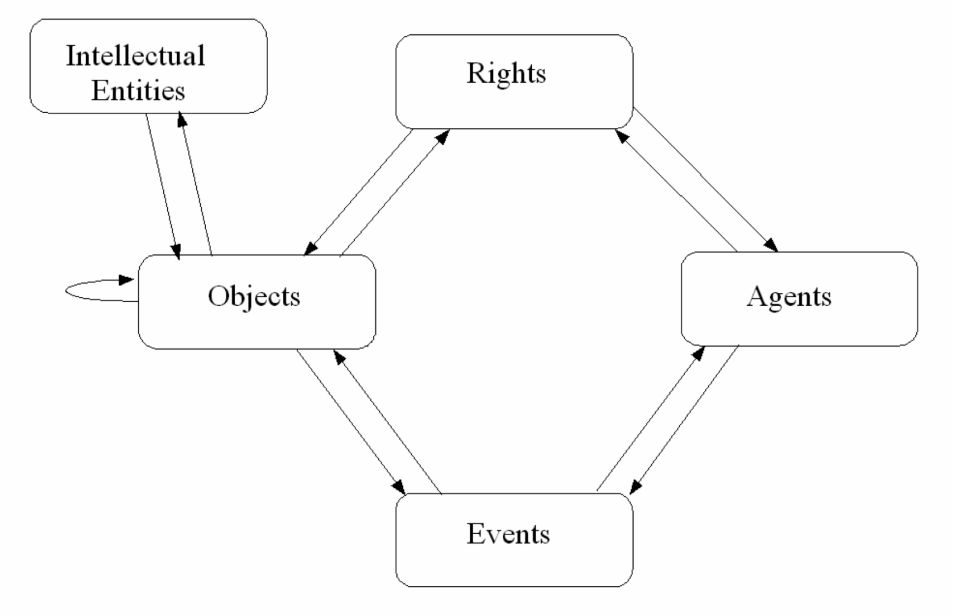

Abb. 4: The PREMIS data model

wichtig angesehen werden: Werke (*Intellectual Entities*), Objekte, Ereignisse, Rechte und Akteure. Jeder Begriff im PREMIS-Verzeichnis ist eine Eigenschaft einer dieser Einheiten (siehe Abb. 4).

> Die Einheiten im PREMIS-Datenmodell werden wie folgt definiert:
>
> **Werk:** ein Objekt das für die Verwaltung und Beschreibung als gedankliche Einheit betrachtet wird, z.B. ein bestimmtes Buch, ein Foto oder eine Datenbank. Werke können andere Werke enthalten, z.B. kann eine Website eine Webseite enthalten, und die Webseite ein Bild. Werke können verschiedene digitale Darstellungen haben.
>
> **(Digitales) Objekt):** eine eigenständige Informationseinheit in digitaler Form.
>
> **Ereignis:** eine Aktion, die Objekte oder Akteure betrifft oder beeinflusst, die mit dem Repositorium verbunden oder ihm bekannt sind.
>
> **Akteur:** Eine Person, Organisation oder Software, die einem Ereignis im Leben eines Objektes oder mit ihm verbundenen Rechten zugeordnet ist.
>
> **Rechte:** die Behauptung von Rechten oder Genehmigungen, die sich auf ein Objekt und/oder einen Akteur beziehen. (S. 6)

Die von PREMIS-Verzeichnis benutzten Begriffe können als Eigenschaften solcher Einheiten angesehen werden.

Auf die verschiedenen Einheiten wird im Verzeichnis näher eingegangen:

> Objekte haben drei Untergruppen: Datei, Datenstrom und Darstellung

Ereignisse haben Ergebnisse, manche auch eine Ausbeute; sie sind typischerweise auf spezielle Weise mit Objekten verbunden

Akteure bilden keinen Schwerpunkt der PREMIS-Verzeichnisses, sie beeinflussen Objekte nur indirekt durch Aktionen

Rechte werden in Version 2 des PREMIS-Verzeichnisses stärker betont und in die Bereichen Urheberrecht, Lizenzen und Gesetzen eingeteilt. Damit wurde auch ein Begriff eingeführt, der die Rechte des Repositoriums zur Bearbeitung von Werken fassen soll.

Die Beziehungen zwischen Objekten werden von PREMIS auf drei Basistypen zurückgeführt:

Strukturelle Beziehungen bestehen zwischen Teilen von Objekten, die erst die Darstellung des Objektes aus seinen Bestandteilen ermöglichen,

Ableitungen ergeben sich aus Reproduktionen oder Umwandlungen von Objekten, die deren Gehalt aber unverändert lassen,

Abhängigkeiten bestehen, wenn ein Objekt andere Objekte für sein Funktionieren benötigt.

Im PREMIS-Konzept werden Objekte als unveränderbar angesehen, damit beziehen sich Metadaten (getreu dem 1:1-Prinzip) immer auf genau ein Objekt. Daher können mit den PREMIS-Begriffen auch keine Veränderungen von Objekten, sondern nur die Erzeugung neuer Objekte als Ableitungen bestehender Objekte beschreiben werden. Die Erzeugung ist dann Resultat eines Ereignisses.

Für die Realisierung des PREMIS-Systems in einem Repositorium werden keine Vorgaben gemacht, es wird aber davon ausgegangen, dass die Basiseinheiten auf irgendeine Weise behandelt werden müssen. Darauf aufbauend sollte ein eigenes Datenmodell definiert werden, das die spezifischen Bedürfnisse des gegebenen Repositoriums berücksichtigt.

Da Repositorien im allgemeinen große Datenmengen verarbeiten müssen, erscheint die automatische Erzeugung von Metadaten wünschenswert, wenn immer dies möglich ist; viele Bergriffe können aus eingelieferten Dateien durch Programme ausgelesen oder bei Aufnahme ans Konstanten gesetzt werden. Falls menschliche Intervention notwendig ist, so wird auf Paare von Begriffen gesetzt, die jeweils aus einer Notation und einer verständlichen Beschreibung bestehen und bei Eingabe möglichst verifiziert werden sollten. Außerdem werden für einige Begriffe die Beschränkung auf kontrollierte Listen vorgeschlagen, um die automatisierte Verarbeitung u erleichtern.

Grundsätzlich erlaubt das PREMIS-Modell individuelle Erweiterungen, mit der Version 2.0 des Verzeichnisses wurde dazu auch ein formaler Mechanismus für die Erweiterung einiger Begriffe eingeführt, für die dies besonders wünschenswert erschien.

Das „PREMIS Data Dictionary“ in der Version 2.0 wird als eigenes Dokument bereitgestellt (http://www.loc.gov/standards/premis/v2/premis-dd-2-0.pdf, 1,1 MB, 184 Seiten) und umfasst das umfangreiches Vokabular zur Beschreibung von Objekten, Ereignissen, Akteuren und Rechte. Für die Nutzung ist die genaue Inspektion der Begriffe notwendig, auf die in dieser kurzen Zusammenfassung nicht eingegangen werden kann.
Im einzelnen gibt es

- 94 Objektbegriffe,
- 19 Ereignisbegriffe,
- 5 Akteurbegriffe,
- 36 Rechtsbegriffe.

6.1 Formate, Umgebungen und Zusammensetzung

Für ein digitales Object ist das *Format* eine zentrale Eigenschaft. Grundsätzliche Fragen dazu sind:

- Was ist ein Format?
- Welche Arten von Objekten haben ein Format?
- Wie identifiziert man ein Format?
- Gibt es einen Unterschied zwischen einem Format und einem Profil?

> Das Konzept von Format erscheint fast intuitiv, wegen der Bedeutung der Formatinforamtion wollte die PREMIS-Gruppe seien Bedeutung aber sehr konkret festlegen. . . . Die Gruppe formulierte ihre eigene Definition: eine spezifische vorgegebene Struktur für die Organisation einer digitalen Datei oder eines Datenstroms.
> . . .
> Ein Repositorium muss die Formatinformationen so genau wie möglich aufzeichnen, idealerweise würde ein Format durch einen direkten Link zu seiner Formatbeschreibung identifiziert. (S. 195)

Derzeit sind die Formatbeschreibungen, z.B. für MIME-Typen und Dateiendungen, ohne Versionsangaben nicht präzise genug. Für die Zukunft sind zentrale Formateregister eine bessere Lösung, aber dann muss auch klar sein, welches Register benutzt wird.

Digitale Materialien benötigen für ihre Darstellung eine komplexe technische *Umgebung*. Wird die Verbindung mit dieser Umgebung gestört, so kann das digitale Material unbrauchbar werden, deren Dokumentation bildet daher einen unverzichtbaren Teil der Bewahrungsmetadaten. Solche Umgebungen bestehen meist aus verschiedenen geschachtelten Komponenten, deren Beschreibung sehr komplex werden kann. Andererseits benötigen oft ganze Klassen von Objekten alle dieselbe Umgebung, so dass diese Verwaltung von Umgebungsmetadaten am besten durch ein Zentralregister geleistet werden kann.

Wird ein Objekt komprimiert oder verschlüsselt, so wird das dann entstandene Format durch diesen Vorgang bestimmt; gleichzeitig hat das Objekt aber noch ein anderes darin eingeschlossenes Format. Damit entstehen komplexe Schichten von Kodierungen, die in der richtigen Reihenfolge wieder rückgängig gemacht werden müssen, um schließlich bei dem korrekten Objekt zu landen. Die PREMIS-Gruppe beschreibt dies mit der *Zwiebel*-Metapher und jede Schicht als Zusammensetzungsstufe, für die jeweils eigene Metadaten erhoben werden.

Ein verwandtes, aber anderes Problem entsteht bei Formaten wie ZIP oder TAR, die mehrere Dateien mit verschiedenen Formaten in einen Behälter packen können; die Inhalte dieser Behälter müssen dann als davon verschiedene eigene Objekte erfasst werden.

6.2 Beständigkeit, Integrität, Echtheit

Bei der Definition der Kernelemente der Bewahrungsmetadaten kommen der Beständigkeit, Integrität und Echtheit der digitalen Objekte zetrale Bedeutung zu, da ohne diese Eigenschaften diese Objete nur einen geringen Wert besäßen, wenn es z.B. auf ihre Beweiskraft ankäme oder sie das kulturelle Erbe bewahren sollen.

Zum Test der *Beständigkeit* (dass ein Objekt sich seit seiner Einlieferung nicht verändert hat) werden im digitalen Kontext Hash-Algorithmen eingesetzt, zur erhöhten Sicherheit können zwei verschiedene eingesetzt und ihre Ergebnisse gespeichert werden.

Die *Integrität* einer Datei wird hingegen durch die Identifikation ihres Formates und dessen Validierung garantiert, dabei müssen eventuell auch komplexe Formate in ihren Einzelbestandteilen abgearbeitet werden.

Die *Echtheit* soll schließlich garantieren, dass das Dokument tatsächlich ist, was es zu sein behauptet. Dazu werden sowohl technische Aspekte (z.B. die detaillierte Beschreibung der digitalen Herkunft), die Bewahrung von unveränderten Originalen als auch digitale Signaturen herangezogen.

6.3 Nicht berücksichtige Metadaten

Die PREMIS-Gruppe hat auch entschieden, einige Metadatenkonzepte nicht in ihre Liste von Kernmetadaten aufzunehmen, die nur in spezifischen Zusammenhängen von Bedeutung sein können. Die aus verschiedenen Gründen ausgeschlossenen Konzepte werden in einer Liste aufgeführt, das PREMIS-Dokument gibt die Gründe für die jeweilige Nichtberücksichtigung an.

7 Open Provenance Model

Wie der Name nahelegt, liefert das „Open Provenance Model“ ([Luc Moreau (Editor), 2009]: http://eprints.ecs.soton.ac.uk/18332/1/opm.pdf ein Modell zur Beschreibung von Herkunftsinformationen. Das Modell liegt hier in Version 1.1 vor, Verweise beziehen sich auf dieses Dokument. In diesem Modell wird die Provenienzinformation durch einen ausgezeichneten gerichteten Graphen abgebildet.

7.1 Anforderungen

> The Open Provenance Model is a model of provenance that is designed to meet the following requirements (S. 3):
>
> 1. To allow provenance information to be exchanged between systems, by means of a compatibility layer based on a shared provenance model.
> 2. To allow developers to build and share tools that operate on such a provenance model.
> 3. To define provenance in a precise, technology-agnostic manner.
> 4. To support a digital representation of provenance for any „thing“, whether produced by computer systems or not.
> 5. To allow multiple levels of description to coexist (added in v.1.1).
> 6. To define a core set of rules that identify the valid inferences that can be made on provenance representation.

Wichtig ist zu beachten, dass das „Open Provenance Model“ gleichzeitig einiges nicht leistet, was man sich unter dem Namen vorstellen könnte (S. 3):

> While specifying this model, we also have some non-requirements:
>
> – It is not the purpose of this document to specify the internal representations that systems have to adopt to store and manipulate provenance internally; systems remain free to choose internal representations that are fit for their purpose.
> – It is not the purpose of this document to define a computer-parsable syntax for this model; realisations of OPM in XML, RDF or others are being specified in separate documents.
> – We do not specify protocols to store such provenance information in provenance repositories.
> – We do not specify protocols to query provenance repositories.

Das OPM erfüllt diese Anforderungen, und im Rahmen des „Third Provenance Challenge“ haben vierzehn Teams gezeigt, dass das OPM den Austausch von Provenienzinformationen ermöglich.

7.2 Kausalitätsgraph

„Wir gehen davon aus, dass die Provenienz von Objekten (digital oder nicht) durch einen ausgezeichneten Kausalitätsgraphen dargestellt wird, das ist eine gerichteter azyklischer (schleifenfreier) Graph, der durch Auszeichnungen (‚annotations') ergänzt wird, die weitere Informationen zur Ausführung enthalten."

Das Modell enthält drei Arten von Knoten:

Definition 7.1. (Objekt („Artifact")) Unveränderliche Zustandsform, die eine physische Verkörperung für ein materielles Objekt oder eine digitale Darstellung in einem Computersystem sein kann.

Definition 7.2. (Prozess) Aktion oder Reihe von Aktionen, die an Objekten oder durch sie verursacht ausgeführt wird und zu neuen Objekten führt

Definition 7.3. (Akteur („Agent")) Kontextbezogenes Element, das als Katalysator eines Prozesses wird, indem es ihn ermöglicht, erleichtert kontrolliert oder seine Ausführung beeinflusst

Damit besteht ist ein Provenienzgraph als gerichteter Graph definiert, dessen Knoten Objekte, Prozess und Akteure sind. Die Kanten des Graphen beschreiben Beziehungen zwischen den Knoten und werden im folgenden näher betrachtet

7.3 Beziehungen

Definition 7.4. (Kausalbeziehung) Eine Kausalbeziehung wird durch eine Kante dargestellt und bezeichnet eine Abhängigkeit zwischen dem Anfang der Kante (der Wirkung) und deren Ende (der Ursache).[5]

In dem Modell werden fünf Kausalbeziehungen benutzt:

1. ein Objekt wird von einem Prozess benutzt (used): der Prozess erfordert die Verfügbarkeit des Objektes; sind mehrere Objekte auf diese Weise mit dem Prozess verbunden, so sind alle diese Objekte erforderlich.
2. ein Objekt wird von einem Prozess erzeugt (was generated by): der Start des Prozesses war ursächlich für die Erzeugung des Objektes[6].
3. ein Prozess wurde durch einen Prozess ausgelöst (was triggered by): der zweite Prozess erfordert für seine Beendigung den Start des ersten Prozesses[7].
4. ein Objekt wurde von einem Objekt abgeleitet (was derived from): das erste Objekt erfordert für seien Erzeugung die Erzeugung des zweiten Objektes.

[5] Dies steht im Widerspruch zu meiner Intuition: der Pfeil zeigt von der Wirkung zur Ursache! Das führt zu einer durchgängig passivischen Formulierung des Modells. (ThF)

[6] Im Original ziemlich kompliziert und unverständlich: „the process was required to initiate its execution for the artifact to have been generated" (S. 7).

[7] Im Original: „An edge „was triggered by" from a process P2 to a process P1 is a causal dependency that indicates that the start of process P1 was required for P2 to be able to complete" (S. 7).

5. und ein Prozess wurde durch einen Akteur kontrolliert (was controlled by): Start und Ende des Prozesses wurden durch den Akteur kontrolliert.

Zur Erhöhung der Übersichtlichkeit werden zusätzlich mehrstufige Beziehungen eingeführt, in denen mehrere Beziehungen, eventuell auch von verschiedenem Typ, zusammengefasst werden; diese können sukzessive auch weiter zusammengefasst werden, so dass „mehrstufige Ableitungen“ und darauf aufbauend auch „sekundäre mehrstufige Kanten“ definiert werden, die dann Benutzungs-, Erzeugungs- oder Auslösungsbeziehungen darstellen.

Wenn verschiedene Knoten mit demselben Objekt verbunden sind, können die entsprechenden Abhängigkeiten für Kanten vom Typ „used“, „was generated by“ oder „was controlled by“ als Rollen aufgefasst werden:

Definition 7.5. (Rolle) Eine Rolle bezeichnet die Funktionen eines Objektes oder eines Akteurs in einem Prozess.

Damit wird schließlich eine „Provenance Graph Definition“ in 19 Punkten gegeben. Definiert werden unter anderem:
Entity, Accounts, Artifacts, Processes, Agents, Edges, Roles, Effective account membership, OPM graph, Account views and their legality, legal OPM graphs, overlapping account views and refinements, Legal Account relation assertions.

7.4 Auszeichnungen

Die Erfahrungen mit der „Third Provenance Challenge“ haben darüber hinaus gezeigt, dass Bedarf an Zusatzinformationen besteht, die diesen Elementen hinzugefügt werden; dafür wird ein Auszeichnungsrahmen definiert. Auszeichnungen erlauben, den Kanten zusätzliche Informationen zuzuordnen (siehe S. 22):

> Der OPM-Auszeichnungsrahmen wird durch die folgenden Regeln definiert:
> 1. Eine OPM-Auszeichung ist eine von den anderen OPM-Elementen verschiedene Klasse von Objekten.
> 2. Mit einer Auszeichnung kann ein OPM-Graph, ein OPM-Knoten, eine OPM-Kante, eine OPM-Beschreibung, eine OPM-Rolle oder eine OPM-Auszeichnung versehen werden.
> 3. Ein ausgezeichnetes Element ist eine auszeichenbares Element mit einer oder mehreren Auszeichnungen.
> 4. Jedes ausgezeichnete Element muss innerhalb eines OPM-Graphen eindeutig durch einen Identifikator identifiziert werden können.
> 5. Eine Auszeichnung ist ein Objekt der Klasse der OPM-Auszeichnungen und besteht aus:
>   - einem Subjekt: ein auszeichenbares Element (durch seinen Identifikator identifiziert) an das die Auszeichnung angeheftet ist;
>   - eine nichtleere Menge von Paaren von Eigenschaften und Werten:
>   - die Eigenschaft gibt einen Namensraum an, der ihren Geltungsbereich beschreibt,

  - der Wert muss einer Typenliste entsprechen;
- eine Liste von Beschreibungen aus den möglichen Beschreibungen der des ausgezeichneten Elementes.

Die beabsichtigte Bedeutung der Eigenschaft-Werte-Paare ist, dass das ausgezeichnete Element (das Subjekt) mit zusätzlichen Beschreibungen versehen wird, deren jede aus einer Eigenschaft des Subjekts und dem Wert für diese Eigenschaft im Kontext einer Beschreibung besteht. Innerhalb einer Auszeichnung sind mehrfache Eigenschaft-Werte-Paare erlaubt. Dieselbe Eigenschaft kann auch mehrfach mit verschiedenen Werten auftreten.

6. Auszeichnungen können selbst wieder ausgezeichnet oder einer Untergruppe zugeordnet werden. (S. 22)

Außerdem wird ein Satz von Eigenschaften definiert, die den Austausch von Provenienzinformationen erleichtern sollen: type, pname, label, value, encoding, profile.

7.5 Profile

Für die individuelle Anpassung an spezifische Bedürfnisse wird noch ein Konzept von Profilen bereitgestellt, in denen „best practices“ und eigene Richtlinien niedergelegt werden können. Dabei wird auch für Vorschläge für die Strukturierung der Provenienzgraphen gemacht.

8 DELOS Digital Library Reference Mode

DELOS (http://www.delos.info/) war eine Exzellenznetz, das bis 2009 im Zuge des siebten Rahmenprogramms von der EU gefördert wurde und dessen Arbeit seitdem von der „DL.org Community“ (http://www.dlorg.eu/)fortgesetzt wird. Im Rahmen von DELOS wurde ein „Digital Library Reference Model“ entwickelt ([DELOS Projekt, 2007]: Version 0.98 von 2007), das von DL.org weiter gepflegt wird ([Athanasopoulos et al., 2010]: Version 1.0 von 2010), siehe auch http://www.dlorg.eu/uploads/Booklets/booklet21x21_nutshell_web.pdf)
Wir orientieren uns hier an der Version 1.0, Verweise beziehen sich auf dieses Dokument.

DELOS behandelt Provenance als „Content Quality Parameter“ (S. 76f):

Definition 1. C164 Content Quality Parameter) Ein Qualitätsparameter, der den Inhalt des Hauptkonzepts betrifft

Die zugehörigen Aspekte werden aufgelistet:

- C165 Authenticity
- C166 Trustworthiness
- C167 Freshness
- C168 Integrity
- C169 Preservation Performance
- C170 Provenance
- C171 Scope
- C172 Size
- C173 Fidelity
- C174 Perceivability
- C175 Viability
- C176 Metadata Evaluation

Der uns interessierende Parameter ist „C170 Provenance“, er wird wie folgt näher spezifiziert und seine Relationen zu anderen Parametern erläutert (S. 152):

Definition 2. (C170 Provenance) Ein Inhaltsqualitätsparameter, der angibt, wie genau der Ursprung und die Geschichte des Informationsobjektes bekannt sind und nachgezeichnet werden können.

In dem DELOS Referenzmodel ist die Provenienz in weitere Relationen eingebunden:

- Provenance <isa> Content Quality Parameter
- Provenance <affectedBy> Metadata
- Provenance <affectedBy> Annotation

- Provenance <affectedBy> Preservation Policy
- Provenance <affectedBy> Information Object

Begründung: Dieser Qualitätsparameter zielt darauf zu bestimmen, in wie weit es möglich ist, die Geschichte und Entwicklung eines Informationsobjektes nachzuvollziehen, um herauszufinden, ob es seinen Zweck erfüllt. Ein Informationsobjekt kann von anderen Informationsobjekten abgeleitet sein (z.B. durch Verschmelzen oder eine Transformation), seine Herkunft zu verfolgen ist nicht immer trivial.
Wenn wir es insbesondere mit wissenschaftlichen Daten zu tun haben, so muss die Herkunft zurückverfolgt werden, da Wissenschaftler wissen müssen, wo die Daten herkamen und welcher Bereinigung, Skalierung oder Modellierung sie unterworfen wurden, bis sie ihren jetzigen Zustand erhielten.

Provenance <affectedBy> Metadata: denn die Metadaten enthalten die zusätzlichen informationen

Provenance <affectedBy> Annotation: denn Annotationen erlauben uns, die Herkunft und den Datenfluss nachzuverfolgen, melden Fehler oder Bemerkungen zu einem Datenobjekt und beschreiben die Qualität und die Sicherheitsstufe eines Datenobjekts [. . .]

Provenance <affectedBy> Preservation Policy: da diese die Art von Metadaten beeinflussen kann, die zu einem Informationsobjekt zugeordnet werden.

Es gibt auch umgekehrte Beziehungen, z.B. (S. 150):

Trustworthiness ¡affectedBy¿ Provenance

Begründung: Vertrauenswürdigkeit betrifft die Verlässlichkeit und Glaubwürdigkeit einer Ressource, was sowohl das Vertrauen des Nutzers in die Ressource selbst betrifft als auch darin, dass dieses Vertrauen nicht missbraucht wird. Es könnte hilfreich sein, digitale Bibliotheken mit ähnlicher oder identischer Aufgabenstellung zu vergleichen, wobei die eine vertrauenswürdiger als die andere ist.

‚Provenance' kann die Vertrauenswürdigkeit beeinflussen, denn Kenntnis der Herkunft und Geschichte einer Ressource können ihre Verlässlichkeit und Glaubwürdigkeit erhöhen.

Insgesamt wird Provenienz in dem „Digital Library Reference Model" von DELOS aber noch zu knapp gefasst:

„Der Begriff von Provenienz
Die aktuelle Version des Referenzmodells erfasst explizit nur den Begriff von Provenienz als Qualitätsparameter. Der Begriff von Provenienz ist weiter. Provenienz, auch als Herkunft bezeichnet, bezieht sich auf die Entwicklungsgeschichte eines Informationsobjektes (einer Ressource) von den Originalquellen an und beschreibt den Prozess, der das Objekt in seinen jetzigen Zustand versetzt hat. In den letzten zehn Jahren wurde die Nachverfolgung der Herkunft entscheidend für die korrekte Auswertung von Daten in einer Vielzahl von Anwendungsbereichen, einschließlich digitaler Bibliotheken." (S. 218)

9 Core Scientific Metadata Model

In diesem Artikel wird eine Metadatenmodell für die wissenschaftliche Forschung beschrieben: „Using a Core Scientific Metadata Model in Large-Scale Facilities“ ([Matthews et al., 2010], Verweise beziehen sich auf dieses Dokument).
Die Grundeinheiten sind bei diesem Modell *Untersuchungen* (Investigations):

> Investigations are studies or parts of studies that have links directly to data holdings, as described above. More specific types of investigations may include the following.
>
> – Experiments. Investigations into the physical behaviour of the environment usually to test a hypothesis, typically involving an instrument operating under some instrumental settings and environmental conditions, and generating datasets in files. E.g., the subjection of a material to bombardment by X-Rays of known frequency recording the resulting diffraction pattern.
> – Measurements. Investigations that record the state of some aspect of the environment over a sequence of points in time and space, using some passive detector, e.g., the measurement of temperature at a point on the earth surface taken hourly using a thermometer of known accuracy.
> – Simulations. Investigations that test a model of part of the world, and a computer simulation of the state space of that model. This will typically involve some simulation package with some initial parameters, and generate a dataset representing the result of the simulation. (S. 110)

Damit richtet sich das Modell eher an experimentelle Natur- oder quantitative Sozialwissenschaften. Für Geisteswissenschaften (Philosophie, Sprachwissenschaften, Mathematik), (qualitative) Sozialwissenschaften wäre es nicht direkt anwendbar: „The CSMD thus can be seen as more suited to experimental science, typically an analysis of a sample in a laboratory or facility.“ (S. 115)

Die wesentliche Begrifflichkeit wird wie folgt zusammengefasst:

Untersuchung: Die Basiseinheit der Studie, eindeutige Kennungen der gegebenen Studie sowie Titel, Beschreibung, Daten etc.

Forscher: Die an der Studie beteiligten Personen, verbunden mit ihren Institutionen und ihrer Rolle bei der Untersuchung (z.B. Leitung, Forschungsassistent).

Thema und Schlagwort: Kontrollierte und unkontrollierte Bezeichnungen als Referenzen der Untersuchung.

Veröffentlichung: Verweise auf Veröffentlichungen, die mit der Untersuchung zusammenhängen (sie motivieren oder auf ihr aufbauen).

Probe: Informationen über die Materialprobe, die untersucht wird, wie Bezeichnung, chemische Formel oder weitere Informationen wie z.B. Gefährlichkeit.

Datensatz: Mit der Untersuchung verbundene Datensätze, die erste Erhebungen und verschiedene Analysen oder Durchläufe der Probe darstellen.

Datei: Datensätze können ineinander geschachtelt sein und mit gespeicherten Informationen – typischerweise Dateien – verbunden sein. Die Dateien enthalten detailliertere Informationen wie Name, Version, Ort, Datenformat, Zeitangaben und Prüfsummen. T

Parameter: Parameter beschreiben messbare Größen der Untersuchung wie Temperatur, Druck, Streuungswinkel, die sich entweder auf die Probe, die Messumstände oder die Messung selbst beziehen. Parameter können sich auf verschiedene Stufen beziehen (Probe, Datensatz, Datei) und mit Namen, Einheiten, Werten und Wertebereichen versehen sein.

Berechtigung: Hier kann angegeben werden, welche Nutzer in welcher Rolle Zugriff auf die Daten haben.

Während hier zwar einige Aspekte von Herkunft abgedeckt sind (insbesondere unter Probe, Datensatz, Datei, Parameter, Berechtigung) scheint kein durchgängiger Ansatz zur Beschreibung von Herkunft vorhanden zu sein.

10 Provenance Vocabulary

Das „Provenance Vocabulary“ ([Hartig and Zhao, 2010b]: http://sourceforge.net/apps/mediawiki/trdf/index.php?title=Provenance_Vocabulary) beschreibt Herkunftsmetadaten im Kontext von „Linked Data“ unter Benutzung von RDF. Die Autoren stellen es sowie die dahinter stehen Überlegungen in einen Artikel näher vor: Publishing and Consuming Provenance Metadata on the Web of Linked Data ([Hartig and Zhao, 2010a]: http://olafhartig.de/files/HartigZhao_Provenance_IPAW2010_Preprint.pdf), auf diesen Artikel beziehen sich die folgenden Zitate.

Herkunftsinformationen darüber, wer Daten erzeugte und veröffentlichte, erlauben Hinweise auf deren Qualität. Hier wird ein Begriffsverzeichnis vorgestellt, um die Provenienz von Daten im Internet als Metadaten zu beschreiben; diese Daten sollen im Rahmen des „Web of Data“ zugänglich und nützlich sein. Informationen über ein spezifisches Objekt können zu verschiedenen Webadressen führen, die dieses Objekt bezeichnen oder es mit verbundenen Daten („Linked Data“) verknüpfen. Welches dieser verbundenen Datenobjekte verdient die größte Glaubwürdigkeit oder bietet die besten Informationen?

„[However,] in a recent study we discovered a general lack of provenance-related metadata about data on the Web. Reasons are the lack of suitable vocabularies to describe Web data provenance and a lack of tools to generate and provide provenance metadata.
. . . we extended several Linked Data publishing tools, enabling them to automatically provide provenance metadata.“ (S. 3)
„Due to our extensions data publishers can easily enrich their data with provenance metadata by simply configuring a few parameters, such as the name and the URI identifying the publisher or the URI of the dataset.“ (S.9)

Das Provenienz-Begriffsverzeichnis wird als OWL-Ontologie definiert und besteht aus einem Kern und ergänzenden Modulen, von denen zunächst werden drei angeboten werden: Typen, Dateien und Integritätstest:

Kern http://trdf.sourceforge.net/provenance/ns# (mit Erklärung)

Typen http://trdf.sourceforge.net/provenance/types# (nur RDF)

Dateien http://trdf.sourceforge.net/provenance/files# (nur RDF)

Integritätstest http://trdf.sourceforge.net/provenance/integrity# (nur RDF)

Datenerzeugung und Datenzugriff werden als zwei Dimensionen der Provenienz identifiziert, die in diesem Zusammenhang typisch sind, und die benutzen Begriffe werden dementsprechend in drei Bereiche eingeteilt: allgemeine Begriffe und solche, die sich auf die Erzeugung von bzw. den Zugriff auf Daten beziehen.

Allgemeine Begriffe sind Akteur („Actor“, unterteilt in menschlich oder nicht-menschlich), Ausführung („Execution“) und Objekt („Artifact“, unterteilt in Datenobjekt und Datei). Damit wird ein allgemeiner Rahmen für die Beschreibung von Provenienz in einer Webumgebung bereitgestellt, der jedoch nicht den Anspruch hat, alle möglichen Aspekte zu beschreiben;

dazu können eigene Begrifflichkeiten gebildet und mittels geeigneter Konstruktionen an die gegebene Beschreibung angebunden werden.

Solche Herkunftsmetadaten können als RDF-Graphen veröffentlicht werden:
„The primary location of metadata about a linked dataset is its voiD description, that is, an RDF document on the Web which describes the dataset based on the Vocabulary of Interlinked Datasets (voiD).“ (S. 7)

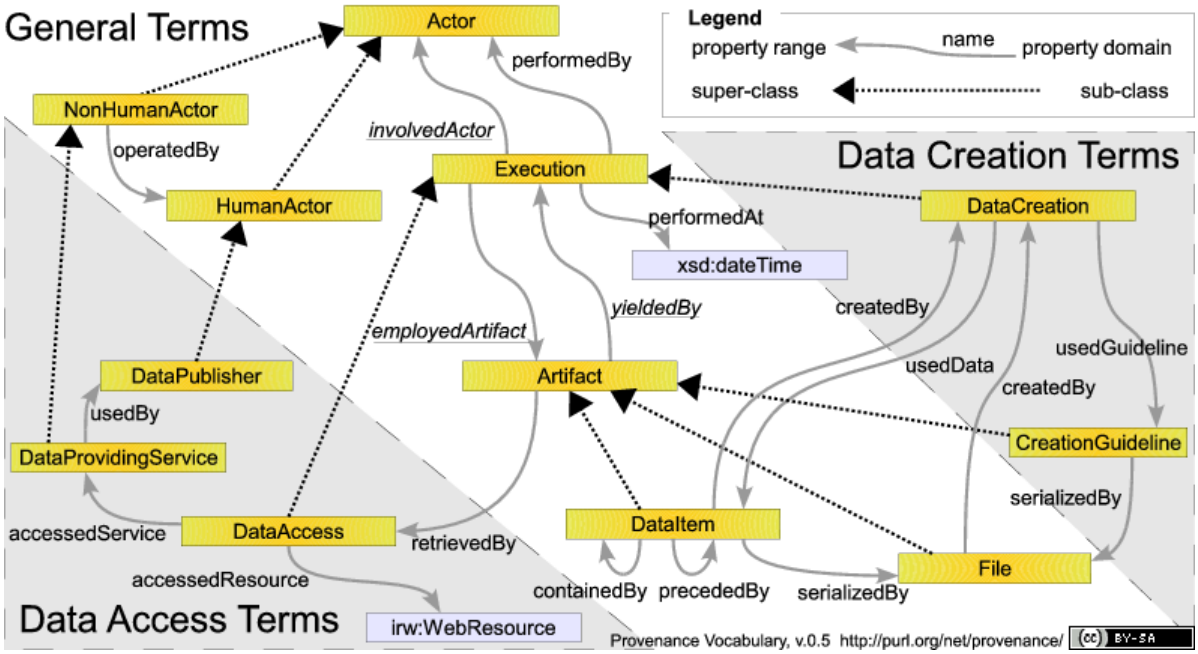

Abb. 5: Classes and properties defined by the Provenance Vocabulary core ontology (S.5)

Diese Daten können dann abgefragt und anschließend analysiert werden. Dies wird an einem Beispiel aus dem Feld der Genom-Analyse gezeigt: Mittels SPARQL können Informationen gefunden und ausgewertet werden. So können die aktuellsten Daten herausgefiltert und veraltete Ergebnisse automatisiert ausgeschlossen werden.

Hier ist eine kurze Auflistung der benutzten Klassen und Eigenschaften:

Allgemeine Klassen: Actor, HumanActor, NonHumanActor, Execution, Artifact, DataItem, File, Representation

Abstrakte Eigenschaften: yieldedBy, involvedActor, employedArtifact

Allgemeine Eigenschaften: containedBy, deployedSoftware, serializedBy, performedBy, operatedBy, performedAt, encodedBy

Datenerzeugungsklassen: DataCreation, CreationGuideline

Datenerzeugungseigenschaften: createdBy, usedGuideline, usedData, precededBy

Datenzugriffsklassen: DataAccess, DataProvidingService, DataPublisher

Datenzugriffseigenschaften: retrievedBy, accessedResource, accessedService, usedBy

Eine umfassende Einführung gibt der „Guide to the Provenance Vocabulary“ (http://sourceforge.net/apps/mediawiki/trdf/index.php?title=Guide_to_the_Provenance_Vocabulary).

11 The Foundations for Provenance on the Web

Der Übersichtsartikel von Luc Mureau „The Foundations for Provenance on the Web“ ([Moreau, 2010a], hier nach [Moreau, 2010b]: http://eprints.ecs.soton.ac.uk/21691/1/survey.pdf, darauf beziehen sich die Verweise) liefert einen umfassenden Überblick über die Diskussion von Provenienz bis etwa Mitte 2009.

Provenienz wird hier als wissenschaftlicher Qualitätsstandard gesehen:
„As the e-science vision becomes reality, researchers in the scientific community are increasingly perceived as providers of online data, which take the form of raw data sets from sensors and instruments, data products produced by workflow-based intensive computations, or databases resulting from sophisticated curation. While science is becoming computation and data intensive, the fundamental tenet of the scientific method remains unchanged: experimental results need to be reproducible.“ (S. 3)

Provenienz besitzt sowohl technische Aspekte, die z.B. Speichersysteme, Gridstruktur, Sicherheit betreffen, als auch soziale Aspekte, z.B. Transparenz oder Kontrolle von Prozessen oder Fragen der Privatsphäre.

Die Frage der Provenienz wird etwa seit den 80er Jahren des letzten Jahrhunderts diskutiert, das Interesse hat aber in der letzten Zeit stark zugenommen: die Hälfte der in diesem Artikel betrachteten 400 Veröffentlichungen stammt aber aus den vorherigen zwei Jahren. Der Autor unterstreicht die Bedeutung der Provenienz von Informationen im Internet und plädiert für eine Integration der verschiedenen Herangehensweisen und schlägt dafür einen grundlegenden Rahmen vor.

11.1 Analyse der Provenienz-Veröffentlichungen

Der Autor untersucht zunächst mittels einer Zitaten-Analyse die Entwicklung der Provenienz-Forschung und versucht ihre Richtung zu bestimmen: „Six clusters have been identified and positioned in time, covering topics as varied as database, workflows, eScience, „Provenance Challenge“, Open Provenance Model, Semantic Web and electronic notebooks.“ (S. 7) „Figure 6 contains a histogram displaying the number of publications on provenance per year. The bibliography contains papers that were known to the author up to summer 2009. A total of 425 papers have been identified. The first publication dates back from 1986 and describes an auditing technique to assist analysts in understanding and validating data results.“ (S. 8)

Die Entwicklung der Grid-Technologie für wissenschaftliche Anwendung und das britische E-Science-Programm haben die Beschäftigung mit Provenienz entscheidend gefördert, mutmaßt der Autor.

11.2 Definitionen

Der Autor betrachtet verschiedene Wörterbuch-Definitionen von „Provenance“ und schlägt dann vor:

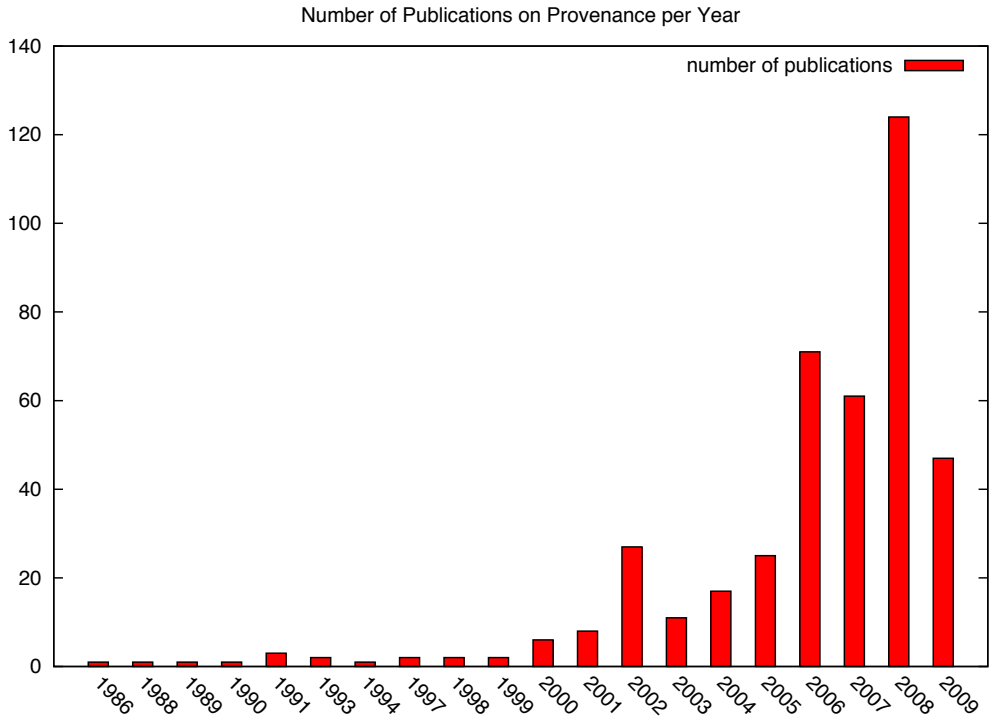

Abb. 6: Number of Provenance Publications

Definition 11.1. (Provenance as Process) The provenance of a piece of data is the process that led to that piece of data.

Definition 11.1 is concerned with provenance as a concept since potentially many things pertaining to execution may be captured under „process“, including the executed program, input data, configuration, computer, electricitiy powering it, users, etc. From a Computer Science perspective, the goal is to conceive a computer-based representation of provenance that permits useful analysis and reasoning. (S. 19)

Der Autor betrachtet dann weitere Definitionen bzw. Herangehensweisen an Herkunftsinformationen (S. 22ff):

- Provenance as Process
- Provenance as a Directed Acyclic Graph
- Why-Provenance
- Where-Provenance
- How-Provenance
- Provenance as Annotations
- Other definitions

Da Provenienz hier weit gefasst wird, gibt es auch eine große Spannbreite der Herangehensweisen in verschiedenen Gemeinschaften. Den Herangehensweisen liegen verschiedene Komplexe von Annahmen zugrunde, die der Autor strukturiert zusammenfasst (siehe S. 25ff):

- **Systembasiert** Hier werden Herangehensweisen gefasst, bei denen das System den Informationsfluss steuert und die Provenienz innerhalb dieses Systems erfass werden muss.
- **Programmbasiert** Wenn sowohl die Programmiersprache als auch das ausführende Programm bekannt sind, können diese benutzt werden, um auf die Provenienz zurückzuschließen oder diese im Prozess der Verarbeitung aufzuzeichnen. Andere Verfahren setzen auf strukturierte Beschreibungen der Ereignisse.
- **Vertrauensbasis** Die Vertrauensbasis sprachorientierter Systeme ist der Compiler und die Laufzeitumgebung, die der Sprachdefinition gehorcht (z.B. SQL oder JAVA), die der strukturierten Beschreibungen die Bibliotheken, Dienste oder Arbeitsabläufe, die durch die Struktur („Ontologie“) beschrieben werden.
- **Granularität** Die Verfolgung von Provenienz kann bezüglich der betrachteten Objekte stark variieren: Daten, Datensätze, Dateien, Verzeichnisse, Tabellen, Zeilen etc.
- **Was wird von Provenienz erfasst?** Hier können sowohl umfassend Rahmen- und Umgebungsbedingungen der Verarbeitung gefasst werden als auch spezifischer die Rohdaten, Daten die ein Ergebnis erzeugt haben oder nur eine Zusammenfassung der Benutzung der Daten.
- **Wessen Provenienz?** Neben der Provenienz von Daten wird auch eine Konzept der Provenienz von Diensten oder von Arbeitsabläufen betrachtet.
- **Zeit** Im allgemeinen wird Zeiterfassung als für Provenienz nicht für notwendig erachtet, aber manchmal nützlich gefunden, wenn die Nutzer sich auf die Zeit beziehen können.

Ein beliebiges System kann ohne seine Mithilfe nicht in seinem ganzen Verhalten erfasst werden, vielmehr muss man auf seine Teilsystem zugreifen können und deren Rückmeldungen zugreifen können. So betrachtet erscheint Provenienz als das Ergebnis einer Abfrage an die verschiedenen Anwendungen, die an einer Berechnung beteiligt waren und die sich daraus ergebende Rekonstruktion des Ablaufs.

Damit nicht jede Anfrage bis an den Urknall zurückführt, muss Provenienz muss allerdings auch an die Nutzerbedürfnisse angepasst sein; ein Endpunkt der Rückverfolgung oder die Art der zu berücksichtigenden Daten müssen festgelegt werden.

Im PASOA-Projekt (Provenance Aware Service Oriented Architecture) wird zwischen Prozessaussagen („p-assertions“) und Provenienz als Abfrage über Prozessaussagen unterschieden. Daraus ergibt sich für Provenienz ein Lebenszyklus: Prozessaussagen werden unabhängig von möglichen Endprodukten gesammelt und können im Nachhinein nach der Provenienz der Daten befragt werden.

11.3 Provenienz in Arbeitsabläufen und Datenbanken

Der Umgang mit Arbeitsabläufen („Workflows“) unterscheidet sich typischerweise in Wissenschafts- und Geschäftsbetrieb: während im ersten Bereich die Abläufe of wiederholt und dabei zunehmend an die Erfordernisse angepasst und damit sehr dynamisch sind, sind sie im zweiten

Bereich eher statische Einrichtungen, die gewisse Geschäftsabläufe festschreiben. Wie viel dabei von den jeweils eingebundenen Komponenten über die inneren Abläufe preisgegeben wird, entscheidet über die dabei erfassbare Granularität der Provenienz.

Im einzelnen wird auf dei folgenden Aspekte eingegangen (S.31):

(i) For users to deal with the amount of information contained in provenance, mechanisms are required to abstract and synthesize information in views customized to users.
(ii) A specific aspect of abstraction is concerned with collections of data.
(iii) If the provenance of everything is to be tracked, consideration should be given to storage requirements.
(iv) The means to actually query provenance need to be provided.
(v) Tracking the evolution of workflows is a special kind of provenance tracking.
(vi) Formal properties of provenance are now emerging.
(vii) Finally, many activities involve humans in the loop, who impact on decisions and processes, and therefore need to be made explicit in provenance representations.

Das Open Provenance Model (siehe Abschnitt 7, Seite 34) führt den Begriff der Beschreibung („account“) ein, damit kann unabhängig vom Arbeitsablauf verschiedene Sichten auf eine Provenienzspur beschrieben werden, die sich für verschiedenen Beobachter sogar widersprechen können.

In wissenschaftlichen Untersuchungen treten oft Sammlungen von Objekten auf, solche Kollektionen werden oft als eigentliche Objekte mit eigener Provenienz betrachtet, das kann aber kompliziert werden, da alle Teile der Sammlung ihre je eigene Geschichte haben. Das führt zu Fragen der Stabilität der Sammlung, der Granularität der Beschreibung oder effizienten Darstellung.

Nicht nur deswegen kann der Umfang der Provenienzdaten ausufern, Moreau zitiert (S. 35): „Provenance can become huge: in the public database Gene Ontology, the provenance of a single tuple has been observed to be 10Mb; likewise, a 250Mb database of biological data is associated with 6Gb of provenance. The size of provenance matters; because this is a multi-dimensional challenge, it has to be acknowledged that there is a trade-off between compact representation (reducing recording/upload time), compact storage (reducing storage requirements) and query time.“

11.4 Die Vision der Offenen Provenienz

Provenienz wird im Zusammenhang mit Datenbanken und Arbeitsabläufen typischerweise als geschlossenes System gesehen. Wenn aber die Provenienz verschiedener Systeme zusammengeführt werden soll, so ist ein breiter angelegtes Herangehen erforderlich, das hier als *Vision der Offenen Provenienz* beschrieben wird. Dabei tritt sofort die Frage auf, was denn erfasst werden soll oder darf: darf die gesamte Nutzerkommunikation mit dem System oder seinem Desktoprechner erfasst werden?

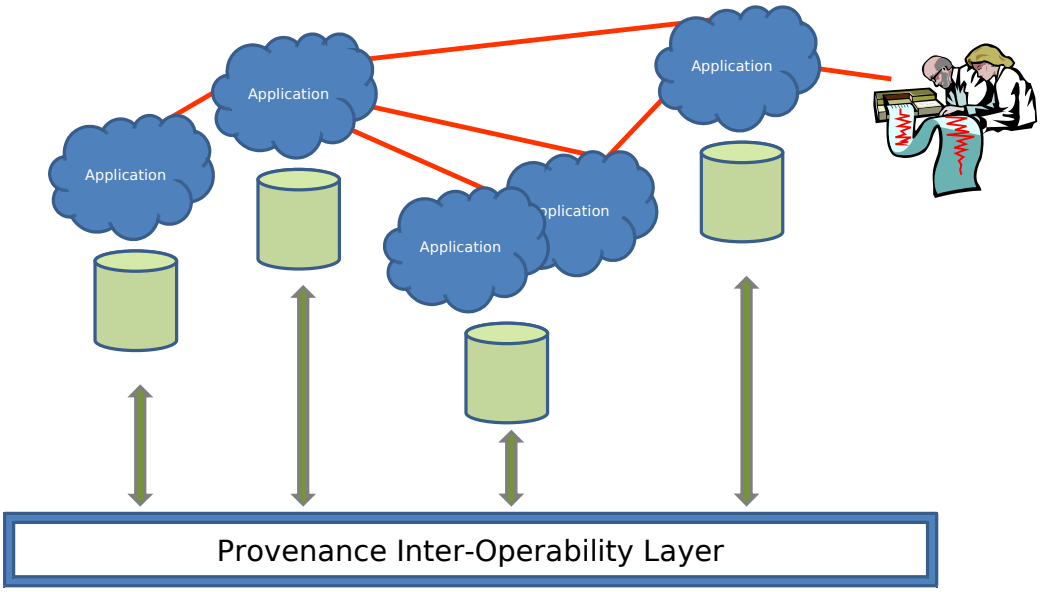

Abb. 7: The Open Provenance Model (OPM)

Die Erfassung der Abläufe in geschlossenen monolithischen Anwendungen wird mittlerweile weitgehend verstanden, solche Anwendungen sind dann „provenienzbewusst“, wenn sie in der Lage sind, die interne Verarbeitung ihrer Daten abrufbar aufzuzeichnen. Im weiteren Verlauf müssen dann Prozessaussagen („p-assertion“) erfasst werden, die in einem externen „provenance store“ gespeichert werden, der eine sichere Langzeitarchivierung der Daten bieten soll.

Werden solche Anwendungen miteinander gekoppelt, so stellt sich die Frage nach einem gemeinsamen Provenienzmodell. Darauf zielt die Vision: „With the Open Provenance Vision, the provenance from individual systems or components can be expressed, connected in a coherent fashion, and queried seamlessly.“ (S. 46)

Um einheitliche Abfragen zu erlauben, muss die Provenienz in den verschiedenen Teilsystemen durch strukturierte Kategorien („Ontologien“) beschrieben werden, die von den verwendeten Technologien unabhängig sind. Die verschiedenen Projekte haben verschiedene solche Begriffsstrukturen entwickelt, mit dem „Open Provenance Model“ entsteht aber ein universelles Modell zur Darstellung und zum Austausch von Provenienzinformationen (zu dessen weiterer Beschreibung siehe 7).

> The Open Provenance Vision postulates that all systems/components should be able to:
> 1. keep a record of provenance for any important data they produce (in their formats and repositories of choice);
> 2. follow conventions when exchanging data so that provenance can be traced across systems;
> 3. export provenance of such data using a common data model, such as the Open Provenance Model;

4. answer provenance queries, structured over the common data model.

(S. 47)

11.5 Provenienz, das Web und das Semantic Web

Der technologische Hintergrund der „Open Provenance Vision“ ist das Internet mit seinen Mashups, RSS Feeds und anderen Verbindungen diverser Datenquellen zu kombinierten Objekten. Hier ist die Provenienz das zentrale Kriterium, um über die Glaubwürdigkeit von Inhalten zu entscheiden. Moreau betrachtet dazu die folgenden Aspekte (S. 55ff):

> Issues in this area can be categorized in the following separate strands.
> (i) Given the importance of provenance, it is to be regarded as first-class data, itself to be exposed on the Web.
> (ii) Semantic Web technologies are themselves being used, not only to represent provenance information, but also to query and reason over it.
> (iii) Given the importance of metadata in the information discovery process, and the ease by which such metadata can be published on the Web, tracking the provenance of RDF-based information has also become a focus of investigation.
> (iv) In the Semantic Web, not only can triples be asserted, but also they can be inferred. In such case, special techniques need to be devised to track their provenance.

Für die Beschreibung von Provenienz in diesem Zusammenhang wird auf dem Webstandard von URIs und HTTP aufgebaut. Für das semantische Web wird mittels RDF darauf eine Informationsstruktur aufgesetzt, die auch Graphen darstellen kann. Die Abfragesprache SPARQL erlaubt darauf Recherchen, und mit der Ontologiesprache OWL können Definitionen und abstrakte Schlüsse formuliert werden.

Die Hoffnung ist, dass diese Technologien es ermöglichen, Forschungsprozesse so zu erfassen, dass bei der Veröffentlichung ihrer Ergebnisse auch ein Zugriff auf die zugrundeliegenden Daten möglich ist und ihre Verarbeitung nachvollzogen werden kann.

11.6 Verlässlichkeit

Digitale Kommunikation hat die Bildung komplexer Kooperationsstrukturen erleichtert, die oft nur kurzzeitig gemeinsame Probleme bearbeiten oder Dienste bereitstellen. In wissenschaftlichen Prozessen berühren sich dann Fragen der Provenienz und der Verantwortung: die Nachverfolgung der Geschichte von Daten und Objekten klärt die Frage der Beteiligung erlaubt die Rechenschaftslegung. Umgekehrt kann man nicht davon ausgehen, dass in einem komplexen System alle Provenienzinformationen für alle Beteiligten einsehbar sind, daher muss man Zugangsbeschränkungen berücksichtigen. Das wird im allgemeinen nicht mit einfachen rollenbasierten Zugangssystemen möglich sein, dazu gibt es Vorschläge, wie OPM-Elemente mit Zugangsrechten ausgezeichnet werden könnten. „Privacy and accountability are both legitimate goals, but they can be at odds.“ (S. 72)

Wenn die Provenienz die Echtheit von Daten garantieren soll, so muss naturgemäß auch die Echtheit der Provenienzinformationen gesichert sein. Dazu werden z.B. kryptographische Prüfsummen vorgeschlagen. Andererseits erfordert eine langfristige Bewahrung digitaler Informationen typischerweise auch deren Transformation; dieses Paradoxon erfordert zu seiner Auflösung die Aufzeichnung der Veränderung der Provenienzinfomamtionen: deren Provenienz.

Das alles beruht natürlich darauf, dass die Originärinformationen, die von dem Provenienzsystem aufgezeichnet wurden, korrekt und vollständig sind, Zertifizierungen von Provenienzsystemen könnten versuchen, dies sicherzustellen.

„A system is accountable if it can provide explanations for its actions, if its past actions are accountable, and if it can be demonstrated that its processes and decisions are compatible with rules, policies, or broadly regulations. . . . [W]ith explicit representation of provenance, one can make systems accountable: provenance provides the necessary *evidence* which makes systems *transparent* and allows an auditor to determine whether *policies* are satisfied.“ (S. 68)

11.7 Schlussfolgerungen

> Given that information flows across multiple services over the Web, being transformed, filtered, processed and repackaged in many different ways, a representation of provenance has to be assembled by bringing evidence of local transformations and derivations into a coherent whole. This is the purpose of the Open Provenance Vision, and the community-driven Open Provenance Model. (S. 74). . .
> For information provenance to be traceable over the Web, each information system or service involved in a global information flow has to track provenance in its local activities. (S. 75)
> . . .
>
> > To make the Web provenance-aware, mentalities have to change: it is no longer sufficient to publish data, but associated provenance must also be made available. While tools may assist in this task, this inevitably increases the human effort involved. (S.77)

Dieser Artikel enthält eine 46-seitige **Provenance Bibliography** mit 461 Einträgen.

12 DCMI Metadata Provenance Task Group

Die „DCMI Metadata Provenance Task Group“ ([Eckert and Panzer, 2010]) hat sich im Juni 2010 gegründet und hatte sich zum Ziel gesetzt, bis zur DCMI-Tagung 2011 einen Entwurf für die Herkunftsinformationen von Metadaten vorzulegen[8], siehe auch deren „DCMI Metadata Provenance Task Group Wiki“ ([DCMI Metadata Provenance Task Group, 2011]). Nähere Informationen liefert der Artikel „Towards Interoperable Metadata Provenance“ ([Eckert et al., 2010])

Metadaten liefern Informationen über Objekte, Herkunftsinformationen oder „Provenance“ für Metadaten sollen angeben, wer für die Metadaten verantwortlich ist. Damit können dann Fragen der Zuverlässigkeit der Informationen bearbeitet werden. Vorgeschlagen werden dafür „Metametadaten“, Informationen über Metadaten. Dahinter steht unter anderem die Vorstellung, dass Metadaten auf unterschiedliche Weise, teils intellektuell, teils automatisiert erzeugt werden können, und solche Metametadaten diese Qualitätsunterschiede ausdrücken können.

Nicht klar ist mir, ob es eine stringente Begründung dafür gibt, nach dieser Stufe aufzuhören, und nicht noch „Metametametadaten“ zu betrachten, die Auskunft über die Verlässichkeit der Aussagen der Metametadaten machen: Klaus sagt, die Klassifikation als „Militär“ kommt von dem Bibliothekar Walter, aber Dieter meint, sie käme von Gerhard, und das sei eine Maschine. . . Und das kann man dann natürlich noch weiter treiben.

1) Arbitrary metametadata statements about a set of statements. 2) Arbitrary metametadata statements about single statements. 3) Metametadata on different levels for each state- ment or sets of statements. 4) Applications to retrieve, maintain and republish the metametadata without data loss or corruption. 5) Data processing applications to store the metametadata about the original RDF data. Dataprocessingapplicationstostorethemetametadata about the original RDF data.

Es werden die folgenden Anforderungen an das System zur Erfassung der Metametadaten aufgestellt(Eckert et al., S. 2):

1. Arbitrary metametadata statements about a set of statements.
2. Arbitrary metametadata statements about single statements.
3. Metametadata on different levels for each statement or sets of statements.
4. Applications to retrieve, maintain and republish the metametadata without data loss or corruption.
5. Data processing applications to store the metametadata about the original RDF data.

Als Format schwebt der Arbeitsgruppe RDF vor, eventuell erweitert um die Möglichkeit bezeichneter Graphen. Als erstes soll ein Entwurf für Standardverfahren und ihre Realisierung geliefert werden.

[8] Präsentationen auf der DCMI Tagung sind unter http://dcevents.dublincore.org/index.php/IntConf/dc-2011/paper/view/43/29 und http://dcevents.dublincore.org/index.php/IntConf/dc-2011/paper/view/44/3 erhältlich.

13 W3C Provenance (Incubator) Group

Im Jahre 2009 formierte sich im Rahmen des W3C die „W3C Provenance Incubator Group“ ([W3C Provenance Incubator Group, 2009]: http://www.w3.org/2005/Incubator/prov/): „The mission of the Provenance Incubator Group, part of the Incubator Activity, was to provide a state-of-the art understanding and develop a roadmap in the area of provenance for Semantic Web technologies, development, and possible standardization.“

Die Gruppe war vom 22. 9. 2009 bis zum 14. 12. 2010 aktiv und veröffentlichte zum Abschluss ihrer Arbeit einen „Final Report“ ([Gil et al., 2010]: http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/), der den folgenden Ausführungen zugrunde liegt. Außerdem wurden von der Gruppe weitere Dokumente vorgelegt, die für die Betrachtung hier interessant sind.

2011 wurde im Anschluss an die Incubator Group eine vollwertige W3C-Arbeitsgruppe gründet ([W3C Provenance Group, 2011]: http://www.w3.org/2011/prov/wiki/Main_Page), die die Arbeit an diesem Thema sehr aktiv weitertreibt (s.u.).

„Provenance“ wird dabei so verstanden:
„Provenance refers to the sources of information, such as entities and processes, involved in producing or delivering an artifact. The provenance of information is crucial in deciding whether information is to be trusted, how it should be integrated with other diverse information sources, and how to give credit to its originators when reusing it.“

13.1 Provenance Vocabulary Mappings

In einem vorbereitenden Studie wurden verschiedene Begrifflichkeiten zu Herkunftsinfomationen verglichen ([Sahoo et al., 2010]: http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings, die Zitate in diesem Abschnitt enstammen diesem Dokument):

> The core set selected for mapping includes:
>
> – Open Provenance Model
> – Provenir ontology
> – Provenance Vocabulary
> – Proof Markup Language
> – Dublin Core
> – PREMIS
> – WOT Schema
> – SWAN Provenance Ontology
> – Semantic Web Publishing Vocabulary
> – Changeset Vocabulary

Dabei wird das „Open Provenance Model“ als Referenzmodell gewählt. In der nächsten Stufe wurde versucht, formale Abbildungen zwischen diesem Modell und den betrachteten Systemen zu finden, dazu werden Begriffe aus dem SKOS-Vokabular[9] (skos:broader skos:narrower skos:related) und OWL2[10] (owl:equivalentClass, owl:equivalentProperty) benutzt, um die Ab-

[9] http://www.w3.org/2004/02/skos
[10] http://www.w3.org/TR/owl2-overview

bildung zu beschreiben.

In einer (großen) Tabelle werden den OPM-Begriffen Process, Artifact, Agent, Account, wasDerivedFrom, used, wasGeneratedBy, wasControlledBy, wasTriggeredBy entsprechende Begriffe der anderen Terminologien zugeordnet. In einzelnen Abschnitten werden für jede dieser Terminologien die Zuordnungen in eigenen Tabellen begründet (z.T. auch in Frage gestellt) sowie kurz die Begriffe erwähnt, die nicht abgebildet werden konnten. Die Ergebnisse dieser Analyse werden dort zusammengefasst:

> Our first finding is that many of the considered models and vocabularies have a set of core concepts that correspond to the notion of processes, artifacts, and agents as defined in OPM. These concepts can be mapped quite naturally between the models. While the modeling of these concepts indicates a process-centric view, several vocabularies take a resource-centric view. Specifying the mappings for these resource-centric vocabularies was more difficult...
> Several vocabularies provide non-causal relationships, something explicitly left out of OPM...
> While many vocabularies provide time related terms, the time dimension is not represented in our mapping. The main reason for this lack was that OPM does not represent time related properties explicitly...
> Further aspects of provenance that are not well captured by OPM and, thus, missing from the core provenance terms are:
> - versioning
> - a notion of artifact identity that persists across transformations,
> - containment relationships and collections, and
> - cryptographic hashes and digital signatures.

13.2 Weitere Beiträge

Die Provenance Incubator Group stellte neben diesem Vergleichsdokument eine Reihe weiterer Untersuchungen an, um die Frage der Provenienz zu strukturieren. Auf einige soll hier direkt verwiesen werden, eine umfassendere Liste findet sich weiter unten.

Provenance Dimensions
[W3C Provenance Incubator Group, 2010a]: Es werden der Herkunftsinformation 17 Dimensionen[11] zugeordnet, das soll die Auswahl der Anwendungsfälle unterstützen (siehe 13.3.2).

User Requirements
[Cheney et al., 2010]: Hier werden Nutzungsanforderungen an das Provenienzsystem formuliert, dazu werden drei Szenarien betrachtet:

Nachrichtensammler: Hier werden Nachrichten aus verschiedenen Quellen (Websites, Blogs, Tweets) zusammengestellt, und Herkunftsinformationen sind für die Beurteilung von deren Qualität wichtig.

[11] Der Begriff der „Dimension“ wird hier etwas unklar genutzt.

Krankheiten: Hier wird das Auftreten einer Krankheit und ihre Ausbreitung dokumentiert: sehr unterschiedliche Quellen müssen erfasst und beurteilt, außerdem wissenschaftliche Forschungsergebnisse und ihnen zugrundeliegende Daten festgehalten werden.

Verträge: Hier wird eine technische Vereinbarung so dokumentiert, dass die Erfüllung der Vertragsvereinbarung und der Ablauf der Entwicklung nachvollzogen werden können.

Entlang dieser Szenarien werden Anforderungen an das Provenance-System formuliert.

Requirements
([W3C Provenance Incubator Group, 2010b]: Aus der Betrachtung der Anwendungsfälle werden detaillierte Anforderungen an das Provenance-System abgeleitet.

State of the Art Report
([Groth et al., 2010]: Entlang der untersuchten drei Anwendungsfälle werden bestehende Möglichkeiten und Lücken der vorhandenen Systeme analysiert:

„There are many proposed approaches and technology solutions that are relevant to provenance, as we have illustrated with three very different scenarios. Despite this large body of growing work, there are several major technology gaps to realizing the requirements of these scenarios."

13.3 Endbericht

Zitate in diesem Abschnitt sind dem „Final Report" ([Gil et al., 2010]) entnommen.
Im dem Endbericht werden die Ergebnisse der Untersuchungen der „Incubator Group" zusammengefasst. Dem Endreport wird die folgende Definition von „Provenance" zugrunde gelegt:

> **Definition** (Provenance): Provenance of a resource is a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource. Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility. Provenance assertions are a form of contextual metadata and can themselves become important records with their own provenance.

Abgegrenzt wird dieser Begriff von den verwandten Begriffen „Metadata" und „Trust".

13.3.1 Use Cases

Zur Analyse der mit Herkunftsinformationen zusammenhängenden Probleme sammelte die Gruppe 33 Anwendungsfälle, daraus gingen die drei „Flagship Scenarios" hervor, die in 13.2 beschrieben wurden. Aus deren Analyse ergibt sich:

13.3.2 Anforderungen an Provenienz

> As seen in the previous section, provenance touches on many different domains and applications. Each of these has different requirements for provenance. Here,

we present the requirements extracted by the group from the collected use cases. The group produced several detailed documents:

- To help organize these requirements, the group developed a set of key dimensions for provenance[12].
- The requirements presented here are a summarized version of the user requirements discussed in the incubator group´s report on Requirements for Provenance on the Web[13].
- The incubator group also collected 140 technical requirements[14]. We refer to these technical requirements indirectly, but we do not include them here.
- In the context of the W3C Semantic Web Interests Group and the RDF Next Steps workshop[15], the Provenance Incubator Group investigated which requirements could impact any future version of RDF. These requirements were documented in a paper titled „Provenance Requirements for the Next Version of RDF[16]“ that was presented[17] at that workshop.

Die Aspekte des Provenienzbegriffs werden in drei Kategorien zusammengefasst, denen die 17 Dimensionen zugeordnet werden:

1. **Content**

   Object: The artifact that a provenance statement is about.

   Attribution: The sources or entities that contributed to create the artifact in question.

   Process: The activities (or steps) that were carried out to generate or access the artifact at hand.

   Versioning: Records of changes to an artifact over time and what entities and processes were associated with those changes.

   Justification: Documentation recording why and how a particular decision is made.

   Entailment: Explanations showing how facts were derived from other facts.

2. **Management**

   Publication: Making provenance available on the web.

   Access: The ability to find the provenance for a particular artifact.

   Dissemination: Defining how provenance should be distributed and its access be controlled.

   Scale: Dealing with large amounts of provenance.

3. **Use**

   Understanding: How to enable the end user consumption of provenance.

[12] http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Dimensions
[13] http://www.w3.org/2005/Incubator/prov/wiki/User_Requirements
[14] http://www.w3.org/2005/Incubator/prov/wiki/Requirements
[15] http://www.w3.org/2009/12/rdf-ws/
[16] http://www.w3.org/2005/Incubator/prov/wiki/images/3/3f/RDFNextStep_ProvXG-submitted.pdf
[17] http://www.slideshare.net/junzhao/2010-06-rdfnext

Interoperability: Combining provenance produced by multiple different systems.

Comparison: Comparing artifacts through their provenance.

Accountability: Using provenance to assign credit or blame.

Trust: Using provenance to make trust judgments.

Imperfections: Dealing with imperfections in provenance records.

Debugging: Using provenance to detect bugs or failures of processes.

An den **Inhalt** eines „provenance record“ werden die folgenden Ansprüche formuliert:

- **Identifikation** Man muss festlegen können, auf welches Objekt sich Provenienzaussagen beziehen und auf dieses Objekt verweisen können.
- **Zuordnung** bezieht sich auf Quellen und Elemente, die bei der Erzeugung des getroffenen Objektes beteiligt waren.
- **Prozesse** bezieht sich auf die Aktionen, die durchgeführt wurden, um das Objekt zu erzeugen.
- **Entstehung und Versionierung** müssen bei der Darstellung von Provenienz beachtet werden.
- Die Dokumentation von **Rechtfertigungen** erlaubt es, Beschlüsse zu verstehen und in Frage zu stellen.
- **Schlussfolgerungen** erlauben es, Informationen aus der gegebenen Provenienzinformation abzuleiten.

Die Ausarbeitung dieser Anforderungen zeigt schon ein sehr reiches Feld an Forderungen, die das Proveienzsystem erfüllen muss.

Die **Verwaltung** der Herkunftsinformationen soll die folgenden Ansprüche erfüllen:

- **Veröffentlichung:** Provenienz muss im Internet verfügbar sein.
- Wenn sie verfügbar ist, muss sie auch **auffindbar** sein.
- Gleichzeitig muss ihre Verbreitung **kontrollierbar** sein.
- Der **Umfang** der Provenienzinformationen muss beherrschbar bleiben.

Aus der **Nutzersicht** ergeben sich weitere Anforderungen:

- Provenienzinformationen müssen für die Interessierten **verständlich** und nutzbar gemacht werden
- Da diese Informationen verschiedenartigen Systemen mit unterschiedlichen Darstellungen entstammt und in vielfältigen Anwendungen genutzt werden soll, ist **Interoperabilität** eine wesentliche Anforderung.
- Provenienz muss zum **Vergleich** von Objekten aufgrund ihrer Herkunft genutzt werden können.
- Provenienz kann zur **Rechenschaftslegung** genutzt werden.

- **Vertrauen** ist ein wesentlicher Aspekt von Provenienz.
- Provenienz kann die Behebung von **Mängeln** unterstützen.
- Provenienz kann der **Fehlersuche** dienen.

Schließlich wird festgehalten, dass sich aus den verschiedenen Ansprüchen sich auch Forderungen an die Weiterentwicklung des RDF-Standards ergeben:

„[T]he group argued for additional desirable capabilities that the current standard RDF model does not offer, including proper identification of RDF statements and an annotation framework permitting a standard approach for linking meta-information like provenance with sets of RDF triples. The group also argued for the development of a common approach to exchange provenance information between systems and publish it on the Web.“

> To obtain a comprehensive understanding of existing work on provenance the group carried out several activities, including:
> - A series of presentations on current work on provenance[18]. The topics covered included applications of provenance (in eGoverment, eScience, multimedia), technical areas (security, database provenance), and published provenance vocabularies.
> - A compilation of surveys of existing research on provenance already available in the literature[19] that cover provenance work from a variety of perspectives
> - An organized collection of references on prior work on provenance annotated with tags that referred to the flagship scenarios and provenance dimensions described earlier
> - An compilation of relevant technologies and standards[20] in relevant areas
> - An analysis of existing vocabularies[21] for provenance, including mappings across terms in the different vocabularies with extensive documentation
> - A State of the Art Report[22] centered on the three flagship scenarios described above including an analysis of existing technology gaps

Aus dieser Analyse ergab sich eine Liste verwandter Technologien und Herangehensweisen:

> - Open Provenance Model[23]: Outcome of the Provenance Challenge series[24] (initiated in 2005), after discussion and consensus of part of the community. For this reason, and for being general and broad enough, it was selected as the model to which the rest of the provenance vocabularies have been mapped.

[18] http://www.w3.org/2005/Incubator/prov/wiki/Presentations_on_State_of_the_Art
[19] http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Survey
[20] http://www.w3.org/2005/Incubator/prov/wiki/Relevant_Technologies
[21] http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings
[22] http://www.w3.org/2005/Incubator/prov/wiki/State_of_the_Art_Report
[23] http://openprovenance.org/
[24] http://twiki.ipaw.info/bin/view/Challenge/

- Provenir Ontology[25]. Common provenance model, which forms the core component of a modular approach to provenance management framework in eScience. Three base classes in the Provenir Ontology are used for representing the primary components of provenance, that is, „data“, „agent“ and „process“.
- Provenance Vocabulary[26]. Developed to describe provenance of Linked Data on the Web. The Provenance Vocabulary is defined as an OWL ontology and it is partitioned into a core ontology and supplementary modules.
- Proof Markup Language[27]. PML is an interlingua for representing and sharing explanations generated by various intelligent systems such as hybrid web-based question answering systems, text analytic components, theorem provers, task processors, web services, rule engines, and machine learning components. The interlingua is split into three modules (provenance, justification, and trust relations) to reduce maintenance and reuse costs.
- Dublin Core[28]. Dublin Core Metadata Terms provide a means to describe resources such that others will be able to interpret those descriptions.
- PREMIS[29]. Stands for „PREservation Metadata: Implementation Strategies“. It is a data dictionary for supporting long-term preservation: defines a core set of semantic units that repositories should know in order to perform their preservation functions.
- WEB OF Trust Schema (WOT)[30]. Schema designed to facilitate the use of Public Key Cryptography tools such as PGP or GPG to sign RDF documents and document these signatures.
- Semantic Web Applications in Neuromedicine (SWAN) Ontology[31]. Ontology for modeling scientific discourse, developed in the context of building a series of applications for biomedical researchers, as well as extensive discussions and collaborations with the larger bio-ontologies community.
- Semantic Web Publishing Vocabulary[32]. An RDF-Schema vocabulary for expressing information provision related meta-information and for assuring the origin of information with digital signatures.
- Changeset Vocabulary[33]. Describes changes to RDF-based resource descriptions.

„We carried out detailed analyses of the requirements and provenance dimensions exhibited by the flagship scenarios, and the extent to which current practice and ongoing research addresses these needs. The State of the Art Report[34] presents these detailed analyses along with

[25] http://wiki.knoesis.org/index.php/Provenir_Ontology
[26] http://purl.org/net/provenance/
[27] http://inference-web.org/2007/primer/
[28] http://dublincore.org/
[29] http://www.loc.gov/standards/premis/
[30] http://xmlns.com/wot/0.1/
[31] http://www.w3.org/TR/hcls-swan/
[32] http://www4.wiwiss.fu-berlin.de/bizer/WIQA/swp/SWP-UserManual.pdf
[33] http://vocab.org/changeset/schema.html
[34] http://www.w3.org/2005/Incubator/prov/wiki/State_of_the_Art_Report

references to the bibliography. Here, we summarize the discussion of how current technology is used to address these kinds of problems.“

Auf der Basis dieser Untersuchungen wurde die Frage bearbeitet, welche technologischen Lücken derzeit die Entwicklung von Provenienz im Internet behindern. Daraus wurde eine umfangreiche Liste erstellt, die sich an den oben aufgeführten Ansprüchen und Nutzungsszenarien orinetiert.

13.3.3 Architektur der Provenienz im Web

Bemängelt wird das Fehlen eines Webstandards zur Veröffentlichung und Nutzung von Provenienz. Nötig sind eine Sprache, um Provenienz darzustellen und ein Mechanismus, um diese Informationen aufzufinden. Dazu wird vorgeschlagen, die HTTP-Server des Internet dahingehend anzureichern, dass sie – neben Medientyp, Textkodierung und Sprache – zusätzlich Provenienzinformationen über die Webressource und ihre Darstellung verhandeln können.

Im Webkontext können sich Herkunftsinformationen auf verschiedene Objekte beziehen; jede Möglichkeit bringt eigenen Probleme mit sich:

- auf die Internetquelle: dies könnte Schwierigkeiten bereiten, da Webseiten sich mit der Zeit ändern; OPM kann derzeit diese Art von Provenienzinformationen nicht liefern, da es sich auf *unveränderbare* Objekte bezieht.
- auf die Darstellung der Internetquelle: damit hätte jede von Server bereitgestellte Darstellung ihre je eigene Provenienz, auch die „on the fly“ produzierten.
- auf den Zustand der Internetquelle.

Außerdem werden drei Übertragungsmodi unterschieden, mit denen Provenienz transportiert werden kann: als Wert, als Referenz und gemischt.

Übertragung als Wert: Die gesamte bekannte Provenienzinformation wird der Übertragung durch den HTTP-Server (im Header oder in der Antwort selbst) hinzugefügt.

Übertragung als Referenz: Hier wird die Provenienzinformation als eine verbundene Internetquelle angesehen, deren URI übertragen wird.

Übertragung teilweise als Wert bzw. Referenz: Hier werden einige Provenienzinformationen direkt übertragen, während für andere Referenzen angegeben werden, die entweder auf vollständige oder vertiefende Provenienzinformationen verweisen.

13.3.4 A Roadmap for Provenance on the Web

Die Analyse der Provenienzanforderungen der betrachteten Anwendungsfälle zeigte größere Lücken in technologischer und begrifflicher Hinsicht. Dazu wurde ein eigener Bericht erstellt: Recommendations for scenarios[35].

Speziell ergeben sich acht allgemeine Empfehlungen:

[35] http://www.w3.org/2005/Incubator/prov/wiki/Recommendations_for_scenarios

1. There should be a standard way to represent at a minimum three basic provenance entities:
   (a) handle (URI) to refer to an object (resource)
   (b) a person/entity that the object is attributed to
   (c) a processing step done by a person/entity to an object to create a new object
2. A provenance framework should include a mechanism to access provenance-related information addressed by other standards, such as:
   – licensing information of the object
   – digital signature for the object
   – digital signature for provenance records
3. A provenance framework should include a standard way for sites to make provenance information about their content available to other parties in a selective manner, and for others to access that provenance information
4. A provenance framework should include a standard way to express the provenance of provenance assertions, as there can be several accounts of provenance and with different granularity and that may possibly conflict
5. A provenance framework should include a representation of provenance that is detailed enough to enable reapplying the process to reproduce it
6. A provenance framework should allow referring to versions of objects as they evolve over time, or to temporal information statements of when the object was created, modified, or accessed. In particular it should provide for a representation of how one version (or parts thereof) was derived from another version (or parts thereof).
7. A provenance framework should include a standard way to represent a procedure which has been enacted (in the scenario, this is to compare that procedure with what was required to be done)
8. A provenance framework should include a way to determine commonality of derivation in two resources (in the scenario, this is needed to judge the independence or otherwise of two reports)

[...] The group agreed that recommendations 1 to 3 from the group's effort were the highest priorities and that they should be addressed within a provenance standardization effort. While acknowledging that the priorities of the recommendations depend on the context, those three recommendations are considered to be the most common and to represent the core set of issues to be addressed.

Der Endbericht schließt mit Hinweisen auf die zu bearbeitenden Probleme und liefert schon die Ansätze eines Arbeitsplans für die zu gründende W3C-Provenienzarbeitsgruppe, die mittlerweile eingerichtet worden ist. Die „W3C Provenance Working Group“ ([W3C Provenance Group, 2011]) formuliert als Ziel:
„Mission: to support the widespread publication and use of provenance information of Web documents, data, and resources. The Working Group will publish W3C Recommendations that define a language for exchanging provenance information among applications.“

Die Gruppe hat bereits eine Reihe von Arbeitspapieren veröffentlicht (siehe http://www.w3.org/2011/prov/wiki/WorkingDrafts), insbesondere ein Provenienzdatenmodell ([PRO, 2012] bzw. den „Editor's Draft" [Belhajjame et al., 2012]), das weitgehend auf dem OPM aufbaut. Hier sind die zentralen Konzepte „Entities", „Activities" und „Agents" zusammen mit Beziehungen, die als „Generation", „Usage", „Attribution", „Association", „Responsibility" und „Derivation" beschrieben werden. Die Arbeitsgruppe ist sehr aktiv und die Dokumente sind noch im Fluss, der Arbeitsplan (siehe http://www.w3.org/2011/prov/wiki/Deliverables) wirkt ambitioniert und ist auf 18 Monate angelegt. Neben einem konzeptionellen und einem formalen Model der „Provenance Information Language" (PIL) sind zusätzliche informelle Dokumente geplant, wie eine Einführung in die Sprache und in beste Vorgehensweisen.

14 Provenienz und Workflow

In der letzten Zeit ist im Zusammenhang mit der Erzeugung von Provenienzinformation zunehmend auf Workflow-Engines geschaut worden. Aus der immer größer werdenden Zahl der entsprechenden Veröffentlichungen werden hier einige vorgestellt, um einen Eindruck von dem aktuellen Stand der Entwicklung zu vermitteln.

14.1 Provenienz und Workflow

Provenienzinformation kann sehr umfangreich werden[36]. Damit ist es aus technischen (viele Messungen in kleinen Zeitabständen, Auswertung komplexer Zusammenhänge) wie aus praktischen (möglichst wenige menschliche Eingriffe) Gründen sinnvoll, die Daten automatisch zu erheben und nur falls nötig direkte Informationen von Menschen eintragen zu lassen. Zu diesem Zweck wird eine zusätzliche Protokollschicht in die Versuchsanordnung oder die Datenmessung eingeführt, in der die Daten gesammelt und an ein Speichersystem weitergereicht werden, „provenance aware“ wird ein solches System im Umfeld des „Open Provenance Model“ genannt. Dafür kann es naturgemäß keine einheitliche Lösung geben, das muss vielmehr, in Abhängigkeit von der geforderten Granularität, an den einzelnen Modulen des System eingerichtet werden.

Die Arbeiten innerhalb eines Projektes können oft zu einzelnen Arbeitsabläufen (Workflows) zusammengefasst werden, in denen vordefinierte Daten durch verschiedene Stufen bei Erfolg zu einem Ergebnis verarbeitet werden bzw. bei Störungen oder Misserfolg ein Abbruch mit aussagekräftigen Meldungen erzeugt wird. Dazu wird normalerweise eine eigenes System des Workflow-Managements mit einer geeigneten „Workflow Engine“ eingesetzt. Solche gibt es reichlich als Open-Source-Software (siehe z.B. für Java: `http://java-source.net/open-source/workflow-engines`), und sinnvollerweise sollte die Erhebung von Provenienz-Informationen in dieses System eingebunden werden.

In einer Grid-Umgebung wird natürlicherweise mit solchen Workflow-Engines gearbeitet, da nur sie die Verteilung der Aufgaben auf die verschiedenen Speicher- und Rechenressourcen des Grid leisten können, die notwendig ist, um die einzelnen Aufgaben („Jobs“) effektiv zu bearbeiten. Insofern ist im allgemeinen und besonders in Grid-Umgebungen die Workflow-Engine der natürliche Kandidat für das Sammeln von Provenienzinformationen[37]. Diese muss die eingegebenen Daten, die darauf angewandten Prozesse und deren Ergebnisse aufzeichnen und so die Entstehung der Ergebnisse nachvollziehbar und reproduzierbar machen.

14.2 Workflow- und Datenbankprovenienz

In einigen Artikeln wird zwischen Workflow- und Datenbankprovenienz unterschieden, z.B. [Acar et al., 2010]: „Provenance has been studied extensively in both database and workflow

[36] Moreau zitiert entsprechende Quellen: „. . . in the public database Gene Ontology, the provenance of a single tuple has been observed to be 10Mb; likewise, a 250Mb database of biological data is associated with 6Gb of provenance.“ [Moreau, 2010b], S. 35

[37] „Scientific workflow systems, however, are ideally positioned to record critical provenance information that can authoritatively document the lineage of analytical results.“, [Davidson et al., 2007], S. 2

management systems, so far with little convergence of definitions or models. Provenance in databases has generally been defined for relational or complex object data, by propagating fine-grained annotations or algebraic expressions from the input to the output. [...] In contrast, workflow provenance aims to capture a complete description of evaluation – or enactment – of a workflow, and this is crucial to verification in scientific computation. Workflows and their provenance are often presented using graphical notation, making them easy to visualize but complicating the formal semantics that relates their run-time behavior with their provenance records."

Der Unterschied liegt damit einerseits im Format (Relationen oder Graphen), andererseits und vor allem in der Nutzung: wissenschaftliche Workflows werden als veränderlich angesehen und müssen den Bedürfnissen der Forschung angepasst werden, der Provenienz wird dabei eine unterstützende Rolle zugewiesen: „The use of provenance in workflow systems also differs from that in database systems. Provenance is not only used for interpreting data and providing reproducible results, but also for troubleshooting and optimizing efficiency. Furthermore, the application of a scientific workflow specification to a particular data set may involve tweaking parameter settings for the modules, and running the workflow many times during this tuning process." ([Davidson et al., 2007], S. 3)

Insgesamt wird versucht, die verschiedenen Aspekte einander näher zu bringen, z.B. beschreibt [Acar et al., 2010] eine „dataflow language", die den Bedürfnissen beider Bereiche gerecht werden soll.

14.3 Workflow und Provenienzinformationen

Die Ansprüche an „scientific workflow management systems" (SWFMSs) sind hoch, und es existieren mittlerweile verschiedene Workflow-Systeme, die die Erzeugung der Provenienzinformationen unterstützen: „Provenance models have also been developed for a variety of workflow systems, such as Chimera, Taverna, Kepler, Karma, and ZOOM; also, many other systems such as PASS and PASOA employ similar ideas." ([Acar et al., 2010], S. 1) Allerdings werden dabei nicht notwendig universelle und austauschbare Provenienzinformationen erzeugt, sondern systemabhängige: „While existing SWFMSs support a wide range of proprietary provenance modeling, collection, representation, and management techniques, the Open Provenance Model (OPM) community initiative provides a basis for provenance exchange across those systems, and a W3C Provenance Working Group is actively pursuing specification of provenance for online systems1. A number of SWFMSs today support mappings between their proprietary provenance models and the OPM." ([Lim et al., 2010], S. 130)

Als Alternativen stehen allerdings auch selbständige Provenienzsystem wie Karma 2 (see [Simmhan et al., 2008]) oder OPMProv (see [Lim et al., 2010]) bereit.

14.4 Provenienzinformation und wissenschaftliche Workflows

Die Bereiche, die intensiv mit Workflows arbeiten, sind typischerweise wissenschaftsorientiert und oft auf verteilte Systeme ausgerichtet. Für einige der hier benutzten Systeme haben sich auch Interessensgruppen gebildet, die die Provenienzproblematik bearbeiten wollen, z.B.

für Taverna (http://www.mygrid.org.uk/dev/wiki/display/provenance/Home) oder Kepler (https://kepler-project.org/developers/interest-groups/provenance-interest-group

Außerdem haben sich mittlerweile einige Tagungen mit dem Thema „Provenance and scientific workflows“ befasst, und es ist ein (schon erwähntes) Sonderheft dazu erschienen:

- Symposium on Provenance in Scientific Workflows, October 13-17 2008, University of Utah, Salt Lake City, USA (http://wiki.esi.ac.uk/ProvenanceInWorkflows)
- Semantic Web in Provenance Management Series, SWPM-2009 und SWPM-2010 (http://ceur-ws.org/Vol-526/,http://ceur-ws.org/Vol-670/), SWPM-2012 geplant für 27. oder 28. Mai in Heraklion, Griechenland
- USENIX Workshop Series on the Theory and Practice of Provenance, 2009 – 2011 (http://www.usenix.org/events/tapp09/-/tap11), der vierte Workshop wird für den 14.-15. Juni 2012 in Boston, MA, USA geplant.
- International Journal of Computers and Their Applications, Volume 18, Number 3, September 2011: Special Issue on Scientific Workflows, Provenance and their Applications

14.5 Provenienzinformation und Sicherheit

Mit zunehmender praktischer Realisierung von Provenienzsystemen muss der Sicherheit dieser Systeme größere Aufmerksamkeit geschenkt werden. Dabei geht es weniger um die „klassische“ Frage, dass die Provenienzinformationen selbst nicht verfälscht werden dürfen, sondern um die Behandlung sensibler Daten in einem Provenienzsystem: „In many cases, both data products and their provenance can be sensitive and effective access control mechanisms are essential to protect their confidentiality.“ ([Chebotko et al., 2008], S. 349) Eine mögliche Lösung des Problems besteht darin, verschiedene Sichten auf die Provenienzdaten zu erzeugen, die den einzelnen Nutzern nur die ihnen erlaubten Informationen zugänglich machen soll, das erfordert natürlich dann ein ausgefeiltes System der entsprechenden Rollen.

Es geht aber auch um die Prozesse selbst: „[I]f intermediate data is revealed with any meaningful precision over a large number of executions of the workflow, the user may learn the behavior of private modules in the workflow.“ ([Davidson et al., 2010], S. 1) Damit würde es möglich, die technischen Geheimnisse des benutzten Moduls auszuspionieren, hier wird also zwischen „Module Privacy“ und „Data Privacy“ unterschieden, zusätzlich wird noch „Provenance Privacy“ betrachte, in der es darum geht, dass der gesamte Arbeitsablauf nicht allgemein sichtbar gemacht werden soll.

15 Zusammenfassung

Hier wird versucht, einen Überblick über die verschiedenen Herangehensweisen zu geben.

15.1 Metadata Framework

„Provenance Information“ wird als „Event“ mit den Teilfeldern erfasst:
Designation, Procedure, Date, Responsible agency, Outcome, Note.
Das „Metadata Framework“ wird hier im wesentlichen als Vorarbeit zu PREMIS erwähnt.

15.2 Open Provenance Architecture

Das „Open Provenance Project“, das im Rahmen der 6. Rahmenprogramms der EU gefördert wurde, hat grundlegende Begriffe und Verfahrensweisen für den Umgang mit Herkunftsinformationen im Grid bereitgestellt, mit dem Ziel „Enabling and Supporting Provenance in Grids for Complex Problems“. Es wurden umfangreiche Anforderungsprofile erstellt und sowohl Verfahren und Protokolle als auch Software entwickelt, mit denen Herkunftsmetadaten erstellt und verwaltet werden können.
Insbesondere wird auch ein Datenmodell entwickelt, der „*p-structure* data type“, das in verschiedenen Dokumenten weiterverwendet und entwickelt wird. Das Modell wird als XML-Schema präsentiert.

15.3 OGF Usage Record

Grundsätzliches Object der Beschreibung ist die „resource“, dazu gibt es „job on that resource“, die auch zusammengesetzt („be batch (i.e. queued) or interactive“) betrachtet werden können.
Es gibt 18 “Base Properties„:
RecordIdentity, GlobalJobId, LocalJobId, ProcessId, LocalUserId, GlobalUsername, JobName, Charge, Status, WallDuration, CpuDuration, EndTime, StartTime, MachineName, Host, SubmitHost, Queue, ProjectName
sowie zusätzliche Erweiterungen:
„Differentiated Properties, Meta Properties, Extensions, Aggregation“,
insgesamt ergeben sich etwa 100 Felder.

15.4 DELOS Provenance

Für DELOS gilt: Provenance ist ein „Content Quality Parameter“, der beeinflusst wird durch „Metadata“, „Annotation“, „Preservation Policy“ und das „Information Object“

15.5 The Origin of Data

In dieser Dissertation fließen Vorarbeiten aus dem PASOA und dem EU Provenance-Projekt zusammen und werden zu einem einheitlichen System zur Erfassung, Verarbeitung, Speicherung und Recherche von Herkunftsinformationen zusammengefasst. Die Benutzung des „*p-structure* data type“ erscheint allerdings recht kompliziert.
Es wäre sinnvoll, dieses mit dem „Open Provenance Model“ zu vergleichen.

15.6 PREMIS

Premis benutzt „Object“, „Event“, „Agent“ und „Rights“. Vorgänge werden durch entsprechende „Entity semantic units“ beschrieben, die für die verschiedenen Einheiten unterschiedlich stark ausdifferenziert sind: Es gibt für Object 95, für Event 19, für Agent 5 und für Rights (vor allem in der neuen Version ergänzt) 36 semantische Einheiten.

15.7 Open Provenance Model

Das „Open Provenance Model“ kodiert Herkunftsinformationen in einem gerichteten annotierten Graphen. Der Graph besitzt drei Arten von Knoten: Artifact, Process, Agent. Beziehungen („Causal Relationships“) werden als Kanten abgebildet: used, was generated by, was triggered by, was derived from, was controlled by.

Die Zusammenhänge werden wie folgt beschrieben: process used an artifact, an artifact was generated by a process, a process was triggered by a process, an artifact was derived from an artifact, and a process was controlled by an agent.

15.8 Core Scientific Metadata Model

Das „Core Scientific Metadata Model“ beschreibt „Untersuchungen“ („Investigations“): Experiments, Measurements, Simulations. Nur Teilaspekte von Provenienz werden unter allgemeineren Begriffen (z.B. „Dataset“) erfasst.

15.9 Provenance Vocabulary

Das „Provenance Vocabulary“ stellt einen umfassenden Begriffsapparat zur Erfassung von Herkunftsinformationen bereit, insbesondere im Kontext von „Linked Data“.
„The general terms include classes for the general types of provenance elements: Actor, Execution and Artifact. Actor has sub-classes HumanActor and NonHumanActor; Artifact has sub-classes DataItem and File.“
Als Modell wird ein RDF-Graph vorgeschlagen.

15.10 The Foundations for Provenance on the Web

Der Übersichtsartikel „The Foundations for Provenance on the Web“ gibt einen Überblick über die Literatur zu Herkunftsinformationen und fasst deren Ergebnisse zusammen. Er empfiehlt den „Open Provenance View“ und als Modell das „Open Provenance Model“.

15.11 DCMI Metadata Provenance Task Group

Dies ist ein relativ neuer Versuch, zu (Dublin Core-) Metadaten Herkunftsinformationen hinzuzufügen, diese werden unter dem Begriff „Metametadaten“ behandelt.

15.12 W3C Provenance (Incubator) Group

Die „W3C Provenance Incubator Group“ hat ein Jahr lang über Herkunftsinformationen „gebrütet“ und relevante Dokumente zum „State of the Art“ und um Vergleich verschiedener Datenmodelle vorgelegt. Insbesondere hat sie herausgearbeitet, dass das „Provenance Problem“ innerhalb der Internetstruktur nicht gelöst ist und eine „Roadmap“ von acht Punkten vorgelegt, die zur Erstellung eines „Provenance Frameworks“ abgearbeitet werden müssen, deren erste drei Punkte die höchste Priorität verdienen. Als „W3C Provenance Group“ wird die Arbeit aktiv fortgesetzt. Es wurde ein Arbeitsplan erstellt und derzeit werden die verschiedenen dort formulierten Deliverables abgearbeitet.

15.13 Abschlussbemerkungen

Trotz mehrerer Projekte, Wettbewerbe und einer Vielzahl von Veröffentlichungen hat sich in der Praxis noch kein allgemeiner Standard für die Aufnahme, Verarbeitung, Speicherung und den Austausch von Herkunftsinformationen herausgebildet. Klar sind elementare Standards:

- Jeder Prozess erzeugt lokale Informationen zum Fluss von Daten und den damit verbundenen Akteuren und Prozessen.
- Die Prozesse werde zu Jobs gebündelt (bzw. als solche aufgerufen), diese Jobs müssen gleichzeitig die Herkunftsinformationen der Prozesse verarbeiten (ist aber in der JSDL noch nicht vorgesehen).
- Jobs werden innerhalb von Arbeitsabläufen („Workflows“) von „Workflow Enactment Engines“ ausgeführt, diese müssen die Informationen der unteren Ebenen zusammenfassen.
- Herkunftsinformationen werden in einem dafür bereitgestellten spezifischen „Provenance Store“ gespeichert und können von dort mit unterschiedlicher Granularität und unterschiedlichen Sichten abgerufen werden.

Zunehmend wird dabei allerdings auf dem „Open Provenance Model“ ([Luc Moreau (Editor), 2009]) aufgebaut.

Zur Verwirklichung dieser Vorstellung fehlen einige wesentliche Bausteine:

- Ein allgemein akzeptiertes Modell zur Speicherung oder zumindest zum Austausch von Herkunftsdaten, allerdings wird zunehmend mit OPM gearbeitet.
- Regeln und Methoden zur Erfassung der Herkunftsdaten auf allen Stufen der Prozessverarbeitung.
- Ein stabiles und anerkanntes System für den „Provenance Store“ inklusive der notwendigen Zugriffsteuerung.

In Abhängigkeit von den Schwerpunkten verschiedener Communities (Grid, Database, Provenance) werden derzeit unterschiedliche Herangehensweisen und Datenmodelle vorgeschlagen und Wege in verschiedenen Richtungen verfolgt, verschiedene Daten- und Datenbanksysteme auf ihre Brauchbarkeit hin abgeklopft und Lösungen für einzelne Probleme vorgeschlagen.

Der Übersichtsartikel von Luc Moreau und die aktuellen Arbeiten der W3C Provenance Group zeigen, dass einerseits aktiv auf diesem Feld gearbeitet wird, sich aber andererseits noch keine stabilen einheitlichen Herangehensweisen herausgebildet haben. Andererseits zeigen die Veröffentlichungen im Bereich Provenienz und Workflows, dass zunehmend an der Realisierung von Provenienzsystemen gearbeitet wird, die in der wissenschaftlichen Praxis eingesetzt werden können.

Literatur

Umut Acar, Peter Buneman, James Cheney, Natalia Kwasnikowska, Jan Van den Bussche, and Stijn Vansummeren. A graph model of data and workflow provenance, 2010. URL http://www.usenix.org/event/tapp10/tech/full_papers/buneman.pdf.

Árpád Andics (editor). User requirements document, 2005a. URL http://www.gridprovenance.org/deliverables/GRID_PROVENANCE-UserRequirements-D211-Month6.pdf.

Árpád Andics (editor). Software requirements document, 2005b. URL http://www.gridprovenance.org/deliverables/GRID_PROVENANCE-SoftwareRequirements-D221-Month6.pdf.

G. Athanasopoulos, L. Candela, D. Castelli, P. Innocenti, Y. Ioannidis, A. Katifori, A. Nika, G. Vullo, and S. Ross. The digital library reference model, 2010. URL http://www.dlorg.eu/uploads/DLReferenceModels/TheDigitalLibraryReferenceModel_v1.0.pdf.

Khalid Belhajjame, Reza B'Far, Stephen Cresswell, Yolanda Gil, Paul Groth, Graham Klyne, Jim McCusker, Simon Miles, James Myers, Satya Sahoo, and Curt Tilmes. Prov-dm: The prov data model, 2012. URL http://dvcs.w3.org/hg/prov/raw-file/default/model/prov-dm.html.

Artem Chebotko, Seunghan Chang, Shiyong Lu, Farshad Fotouhi, and Ping Yang. Secure scientific workflow provenance querying with security views. In *The Ninth International Conference on Web-Age Information Management, 2008. WAIM '08.*, pages 349–356. IEEE, IEEE Conference Publications, 2008. doi: 10.1.1.155.4249.

James Cheney, Yolanda Gil, Paul Groth (Editor), and Simon Miles. Requirements for provenance on the web, 2010. URL http://www.w3.org/2005/Incubator/prov/wiki/User_Requirements.

Susan Davidson, Sarah Cohen-Boulakia, Anat Eyal, Bertram Ludäscher, Timothy McPhillips, Juliana Freire, Shawn Bowers, Manish Kumar Anand Susan Davidson, Sarah Cohen-Boulakia, Anat Eyal, Bertram Ludäscher, Timothy McPhillips, Juliana Freire, Shawn Bowers, and Manish Kumar Anand. Provenance in scientific workflow systems, 2007. URL http://sites.computer.org/debull/a07dec/susan.pdf.

Susan B. Davidson, Sanjeev Khanna, Sudeepa Roy, and Sarah Cohen Boulakia. Privacy issues in scientific workflow provenance. In *Proceedings of the 1st International Workshop on Workflow Approaches to New Data-centric Science*. ACM, 2010. ISBN 978-1-4503-0188-6. URL http://www.cis.upenn.edu/~sudeepa/wands10-wf-privacy.pdf.

DCMI Metadata Provenance Task Group. DCMI metadata provenance task group wiki, 2011. URL http://wiki.bib.uni-mannheim.de/dc-provenance/doku.php.

DELOS Projekt. Delos digital library reference model, 2007. URL http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf.

Kai Eckert and Michael Panzer. DCMI metadata provenance task group, 2010. URL http://www.dublincore.org/groups/provenance/.

Kai Eckert, Magnus Pfeffer, and Johanna Völker. Towards interoperable metadata provenance, 2010. URL http://ki.informatik.uni-mannheim.de/fileadmin/publication/eckert10provenance.pdf.

EU Grid Provenance Project. Open provenance specification, 2005. URL http://www.gridprovenance.org/openSpecification/.

Yolanda Gil, James Cheney, Paul Groth, Olaf Hartig, Simon Miles, Luc Moreau, and Paulo Pinheiro da Silva. Provenance xg final report, 2010. URL http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/.

Paul Groth, Sheng Jiang, Simon Miles, Steve Munroe, Victor Tan, Sofia Tsasakou, and Luc Moreau. An architecture for provenance systems, 2006. URL http://eprints.ecs.soton.ac.uk/13216/1/provenanceArchitecture10.pdf.

Paul Groth, James Cheney, Simon Miles, James Myers, and Yolanda Gil. State of the art report, 2010. URL http://www.w3.org/2005/Incubator/prov/wiki/State_of_the_Art_Report.

Paul T. Groth. The origin of data. Enabling the determination of provenance in multi-institutional scientific systems through the documentation of processes, 2007. URL http://eprints.ecs.soton.ac.uk/14649/.

Olaf Hartig and Jun Zhao. Publishing and consuming provenance metadata on the web of linked data, 2010a. URL http://olafhartig.de/files/HartigZhao_Provenance_IPAW2010_Preprint.pdf.

Olaf Hartig and Jun Zhao. Provenance vocabulary core ontology specification, 2010b. URL http://sourceforge.net/apps/mediawiki/trdf/index.php?title=Provenance_Vocabulary.

Chunhyeok Lim, Shiyong Lu, Artem Chebotko, and Farshad Fotouhi. Storing, reasoning, and querying opm-compliant scientific workflow provenance using relational databases. *Future Generation Computer Systems*, 2010. doi: 10.1016/j.future.2010.10.013. URL http://dev.viewsystem.org/~chlim/publications/opmprov.pdf.

Luc Moreau (Editor). The open provenance model core specification (v1.1), 2009. URL http://eprints.ecs.soton.ac.uk/18332/1/opm.pdf.

Brian Matthews, Shoaib Sufi, Damian Flannery, Laurent Lerusse, Tom Griffin, Michael Gleaves, and Kerstin Kleese. Using a core scientific metadata model in large-scale facilities. *The International Journal of Digital Curation*, 5(1):106–118, 2010. URL http://www.ijdc.net/index.php/ijdc/article/viewFile/149/211.

Luc Moreau. The foundations for provenance on the web. *Foundations and Trends in Web Science*, 2(2-3):99–241, 2010a. doi: 10.1561/1800000010. URL http://www.nowpublishers.com/product.aspx?product=WEB&doi=1800000010.

Luc Moreau. The foundations for provenance on the web. 2010b. URL http://eprints.ecs.soton.ac.uk/21691/1/survey.pdf.

Luc Moreau and Paolo Missier (edts.). The PROV data model and abstract syntax notation, 2012. URL http://www.w3.org/TR/prov-dm/.

Steve Munroe, Paul Groth, Sheng Jiang, Simon Miles, Victor Tan, Luc Moreau, John Ibbotson, and Javier Vazquez. Data model for process documentation, 2006. URL http://eprints.soton.ac.uk/263200/1/ws-prov-dm.pdf.

OCLC/RLG Working Group on Preservation Metadata. Preservation metadata and the oais information model. a metadata framework to support the preservation of digital objects, 2002. URL http://www.oclc.org/research/projects/pmwg/pm_framework.pdf.

Open Grid Forum. URL http://www.ogf.org/.

Open Grid Forum: Usage Group. Usage record – format recommendation, 2006. URL http://www.ogf.org/documents/GFD.98.pdf.

PREMIS Maintenance Activity (Library of Congress). Premis data dictionary for preservation metadata (version 2.0), 2008. URL http://www.loc.gov/standards/premis/v2/premis-2-0.pdf.

Satya Sahoo, Paul Groth, Olaf Hartig, Simon Miles, Sam Coppens, James Myers, Yolanda Gil, Luc Moreau, Jun Zhao, Michael Panzer, and Daniel Garijo. Provenance vocabulary mappings, 2010. URL http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings.

Yogesh L. Simmhan, Beth Plale, and Dennis Gannon. Karma2: Provenance management for data driven workflows. *International Journal of Web Services Research*, 5(2), 2008. URL http://d2i.indiana.edu/sites/default/files/Simmhan-JWSR-07.pdf.

W3C Provenance Group. W3C provenance group wiki, 2011. URL http://www.w3.org/2011/prov/wiki/Main_Page.

W3C Provenance Incubator Group, 2009. URL http://www.w3.org/2005/Incubator/prov/.

W3C Provenance Incubator Group. Provenance dimensions, 2010a. URL http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Dimensions.

W3C Provenance Incubator Group. Requirements, 2010b. URL http://www.w3.org/2005/Incubator/prov/wiki/Requirements.