Analyse von technischen Provenance-Modellen und Evaluation der Auswirkungen auf die Interoperabilität von Werkzeugen

(R 1.3.3)

Version 20. Februar 2013
Arbeitspaket 1.3.3
Verantwortlicher Partner SUB, GWDG, MPDL

DARIAH-DE
Aufbau von Forschungsinfrastrukturen für die e-Humanities

Dieses Forschungs- und Entwicklungsprojekt wird / wurde mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF), Förderkennzeichen 01UG1110A bis M, gefördert und vom Projektträger im Deutschen Zentrum für Luft- und Raumfahrt (PT-DLR) betreut.

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Projekt: DARIAH-DE: Aufbau von Forschungsinfrastrukturen für die e-Humanities
BMBF Förderkennzeichen: 01UG1110A bis M
Laufzeit: März 2011 bis Februar 2014

Dokumentstatus: <**Entwurf**, Final>
Verfügbarkeit: <öffentlich, **DARIAH-DE-intern**>
Autoren:

Stefan E. Funk (SUB Göttingen), Daniel Kurzawe (GWDG), Bastien Saquet (MPDL), Stefan Schmunk (SUB Göttingen)

Revisionsverlauf:

Datum	**Autor**	**Kommentare**
22.11.2012	Stefan E. Funk	Erste Version
17.12.2012	Stefan E. Funk	Thomas Fischers Arbeiten eingefügt, in Google Docs formatiert
19.12.2012	Stefan Schmunk, Stefan E. Funk, Bastien Saquet	Outline
17.01.2013	Bastien Saquet	Modelle Beschreibung hinzugefügt
28.01.2013	Bastien Saquet	Service-Konzept hinzugefügt
30.01.2013	Stefan E. Funk	Erweiterung Service-Konzept und Interoperabilität
01.02.2013	Stefan E. Funk, Bastien Saquet, Daniel Kurzawe	Überarbeitung
05.02.2013	Stefan E. Funk, Bastien Saquet, Daniel Kurzawe, Stefan Schmunk	Überarbeitung, Erstellung V0.9

15.02.2013	Stefan E. Funk, Harald Lordick, Bastien Saquet, Wibke Kolbmann, Daniel Kurzawe, Stefan Schmunk	Revision und Fertigstellung 1.0
20.02.2013	Martina Kerzel, Bastien Saquet, Stefan Schmunk	Blumenbach und letzte Revision

Inhalt

1. Einleitung

Dieser Report betrachtet die technischen Anforderungen bzgl. der Provenienz von geisteswissenschaftlichen Daten innerhalb der DARIAH-DE Infrastruktur. Dabei wird insbesondere die Provenienz bei der Erhebung und Verwaltung von Daten analysiert, um exemplarisch die Frage beantworten zu können, welche Bedeutung Provenienz besitzt und welche Anforderungen in technischer und informationswissenschaftlicher Hinsicht von Geisteswissenschaftlern an Provenienzkonzepte gestellt werden. Der Fokus dieses Dokuments liegt deshalb auf folgenden Punkten:

- Definition von Provenienz im Kontext von DARIAH
- „State-of-the-art“ von Provenienzmodellen
- Konzept eines DARIAH Provenienz-Dienstes

1.1 Definition: Provenienz

Die generische Definition von Provenienz lautet:

> Ein „Bereich, aus dem jemand, etwas stammt; Herkunft“[1].

Im Informationssystem betrifft die Provenienz alles, das etwas über die Herkunft eines Objekts zu einem bestimmten Zeitpunkt aussagen kann, wie z.B.:

- die vorherige Version dieses Objektes
- die Transformation, die informationstechnologisch eingesetzt wurde
- die Personen und ihre Rollen im Herkunfts- bzw. Herstellungsprozess
- die Quellen, Daten, Objekte, die involviert waren

Alle diese Daten werden als Provenienzmetadaten begriffen. Die Aufgaben und Funktionen solcher Daten sind vielseitig. Sie werden aus rechtlichen Gründen genutzt, um die Verantwortung und die Rechte von Personen an Daten, so z.B. Forschungsdaten, dokumentieren zu können. Provenienzmetadaten ermöglichen die Nachnutzung von Objekten und Daten. In der Tat ist das Wissen über die Herkunft von Daten notwendig, um die Qualität und den Wert dieser Daten evaluieren und beurteilen zu können. Dieser Aspekt spielt eine immer größer werdende Rolle in der Langzeitarchivierungsinfrastruktur. Provenienzdaten können den Erstellungsprozess von Objekten beschreiben und damit

[1] Vgl. Duden. Provenienz. http://www.duden.de/rechtschreibung/Provenienz

sowohl die Validierung als auch die Authentizität von Ergebnissen sichern. Forschungseinrichtungen und Förderorganisationen legen im Rahmen der Empfehlungen zur "guten wissenschaftlichen Praxis"zunehmend Wert auf Provenienzangaben, etwa im Hinblick auf eine grundsätzliche Qualitätssicherung.[2]

Diese Daten sind also kritische Daten. Die Authentizität der Provenienzmetadaten soll sicher gestellt bleiben, um sie vor Verfälschung zu schützen. Außerdem sollen diese Daten vor einer fremden Nutzung und daraus resultierenden Rechtsverstößen geschützt werden (z.B. Benutzung von Personendaten).

1.2 Provenienz in den Geisteswissenschaften

Für Geisteswissenschaftler ist das Thema Provenienz von zentraler Bedeutung: Es betrifft Informationen, Quellen, Forschungs- und Metadaten, die beliebigen Quellen entstammen können und die zugleich die gesamte Bandbreite an geisteswissenschaftlichen Forschungsdaten mit all ihrer semantischen Vielfalt abdecken können: geisteswissenschaftliche analysierbare Texte, Handschriften, Fotografien und Gemälde, aber auch statistisch-empirisches Datenmaterial, soziologisch interpretierbare Umfragen und Interviews, medizinische und naturwissenschaftliche Messungen, technische Zeichnungen – die Liste ließe sich beliebig erweitern. Wir bezeichnen einen Komplex aus Speichermedien und Software dann als "Provenienzsystem", wenn es in der Lage ist, Provenienzinformationen zu diesem Projekt bzw. zu Forschungsfragen aufzunehmen, zu verwalten und entsprechende Recherchen zu unterstützen bzw. zur Beantwortung diesbezüglicher Fragen zu führen. Dieses System muss zugleich die Sicherheit gewährleisten, die für den gegebenen Arbeitsbereich sowohl von Entwicklern als auch geisteswissenschaftlichen Nutzern gefordert wird.

Es kann rechtliche Auflagen, ethische Gründe oder die Forderung nach wissenschaftlicher Überprüfbarkeit geben, die eine Dokumentation der Provenienz erforderlich machen. Gerade der zuletzt genannte Aspekt ist für geisteswissenschaftliche Forschung, unabhängig von der disziplinären Verortung, von elementarer Bedeutung, da nur auf diese Art und Weise die Validität von Forschungsergebnissen überhaupt überprüft werden kann. In Abhängigkeit von solchen Forderungen ergeben sich unterschiedliche Anforderungen an das System, was z.B. die Zugriffskontrolle oder die Granularität der Aufzeichnung betrifft. Zudem kann es auch für die Untersuchung wissenschaftlicher Fragestellungen von Bedeutung sein, Informationen über die Herkunft von Forschungsdaten zur Verfügung zu

[2] Deutsche Forschungsgemeinschaft. Sicherung guter wissenschaftlicher Praxis. http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf (7. Februar 2013)

stellen. Gerade bei geistes- und sozialwissenschaftlichen Fragestellungen hat die Kenntnis der Herkunft der erhobenen Forschungsdaten durchaus Konsequenzen für die wissenschaftliche Arbeit.

In den Geisteswissenschaften ist nicht nur der Forschungsgegenstand, sondern die Entstehungsgeschichte von Quellen und Forschungsdaten von zentraler Bedeutung für den Forschungsprozess. Quellenkritik ist etwa in der Geschichtswissenschaft elementarer Bestandteil eines jeden Forschungsprozesses, unabhängig von Forschungsinteresse und individueller Forschungsfrage. Hierbei werden – und dies gilt generisch für alle geisteswissenschaftlichen Fachdisziplinen – Forschungsdaten nicht nur als Forschungsobjekte betrachtet, sondern auch Aspekte der Entstehung und Bearbeitung kontextualisiert, hinterfragt und in den Forschungsprozess integriert. Gerade aus diesen Gründen hat Provenienz in den Geisteswissenschaften eine zentrale methodische Bedeutung, die fachgerechte Analyse der Forschungsdaten bedarf zwingend dieses Hintergrunds. Das bekannte und augenfällige Beispiel soll hier diese besondere geisteswissenschaftliche Relevanz verdeutlichen:

Alexander Maltschenko, einer der Weggefährten Lenins und russischer Revolutionär, wurde im Zuge der Stalinistischen Säuberungen Ende der 1920er Jahre inhaftiert und danach systematisch aus dem kollektiven Gedächtnis entfernt. Hierzu wurden Einträge in Enzyklopädien gelöscht und zugleich wurden Fotografien retuschiert, die ihn beispielsweise gemeinsam mit Lenin gezeigt hatten. Ähnliches geschah mit Leo Trotzki, dem zwar 1929 die Flucht gelang, aber dessen Beteiligung an der russischen Revolution systematisch im kollektiven Gedächtnis der russischen Bevölkerung unterdrückt wurde. Für geisteswissenschaftliche Forscher ist es deshalb von großer Bedeutung, sich zu vergegenwärtigen, dass Quellen “gemacht” sind, im Extremfall, wie in diesem Beispiel, zensiert und manipuliert sein können. Das Wissen um Enstehungs- und Bearbeitungszusammenhänge von Quellen ist deshalb unverzichtbar, und stellt sich im digitalen Workflow um so dringlicher.

1.3 Provenienz in DARIAH-DE

DARIAH unterstützt die Geistes- und Kulturwissenschaften durch den Aufbau einer Infrastruktur für digitale Forschungsobjekte und Verfahren. Dazu soll ein generisches Provenienzmodell geschaffen werden, das Provenienzinformationen sowohl von traditionellen als auch digitalen Forschungsdaten umfasst. Der Provenienz-Dienst soll sich gleichzeitig durch Kompatibilität zu weiteren DARIAH- und externen digitalen Diensten auszeichnen (siehe den Absatz zur Interoperabilität).

1.4 Out of scope

Dieses Dokument skizziert eine technische Spezifikation für einen DARIAH-Provenienz-Dienst. Es beschreibt ein Konzept, das, um implementiert werden zu können, eine zusätzliche technische Evaluierung und Umsetzung benötigt. Eine Technologie (Provenienzmodell, Datenbank, etc.) wird in diesem Report nicht festgelegt. Deshalb wurden nur technologische Aspekte berücksichtigt, die spezifisch von der DARIAH-Infrastruktur benötigt werden. Auf diesem hier beschriebenen DARIAH-Provenienz-Dienst könnte auch eine Benutzeroberfläche realisiert werden, um die Provenienzmetadaten zu verwalten. Dies ist aber nicht Bestandteil des Konzepts, das sich auf technische Schnittstellen beschränkt, die für andere Komponenten der DARIAH-Infrastruktur zur Verfügung stehen.

2 Nutzungs-Szenarien

Im DARIAH-DE Antrag wird ein mögliches Nutzungsszenario wie folgt beschrieben:

> *"Thanks to the technical environment developed in the VCC e-Infrastructure of DARIAH, Daniela benefits from both a single sign-on environment for accessing all the necessary digital assets, as well as a virtual portfolio where she can gather all selected sources. The metadata and Provenienz-Metadaten associated with an item allows her to evaluate the authenticity and value of an asset, and to link to related items (e.g. other sources for the same asset, other formats, research addressing that asset) once she has found something of interest. From this portfolio, she can publish geographical views on her data, which she exchanges and discusses with other colleagues in Europe."* [3]

2.1 Provenance Data Repository

Provenienzdaten können in DARIAH-Anwendungen erstellt und innerhalb eines DARIAH Provenienz-Repositorys aufgenommen werden. Über dieses Repository können Daten in die Langzeitarchivierung überführt werden. Das Repository bietet dabei Standardschnittstellen, um Provenienzdaten zu erzeugen, zu visualisieren, zu suchen und zu verteilen.

Daten dieser Anwendung werden von anderen DARIAH-Diensten verwendet: Die Collection Registry registriert eine neue Sammlung, die DARIAH-Suche indexiert die dort erschlossenen Daten, PIDs werden vergeben. Alle diese Dienste benutzen das Provenienz Repository, so dass "DARIAH Provenienz Metadaten" für alle Objekte zentral verwaltet

[3] Vgl. DARIAH-DE-Antrag, Seite 24f., The Digital Postgraduate.

sind. Alle Transformationen dieser Objekte lassen sich so über die gesamte DARIAH-Infrastruktur nachverfolgen.

2.2 Provenance Data Extractor

Ein Service der DARIAH Infrastruktur will Provenienzmetadaten aus einem File (Bild, Video, PDF, etc.) extrahieren. Der DARIAH Provenance Data Extractor bietet eine entsprechende Schnittstelle an, die die Provenienzmetadaten aus der Datei ausliest und im angeforderten Format (XML, RDF, JSON, etc.) an den Service zurückliefert. Zur Speicherung der Provenienzmetadaten kann der Service das Provenance Repository benutzen.

2.3 Johann Friedrich Blumenbach – online

In dem wissenschaftshistorischen Forschungsprojekt Blumenbach-online[4] geht man derzeit davon aus, dass sich anhand der Visualisierung von Provenienzdaten von Sammlungsobjekten neue Forschungsfragen ergeben. Johann Friedrich Blumenbach gehörte im 18. Jh. zu denjenigen Göttinger Forschern, die das Forschen am Objekt zu einem integralen Bestandteil der Lehre und der eigenen Forschungsarbeit machten. J.F. Blumenbach baute während seiner Lehrtätigkeit an der Göttinger Universität eine mehrere Tausend Stück umfassenden Naturaliensammlung auf, die im Zuge der Diversifikation der Wissenschaftsdiziplinen über nicht nur über den Göttinger Campus, sondern auch auf weitere Universitäten im internationalen Raum verteilt wurden[5]. Von diesen Sammlungsobjekten konnte das Projekt mittlerweile rund 6.100 Stücke identifizieren.
Einst gelangten diese Stücke über das zeitgenössische Gelehrtennetzwerk in den Besitz Blumenbachs bzw. in den Besitz der Göttinger Universität. So waren es z.B. Schüler wie Alexander von Humboldt, Freunde wie Johann Wolfgang Goethe oder Kollegen wie Baron von Asch, die von ihren Reisen Gegenstände, wie z.B. getrocknete Pflanzen, Kultgegenstände, Gesteinsproben, versteinerte Knochen, Tiere, Zeichnungen und vieles mehr mitbrachten oder nach Göttingen sendeten. Andere Gegenstände fanden erst über verschiedene Stationen ihren Platz innerhalb der Blumenbachschen Naturaliensammlung. Auskunft darüber geben die aufbewahrten Etikettenserien der Sammlungsobjekte sowie Kataloge und der Schriftwechsel Blumenbachs.
Im Rahmen der Digitalisierung von Blumenbachschen Schriften und Sammlungsobjekten

[4] http://www.blumenbach-online.de/

[5] Vgl. Wolf Lepenies, *Das Ende der Naturgeschichte. Wandel kultureller Selbstverständlichkeiten in den Wissenschaften des 18. und 19. Jahrhunderts*, 1976; Dietrich von Engelhardt, *Historisches Bewußtsein in der Naturwissenschaft von der Aufklärung bis zum Positivismus*, 1979; Nicolaas Rupke, *The great chain of history*, 1983.

werden neben den naturwissenschaftlichen Metadaten, Angaben zu Nennungen von Objekten in Blumenbachs Schriften, Personen- und Ortsangaben ebenfalls die Provenienzdaten der Sammlungsobjekte - soweit vorhanden - erfasst. Auf diese Weise schafft das Projekt die Voraussetzung dafür, mit Hilfe des DARIAH-Geobrowsers die Wege, die ein oder mehrere Fundstücke (z.B. einer Klasse, wie etwa Meteoriten) in einer bestimmten Zeitspanne zurückgelegt haben, zu visualisieren.
Da die einheitliche Bumenbachsche Sammlung an der Schwelle des 19. Jhs. mit der Diversifikation der naturwissenschaftlichen Disziplinen in die jeweiligen Institute verteilt wurden und darüber hinaus Fundstücke z.T. ins Ausland verkauft oder abgegegeben wurden, besteht darüber hinaus der Bedarf die Daten auch zu einem späteren Zeitpunkt zu ergänzen oder sogar zu verändern, sollte sich anhand später entdeckten Materials eine andere bzw. zu erweiternde Provenienzkette ergeben als zuvor angenommen.
Darüber hinaus ist die Community der Blumenbachforscher international aufgestellt, so dass eine Bearbeitung der Daten im Rahmen einer auf TextGrid[6] basierenden Virtuellen Forschungsumgebung realisiert wird und die Nachverfolgung der Datenhistorie aus fachwissenschaftlicher Sicht eine bedeutende Rolle spielt.

2.4 Datenerhebung – Datensammlung

Daten als Grundlage geisteswissenschaftlicher Forschung werden in den seltensten Fällen automatisch gesammelt bzw. generiert. Die Datenlage ist in den meisten Fällen heterogen, sowohl hinsichtlich der Datei- und Datenformate (Bilder, Fotografien, Karten, Noten, Musik, Beschreibungen von Sammlungsobjekten, handschriftliche oder gedruckte Texte – nur in den seltensten Fällen bereits maschinenlesbar semantisch ausgezeichnet) als auch hinsichtlich der Provenienz – aus Beständen von Archiven, Bibliotheken, privaten oder öffentlichen Sammlungen, Museen, aber auch aus eigenen Arbeiten, Erhebungen oder Messungen des Forschers (Umfragen, Interviews, Ausgrabungen etc.). Um diese Heterogenität erfassen zu können, sollte eine zusätzliche Protokollschicht in die Datenerhebung eingeführt werden, 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 keine einheitliche Lösung geben, da diese Eigenschaft, in Abhängigkeit von der geforderten Granularität, an den einzelnen Modulen des Systems ausgerichtet werden muss.

2.5 Workflow-Management

Die Arbeiten innerhalb eines Projektes können in der Regel zu Arbeitsabläufen (Workflows) zusammengefasst werden, in denen Daten durch verschiedene

[6] http://www.textgrid.de/

Bearbeitungsschritte zu einem Ergebnis verarbeitet werden (bzw. bei Störungen oder Fehlern ein Abbruch mit geeigneten Meldungen erzeugt wird). Hier kann ein eigenes System für ein Workflow-Management eingesetzt werden. Solche Workflow-Engines sind als Open-Source-Software[7] verfügbar, und sinnvollerweise sollte die Erhebung von Provenienzinformationen in ein solches Paket eingebunden werden. Systeme wie Taverna [8] oder Kepler[9] entsprechen dieser Anforderung.

3 Beschreibung und Bewertung einiger Modelle

3.1 PREMIS Data Dictionary

Das "PREMIS Data Dictionary for Preservation Metadata"[10] ist der Standard für die Beschreibung von Preservation Metadaten und enthält auch ein Modell für Provenienzmetadaten. Es wurde im Juli 2012 in Version 2.2 veröffentlicht.

Diese Fassung ist sehr ausgereift (Version 1.0 wurde im Mai 2005 veröffentlicht), zahlreiche Erfahrungen und Feedbacks wurden darin berücksichtigt. PREMIS-Tools haben meist entweder die Funktion von Repositorien oder Metadaten-Extraktoren.

3.2 W3C PROV

Das "World Wide Web Consortium"[11] (W3C) ist die Organisation zur Enwicklung und Standardisierung von Web-Technologien wie HTML, XML, RDF oder CSS. Als Standard für Provenienzmetadaten ist W3C PROV[12] entstanden.

PROV ist noch in Entwicklung, als Termin für die Fertigstellung des Modells als „Recommendation" wurde der 28. Februar 2013 genannt.[13] Der W3C PROV wird aktiv weiter entwickelt. Die aktuelle Spezifikation ist momentan erst als „Draft" bezeichnet. Wenn man die Entwicklung anderer W3C Recommendations betrachtet, ist anzunehmen, dass

[7] Open Source Workflow Engines in Java. http://java-source.net/open-source/workflow-engines
[8] Home - Provenance - myGrid developer wiki.
http://www.mygrid.org.uk/dev/wiki/display/provenance/Home
[9] Provenance Interest Group.
https://kepler-project.org/developers/interest-groups/provenance-interest-group/
[10] PREMIS Data Dictionary for Preservation Metadata.
http://www.loc.gov/standards/premis/v2/premis-2-2.pdf
[11] http://www.w3.org/
[12] http://www.w3.org/2011/prov/wiki/Main_Page
[13] http://www.w3.org/2011/prov/wiki/TimetableToRec

sich auch der W3C PROV zu einem akzeptierten Standard entwickelt. Eine Liste von Tools, die den W3C PROV bereits implementieren, steht schon zur Verfügung.[14]

3.3 Open Provenance Model (OPM)

Das „Open Provenance Model“[15] ist ein Community-Projekt zur Erstellung eines Modells für Provenienzinformationen. Die letzte Fassung ist Version 1.1, die am 27. Juli 2010 veröffentlich geworden ist[16]. Version 1.2 wurde angekündigt, jedoch bisher nicht umgesetzt [17]. Der hauptsächliche OPM-Entwickler arbeitet mittlerweile in der Arbeitsgruppe des W3C PROV. Wir können daher vermuten, dass OPM vorerst nicht weiterentwickelt wird.

3.4 Technische Bewertung

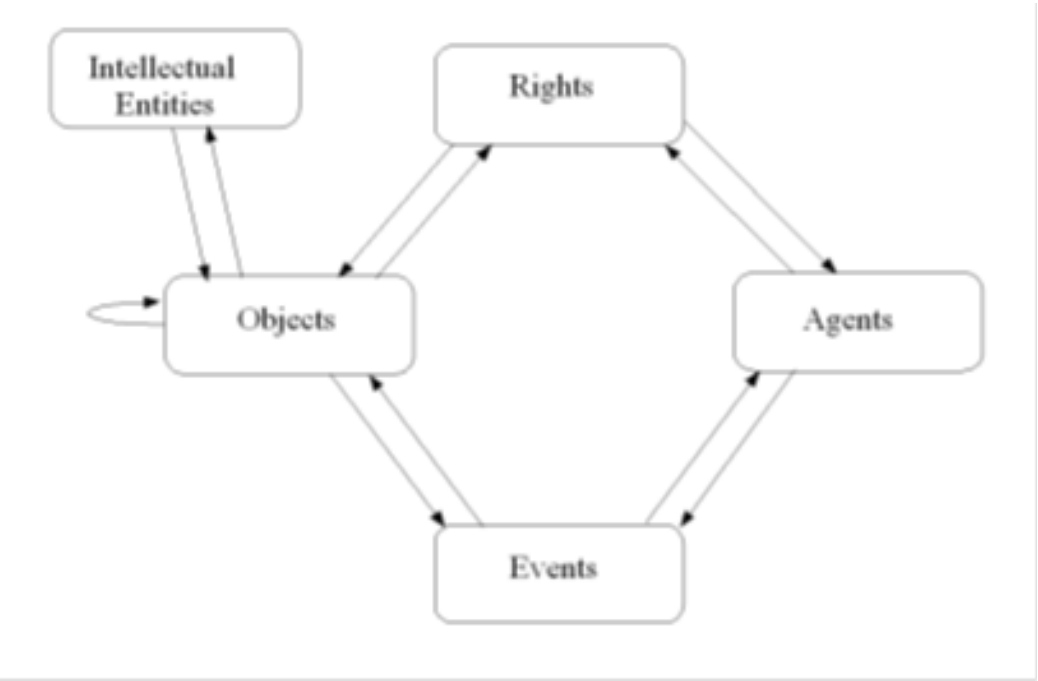
Sowohl PREMIS als auch W3C PROV – OPM wird hier aufgrund der unsicheren Entwicklungslage nicht weiter berücksichtigt – basieren auf detaillierten Datenmodellen.

3.4.1 Das PREMIS Data Dictionary

Das Provenienz-Datenmodell von PREMIS enthält die folgenden Objekte:

- **Intellectual Entity***:* Ein Set von Inhalten, die zum Management und zur Beschreibung bestimmt sind (z.B. Buch, Karte, Datenbank). Eine *Intellectual Entity* kann weitere *Intellectual Entities* beinhalten (z.B. eine Webseite mit Bildern) sowie mehrere digitale Instanzen besitzen.
- **Object**: Ein abstraktes digitales Objekt als eine Informationseinheit.
- **Event**: Eine Aktivität, die sich auf ein *Object* auswirkt oder es benutzt.
- **Agent***:* Eine Person, Organisation oder eine Software, die mit einem *Object Event* oder mit *Object Rights* verbunden ist.
- **Rights**: Eine Aussage über ein oder mehrere Rechte, die ein *Object* und/oder einen *Agent* betreffen.

[14] http://www.w3.org/2011/prov/wiki/ProvImplementations
[15] http://openprovenance.org/
[16] http://eprints.soton.ac.uk/271449/
[17] http://twiki.ipaw.info/bin/view/OPM/WorkInProgressV1pt2

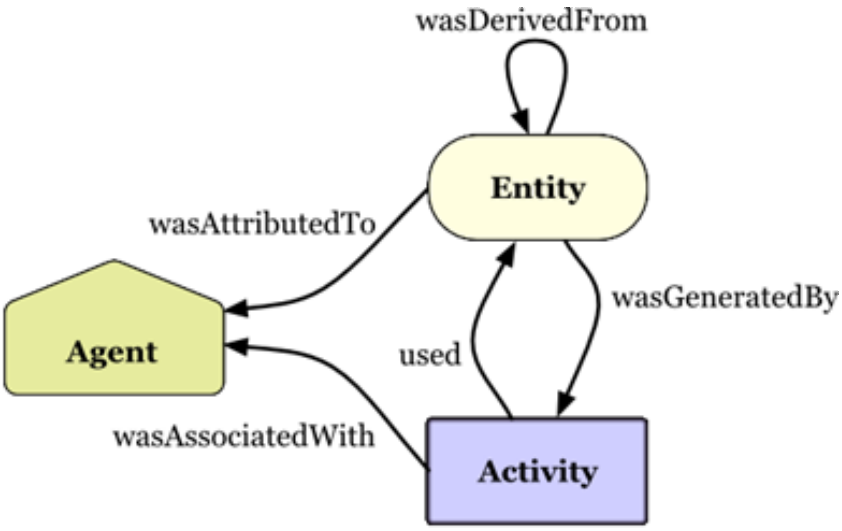


Das PREMIS Data Directory

3.4.2 Das Datenmodell W3C PROV

Das Modell des W3C PROV enthält folgende Objekte:

- **Entity**: Physische, konzeptuelle, digitalisierte und alle anderen Dinge mit festen Aspekten. *Entities* können real oder imaginär sein. Beispiel: Webseite, Dokument, Artikel, etc.
- **Activity**: Beschreibt die Erstellung bzw. die Änderung einer Eigenschaft einer *Entity*, meistens unter Benutzung einer weiteren *Entity*.
- **Agent**: Nimmt eine Rolle ein bei einer *Activity*. Ein *Agent* kann Verantwortung für eine *Activity* haben. Alle *Entities*, die Verantwortung haben können, können ein *Agent* sein.
- **Role**: Ist die Funktionsbeschreibung, die eine *Entity* oder ein *Agent* bei einer *Activity* spielt. Sie spezifiziert die Beziehung zwischen einer *Entity* und einer *Activity* und kann die Verantwortung eines *Agent* bei einer *Activity* beschreiben.
- **Derivation**: Wenn die Existenz einer *Entity* zum Teil oder ganz aus einer anderen *Entity* hervorgeht, ist die erstere eine *Derivation* der letzteren.
- **Revision**: Ist eine spezielle *Derivation*. Sie beschreibt, dass eine *Entity* mehrere Versionen hat. Jede *Revision* ist eine *Entity* und ist mit ihrem Vorgänger verlinkt.
- **Plans**: Vordefiniert Maßnahmen (Workflow).
- **Time**: Zeit.



Das Modell W3C PROV

Beide beschriebenen Datenmodelle sind generisch strukturiert, um alle Arten von Provenienzmetadaten beschreiben zu können. Das PREMIS-Modell betont rechtliche Aspekte, die aber auch mit Hilfe des W3C PROV Datenmodells beschrieben werden können. Für beide Modelle gibt es mehrere Implementationen.

Für PREMIS existiert ein XML Schema[18] und eine OWL Ontologie[19], für das W3C PROV existiert eine OWL Ontologie[20] (PROV-O) und ebenfalls ein XML Schema[21] (PROV-XML).

Für PREMIS liegt die Priorität auf dem XML Schema, während OWL hier neu ist und erst als Entwurf vorliegt. Für W3C PROV liegt die Priorität auf der OWL Ontologie (PROV-O ist eingestuft als „Recommandation“, PROV-XML als „Note“). PREMIS hat den Vorteil, dass dessen zwei Implementationen bereits als XSD- und OWL-Dateien vorliegen, während W3C PROV nur als Spezifikation angeboten wird[22].

3.5 Kompatibilität der Modelle mit DARIAH

Mit ihrem generischen Datenmodell sind beide Modelle kompatibel mit dem geplanten DARIAH Provenance Service. Zwei Dinge können nun die Entscheidung für eines der beiden Modelle beeinflussen:

[18] http://www.loc.gov/standards/premis/v2/premis.xsd
[19] http://premisontologypublic.pbworks.com/w/file/fetch/58521655/premis2.2_v0.1.owl
[20] http://www.w3.org/TR/2012/CR-prov-o-20121211/
[21] http://www.w3.org/TR/2012/WD-prov-xml-20121211/
[22] Das Projekt ProvToolBox(https://github.com/lucmoreau/ProvToolbox) vom PROV Co-Chair Luc Moreau bietet Implementierungen als XML-, RDF-, Json- und DOT-Dateien. Diese sind jedoch keine offiziell implementierten, so dass eine Evaluation angebracht ist.

- Die Reife: Zur Zeit der Implementierung des DARIAH Provenance Service wird die erste Version von W3C PROV veröffentlicht sein. Dennoch wird PREMIS dann reifer sein und mehr Funktionalitäten bieten.
- Die Organisationen: Beide Modelle kommen aus unterschiedlichen Zusammenhängen. PREMIS stammt aus der Bibliothekswelt und ist etwas mehr auf digitale Langzeitarchivierung ausgerichtet. PROV kommt vom W3C-Konsortium – das viel Erfahrung und Expertise im Bereich Web-Standards besitzt, und ist eher fachübergreifend konzipiert. Deswegen wird es eventuell eher als Standard im Bereich der Provenienzmetadaten akzeptiert werden.

Ausgereiftheit, funktionale Vielfalt und breite Akzeptanz der Modelle sind Kriterien, die bei Beginn der praktischen Entwicklung des DARIAH Provenance Service zur Entscheidung herangezogen werden müssen.

4 Provenienz in der DARIAH-Infrastruktur

4.1 Persistenz

Die Persistenz der Provenienzmetadaten ist mit der nachhaltigen Speicherung der Objektmetadaten gegeben und muss an dieser Stelle nicht gesondert und aus technischer Sicht betrachtet werden. Wichtig ist es zu beachten, dass die Provenienzmetadaten in den meisten Fällen eine längere Lebensdauer als das Objekt selbst haben sollten. In den Provenienzmetadaten könnte zum Beispiel auch die Löschung eines Objekts vermerkt werden und so dokumentiert werden, dass und aus welchem Grund das Objekt nicht mehr vorhanden oder zugänglich ist. Neben der Persistenz der Daten müssen auch die Verweise auf die Provenienzmetadaten langfristig stabil bleiben. Hierfür können persistente Identifikatoren genutzt werden, welche als Vermittler zwischen Referenz, Objekt und Metadaten dienen.

Die Metadaten zur Provenienz könnten als solche getrennt vom betroffenen Objekt gespeichert werden, solange sie in der Bearbeitungsphase und damit änderbar sind, bzw. es abzusehen ist, dass weitere Provenienz-Metadaten hinzugefügt werden sollen. Ist die Bearbeitung abgeschlossen, können die Metadaten zur Provenienz mit dem Objekt zusammen archiviert werden, beispielsweise in der DARIAH Bit-Preservation.

Die Provenienzmetadaden können auch nachträglich geändert werden. Zum Beispiel,

wenn nach dem Sichern des Objektes erkannt wird, dass es zusätzliche Quellen zum Objekt gab, oder das fälschlicherweise Quellen genannt wurden, die unzutreffenden waren. In diesem Fall müssen die Provenienzdaten verändert werden. Dafür sollen die Provenienzmetadata mit Version persistiert werden.

4.2 Interoperabilität

Zu diesem Thema wird auf den Report “DARIAH Interoperabilität für Werkzeuge”[23] verwiesen.

4.2.1 AAI

Die Metadaten zur Provenienz eines DARIAH-Objekts haben zunächst dieselben Zugriffsrechte wie das Objekt selbst – sofern nicht vom Besitzer des jeweiligen Objekts anders definiert. Ein Service, der Metadaten zur Provenienz speichert, muss also die DARIAH AAI implementieren bzw. bedienen.[24]

4.2.2 Bit Preservation / LZA

Weiterhin könnte ein Provenienz-Service die Metadaten – sofern sie absehbar nur kurzfristig gespeichert werden sollen – in einer dafür vorgesehenen Datenbank sichern und diese schließlich mit dem Objekt selbst als abgeschlossenes Objekt in der DARIAH Bit-Preservation ablegen. Eine weitere Option ist die sofortige Sicherung der Provenienz-Metadaten seitens des Provenienz-Services in der DARIAH Bit-Preservation – zum Beispiel als XML-Datei in einem entsprechenden Provenienz-Datenformat.

5. DARIAH Provenance Service Konzept

5.1. Service-Architektur

5.1.1. Überblick

Der DARIAH Provenance Service besteht grundsätzlich aus drei Funktionalitäten:

- dem Provenance Repository,

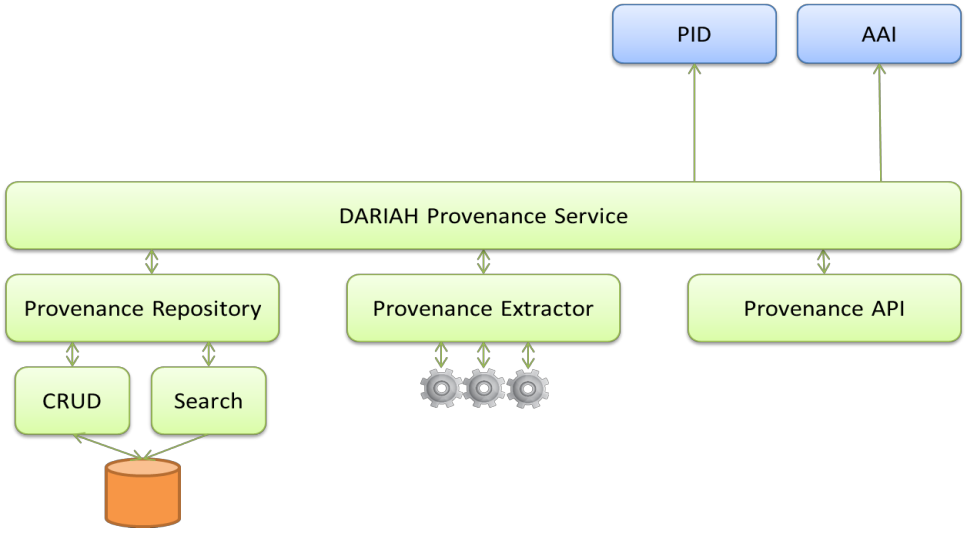
[23] Preparing for the construction of the Digital Research Infrastructure for the Arts and Humanities, Technical Report, Absatz 3.6

[24] Vgl. DARIAH Authorization and Authentication Infrastructure. https://dev2.dariah.eu/wiki/download/attachments/6783645/DARIAH-AAI-Concept-v0.3a.pdf?version=1&modificationDate=1328883126670

- dem Provenance Extractor und
- der Provenance API.

Als DARIAH Service benutzt er die folgenden DARIAH Services:

- Den DARIAH AAI Service für die Authentifikation und Autorisierung und
- den DARIAH PID Service für die Identifikation der Daten.



Architektur des DARIAH Provenance Service

5.1.2. Provenance Repository

Das Provenance Repository archiviert die Provenienzmetadaten. Es bietet zwei Schnittstellen:

- eine CRUD Schnittstelle und
- eine Search Schnittstelle

Die Beschreibung der darunter liegenden Technologie (Datenbank), um die Provenienzmetadaten zu persistieren, ist “Out-of-Scope” dieses Dokuments. Es könnte entweder eine eigene Datenbank dafür genutzt werden, oder aber auch der DARIAH Bit-Preservation Service.

CRUD

Das CRUD-Modul implementiert grundlegende Methoden für den Zugriff auf das Provenienz-Repository (Create, Retrieve, Update, Delete). Diese Methoden werden über

die unten beschriebene Provenance API bedient.

Search

Dieses Modul implementiert eine Suchfunktion über die gesamten Provenienzdaten der Repositories – mit Rücksicht auf die Zugriffsrechte. Das Modul soll ein standardisiertes Such-Protokoll implementieren, zum Beispiel Search/Retrieve via URL (SRU)[25].

Wichtige Anmerkung

Das Provenance Repository ist eine Archiv für Provenienzmetadaten in der DARIAH Insfrastruktur. Es ist insbesonderes für Nutzungsszenarien geeignet, wie im Kapitel 2.1, 2.4 und 2.5 beschrieben. Es soll aber nicht als LangzeitArchivierung Repository für Provenienzmetadaten benutzt werden. Im Fall von Langzeitsarchivierung sollen diese Metadaten mit dem Daten zusammen archiviert (z.B. im DARIAH Bit Preservation Dienst), wie im Kapitel 4.1 und 4.2.2 beschrieben.

5.1.3. Provenance Extractor

Der Provenance Extractor bietet eine Reihe von Funktionen, um aus digitalen Daten (Bilder, Videos, PDFs etc.) automatisiert Provenienzmetadaten zu extrahieren. Es gibt eine Reihe von Werkzeugen, die einen solchen Service anbieten, zum Beispiel sind für PREMIS einige Tools auf den Seiten der Library of Congress zu finden[26]. Der Provenance Extractor bietet eine REST-Schnittstelle, über die der Nutzer die Extraktion vorhandener Provenienz-Metadaten anstoßen kann. Extrahierte Metadaten können automatisiert im Provenance Repository abgelegt werden.

5.1.4. Provenance API

Die Provenance API bietet Funktionalitäten an, um Provenienzmetadaten zu erstellen oder zu manipulieren. Diese API soll – wie alle anderen Methoden des DARIAH Provenance Services auch – das REST-Protokoll für die Kommunikation zwischen Service und Nutzer implementieren. Die gesamte API-Spezifikation ist „Out-of-Scope“ dieses Dokuments, insbesondere, weil vor einer Definition der API das Datenmodell vollständig definiert sein muss. Im folgenden werden einige denkbare Methoden der API aufgelistet, die auf dem PROV-Modell basieren:

[25] SRU: Search/Retrieval via URL – SRU, CQL and ZeeRex (Standards, Library of Congress). http://www.loc.gov/standards/sru/

[26] Tools for preservation metadata implementation: PREMIS (Preservation metadata) – PREMIS: Preservation Metadata Maintenance Activity (Library of Congress). http://www.loc.gov/standards/premis/tools_for_premis.php

- **CreateAgent**: Erstellt einen neuen “Agent”
- **CreateEntity**: Erstellt eine neue “Entity”
- **CreateActivity**: Erstellt eine neues “Activity”
- **AddAgent**: Fügt einen Agent zu einem Provenienzobjekt hinzu (z.B. für die Verbindung zweier Objekte)
- **AddEntity**: Fügt eine Entity zu einem Provenienzobjekt hinzu (z.B. für die Verbindung zweier Objekte)
- **AddActivity**: Fügt eine Activity zu einem Provenienzobjekt hinzu (z.B. für die Verbindung zweier Objekte)

5.1.5. Content Negotiation

Der Provenance Service implementiert HTTP Content Negotiation für Formate wie XML, RDF, JSON oder DOT. Alle DARIAH Provenance Service-Schnittstellen sind in der Lage, diese Formate zu bedienen.

5.1.6. Transformation

Der Provenance Service transformiert Provenienzmetadaten von einem Format in ein anderes. Insbesonders transformiert er Metadaten vom internen DARIAH Provenance Service Format (im noch zu bestimmenden Standard) in alle vom Content Negotiation Module bedienten Formate, und umgekehrt. Diese Module ist über eine API nutzbar, und kann deshalb nicht nur für die Content Negotiation genutzt werden, sondern auch extern.

5.1.7. PID Resolver

Dieses Modul benutzt den DARIAH PID Service, um die Provenienzmetadaten zu verwalten bzw. auf die Datenobjekte zuzugreifen (zum Beispiel für die Extraktion der Provenienzmetadaten). Ob die Provenienzmetadaten einen eigenen PID bekommen oder ob sie an den PID des Objekts gebunden werden, hängt von der Implementation des Services ab.

5.1.8. AAI

Diese Modul implementiert die DARIAH AAI für den Provenance Service. Siehe hierzu Abschnitt 4.2.1.

5.1.9 Validierung

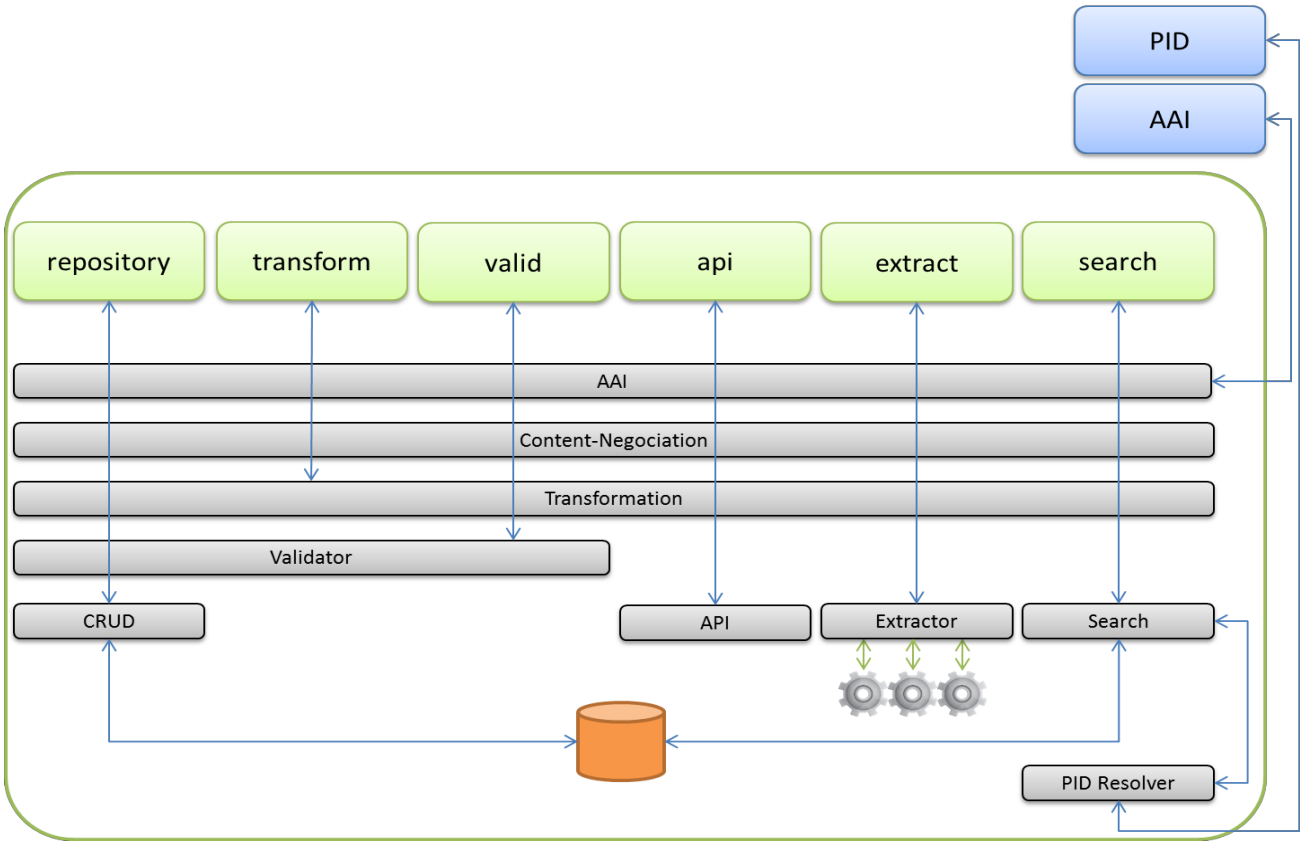
Die Validierung von Provenienzmetadaten soll auf zwei Ebene achten:

1. Die Gülitigkeit den Metadaten gegen dem Provenienz Datenmodell.
2. Die Gülitgkeit den Metadaten gegen den Provenienz Model Regeln.

Der erste Punkt ist trivial. Es handelt sich um die Validität gegen ein einfache Daten Schema. Wie zum Beispiel XML Daten die mit einem XML Schema validiert werden.
Der zweite Punkt ist komplizierter. Es soll damit möglich sind, ob die Daten nicht nur formal gültig sind, sondern auch inhaltlich. Beispielsweise, wenn eine neue Version eines bereits hinterlegten Datenobjekts gespeichert wird, muss dieses versionierte Datenobjekt auch einen neuen Provenienz Datensatz erhalten. Jedem Datenbjekt (und jeder Version eines Datenobjekts) sollte also ein eigener Satz an Provenienzdaten zugeordnet sein. Die Validierung soll, so weit möglich ist, solche Fälle validieren können. Solche Regel sind bei den Modelle definiert[27] [28].

5.2. Schnittstellen

Alle Schnittstellen des DARIAH Provenance Service implementieren das REST-Protokoll sowie HTTP Messages[29]:



Das Messaging-Konzept des DARIAH Provenance Services

[27] http://www.w3.org/TR/2012/WD-prov-overview-20121211/
[28] http://www.loc.gov/standards/premis/v2/premis-2-2.pdf
[29] http://www.w3schools.com/tags/ref_httpmessages.asp

5.2.1. Repository

Methode	Verb	Beschreibung
GET	Retrieve	Holt ein Provenienzdatum ab
POST	Create	Erstellt ein neues Provenienzdatum
PUT	Update	Ändert ein Provenienzdatum
DELETE	Delete	Löscht ein Provenienzdatum

5.2.2. Search

Methode	Verb	Beschreibung
GET	Search	Sucht und gibt das Ergebnis zurück (das Format ist via Content Negotiation definiert)
POST	-	-
PUT	-	-
DELETE	-	-

5.2.3. Extract

Methode	Verb	Beschreibung
GET	-	-
POST	Extract	Extrahiert den als Parameter übergebenen digitalen Inhalt und gibt ein Provenienzdatum zurück (das Format ist via Content Negotiation definiert)

PUT	-	-
DELETE	-	-

5.2.4. Provenance, Valid und Transform

Diese Schnittstellen können erst dann definiert werden, wenn das Datenmodell festgelegt wurde.

6. Fazit

Mit diesem Report wird ein generelles Konzept zur Erfassung und Verwaltung von Provenienzmetadaten vorgestellt und seine mögliche Anwendung eruiert. Der Report ist Grundlage für eine ausführlichere Architekturskizze und kann in Kombination mit noch zu erstellenden detaillierten geisteswissenschaftlichen Anforderungsprofilen zur Implementierung eines Provenienz-Systems führen.
Da es jedoch zur Zeit noch keine genauer spezifizierten Anforderungen und Use-Cases gibt, empfehlen wir die Implementierung erst bei entsprechenden Anforderungen zu realisieren. Eine Einschätzung des personellen Aufwands inkl. einer Aufstellung der hiermit verbundenen Aufgaben und Arbeitsschritte ist in Punkt 7 “Roadmap” aufgeführt.

7. Roadmap

Phase	Aufgabe	Personen-monate
1	● Modell auswählen (PREMIS/PROV) ● Gesamt-Architektur in UML formulieren ● Technologie auswählen (Programmiersprache, Datenbank, internes Datenformat etc.)	1
2	● UML in Code transformieren ● Implementierung der CRUD-Funktionen ● Validierungs-Modul	2
3	● Provenance Extractor Modul ● Suche implementieren	2
4	● Transformations-Modul	2

	• Content Negotiation Modul	
5	• PID Resolver Modul • Spezifikation der Provenance API	1
6	• Implementierung der Provenance API	2

8. Literatur

Fischer, Thomas: WissGrid Deliverable 3.4.3. AP 3: Langzeitarchivierung von Forschungsdaten. Metadaten und Provenienz: Eine Übersicht.
http://www.wissgrid.de/publikationen/deliverables/wp3/WissGrid-D3.4.3-Provenienz.pdf

PREMIS Data Dictionary for Preservation Metadata, version 2.2, The Library of Congress.
http://www.loc.gov/standards/premis/v2/premis-2-2.pdf

PROV-DM: The PROV Data Model (W3C Candidate Recommendation 11 December 2012), The World Wide Web Consortium.
http://www.w3.org/TR/2012/CR-prov-dm-20121211/

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, SWPM-2012 , 27. oder 28. Mai in Heraklion, Griechenland.
http://ceur-ws.org/Vol-526/, http://ceur-ws.org/Vol-670/ und http://ceur-ws.org/Vol-856/

USENIX Workshop Series on the Theory and Practice of Provenance, 2009-2012.
https://www.usenix.org/conference/tapp09, https://www.usenix.org/conference/tapp10, https://www.usenix.org/conference/tapp11, https://www.usenix.org/conference/tapp12