D-Grid-Integrationsprojekt 2

Diskussionspapier

A: Status des Dokuments
Version 1.0, Draft.

1 Herkunftsmetadaten

Herkunftsinformation ist für ein beliebiges elektronisches Dokument aus mindestens zwei Gründen von Bedeutung:

- Zur Feststellung der Vertrauenswürdigkeit des Dokumentes: Wer hat es erstellt, von wem ist es verändert worden, wie wird garantiert dass niemand seinen Inhalt unbemerkt verändern konnte?
- Zur Bewahrung seiner Nutzbarkeit: In welchem (digitalen) Format liegt das Dokument vor, mit welchen Programmen wurde es erstellt bzw. bearbeitet, mit welchem Programm kann es präsentiert werden?

Innerhalb einer Grid-Umgebung, die einerseits auf (viele) verschiedene Nutzer und andererseits eventuell auf eine lange Nutzbarkeit angelegt ist, erhöht sich die Bedeutung dieser Informationen, wird aber auch gleichzeitig komplizierter, weil vielfältige Interaktionen mit verschiedenen Teilsystemen des Grid zur Erzeugung, Speicherung, Authentifizierung, Weiterverarbeitung berücksichtigt werden müssen.
Zusätzlich werden in einer Grid-Umgebung 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. Das „DELOS Digital Library Reference Model“
   (http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf)
   des *DELOS* Projektes (2007)

5. „The Origin of Data“, Dissertation von Paul T. Groth (http://eprints.ecs.soton.ac.uk/14649/) (2007)
6. Das „PREMIS Data Dictionary for Preservation Metadata (version 2.0)“ (http://www.loc.gov/standards/premis/v2/premis-2-0.pdf) (2008)
7. Das „Open Provenance Model“ (http://eprints.ecs.soton.ac.uk/18332/1/opm.pdf) als Ergebnis des *Third Provenance Challenge* (2008)
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)
12. Das „W3C Provenance Incubator Group“ (http://www.w3.org/2005/Incubator/prov/XGR-prov/) (2010)

Die Darstellung erfolgt zweisprachig, da es mir nicht notwendig und unnötig aufwendig erscheint, die englischen Zitate ins Deutsche zu übersetzen oder alternativ meine Bemerkungen auf englisch zu machen.

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.

Einen umfassenden Überblick über ein mögliches System zur Erfassung, Verwaltung und Nutzung von Herkunftsinformationen gibt die Dissertation von Paul Groth, die im Umfeld des EU Provenance-Projektes entstanden ist. Erkenntnisse daraus fließen in die umfassenden Arbeiten der W3C Provenance Incubator 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 liefert.
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 Incubator Group“, die auch vielfältige zusätzliche und weiterführende Dokumente bereitstellt.

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

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