UNIVERSITÄTSMEDIZIN GÖTTINGEN UMG

Exposé zur Bachelorarbeit

im Studiengang "Angewandte Informatik"

Nutzung von Persistent Identifiern zur Umsetzung der FAIR-Prinzipien in Datenablageplattformen für die medizinische Forschung

Cornelius Knopp
ORCID: 0000-0002-1505-594X

03. August 2017

Lehrstuhlinhaber:	Prof. Dr. Otto Rienhoff
Erstbetreuer:	Prof. Dr. Ulrich Sax
Zweitbetreuer:	Dr. Harald Kusch

Institut für Medizinische Informatik
Universitätsmedizin Göttingen

1. Einführung und Fragestellung

Im Rahmen medizinischer Forschungsvorhaben steigt die Menge der zu archivierenden Daten stetig an. Diese Datenbestände, bestehend aus erfassten Rohdaten sowie generierten Analyse- und Ergebnisdaten, müssen über einen langen Zeitraum archiviert werden, um die Reproduzierbarkeit der Forschung zu gewährleisten. Bei der Ablage dieser Daten müssen auch die zugehörigen Informationen zur Datenprovenienz gespeichert werden. Hierzu gehören neben den Informationen zu Autorenschaft und Version auch die Zugriffs- und Verwertungsrechte. Doch um die Forschungsergebnisse validieren zu können oder darauf aufzubauen, ist es notwendig, dass Daten und Metadaten gefunden werden und darauf zugegriffen werden kann.
Um diese Herausforderungen an das Forschungsdatenmanagement (FDM) zu lösen, wurden im Rahmen der Joint Declaration of Data Citation Principles (JDDCP) Leitlinien für eine optimierte Archivierung von Daten verabschiedet. [1] Findable, Accessible, Interoperable & Reusable (FAIR) sind diese Prinzipien, welche für alle Stakeholder Vorteile im Bereich der Verfügbarkeit und Nutzbarkeit mit sich bringen. Diese Leitlinien werden auch in den Stellungnahmen der Deutsche Forschungsgemeinschaft (DFG) [2] und des Rat für Informationsinfrastruktur (RfII) [3] im Bezug auf die Langzeitarchivierung von Forschungsdaten gefordert. Das Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit (nestor) zeichnet mit dem gleichnamigen nestor-Siegel auf Grundlage der DIN-Normen 31644 [4] und 31646 [5] vertrauenswürdige digitale Langzeitarchive aus und setzt hierbei eine Implementierung der FAIR-Leitlinien voraus. [6]
Persistent Identifier (PID) sind global eindeutige Identifikatoren, die zur Referenzierung von digitalen Objekten dienen. Zur Gewährleistung der globalen Eindeutigkeit gibt es zentrale Vergabestellen, welche den teilnehmenden Einrichtungen eindeutige Präfixe für ïhre"PIDs zuweisen. Von Hyperlinks unterscheiden sich PIDs vorallem durch die annotierten Metdadaten. Die Domain eines Hyperlinks ist, genau wie der PID ein Zeiger auf den variablen Speicherort der abzurufenden Daten. Durch die bereits erwähnte Metadaten-Annotaion ist der PID mit der International Standard Book Number (ISBN) aus dem Buchhandel zu vergleichen. Beide sind sie jedoch dafür ausgelegt, den Zugriff auch bei verändertem Speicherort weiter zu garantieren, indem eine Weiterleitung von dem unveränderlichen PID zur aktuellen Zugriffsadresse erfolgt. [7]
Daher stellt sich die folgende Frage:

> **(1) Wie sind Persistent Identifier (PID) für die Umsetzung der FAIR-Prinzipien in der medizinischen Forschung verwendbar ?**

Weiterhin beschäftige ich mich in dieser Arbeit mit dem Konzept der PIDs in Zusammenhang mit den beiden Datenablageplattformen openBIS und SEEK.
openBIS ist eine Open-Source Plattform zur Verwaltung und Analyse von Forschungsdaten aus dem Bereich der Systembiologie. Hauptziel ist es, Reproduzierbarkeit von Forschungsdaten zu gewährleisten. Dies geschieht, indem Informationen über Prozesse, Originaldaten sowie deren Derivate, verwendete Analysemethoden und gesammeltes Wissen gesammelt und miteinander verknüpft archiviert werden. [8] Im Gegensatz zu reinen Datenablagen oder Netzlaufwerken liegt der Fokus auf der Metadaten-Annotation.
Die FAIRDOM-Initiative, ein Zusammenschluss der Universität Manchester, des Heidelberger Institut für Theoretische Studien (HITS), der Universität Zürich, der **ETH!** (**ETH!**) und der Universität Leiden, entwickelte eine umfangreiche Serviceeinrichtung für Systembiologie. Diese hat sich zum Ziel gesetzt, Forscher, Studenten, Förderer und Verleger darin zu Unterstützen ihre Forschungsprojekte FAIR zu gestalten. Hierzu wurde die Plattform SEEK entwickelt, welche ähnlich wie openBIS die Forschungsdaten inklusive der annotierten Metadaten speichert. Im Unterschied zu openBIS überspannt SEEK einen deutlich größeren Abschnitt der Forschungsunterstützung. Ein Forschungsprojekt kann, beginnend bei Anträgen und Standard Operating Procedures (SOPs), über Roh-, Analyse- und Ergebnisdaten, bis hin zu (Zwischen-)Berichten und Publikationen, komplett digital archiviert werden. [9, 10] Durch diese vollständige Ablage und die damit einhergehende Verknüpfung der einzelnen Datensätze, sind die FAIR-Prinzipien sowie die Richtlinien für Gute wissenschaftliche Praxis (GWP) erfüllt.
Zu klären bleibt also die Frage,

> **(2) Inwiefern sind PIDs in Verbindung mit den Plattformen openBIS und SEEK nutzbar ?**

Die SEEK-Plattform ermöglicht es den Nutzern bereits, Dokumente und Publikationen mittels eines Representational State Transfer (REST) Webservices mit einem Digital Object Identifier (DOI) zu versehen. [9] Der DOI ist eine Spezialisierung des PID auf dem Gebiet der digital verfügbaren Dokumente. Dieser Identifier erfüllt alle Voraussetzungen, ist jedoch nur auf Dokumente und Publikationen beschränkt. Die Möglichkeit diese Registrierung nativ aus SEEK heraus durchzuführen, bietet die Chance, auch andere Datensätze mit einer PID zu versehen. Der dazugehörige Generator für diese PIDs nutzt ebenfalls einen REST-Webservice und das JavaScript Object Notation (JSON) Format. Dieser Service wird durch die Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen (GWDG) als Teil des European Persistent Identifier Consortium (ePIC) bereitgestellt und generiert einen PID mit den zugehörigen Metadaten zu einem gegebenen Datensatz. Die Auflösung (Resolvierung) dieses PIDs wird durch redundante Datenbanken bei den verschiedenen Mitgliedsinstitutionen dauerhaft sichergestellt. Zu definieren bleibt ein Nutzungskonzept, welches grundlegende Aussagen zur Verwendung der PIDs macht und für einzelne Forschungsprojekte angepasst übernommen werden kann.

Durch diesen Teil der Arbeit soll die folgende Frage beantwortet werden:

> **(3) Wie kann eine konzeptionelle Integration sowie die Nutzung eines Services zur Registrierung und Verwendung von PIDs in einer Datenablageplattform aussehen ?**

Durch die Beantwortung dieser 3 Fragen sollen die Grundlagen einer zuverlässigen Langzeitarchivierung von Forschungsdaten im Bezug auf Auffindbarkeit, Erreichbarkeit, Interoperabilität und Wiederverwendbarkeit (FAIR) erläutert und ein Konzept zur Umsetzung dieser Prinzipien mit Hilfe von PID erarbeitet werden.

2. Material & Methoden

Zur Bearbeitung und Beantwortung dieser Fragen werde ich zunächst mittels Fachbüchern, Journal-Artikeln und Stellungnahmen der einschlägigen Fachgremien eine Übersicht über den aktuellen Forschungsstand im Bereich Datenidentifizierung bei der Langzeitarchivierung (LZA) und Persistent Identifier vermitteln. Hierzu gehören u.a. die Publikationen der Research Data Alliance (RDA) Arbeitsgruppe PID Information Types, des Kompetenznetzwerks Langzeitarchivierung und Langzeitverfügbarkeit (nestor), der FAIRDOM Initiative sowie von DFG, RfII und ePIC.
Des weiteren werde ich die FAIR-Prinzipien erläutern und die Vorteile ihrer Verwendung darlegen.
Die Datenablageplattformen SEEK und openBIS werden in einer Testumgebung auf neu zu konfigurierenden Servern installiert, um ohne Beeinflussung von anderen Prozessen analysiert werden zu können. Diese Plattformen werden auf mögliche Anknüpfungspunkte für die Registrierung von PIDs untersucht und die entsprechenden Konfigurationen der Software werden offengelegt. Durch die damit verbundene Untersuchung der API werden die Kommunikationsmöglichkeiten zwischen Plattform und PID-Service ermittelt und an die lokalen Anforderungen der Testumgebung angepasst.
Dieser PID-Service wird durch die GWDG als Teil des ePIC bereitgestellt und fungiert als neutraler Vermittler zwischen Nutzer und den abzurufenden Daten. (vgl. Abbildung 1) Der PID-Service weist jedem zu identifizierenden Objekt eine eindeutigen PID zu, welcher sowohl die Referenz zum Objekt selbst, als auch die Metadaten enthält. Durch einen ebenfalls bereitgestellten Resolver-Dienst ist es möglich, den PID aufzulösen und die aktuell gültige Zugriffsadresse zu erhalten. [11]
Durch eine redundante Vorhaltung der Datenbanken aller 5 Rechenzentren ist die Auflösung des PIDs auch bei Ausfall der Erreichbarkeit von Rechenzentren möglich.

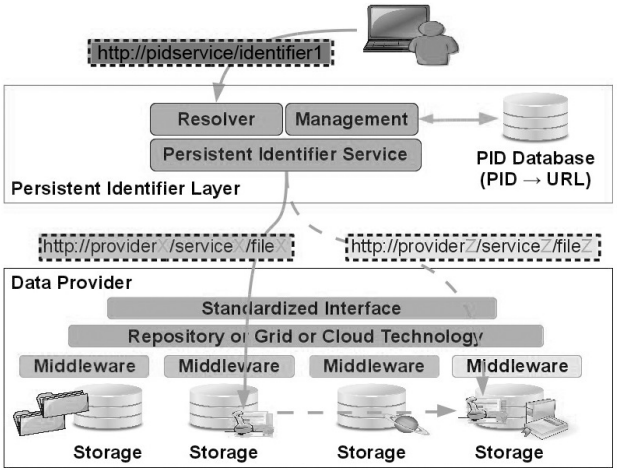

Abbildung 1: Schematische Funktionsweise des ePIC Persistent Identifier Service, aus [11]

3. Prüfer

Prof. Dr. Ulrich Sax Institut für Medizinische Informatik Leiter AG Infrastruktur für Translationale Forschung Universitätsmedizin Göttingen Georg-August-Universität Göttingen Robert-Koch-Straße 40 D-37075 Göttingen	**Dr. Harald Kusch** Institut für Medizinische Informatik Sonderforschungsbereich 1002 - Herzinsuffizienz Universitätsmedizin Göttingen Georg-August-Universität Göttingen Robert-Koch-Straße 40 D-37075 Göttingen

Literaturverzeichnis

[1] Martone ME (Hg.). Joint Declaration of Data Citation Principles. San Diego;. URL: https://www.force11.org/group/joint-declaration-data-citation-principles-final [Eingesehen am: 19.06.2017].

[2] Informationsverarbeitung an Hochschulen: Organisation, Dienste und Systeme. Stellungnahme der Kommission für IT-Infrastruktur für 2016-2020;. URL: http://www.dfg.de/download/pdf/foerderung/programme/wgi/kfr_stellungnahme_2016_2020.pdf [Eingesehen am: 13.06.2017].

[3] Leistung aus Vielfalt: Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland. Göttingen; 2016. URL: http://www.rfii.de/de/category/dokumente/ [Eingesehen am: 13.06.2017].

[4] Deutsches Institut für Normung e V . Information und Dokumentation - Kriterien für vertrauenswürdige digitale Langzeitarchive. Berlin: Beuth; 2012-04.

[5] Deutsches Institut für Normung e V . Information und Dokumentation - Anforderungen an die langfristige Handhabung persistenter Identifikatoren (Persistent Identifier). Berlin: Beuth; 2013-01.

[6] Erläuterungen zum nestor-Siegel für vertrauenswürdige digitale Langzeitarchive. vol. 17 of nestor-materialien. Version 2.0 ed. Frankfurt am Main: nestor - Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland; 2016. URL: http://nbn-resolving.de/urn:nbn:de:0008-2016111106 [Eingesehen am: 13.06.2017].

[7] Tonkin E. Persistent identifiers: Considering the options. Ariadne, Web Magazine for Information Professionals. 2008;(56). URL: http://www.ariadne.ac.uk/issue56/tonkin [Eingesehen am: 13.06.2017].

[8] Bauch A, Adamczyk I, Buczek P, Elmer FJ, Enimanev K, Glyzewski P, et al. openBIS: A flexible framework for managing and analyzing complex data in biology research. BMC bioinformatics. 2011;12:468.

[9] Wolstencroft K, Owen S, Krebs O, Nguyen Q, Stanford NJ, Golebiewski M, et al. SEEK: A systems biology data and model management platform. BMC systems biology. 2015;9:33.

[10] Wolstencroft K, Krebs O, Snoep JL, Stanford NJ, Bacall F, Golebiewski M, et al. FAIRDOMHub: A repository and collaboration environment for sharing systems biology research. Nucleic acids research. 2017;45(D1):D404–D407.

[11] Kálmán T, Kurzawe D, Schwardmann U. European Persistent Identifier Consortium - PIDs für die Wissenschaft. In: Altenhöner R, Oellers C (Hg.). Langzeitarchivierung von Forschungsdaten. Berlin: Scivero Verl.; 2012. p. 151–168.

Abkürzungsverzeichnis

API	Application Programming Interface
DFG	Deutsche Forschungsgemeinschaft 1
DOI	Digital Object Identifier 2
ePIC	European Persistent Identifier Consortium 2
FAIR	Findable, Accessible, Interoperable & Reusable 1
FDM	Forschungsdatenmanagement 1
GWDG	Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen 2
GWP	Gute wissenschaftliche Praxis 2
HITS	Heidelberger Institut für Theoretische Studien 2
ISBN	International Standard Book Number 1
JDDCP	Joint Declaration of Data Citation Principles 1
JSON	JavaScript Object Notation 2
LZA	Langzeitarchivierung 3
nestor	Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit 1
openBIS	Forschungsdatenmanagement System der ETH Zürich
ORCID	Open Researcher and Contributor ID
PID	Persistent Identifier 1
RDA	Research Data Alliance 3
REST	Representational State Transfer 2
RfII	Rat für Informationsinfrastruktur 1
SEEK	Forschungsdatenmanagement System der FAIRDOM Initiative
SOP	Standard Operating Procedure 2