Anna Neovesky

[preprint GfM-Jahrestagung 2016 Mainz;
Beitrag erscheint 2018 im Beitragsarchiv des Internationalen Kongresses der Gesellschaft für Musikforschung, Mainz 2016 – »Wege der Musikwissenschaft«, hrsg. von Gabriele Buschmeier und Klaus Pietschmann, http://schott-campus.com/gfm-jahrestagung-2016]

… und was machen wir nun mit den Daten? Nutzungsszenarien und deren Voraussetzungen am Beispiel von Akademievorhaben

1 Einleitung

Langzeitarchivierung, Nachhaltigkeit und offene Daten sind elementare und viel diskutierte Themen in den Digitalen Geisteswissenschaften und in der Digitalen Musikwissenschaft.[1] Mittlerweile besteht grundsätzlicher Konsens darüber, dass Forschungsdaten frei zugänglich gemacht und archiviert werden, offene Lizenzmodelle sind etabliert. Ein Aspekt hierbei ist die langfristige Archivierung und Bereitstellung von Forschungsdaten und Applikationen und wie diese umgesetzt werden kann. Ein weiterer Aspekt betrifft die Möglichkeiten, die sich dadurch ergeben. Denn erst die Nachnutzung, Verknüpfung und Anreicherung der bereitgestellten Daten erlaubt es, weitergehende Forschung durchzuführen und das Potential zu nutzen, das offene Forschungsdaten bieten. Dieser Artikel behandelt die Vorbedingungen für die Nachnutzung und darauf aufbauende Forschungsmöglichkeiten. Hierzu werden mehrere Beispiele aus den Forschungsvorhaben der Akademie der Wissenschaften und der Literatur | Mainz vorgestellt sowie Ausblicke auf zukünftige Anwendungen gegeben. Zentral vorgestellt wird eine generische Suche für Musiknoten, die sowohl unter Nachnutzung vorhandener Daten und Technologien entwickelt wurde als auch selbst unter einer freien Lizenz steht und nachnutzbare Daten liefert.
Die vorgestellten Projekte wurden an der Digitalen Akademie der Wissenschaften und der Literatur | Mainz durchgeführt. An der Akademie werden aktuell rund 40 Projekte aus dem Bereich der geisteswissenschaftlichen Grundlagen- und Langzeitforschung durchgeführt. Als zentrale Forschungseinrichtung für Digitale Geisteswissenschaften an der Akademie ist die ›Digitale Akademie‹ an den Vorhaben jeweils mit unterschiedlichen Schwerpunkten beteiligt. Kernanliegen der Digitalen Akademie ist es dabei, das ganze Spektrum der Digital Humanities von der Retrodigitalisierung und Kuratierung von Fachdaten über die wissenschaftliche Konzeption und Entwicklung geisteswissenschaftlicher Anwendungen bis hin zur Erschließung informationswissenschaftlich-informatischer Verfahren für die Forschung abzubilden und in die Projekte mit einzubringen. Ein Schwerpunkt liegt hierbei darauf, dass Lösungen die in einzelnen Vorhaben erarbeitet werden auch an anderen Stellen anwendbar sind und daher generisch und nachhaltig entwickelt und bereitgestellt werden. Forschungsdaten und Applikationen werden soweit möglich im Open Access unter freier Lizenzierung bereitgestellt.[2]

---

[1] So stehen 2017 die internationale Tagung der Digital Humanities Community (DH2017) als auch die Tagung der Digital Humanities im deutschsprachigen Raum (DHd 2017) unter dem Zeichen der Digitalen Nachhaltigkeit und des offenen Zugangs.

[2] Zur Lizenzierung an der Akademie siehe das Interview Aline Deicke, Anna Neovesky »Lizenzierung bei der Digitalen Akademie«, in: forschungslizenzen.de (2017), http://forschungslizenzen.de/lizenzierung-bei-der-digitalen-akademie-mainz, 2.2.2017.

2 Forschungsdaten und die Voraussetzungen der Nachnutzung

2.1 Definition ›Forschungsdaten‹

Die Vorbedingung für die Nachnutzbarkeit von Forschungsdaten ist deren offene Bereitstellung. Hierfür gilt es als erstes zu definieren, was unter dem Begriff ›Forschungsdaten‹ verstanden wird. Definitionen des Begriffes beschäftigen sich sowohl mit der Art wie die Daten erhoben werden als auch mit den verschiedenen Datentypen und Datenformaten, in denen sie – auch abhängig von der jeweiligen Wissenschaftsdisziplin – auftreten.[3] In diesem Beitrag folgt der Begriff der Definition der Allianz der deutschen Wissenschaftsorganisation als Daten, »die im Zuge wissenschaftlicher Vorhaben z. B. durch Digitalisierung, Quellenforschungen, Experimente, Messungen, Erhebungen oder Befragungen entstehen«.[4] Außerdem werden in einer Erweiterung des Begriffes auch Software sowie Anwendungen und Werkzeuge, die im Zusammenhang der Forschung entstehen, mit einbezogen, denn auch diese Applikationen sollten im Zusammenhang einer nachhaltigen wissenschaftlichen Tätigkeit und Entwicklung nachnutzbar und frei zugänglich sowie offen lizenziert sein.

2.2 Lizenzierung

Eine Lizenz – unabhängig davon ob es sich um eine offene oder eine restriktive handelt – schafft für Nutzer, die die Daten verwenden möchten Rechtssicherheit. Eine offene Lizenzierung ermöglicht Nachnutzung. Je nach dem Grad der Offenheit sind unterschiedliche Nutzungsformen möglich: von der reinen Verifizierung von Forschungsergebnissen bis hin zur Weiterentwicklung und neuen Kontextualisierungen.[5] Für Texte und Bilder haben sich in der Wissenschaft mittlerweile die Creative Commons Lizenzen als Standard durchgesetzt.[6] Für Software kommen meist MIT-Lizenz und GNU General Public License (GNU GPL) zum Einsatz.

2.3 Bereitstellung

Sobald die Lizenzierung geklärt ist, geht es um die Bereitstellung der Daten und die Frage wo und wie die Forschungsdaten verfügbar gemacht werden. Dies kann zum einen auf der jeweiligen Forschungsplattform geschehen oder auch in einem Repositorium, etwa der ausführenden Forschungsinstitution, der Universität oder in den Infrastrukturangeboten der Digitalen Geisteswissenschaften, wie etwa dem TextGrid Repository.[7] Software oder auch Materialsammlungen werden häufig auf GitHub bereitgestellt. Plattformen wie das DARIAH-DE Collection Registry oder re3data.org verzeichnen Projekte, deren Forschungsdaten verfügbar sind. Eine Dokumentation der Daten gibt einen Überblick über das Datenmodell, die verwendeten

---

[3] Zu einer Übersicht über Begriffsdefinitionen siehe http://www.forschungsdaten.org/index.php/Forschungsdaten, 2.2.2017. Zu verschiedenen Typen von Forschungsdaten und einer Übersicht über Lizenzierungsmodelle vgl. Nikolaos Beer u. a.: Datenlizenzen für geisteswissenschaftliche Forschungsdaten - Rechtliche Bedingungen und Handlungsbedarf, in: *DARIAH-DE Working Papers 6* (2014), http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dwp-2014-6.pdf, 2.2.2017.

[4] Helmholtz-Gemeinschaft »Schwerpunktinitiative Digitale Information der Allianz der deutschen Wissenschaftsorganisationen«, in: *allianzinitiative.de* (2016), http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten, 2.2.2017.

[5] Als Beispiel für Überlegungen zur Lizenzierung von Forschungsprojekten s. Aline Deicke, Anna Neovesky »Lizenzierung von digitalen Editionen am Beispiel von Projekten zur jüdischen Geschichte« in: *Medaon 9, 17* (2015), http://www.medaon.de/pdf/medaon_17_Deicke+Neovesky.pdf, 2.2.2017.

[6] So stehen die meisten Forschungsdaten wie auch Open Access Online-Zeitschriften und wissenschaftliche Blogs überwiegend unter einer CC-Lizenz.

[7] »TextGrid Repository« https://textgridrep.org, 2.2.2017.

Standards, Klassifikationssysteme und Vokabulare, das methodische Vorgehen bei der Erstellung und die Zugriffsmöglichkeiten.[8]

Wichtig ist, dass die Daten in einem standardisierten Format bereitgestellt werden oder bestenfalls mehrere Zugriffsmöglichkeiten existieren, die für verschiedene Anwendungsfälle geeignet sind. Diese sollten den verbreiteten Standards der jeweiligen Disziplinen und den Datentypen folgen. Das heißt zum Beispiel, dass das XML-Format der Charters Encoding Initiative (CEI)[9] für Urkunden, das XMP Format für die Auszeichnung von Metadaten an Bilddaten, RDF für eine semantische Auszeichnung der Daten sowie BEACON-Datein für die Bereitstellung und Vernetzungsmöglichkeit von personenbezogene Daten mittels Normdaten der Gemeinsamen Normdatei (GND) verwendet werden. Austauschformate in einfach lesbarer Textform wie JSON eignen sich besonders gut für den Datenaustausch zwischen Anwendungen. Die Daten können dabei entweder als Download oder über eine Schnittstelle angeboten werden. Weit verbreitete Schnittstellen sind das Simple Object Access Protocol (SOAP) und die Representational State Transfer Architektur (REST).
Um eine langfristige Bereitstellung zu sichern ist eine dauerhafte Erreichbarkeit und Zitierbarkeit der Ressourcen zentral. Hierzu gehört zum einen die persistente Adressierung der Webressourcen zum Beispiel mittels Uniform Resource Name (URN) oder Digital Object Identifier (DOI)[10] und zum anderen eine Sicherung der Verfügbarkeit der Daten mittels Langzeitarchivierung.

3 Beispiele für Nachnutzung aus Forschungsprojekten der Mainzer Akademie

Es gibt zahlreiche Möglichkeiten Forschungsdaten nachzunutzen, von denen hier vier Bespiele aus den Projekten der Mainzer Akademie vorgestellt werden.[11] Sie bilden sowohl verschiedene Disziplinen als auch verschiedene Einsatzmöglichkeiten ab. Hierbei können die Daten in Verbindung mit anderen Repositorien in einen neuen Kontext gesetzt werden oder auf den bestehenden Daten aufgesetzte Forschung durchgeführt werden. Diese einzelnen Beispiele illustrieren die Nachnutzung von Forschungsdaten zur Anreicherung eigener Inhalte, Nachnutzung in Abschlussarbeiten von Studierenden und in experimentellen Projekten, Anschlussforschungen mit den Daten sowie Entwicklung von Applikationen mittels offener Software und frei lizenzierten Daten.

3.1 Mainzer Professorenkatalog: Anreicherung über Normdaten

[8] Als Beispiel siehe die Dokumentation der Daten der Regesta Imperii: http://www.regesta-imperii.de/daten.html, 2.2.2017.

[9] »Charters Encoding Initiative« http://www.cei.lmu.de/index.php, 2.2.2017.

[10] Für eine Gegenüberstellung verschiedener Möglichkeiten der persistenten Adressierung und eine Diskussion des Konzeptes der persistenten Identifizierbarkeit vgl. Eckhart Arnold, Stefan Müller »Wie permanent sind Permalinks?« in: *Informationspraxis Bd. 3, Nr. 1* (2017), DOI: http://dx.doi.org/10.11588/ip.2016.2.33483, 2.2.2017.

[11] Für weitere Beispiele der Nachnutzung vgl. Andreas Kuczera, Yannick Weber, Max Grüntgens, Aline Deicke, Frederic von Vlahovits, Dominik Kasper »Ebenen der Nachnutzung von Forschungsdaten in der Mainzer Akademie« Vortrag im Rahmen der Tagung der AG eHumanities der Union der deutschen Akademien der Wissenschaften, Düsseldorf (2016), http://digicademy.github.io/2016-agehum-ddorf, 2.2.1017.

Eine Möglichkeit der Nachnutzung von Daten ist die Anreicherung eigener Inhalte durch weiterführende Informationen anderer Forschungsplattformen. Dies kann am einfachsten über eindeutige Identifikatoren geschehen. Für Personen, aber auch Körperschaften und Ereignisse ist dies die GND. Wenn Personen mit einer GND versehen sind, können diese über das BEACON-Dateiformat verbunden werden. BEACON ist Format, das in einer Textdatei eine Basis-URL der Plattform und die darin enthaltenen GNDs enthält. Über diese wird dann die Gesamt-URL erstellt, die auf den einzelnen Datensatz referenziert. Viele Forschungsvorhaben bieten für ihre Inhalte bereits ein solches Datenformat an. Eine Auflistung von Repositorien findet sich in der Wikipedia.[12]
Der Mainzer Professorenkatalog[13] verzeichnet die Mainzer Professorinnen und Professoren in den Jahren 1946 bis 1973. Zu den Professoren werden neben Lebensdaten und biographischen Informationen detaillierte Informationen zur akademischen Laufbahn aufgeführt.[14] Über BEACON werden Ressourcen weiterer Portale eingebunden, wie etwa der Deutschen Digitalen Bibliothek, der Wikipedia und verschiedenen Lexika.
Dadurch wird ermöglicht, dass auch Publikationen, Korrespondenzen, Reden und weitere an anderer Stelle gesammelten Materialien in das Angebot eingebunden werden.

3.2 Regesta Imperii: experimentelle Projekte und Abschlussarbeiten von Studierenden

Die Regesta Imperii sind ein Grundlagenwerk zur Forschung des Mittelalters. Das Vorhaben verzeichnet sämtliche Urkunden der römisch-deutschen Könige und Kaiser von den Karolingern bis zu Maximilian I. (ca. 751–1519) sowie der Päpste des frühen und hohen Mittelalters in Form von deutschsprachigen Regesten. Über 130 000 Regesten zu Urkunden sind online verfügbar[15] und mit CC-BY 4.0 lizenziert. Die CEI Daten der Regesten können über eine REST-Schnittstelle oder als Download abgerufen werden.

2015 nahm das Projekt mit seinen Daten am Kulturhackathon ›Coding da Vinci‹[16] teil. Bei der Veranstaltung stellten 30 Kulturinstitutionen ihre Kulturdaten zur Verfügung. Teilnehmer aus der Entwickler- und der Design- Community konnten Teams bilden und über einen Zeitraum von mehreren Wochen Anwendungen und Konzepte mit den Daten entwickeln, die in einer Abschlussveranstaltung vorgestellt wurden. Hintergrund der Veranstaltung war zum einen eine neue Zielgruppe für das kulturelle Erbe zu begeistern und zum anderen Forschungsinstitutionen die Möglichkeiten, die offene Daten bieten, aufzuzeigen und sie für eine freie Bereitstellung ihrer Daten zu gewinnen. Die Regesta Imperii stellten hierfür Bilder und die Regesten-Texte frei lizenziert zur Verfügung. Damit stand die Teilnahme an dem Hackathon auch im Zusammenhang mit der freien Lizenzierung der Forschungsdaten. Im Rahmen des Hackathon entwickelte ein

[12] »BEACON« https://de.wikipedia.org/wiki/Wikipedia:BEACON, 2.2.2017.

[13] Mainzer Professorenkatalog ist ein Gemeinschaftsprojekt der Akademie, der Johannes Gutenberg Universität Mainz und des Instituts für geschichtliche Landeskunde e.V., online unter http://gutenberg-biographics.ub.uni-mainz.de, 2.2.2017.

[14] Die Registeransicht der erfassten Professoren sowieVerweis zur BEACON-Datei: http://gutenberg-biographics.ub.uni-mainz.de/personen.html, 2.2.2017.

[15] »Regesta Imperii« http://www.regesta-imperii.de, 2.2.2017.

[16] »Kulturhackathon Coding da Vinci« https://codingdavinci.de/, 2.2.2017.

Team aus drei EntwicklerInnen eine Visualisierung zu den Ausstellungsorten. Außerdem reicherten sie das Material um biographische Daten der Kaiser aus der Wikipedia an.[17]

Die seither frei lizenzierten Inhalte fanden darüber hinaus mehrfach Anwendung in der Lehre, so etwa an der Universität Heidelberg, wo sich eine Bachelorarbeit im Bereich Computerlinguistik mit den Daten beschäftigte. Mittels Topic Modelling, dem computergestützten Erkennen von inhaltlichen Ähnlichkeiten, wurden verschiedene Themen, die in der Datensammlung verhandelt werden, modelliert und die einzelnen Regesten diesen Themen zugeordnet.[18] Im Rahmen der Vorlesung ›Big Data‹ der Informatik an der Mainzer Johannes Gutenberg Universität werden die Daten genutzt um mittels informatischer Verfahren Itinerare der Herrscher nachzuzeichnen und zu analysieren.

3.3 Controversia et Confessio: Netzwerkanalyse

Forschungsdaten bieten jedoch nicht nur externen Forschern die Möglichkeit weiterführend mit den Daten zu arbeiten, sie können auch eigene darauf aufbauende Forschungen ermöglichen, so etwa in dem Akademievorhaben ›Controversia et Confessio - Quellenedition zur Bekenntnisbildung und Konfessionalisierung (1548-1580)‹. Das Projekt dokumentiert die theologischen Grundsatzdiskussionen um die authentische Bewahrung von Luthers Erbe, die in der zweiten Hälfte des 16. Jahrhunderts aufbrachen und wesentlich zur Identitätsbildung des Protestantismus Wittenberger Prägung beitrugen. Die Diskussion erfolgte über akademische Disputationsthesen, satirische Lieder, illustrierte Flugblätter und Bekenntnisse sowie Streitschriften. Diese Streitschriften sammelt und ediert das Vorhaben.[19]
Die Forschungsplattform des Projektes bietet einen Zugang zu den Streitschriften in ihren jeweiligen thematischen Bezügen sowie den daran in unterschiedlichen Rollen – als Autoren, Herausgeber, Widmungsempfänger und auch als Adressaten und Gegner – beteiligten Personen.[20]
Die Forschungsdaten sind mit CC BY 4.0 lizenziert.
Aufbauend auf diesen Daten wurden mit den Methoden der Sozialen Netzwerkanalyse die Beziehungen der Gegnerschaften der Autoren und Adressaten der theologischen Streitschriften qualitativ untersucht. Dies sollte zunächst einen einen neuen Blick auf die Daten ermöglichen und entwickelte sich dann zu einem eigenständigen weiterführenden Forschungsansatz.[21] Der Fokus lag dabei auf einer Analyse der spezifischen Struktur des Kommunikationsnetzwerkes, das sich aus der Autor-Gegner-Beziehung der durchaus polemischen Streitschriften ergibt. Neben

---

[17] Das Projekt auf GitHub: https://github.com/danielbaak/imperii-viz/blob/master/README.md, 2.2.2017.

[18] Für eine Zusammenfassung der Ergebnisse der Bachelorarbeit »Tracing Players and Themes in the Regesta Imperii by Training SVMs in a (semi-)Supervised Fashion« von Juri Alexander Opitz vgl. Juri Opitz, Anette Frank »Deriving Players & Themes in the Regesta Imperii using SVMs and Neural Networks« in: *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Berlin 2016, S. 74–83, https://aclweb.org/anthology/W/W16/W16-2108.pdf, 2.2.2017.

[19] Das Forschungsvorhaben ist am Leibniz-Institut für Europäische Geschichte (IEG) angesiedelt. Zum Projekt vgl. http://www.controversia-et-confessio.de/einfuehrung.html, 2.2.2017. Die Bände der Edition erscheinen im Druck, ein erster Band ist zudem als Digitale Edition, erarbeitet von der Herzog August Bibliothek Wolfenbüttel, verfügbar http://diglib.hab.de/edoc/ed000211/start.htm, 2.2.2017.

[20] »Controversia et Confessio« http://www.controversia-et-confessio.de/cc-digital.html, 2.2.2017.

[21] Vgl. Aline Deicke, Anna Neovesky »Contextualizing Controversies of the Post-Lutheran Reformation: A Workflow for Network Analytics Involving Relational and Graph Databases« in: *Proceedings of the 3rd HistoInformatics Workshop*, hrsg. von M. Düring, A. Jatowt, J. Preiser-Kapeller, A. van den Bosch, Krakau 2016, S. 72-76, http://ceur-ws.org/Vol-1632, 2.2.2017.

Reziprozität und Aktivität im Netzwerk wurden vor allem die verschiedenen Arten von Kontroversen näher beleuchtet, die die reformatorische Streitkultur prägten.

4 Eine generische Incipit-Suche für die Gluck Gesamtausgabe

4.1. Die Gluck Gesamtausgabe

Die Gluck Gesamtausgabe ist seit 1978 ein Vorhaben im Programm der Mainzer Akademie. Ziel ist eine historisch-kritische Edition des Gesamtwerks Christoph Willibald Glucks für die Wissenschaft und die musikalische Praxis. Das digitale Werkverzeichnis[22] bietet einen Zugang zu den Werken wie auch daran beteiligten Personen und Quellen. Das Werkverzeichnis ist ›work-in-progress‹, aktuell stehen Informationen zu 50 Opern und rund 600 Personen und deren Funktionen bereit. Zu den Werken sind Informationen zur Entstehung, Uraufführung, Aufführungsorten, beteiligten Personen, Quellen und Literatur.
Erschlossen werden können die Inhalte zum einen über verschiedene Registerzugänge, zum anderen über eine Suche in den Werken und der Literatur. Zusätzlich zur Text-Suche, sollen die Werke auch über ihre charakteristischen Melodien auffindbar gemacht werden. Das heißt, Notenincipits sollen an den Werkteilen angegeben werden können, die Incipits sollen auf der Forschungsplattform angezeigt werden und die Incipits sollen durchsucht werden können und zu den jeweils passenden Werkteilen führen.
Die Umsetzung dieser Desiderate erforderte daher die Entwicklung einer Suchfunktionalität auf den Notenincipits nötig, eine zugrundeliegenden Kodierung der Noten und eine Anzeige der Incipits.

4.2 Kodierung von Musiknoten

Es gibt mehrere Kodierungsstandards zur Musikannotation, die jeweils für bestimmte Anwendungsfälle genutzt werden. Neben dem XML-Standard der Music Encoding Initiative (MEI)[23], der vor allem für den Einsatz in Digitalen Musikeditionen geeignet ist, stehen unter anderem mit MusicXML und Plaine & Easie (PAE)[24] weitere Formate zur Verfügung.

PAE, das bei der Gluck-Gesamtausgabe Anwendung findet, ist ausschließlich für die Annotation von Noten ausgelegt, nicht für die Angabe von anderen – textuellen – Daten. Dies entspricht dem Szenario in dem Projekt, da die Kodierung hier zum einen zur Darstellung der Noten im Web und zum anderen für die Suche genutzt werden. Das Format wird zudem von einem anderen Akademievorhaben bereits mit guten Erfahrungen genutzt. Im Répertoire International des Sources Musicales (RISM) sind auf diese Art und Weise über 1 Million Notenincipits an Handschriften erfasst und durchsuchbar.[25]
Ein weiterer Grund ist, dass, wie in vielen anderen Musikprojekten, für die Notenanzeige im Web das Verovio Toolkit[26] zum Einsatz kommt, das neben anderen Formaten auch PAE unterstützt. Außerdem ermöglicht PAE eine intuitive Nutzung und erfordert keine umfassende Einarbeitung, da es leicht menschenlesbar und nah an der Notennotation ist.

---

[22] »Gluck Gesamtausgabe« http://www.gluck-gesamtausgabe.de, 2.2.2017.

[23] »Music Encoding Initiative« http://music-encoding.org, 2.2.2017.

[24] Die Dokumentation von Plaine & Easie Code: http://www.iaml.info/plaine-easie-code, 2.2.2017.

[25] » Répertoire International des Sources Musicales« https://opac.rism.info, 2.2.2017.

[26] Verovio wird durch das Swiss RISM Office von Laurent Pugin entwickelt und steht unter der LGPLv3 Lizenz: http://www.verovio.org, 2.2.2017.

Ferner kann der Code, der lediglich aus Zahlen, Buchstaben und einigen Sonderzeichen besteht, auch ohne komplexe Umwandlung des Formats direkt in den Suchindex aufgenommen und als Grundlage der Suche genommen werden.

4.3 Funktionalität und Umsetzung

Für die Implementierung der Funktionalität, in Notenincipits zu suchen, wurde eine generische Incipitsuche entwickelt.[27] Sie ermöglicht es, über eine REST-Schnittstelle PAE-kodierte Daten in einen Suchindex hinzuzufügen und auf diesen dann zu suchen. Die IncipitSearch funktioniert zum einen als eigenständige Plattform,[28] die Funktionalitäten können aber auch in eine andere Applikation integriert werden. Für die Gluck-Gesamtausgabe ist die Suche aktuell[29] in einer internen Version im Einsatz, bei der Suchfunktionalitäten und Ergebnisranking getestet und evaluiert werden, bevor weitere Notenincipits annotiert werden und die Funktionalität für die Öffentlichkeit freigeschaltet wird.
Die Anwendung nutzt Open Source Technologien nach und steht selbst unter der MIT Lizenz auf GitHub zur Verfügung, kann also bereits nachgenutzt werden.

4.4 Nachgenutzte Komponenten

Nachgenutzte Technologien und Daten waren bereits für die Entwicklung nötig, da zu Beginn keine PAE-kodierten Incipits vorlagen. Da das RISM seine Daten jedoch in den Formaten MARC-XML und RDF sowie über einen SPARQL-Endpoint CC BY lizenziert zur Verfügung stellt[30], konnten diese als Testdaten für die Entwicklung genutzt werden.[31] Für die IncipitSearch wurde das MarcXML Format genutzt, weil die Daten dort in einer ohne größere Nachverarbeitung nutzbarer Form vorliegen. MARC-XML, ein bibliographisches Datenformat, weist die Incipits in PAE aus und enthält außerdem weitere Informationen zum Datensatz wie url und Titel. Die mittlerweile in PAE erfassten Incipits der Gluck-Gesamtausgabe können mittlerweile ebenfalls über eine REST-Schnittstelle bezogen und an die Suchfunktionalität angeschlossen werden.

Die Anzeige der kodierten Musiknoten im Web erfolgt mit Verovio, einer Bibliothek zur Darstellung von Noten. Das ›JavaScript Toolkit‹ bietet Funktionalitäten zur Anzeige von Musiknoten im Web.[32]

---

[27] Die Applikation ist auf GitHub verfügbar und besteht aus den Komponenten »IncipitSearch« https://github.com/annaneo/incipitSearch [2.2.2017] für die Indexierungs- und Suchfunktionalitäten und »PianoKeyboard« https://github.com/annaneo/pianoKeyboard [2.2.2017] für das Suchinterface. Beide Komponenten wurden von Dipl.-Inf. Gabriel Reimers unter Mitarbeit der Autorin entwickelt. Seit 2017 ist das zentrale Repositorium unter https://github.com/digicademy/incipitSearch [17.2.2018] zu finden.

[28] Seit November 2018 ist IncipitSearch unter https://incipitsearch.adwmainz.net/ in einer ersten Version online verfügbar. Neben der Gluck-Gesamtausgabe können auch Teil der Bestände des RISM und des Servizio Bibliotecario Nazionale (SBN) surchsucht werden.

[29] Stand: Februar 2017.

[30] Der Open Data Service von RISM unter https://opac.rism.info/index.php?id=8&L=0, 2.2.2017.

[31] Die Incipits aus RISM, unter denen etliche Incipits zu Inhalten sind, die auch in der Gluck-Gesamtausgabe verzeichnet sind, konnten nicht in die Gluck-Gesamtausgabe übernommen werden, da RISM die Quellen im Fokus hat und dadurch für einzelne Werkteile, je nach Vorgabe durch die Quelle, teils mehrere Incipit-Varianten aufführt. Die Gluck Gesamtausgabe ist hingegen bestrebt eine Normvariante zu erstellen.

[32] »Verovio JavaScript Toolkit« http://www.verovio.org/javascript.xhtml, 2.2.2017.

›ElasticSearch‹[33] kommt als Suchmaschine zum Einsatz. ElasticSearch basiert auf der frei lizenzierten Programmbibliothek zur Volltextsuche ›Lucene‹ und ist eine der meistgenutzten Suchmaschinen. ElasticSearch kommuniziert mit anderen Ressourcen und Applikationen über REST und ist mit einer freien Softwarelizenz lizenziert.

4.5 Systemarchitektur

Die Systemarchitektur der IncipitSearch folgt dem Architekturmuster von ›Microservices‹, um eine möglichst große Nachnutzung der Gesamtapplikation wie auch einzelner Komponenten zu ermöglichen. Microservice-Architektur bedeutet die Kapselung von einzelnen Funktionalitäten.[34] Diese sind in mehrere Microservices unterteilt, die auch unabhängig voneinander voll funktional sind: [1] Abgreifen der Daten aus dem Repositorium über eine definierbare Schnittstelle. [2] Erstellen des Suchindexes und Einbindung einer Suchmaschine [3] Nutzerschnittstelle mit Klaviatur für die Noteneingabe [4] Schnittstelle für Rückgabe der Ergebnisse.

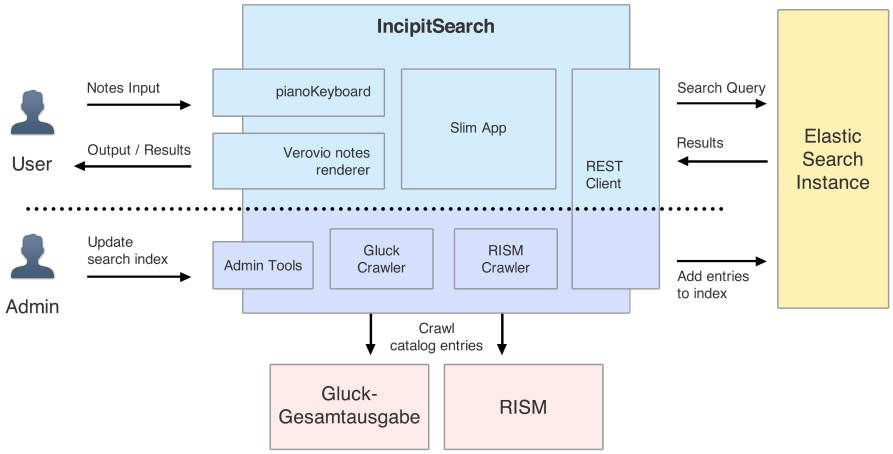

Abb. 1 Systemarchitektur der IncipitSearch.

Die einzelnen Komponenten der Anwendung können in Kombination für ähnliche Repositorien genutzt werden oder einzeln in andere Projekte eingebunden werden. Schnittstellen können entsprechend konfiguriert und zusätzlich benötigte Funktionen ergänzt werden. Entsprechend ist der Einsatz nicht nur auf musikwissenschaftliche Inhalte beschränkt. Diese generische Anwendbarkeit führt nicht nur zu breiteren Einsatzmöglichkeiten, sondern erhöht auch die Zahl der Entwickler, die mit der Software arbeiten und sie weiterentwickeln können. Dadurch kann eine langfristige Nutzbarkeit leichter gewährleistet werden und der entstandene Code nachhaltig in anderen Forschungsapplikationen eingesetzt werden.

5 Fazit: Das machen wir mit den Daten!

Es wurden Beispiele aus verschiedenen Forschungsvorhaben gezeigt, die erst durch die Möglichkeit der Nachnutzung anderer Daten und Applikationen ermöglicht wurden und selbst

[33] »ElasticSearch« www.elastic.co/de/products/elasticsearch, 2.2.2017.

[34] Zur Microservice Architektur vgl. Lars Röwekamp, Arne Limburg »Der perfekte Microservice«, in: *heise online* (2016), https://heise.de/-3091905, 2.2.2017.

wiederum für eine Nachnutzung bereitstehen. Forschungsdaten, die nachhaltig und nachnutzbar sind, müssen zunächst in einem standardkonformen Format vorliegen. Darüber hinaus ist eine offene Lizenzierung notwendig, um Rechtssicherheit über den Grad der Offenheit zu geben. Die Daten können dann bereitgestellt werden. Dies muss, um eine langfristige Nachnutzbarkeit zu ermöglichen, nachhaltig erfolgen.

Die so bereitgestellten Daten können dafür genutzt werden, Ergebnisse von Forschung nachzuprüfen und zu verifizieren. Inhalte können angereichert und vernetzt werden, indem andere Repositorien und Informationen eingebunden werden.
Die Daten können auch zur weitergehenden Forschung genutzt werden und in Studium und Lehre zum Einsatz kommen. Auf dieser Basis neu entwickelte Software sowie Forschung können wiederum unter einer offenen Lizenz frei verfügbar gemacht werden.

Was gibt es noch zu tun, um den Grad an Offenheit und die Möglichkeiten der Nachnutzbarkeit zu erhöhen? Eines ist die bessere Sichtbarkeit und Auffindbarkeit der Daten. So gibt es bereits verschiedene, wenn auch nicht vollständige, Sammlungen von Repositorien zu offenen Daten. Besonders Daten aus kleiner angelegten oder institutionell nicht so stark eingebundenen Forschungsvorhaben müssen bessere Möglichkeiten haben ihre Bestände in Infrastrukturen anzubinden und verfügbar zu machen. Dies betrifft auch im Rahmen von Promotionen entstandene und zum Teil hochspezielle Datenbestände.
Elementar ist auch eine ausführliche Dokumentation der bereitgestellten Daten. Nur so besteht Klarheit über die bereitgestellten Daten und das methodische Vorgehen bei der Ausarbeitung. Auch erleichtert gute Dokumentation den Nutzern den Einstieg in das Material und in die technischen Aspekte bei der Nachnutzung.
Frühzeitige Bereitstellung von Forschungsdaten sollte ebenfalls stärker in den Vordergrund rücken. Meistens werden diese erst zu einem späteren Zeitpunkt des Forschungsvorhabens publiziert. So kommt es dazu, dass viele Daten erst zum Ende des Projektes oder nach Projektabschluss bereitstehen und dann für Nachnutzende zum Beispiel keine Ansprechpartner mehr zur Verfügung stehen. Eine frühzeitige Öffnung – etwa eines Teilbestands als work-in-progress oder in einem niedrigschwelligeren Format wie JSON – ermöglicht bereits eine frühzeitige Nutzung der Daten und kann auch hilfreich für das jeweilige Projekt sein, wenn es zum Beispiel Feedback zu seinen Daten bekommt.
Besonders nützlich ist es, die Daten – wie es von einigen Projekten bereits praktiziert wird – in mehreren Formaten bereitzustellen. Da unterschiedliche Anwendungen verschiedene Formate benötigen und verarbeiten, kann auf das jeweils passende zurückgegriffen werden.

Daher schließt dieser Artikel mit einem Plädoyer für eine frühzeitige Bereitstellung der Forschungsdaten, am besten in verschiedenen Formaten, und für eine generische, gekapselte und wohldokumentierte Ausgestaltung von Software und Applikationen. Gedankt sei allen Institutionen, Forschern und Entwicklern, die ihre Applikationen, Software und Daten verfügbar machen und damit ein Aufbauen auf bereits bestehenden Ideen und darüber hinausgehende Forschung und Neuentwicklungen ermöglichen.

5 Literaturverzeichnis

Eckhart Arnold, Stefan Müller »Wie permanent sind Permalinks?« in: Informationspraxis Bd. 3, Nr. 1 (2017), DOI: http://dx.doi.org/10.11588/ip.2016.2.33483, 2.2.2017.

Aline Deicke, Anna Neovesky »Lizenzierung von digitalen Editionen am Beispiel von Projekten zur jüdischen Geschichte« in: Medaon 9, 17 (2015), http://www.medaon.de/pdf/medaon_17_Deicke+Neovesky.pdf, 2.2.2017

Aline Deicke, Anna Neovesky »Contextualizing Controversies of the Post-Lutheran Reformation: A Workflow for Network Analytics Involving Relational and Graph Databases« in: *Proceedings of the 3rd HistoInformatics Workshop*, hrsg. von M. Düring, A. Jatowt, J. Preiser-Kapeller, A. van den Bosch, Krakau 2016, S. 72-76, http://ceur-ws.org/Vol-1632, 2.2.2017.

Aline Deicke, Anna Neovesky »Lizenzierung bei der Digitalen Akademie«, in: *forschungslizenzen.de* (2017), http://forschungslizenzen.de/lizenzierung-bei-der-digitalen-akademie-mainz, 2.2.2017.

Andreas Kuczera, Yannick Weber, Max Grüntgens, Aline Deicke, Frederic von Vlahovits, Dominik Kasper »Ebenen der Nachnutzung von Forschungsdaten in der Mainzer Akademie« Vortrag im Rahmen der Tagung der AG eHumanities der Union der deutschen Akademien der Wissenschaften, Düsseldorf (2016), http://digicademy.github.io/2016-agehum-ddorf, 2.2.1017.

Juri Opitz, Anette Frank »Deriving Players & Themes in the Regesta Imperii using SVMs and Neural Networks« in: *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Berlin 2016, S. 74–83, https://aclweb.org/anthology/W/W16/W16-2108.pdf, 2.2.2017.

Lars Röwekamp, Arne Limburg »Der perfekte Microservice«, in: heise online (2016), https://heise.de/-3091905, 2.2.2017.

Zitiervorschlag:
Anna Neovesky: …und was machen wir nun mit den Daten? Nutzungsszenarien und deren Voraussetzungen am Beispiel von Akademievorhaben. *Kongresses der Gesellschaft für Musikforschung, Mainz 2016 – »Wege der Musikwissenschaft«*, DOI: 10.5281/zenodo.1175633 [preprint].