SARA-Dienst: Software langfristig verfügbar machen

Franziska Rapp, Stefan Kombrink, Volodymyr Kushnarenko, Universität Ulm
Matthias Fratz, Daniel Scharon, Universität Konstanz

Zusammenfassung:
Software spielt in vielen Disziplinen eine wichtige Rolle im Forschungsprozess. Sie ist entweder selbst Gegenstand der Forschung oder wird als Hilfsmittel zur Erfassung, Verarbeitung und Analyse von Forschungsdaten eingesetzt. Zur Nachvollziehbarkeit der durchgeführten Forschung sollte Software langfristig verfügbar gemacht werden. Im SARA-Projekt zwischen der Universität Konstanz und der Universität Ulm wird ein Dienst entwickelt, der versucht die Einschränkungen bereits bestehender Angebote aufzuheben. Dies beinhaltet u.a. die Möglichkeit, die gesamte Entwicklungshistorie auf einfache Weise mitzuveröffentlichen und für Dritte zur Online-Exploration anzubieten. Zudem bestimmen die Forschenden den Zeitpunkt und Umfang der zu archivierenden/veröffentlichenden Software-Artefakte selbst. Der SARA-Dienst sieht auch die Möglichkeit vor, eine Archivierung ohne Veröffentlichung vorzunehmen. Der geplante Dienst verbindet bereits bestehende Publikations- und Forschungsinfrastrukturen miteinander. Er ermöglicht aus der Arbeitsumgebung der Forschenden heraus eine Archivierung und Veröffentlichung von Software und unterstützt Forschende dabei, bereits prozessbegleitend Zwischenstände ihrer Forschung festzuhalten. Aufgrund seines modularen Aufbaus kann der SARA-Dienst in unterschiedlichen Szenarien zum Einsatz kommen, beispielsweise als kooperativer Dienst für mehrere Einrichtungen. Er stellt eine sinnvolle Ergänzung zu bestehenden Angeboten im Forschungsdatenmanagement dar.

Summary:
Software plays an important role in many scientific disciplines. Whether software itself is the research focus or whether software tools are used to create, process and analyse data – software should be available for the long term to make the research process reproducible. In the context of the SARA project conducted by the University of Konstanz and Ulm University, a service is being developed which aims to avoid restrictions of existing services. This includes the possibility to easily publish the whole change history and make it available for others to explore online. Additionally, the researchers decide when and what they want to archive/publish. The SARA service also allows for archiving of software without making it publicly accessible. It connects existing publication and research infrastructures. Researchers can trigger a publication of software from their research environment and are encouraged to publish software artefacts already during the research process. The new service can be used in various scenarios due to its modular design, for example as a cooperative service for several institutions. The SARA Service is a useful addition to already existing research data management services.

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/2018H2S92-105
Autorenidentifikation: Rapp, Franziska: GND 1131101308; Fratz, Matthias: GND 1097829235; Kombrink, Stefan: GND 1150699604; Scharon, Daniel: GND 1093687045; Kushnarenko, Volodymyr: GND 1147751196
Schlagwörter: Softwarearchivierung, Veröffentlichung von Software, SARA-Service, DSpace, GitLab, Institutionelles Repositorium, Reproduzierbarkeit, Git

1. Motivation

Software spielt seit Jahrzehnten in wissenschaftlicher Forschung in vielen Disziplinen eine essentielle Rolle. Ob zur Erfassung, Analyse und Interpretation von Forschungsdaten oder als eigenständiges Forschungsdatum – Software sollte zur Nachvollziehbarkeit der durchgeführten Forschung längerfristig archiviert und verfügbar gemacht werden.[1] Die Spanne an Software reicht dabei von kleinen Skripten oder Makros über umfangreiche und komplexe Programme bis hin zu kompletten Software-Frameworks. Wir sprechen in diesem Kontext auch von Software-Artefakten.

Um sich im Rahmen von Publikationen auf Software-Artefakte beziehen zu können, sollten diese entsprechend zitierbar sein. Die Software Citation Principles der FORCE 11 „Software Citation Working Group" stellen unter anderem die folgenden Anforderungen an Software-Publikationen:[2]

- Sie sollen durch einen Persistent Identifier zitierbar sein (am besten DOI).
- Diese sollen auf eine jeweilige Software-Version verweisen.
- Eine Landing Page mit zusätzlichen Informationen vor dem eigentlichen Download sollte angezeigt werden.

Um all diesen Anforderungen gerecht zu werden, möchte das Projekt SARA (Software Archiving of Research Artefacts) einen Dienst aufbauen, über den im Rahmen von Forschung entwickelte und angepasste Software-Artefakte nach diesen Vorgaben publiziert werden können. Dabei werden bereits vorhandene Entwicklungs- und Publikationsinfrastrukturen nachgenutzt und über SARA miteinander verbunden.

Der Dienst soll die Forschenden in ihren Workflows begleiten und sie anregen, Zwischenstände ihrer Forschungsarbeit auch bei Software-Artefakten bereits prozessbegleitend festzuhalten. Zur besseren Nachvollziehbarkeit der Forschung sollte die Entwicklungshistorie miterhalten werden. Dies ist insbesondere von Bedeutung, wenn man die übliche Vorgehensweise bei der Softwareentwicklung berücksichtigt, nach welcher komplexere Software-Artefakte nicht komplett eigenständig neuentwickelt werden, sondern auf bestehenden Entwicklungen aufbauen. So verwendet jedes nicht-triviale Programm Programmbibliotheken oder erweitert/ergänzt bestehenden Programmcode.

2. Stand der Technik

In der Software-Entwicklung gilt es seit langem als Best Practice, sämtlichen Quellcode in einem Versionsverwaltungssystem zu hinterlegen, in dem jede Änderung protokolliert und dokumentiert wird. Hierüber werden Änderungen verschiedener Entwickler/innen zusammengeführt und auftretende Konflikte gelöst. Durch diesen Prozess entsteht eine lückenlose Dokumentation des Entstehungsprozesses eines Software-Artefakts, anhand derer sich jede Zwischenversion wiederherstellen

1 Arbeitskreis Open Science, „Helmholtz Open Science: Zugang zu und Nachnutzung von wissenschaftlicher Software," 2017, zuletzt geprüft am 12.02.2018, https://os.helmholtz.de/?id=2766.
2 Arfon M. Smith et al., „Software Citation Principles," *PeerJ Computer Science* 2 (2016): e86, 12-13, https://doi.org/10.7717/peerj-cs.86.

lässt. Im wissenschaftlichen Bereich ermöglicht dies die Reproduzierbarkeit softwaregestützter Datenverarbeitung.

Git bildet unter Versionsverwaltungssystemen mittlerweile den De-facto-Standard, insbesondere bei neuen Projekten, gefolgt von Subversion (SVN), das noch in vielen älteren Projekten genutzt wird. Von diesen beiden Systemen ist Git besser zur Archivierung geeignet: Es ist eine neuere Konzeption mit aktiver Entwicklung und wird damit voraussichtlich länger verfügbar sein. Insbesondere ist Git von vornherein als verteiltes System ausgelegt, mit explizit systemunabhängigen, genau definierten Klartextformaten. SVN-Repositories können zudem leicht in Git-Repositories umgewandelt werden. Jede Arbeitskopie eines Git-Repositories enthält eine vollständige Kopie der gesamten Versionsgeschichte und ist somit, anders als SVN, nicht von einem Server oder Dienst abhängig, auf dem die kanonische Kopie der Versionsgeschichte liegt.

Für Open-Source-Projekte gibt es zahlreiche Anbieter öffentlicher Git-Repositorien. Die bekanntesten sind GitHub[3], SourceForge[4] und Google Code[5] (eingestellt). Die Bereitstellung durch größere Unternehmen bietet eine bessere Persistenz als bei einem lokalen Git-Repository – beispielsweise hat Google bei der Einstellung von Google Code sämtliche Projekte explizit archiviert. Diese Dienste sind aber nicht von vornherein als Archiv ausgestaltet und garantieren in der Regel keine dauerhafte Speicherung, d.h. Projekte könnten grundsätzlich das Schicksal von GeoCities teilen, einem Dienst, der 2009 eingestellt wurde, ohne dass von Anbieterseite eine Archivierung erfolgte.[6]

Mit Zenodo[7] gibt es einen Dienst zur Archivierung von Software, der vom CERN gehostet wird und mit dessen Hilfe beliebige Versionsstände eines Projekts auf GitHub zitierbar (mittels DOI) archiviert und veröffentlicht werden können. Hierüber kann jederzeit der aktuelle Stand der Software zitierbar archiviert werden. Dadurch ist die Reproduzierbarkeit der mit dieser Version verarbeiteten Forschungsdaten gewährleistet. Es wird jedoch nur die jeweilige Version selbst gespeichert. Die Versionsgeschichte kann lediglich im ursprünglichen Projekt auf GitHub nachverfolgt werden. Dieses ist auf Zenodo verlinkt, wird jedoch nicht mitarchiviert – das Projekt kann vom Nutzer nach wie vor jederzeit verändert oder (auch versehentlich) gelöscht werden.

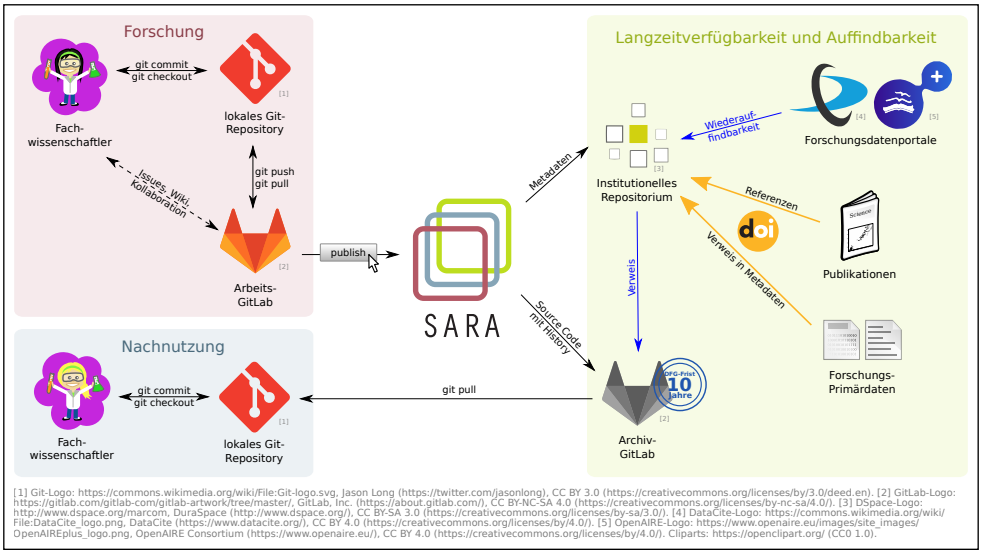
Darüber hinaus existieren zahlreiche weitere, teils fachspezifische Repositorien für Forschungsdaten und Software. Der Schwerpunkt liegt zumeist nicht auf Software. Source Code kann als Ordner oder als ZIP-Archiv desselben hinterlegt werden. Wenn archivierende Nutzer/innen darauf achten, den versteckten git-Ordner mit einzuschließen, bleibt dabei die Versionsgeschichte enthalten und kann nach dem Entpacken genutzt werden. Das Ergebnis, ein ZIP-Archiv einer Working Copy, hat aber den Charakter einer Notlösung. Eine Web-Voransicht, sofern vorhanden, zeigt in der Regel lediglich Dateien und keine Versionen an. Metainformationen wie Commit Messages werden nicht angezeigt.

3 GitHub, https://github.com.
4 SourceForge, https://sourceforge.net/.
5 Google Code Archive, https://code.google.com/archive/.
6 Christian Klaß, „Geocities geschlossen - und für die Nachwelt archiviert," *Golem.de*, 26. Oktober 2009, zuletzt geprüft am 12.02.2018, https://www.golem.de/0910/70706.html.
7 Zenodo, https://zenodo.org/.

Zenodo kommt den gewünschten Eigenschaften am nächsten: Für die archivierten Versionsstände ist eine zuverlässige Speicherung durch das CERN gewährleistet. Zudem kann bei noch aktiven Projekten die Versionsgeschichte im zugehörigen GitHub-Projekt exploriert werden. Der Erhalt dieser Daten ist jedoch nicht garantiert. Insofern fehlt eine Kombination zur garantierten, mittel- bis langfristigen Speicherung der gesamten Versionsgeschichte, kombiniert mit der Möglichkeit diese bequem betrachten und analysieren zu können.

Das Projekt SARA – Software Archiving of Research Artefacts wurde Anfang 2016 als Kooperation der Arbeitsgruppe Verteilte Systeme an der Universität Konstanz, dem Institut für Organisation und Management von Informationssystemen an der Universität Ulm und dem Kommunikations- und Informationszentrum der Universität Ulm gestartet und verfolgt das Ziel, einen neuen Dienst zur Archivierung und Publikation von Software anzubieten, der diese Einschränkungen bestmöglich aufhebt.

3. Der SARA-Workflow: Archivieren & Publizieren von Software



[1] Git-Logo: https://commons.wikimedia.org/wiki/File:Git-logo.svg, Jason Long (https://twitter.com/jasonlong), CC BY 3.0 (https://creativecommons.org/licenses/by/3.0/deed.en). [2] GitLab-Logo: https://gitlab.com/gitlab-com/gitlab-artwork/tree/master/, GitLab, Inc. (https://about.gitlab.com/), CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/). [3] DSpace-Logo: http://www.dspace.org/marcom, DuraSpace (http://www.dspace.org/), CC BY-SA 3.0 (https://creativecommons.org/licenses/by-sa/3.0/). [4] DataCite-Logo: https://commons.wikimedia.org/wiki/File:DataCite_logo.png, DataCite (https://www.datacite.org/), CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). [5] OpenAIRE-Logo: https://www.openaire.eu/images/site_images/OpenAIREplus_logo.png, OpenAIRE Consortium (https://www.openaire.eu/), CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Cliparts: https://openclipart.org/ (CC0 1.0).

Abb. 1:[8] *Die verschiedenen Komponenten des SARA-Service. Der SARA-Dienst ist so konzipiert, dass Arbeits-GitLab, Archiv-GitLab und institutionelle Repositorien mehrfach vorkommen können.*

8 Franziska Ackermann et al., „SARA-Service: Langzeitverfügbarkeit und Publikation von Softwareartefakten“ (Poster auf den E-Science-Tagen in Heidelberg März 2017), https://doi.org/10.11588/heidok.00022887, CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

Der SARA-Workflow in Kürze

Der SARA-Server ist der Dreh- und Angelpunkt für die Archivierung und Veröffentlichung von Software. Er verbindet die verschiedenen Komponenten miteinander. Zu Beginn steht ein Git-Repository auf z.B. einer GitLab-Instanz, in dem die Forschenden einzeln oder kollaborativ an Projekten arbeiten und direkt aus dieser Umgebung heraus eine Archivierung und Veröffentlichung anstoßen können. Wir bezeichnen diese Arbeitsumgebung an dieser Stelle als „Arbeits-GitLab". Der SARA-Server nimmt den Auftrag der Forschenden entgegen, bietet ihnen verschiedene Archivierungs- und Veröffentlichungsoptionen an und fragt weitere Details ab. Anschließend wird die Überführung der Software-Artefakte in ein Archiv-GitLab angestoßen, welches die Langzeitverfügbarkeit sicherstellt. Die URL, unter der die Daten zu finden sind, wird zusammen mit weiteren Metadaten in einem institutionellen[9] Repositorium veröffentlicht. Über die Repositorien erfolgt die Vergabe einer DOI oder eines anderen Persistent Identifier zur dauerhaften und eindeutigen Adressierung der veröffentlichten Software-Artefakte. Der SARA-Dienst berücksichtigt, dass nicht alle Software-Artefakte publiziert werden können (z.B. aus urheberrechtlichen, ethischen, patentrechtlichen Gründen) und bietet daher auch die Möglichkeit, eine reine Archivierung vorzunehmen, bei der die Daten nicht öffentlich zugänglich sind. Über die Repositorien sind Metadaten zu den Software-Artefakten auch in verschiedenen Nachweissystemen wie OpenAIRE[10], BASE[11], Google Scholar[12], DataCite Search[13] etc. zu finden. Dies erhöht die Wiederauffindbarkeit der Daten und ermöglicht es anderen Forschenden, Software-Artefakte bei entsprechender Lizenzierung aus dem Archiv-GitLab in die eigene Forschungsumgebung zu übernehmen und nachzunutzen.

Der große Mehrwert von SARA für Forschende liegt darin, dass sie aus ihrer gewohnten Forschungsumgebung heraus eine Archivierung und optional Veröffentlichung vornehmen können und dabei viele Schritte automatisiert ablaufen. Gleichzeitig liegen relevante Entscheidungen weiterhin bei den Forschenden.

Merkmale des SARA-Service

- Forschende bestimmen Zeitpunkt und Umfang der Veröffentlichung selbst (im Unterschied zu automatisierten Veröffentlichungen ohne Selektionsmöglichkeit)
- Archivierung der Entwicklungshistorie möglich (im Unterschied zu einer Verlinkung auf dynamische, löschbare Projekte)
- Sicherstellung der Langzeitverfügbarkeit über Archiv-GitLab
- Online-Datenexploration im Archiv-GitLab
- Sicherstellung der Auffindbarkeit und Zitierbarkeit über DOI-Vergabe in institutionellen Repositorien

9 Der geplante Dienst kann ebenso zusammen mit fachlichen Repositorien eingesetzt werden.
10 OpenAIRE, https://www.openaire.eu/.
11 BASE Bielefeld Academic Search Engine, https://www.base-search.net/.
12 Google Scholar, https://scholar.google.de/.
13 DataCite, https://search.datacite.org/.

Der SARA-Workflow im Detail

3.1. Forschung

Softwareentwicklung findet zunächst in einem lokalen Git-Repository statt, das sich z.B. auf dem Rechner der Forschenden befindet. Über einen kollaborativ genutzten Git-Server erfolgt die Zusammenarbeit mit anderen Forschenden. Bei lokal an Einrichtungen oder Universitäten betriebenen Diensten kommt in der Regel die Software GitLab zum Einsatz. GitLab bietet ein Web-Frontend, das eine grafische Oberfläche und weitere Funktionalitäten für die Zusammenarbeit bietet (Issues, Wikis etc.). Im Folgenden sprechen wir deshalb von einem „Arbeits-GitLab". Der SARA-Dienst ist so strukturiert, dass eine Anbindung an verschiedene Git-Web-Frontends möglich ist. Eine zusätzliche Anbindung an GitHub wird im Rahmen des Projekts erprobt.

Die Grundlage im SARA-Workflow bildet das Arbeits-GitLab. Darin können einzelne Forscher/innen in einem eigenen Git-Repository arbeiten, z.B. wenn es sich um eine eigenständig durchgeführte Abschlussarbeit handelt, oder es können kollaborativ mehrere Forschende an einem gemeinsamen Projekt arbeiten. Das „Arbeits-GitLab" unterscheidet sich vom „Archiv-GitLab" dadurch, dass es für die gemeinsame Zusammenarbeit genutzt wird, während das „Archiv-GitLab" die Langzeitverfügbarkeit sicherstellt und für Standard-Nutzer nur lesende Rechte beinhaltet.

Für die produktive und effiziente Nutzung eines Versionsverwaltungssystems haben sich in der Softwareentwicklung Best Practices entwickelt. Diese Verhaltensregeln und Workflows sind zunächst besonders für die Softwareentwicklung geeignet, können aber als Grundlage für andere Fachdisziplinen dienen. Sofern einige grundlegende und wenig strittige Konventionen eingehalten werden, kann dadurch auch die Archivierung und Publikation der Software erheblich vereinfacht werden. Dazu zählt insbesondere, Commits mit sinnvollen Metadaten zu versehen – mit dem korrekten Namen der Nutzerin oder des Nutzers oder einem offiziellen Alias, einer offiziellen (bevorzugt institutionellen) E-Mail-Adresse und einer aussagekräftigen Beschreibung der Änderungen. Git unterstützt Nutzer/innen dabei dahingehend, dass ihre Identität gespeichert wird und Commit-Beschreibungen (in sehr eingeschränktem Umfang) generiert werden können.

3.2. SARA-Server

Auf dem SARA-Server können Forschende das GitLab auswählen, in dem sie arbeiten, und ein Projekt selektieren, das ganz oder in Teilen archiviert/veröffentlicht werden soll. Anschließend werden sie auf dem SARA-Server durch verschiedene Schritte geleitet, zu denen die Bestimmung des Umfangs und der Art der Archivierung/Veröffentlichung gehören sowie die Auswahl einer Lizenz und Vergabe von Metadaten.

Forschende nehmen eine Selektion der zu archivierenden/veröffentlichenden Daten vor und entscheiden sich für eine bestimmte Archivierungs- oder Veröffentlichungsoption. Dabei kann einerseits nach Branches selektiert werden, so dass Irrwege, verworfene Ideen oder nicht veröffentlichbare Daten aussortiert werden können, andererseits kann die Versionsgeschichte auch auf die wichtigsten Schritte reduziert werden. Diese Selektion erfolgt auf dem SARA-Server und ist von der verwendeten Repository-Software unabhängig.

Abb. 2: Auswahl von Umfang und Art der Archivierung/Veröffentlichung

Im Bereich der Metadaten werden Informationen automatisch aus Git und GitLab extrahiert, damit Nutzer/innen möglichst wenig selbst eingeben müssen. Orientieren sich Nutzer/innen bei der Arbeit mit Git an bestimmten Best Practices, erfahren sie an dieser Stelle eine Arbeitserleichterung. Eine der Best Practices im Open-Source-Bereich sieht vor, im Wurzelverzeichnis des Git-Repository eine Datei mit dem Namen „LICENSE“ oder „COPYING“ anzulegen, in welcher die Lizenz als Text enthalten ist. Diese Datei wird vom SARA-Dienst automatisch erkannt. Ist keine Datei hinterlegt, müssen Nutzer/innen eine solche Datei anlegen oder während des Archivierungs- /Publikationsprozesses eine Lizenz auswählen, die vom SARA-Dienst als LICENSE-Datei zu den archivierten Daten geschrieben wird. Der SARA-Dienst stellt auf diese Weise sicher, dass eine Lizenz vergeben wurde, die interessierte Dritte als LICENSE oder COPYING-Datei im Wurzelverzeichnis des Git-Repository vorfinden. Somit sind die Möglichkeiten und Grenzen einer Nachnutzung von Anfang an explizit geklärt. Was technisch nicht garantiert werden kann ist, dass Nutzer/innen eine sinnvolle Lizenz vergeben. Dies muss z.B. bei Abschlussarbeiten mit der Betreuerin oder dem Betreuer besprochen werden. Auch Beratungsstellen der Einrichtungen können hier Hilfestellung geben.

Nach erfolgreicher Archivierung der ausgewählten Daten im Archiv-GitLab folgt als nächster Schritt die Veröffentlichung über ein institutionelles Repositorium, sofern anfangs eine Option ausgewählt wurde, die eine Veröffentlichung beinhaltet. In diesem Fall geben die Nutzer/innen ihre E-Mail-Adresse

an, mit welcher sie im Repositorium registriert sind. Dadurch kann der Nachweis in ihrem Namen angelegt werden und die spätere Kommunikation zwischen Repositorien und Nutzer/inne/n wird erleichtert. Für den Nachweis der Software-Artefakte übermittelt der SARA-Dienst die erfassten Metadaten über eine geeignete Schnittstelle ins Repositorium. Da veröffentlichte Software-Artefakte einen DOI erhalten sollen, werden hierfür notwendige Pflichtmetadaten vom SARA-Server als Minimalmetadatensatz abgefragt. Hiervon ausgenommen sind der Identifier, der erst im Repositorium vergeben wird und die Angabe des Publisher, welche die betreibende Einrichtung selbst vornimmt. Zum Minimalmetadatensatz gehören außerdem der Link ins Archiv-GitLab und die Person, welche die Veröffentlichung angestoßen hat.

DataCite Pflichtfelder: [Identifier]; Creator; Title; [Publisher]; Publication Year; Resource Type

Weitere Pflichtfelder: Link ins Archiv-GitLab; Person, welche die Veröffentlichung angestoßen hat

Im Projekt liegt der Fokus auf der Anbindung von DSpace[14], das weltweit als Software für institutionelle Repositorien eingesetzt wird, Open Source ist und eine große und aktive Community besitzt. Sowohl an der Universität Ulm als auch an der Universität Konstanz wird DSpace als institutionelles Repositorium eingesetzt und ist Veröffentlichungsplattform und Universitätsbibliographie zugleich. Das Konzept des SARA-Dienstes sieht einen modularen Ansatz vor, sodass theoretisch auch die Anbindung weiterer Repositorien erfolgen kann, die auf einer anderen Software basieren.

Für DSpace sieht der SARA-Dienst vor, dass verschiedene Varianten der Veröffentlichung unterstützt werden, je nach Philosophie der Einrichtung:

Variante 1 (Workspace): Der SARA-Dienst stößt im Namen der Nutzerin oder des Nutzers eine Submission in DSpace an, die zunächst in den Bereich des Nutzers gelangt (sog. „Workspace“). Die Nutzerin oder der Nutzer loggt sich im institutionellen Repositorium ein, ergänzt die Metadaten ggf. um institutionsspezifische Felder, die ihm in der DSpace-Instanz seiner Einrichtung angeboten werden, stimmt ggf. einem Veröffentlichungsvertrag zu und schließt den Vorgang ab. Sofern in der DSpace-Instanz ein Überprüfungsschritt vor der Freischaltung eingerichtet ist (sog. „Workflow“), ist der Nachweis nicht sofort öffentlich sichtbar, sondern erst nach einer formalen Prüfung durch die Sachbearbeitung. Dieser zusätzliche Schritt ist in DSpace nicht erforderlich, kann jedoch eingerichtet werden. Dies wird von den meisten institutionellen Repositorien mit von Nutzer/inne/n eingereichten Submissions zu Zwecken der Qualitätskontrolle (in der Regel Normierung und Ergänzung von Metadaten) so praktiziert.

Variante 2 (Workflow): Diese Variante setzt voraus, dass in DSpace der Überprüfungsschritt durch die Sachbearbeitung eingerichtet ist. Der SARA-Dienst stößt eine Submission in DSpace an, die direkt in den Bearbeitungspool für die Sachbearbeitung gelangt (sog. „Workflow“). Die Sachbearbeiter/innen prüfen das Item, ergänzen ggf. Metadaten und treten bei Fragen mit der Nutzerin oder dem Nutzer in Kontakt. Eine Ergänzung & Korrektur der Metadaten in DSpace erfolgt durch die

14 DSpace, http://www.dspace.org/.

Sachbearbeiter/innen, da die Nutzerin oder der Nutzer keinen Zugriff auf die Daten hat. Möchte eine Einrichtung, dass Nutzer/innen selbst neben dem Minimalmetadatensatz weitere Felder ausfüllen, so ist vorgesehen, dass die Einrichtung den SARA-Dienst so konfigurieren kann, dass Nutzer/inne/n bereits dort weitere Metadatenfelder ausfüllen können oder müssen. Nach einer formalen Prüfung wird der Nachweis freigeschaltet und ist öffentlich sichtbar.

Variante 3 (Archiv): Wenn weder Variante 1 noch Variante 2 zum Tragen kommen soll, gelangen Nachweise, die vom SARA-Server an DSpace übermittelt werden, direkt in das sog. DSpace-„Archiv", was in der Regel bedeutet, dass sie direkt öffentlich sichtbar sind. Die Entscheidung hierüber liegt beim Betreiber des institutionellen Repositoriums.

3.3. Langzeitverfügbarkeit und Auffindbarkeit

Die Langzeitverfügbarkeit der archivierten bzw. veröffentlichten Software-Artefakte wird über ein Archiv-GitLab sichergestellt. Dort werden die archivierten Daten für mindestens zehn Jahre vorgehalten, entsprechend den Empfehlungen der DFG zur „Sicherung guter wissenschaftlicher Praxis".[15] Als Exitstrategie ist derzeit geplant, im Bedarfsfall sämtliche Dateien in einer ausgecheckten Working Copy des Git-Repository als ZIP-Datei auf einem geeigneten Dienst zur Archivierung abzulegen. Die Versionsgeschichte bleibt dadurch erhalten.

Die Auffindbarkeit wird über institutionelle Repositorien sichergestellt, die zur dauerhaften und eindeutigen Adressierung der veröffentlichten Software-Artefakte DOIs oder andere Persistent Identifier vergeben. Sie können in Publikationen oder anderen Veröffentlichungen für das Zitieren der Software-Artefakte verwendet werden. Über den DOI gelangt man zur Landing Page im Repositorium, auf der beschreibende Metadaten und der Link ins Archiv-GitLab zu finden sind. Über den Link erfolgt der Sprung ins Archiv-GitLab zu den Software-Artefakten. Diese können im Webbrowser exploriert werden und bei Bedarf in die eigene Arbeitsumgebung der Forschenden übernommen werden.

Über die bei Repositorien bereits etablierten Mechanismen zur Verbreitung von Metadaten wird die Wiederauffindbarkeit der veröffentlichten Software-Artefakte erhöht (EU-Portal OpenAIRE, BASE, DataCite Search, Data Citation Index[16], Google Scholar u.a.).

3.4. Nachnutzung

Veröffentlichte Software-Artefakte können im Archiv-GitLab online exploriert werden. Es besteht die Möglichkeit zum Download der archivierten Version sowie der Zwischenversionen (sofern die Entwicklungshistorie mitveröffentlicht wurde). Eine „LICENSE" oder „COPYING" Datei im Git-Repository informiert über die Nutzungsbedingungen. Gegebenenfalls können die Software-Artefakte für weiterführende Forschung in die eigene Arbeitsumgebung übernommen werden.

15 „Deutsche Forschungsgemeinschaft, *Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft"*, Ergänzte Auflage. (Weinheim: WILEY-VCH Verlag GmbH & Co. KGaA, 2013). http://dx.doi.org/10.1002/9783527679188.oth1.

16 „Data Citation Index," Clarivate Analytics, zuletzt geprüft am 15.02.2018, http://wokinfo.com/products_tools/multidisciplinary/dci/.

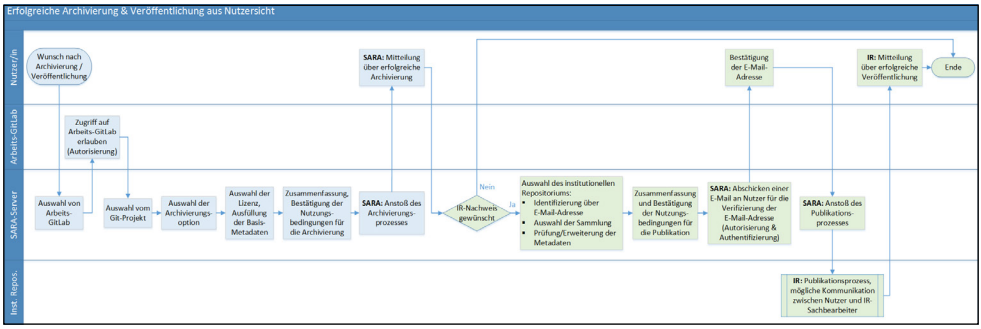


Abb. 3: Der Archivierungs-/Veröffentlichungsprozess aus Nutzersicht

4. Einsatzszenarien für den SARA-Dienst: individuell oder kooperativ

Durch die geplante Bereitstellung der im SARA-Projekt vorgenommenen Entwicklungen als Open Source ergeben sich verschiedene Nachnutzungsmöglichkeiten.

Ein mögliches Szenario für einen SARA-Dienst kann der Einsatz innerhalb nur einer Einrichtung sein: SARA-Server, Arbeits-GitLab, institutionelles Repositorium und Archiv-GitLab werden von derselben Einrichtung betrieben. Der Vorteil liegt darin, dass alles aus einer Hand angeboten wird, die Wege kurz sind und keine Vereinbarungen mit Dritten erfolgen müssen. Dem steht entgegen, dass die Gefahr besteht, dass über die Zeit Einzellösungen entstehen und Synergien durch kooperative Lösungen ungenutzt bleiben.

Die Stärke des Konzepts für einen SARA-Dienst liegt darin, dass der Dienst kooperativ genutzt werden kann. Das E-Science-Projekt SARA fasst einen solchen Dienst für die Universitäten und Hochschulen im Land Baden-Württemberg ins Auge. Der geplante Dienst ist im Kontext der E-Science-Strategie des Landes Baden-Württemberg zu sehen.[17] Ein Arbeits-GitLab kann als Landesdienst angeboten werden, die Authentifizierung würde über bwIDM[18] erfolgen. Alternativ oder zusätzlich können bestehende Arbeits-GitLabs, z.B. einzelner Institute, an den SARA-Dienst angebunden werden. Interessierte Einrichtungen sollen das eigene institutionelle Repositorium (ggf. auch mehrere) an den SARA-Dienst anschließen können. Zunächst ist dies nur für DSpace-basierte Repositorien geplant. Durch den modularen Ansatz kann dies jedoch theoretisch auf weitere Repositorien-Software ausgeweitet werden. Forschende teilnehmender baden-württembergischer Universitäten und Hochschulen können im skizzierten Szenario aus dem Arbeits-GitLab heraus archivieren und veröffentlichen. Der Nachweis der archivierten Software-Artefakte und die Vergabe eines Persistent Identifier erfolgen im institutionellen Repositorium der jeweiligen Einrichtung. Für Forschende der Universität Ulm erfolgen

17 „E-Science: Wissenschaft unter neuen Rahmenbedingungen," Ministerium für Wissenschaft, Forschung und Kunst, zuletzt geprüft am 12.02.2018, https://mwk.baden-wuerttemberg.de/de/forschung/forschungslandschaft/e-science/.

18 „bwIDM | Föderiertes Identitätsmanagement der baden-württembergischen Hochschulen," https://www.bwidm.de/.

Nachweis/Veröffentlichung und DOI-Registrierung beispielsweise in OPARU[19], für Forschende der Universität Konstanz in KOPS[20]. Als Archiv-GitLab soll eine fürs ganze Land bereitgestellte Installation dienen. Der SARA-Server übernimmt die Kommunikation der Komponenten untereinander. Der Vorteil der kooperativen Nutzung liegt darin, dass Arbeits-GitLab, Archiv-GitLab und SARA-Server nicht von jeder Einrichtung einzeln betrieben werden müssten. Der geplante kooperative Dienst orientiert sich an den Infrastruktur-Empfehlungen des Positionspapiers „Zugang zu und Nachnutzung von wissenschaftlicher Software“ des Arbeitskreises Open Science der Helmholtz-Gemeinschaft.[21]

5. Architektur

Die Grafik stellt die Architektur des SARA-Dienstes dar. Da der Dienst als eine Web-Anwendung konzipiert ist, durchläuft ein Nutzer den Archivierungs- und Nachnutzungsvorgang stets mithilfe seines Web-Browsers.

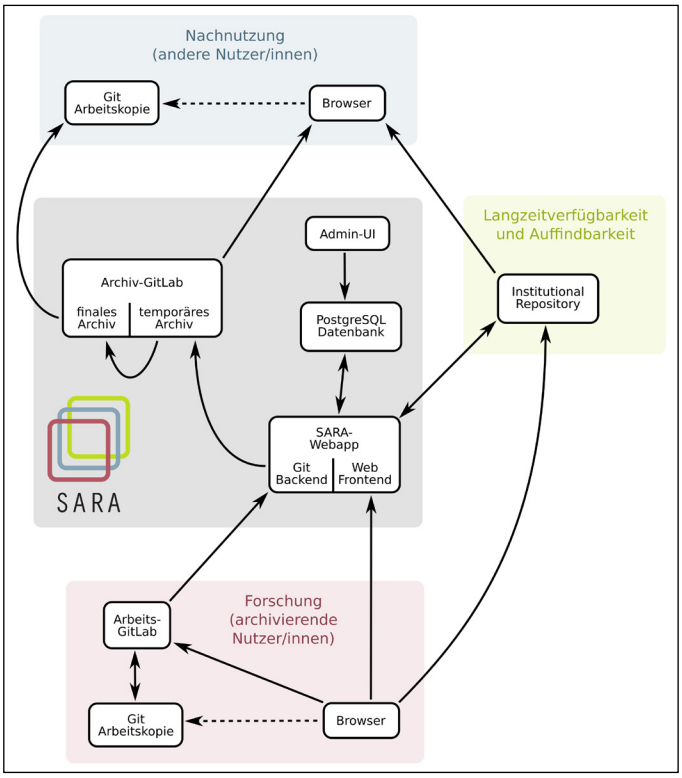


Abb. 4: Die Architektur des SARA-Dienstes

19 OPARU, https://oparu.uni-ulm.de/.
20 KOPS – Das Institutionelle Repositorium der Universität Konstanz, https://kops.uni-konstanz.de/.
21 Arbeitskreis Open Science, „Helmholtz Open Science: Zugang zu und Nachnutzung von wissenschaftlicher Software,“ Punkt 4.4. Infrastrukturen, 2017, zuletzt geprüft am 12.02.2018, https://os.helmholtz.de/?id=2766.

Die Web-Anwendung (SARA WebApp) benötigt zum Betrieb eine Datenbank, die für teilnehmende Einrichtungen zuvor konfiguriert werden muss. Dort lassen sich die Zugangsdaten für die Archiv-GitLabs, institutionellen Repositorien und deren detaillierte Konfiguration wie Default-Kollektionen und Metadaten-Mappings voreinstellen. Zudem werden dort die benutzerdefinierten Metadaten zwischengespeichert, bis der Archivierungs- bzw. Publikationsvorgang abgeschlossen wurde.

Der Web-Dienst wird in Java EE und JavaScript realisiert. Für die Daten wird eine PostgreSQL Datenbank eingesetzt, welche lokal oder remote betrieben werden kann. Vorerst wird die Anbindung an die institutionellen Repositorien mittels SWORD Protokoll und REST Interface umgesetzt, was voraussetzt, dass teilnehmende Einrichtungen für beide Schnittstellen einen Benutzerzugang freischalten müssen.

6. Laufzeitumgebung: Kooperation mit CiTAR

Das SARA-Projekt adressiert die Langzeitverfügbarkeit und Nachweisbarkeit von Software-Artefakten, welche im Rahmen von Forschungsprojekten entwickelt wurden und ermöglicht, dass diese zitiert werden können. Zu Recht ist eine zeitgleiche Archivierung der von den Entwicklern verwendeten Laufzeitumgebung häufig ein Muss-Kriterium, damit die ursprünglichen Ergebnisse der Software verifiziert werden können. Die längerfristige Archivierung von Laufzeitumgebungen wie kompilierte Software, Tool Chains und Betriebssysteminstallationen sprengt die Kapazitäten des SARA-Projektes. Jedoch ist dies eine zentrale Aufgabe des Projektes „Citing and archiving research“ (CiTAR), welches ebenso wie SARA im E-Science-Bereich angesiedelt ist. Für eine mögliche Zusammenarbeit wurden die folgenden Punkte identifiziert:

Forschende können im Git-Projekt eine „Bauanleitung“ für die Entwicklungsumgebung mitführen und -pflegen (Best Practice). Findet SARA beim Publikations- oder Archivierungsvorgang eine solche Anleitung, ist angedacht, dass die Nutzung des CiTAR-Dienstes zur Erstellung und längerfristigen Archivierung der Entwicklungsumgebung angeboten wird. Statt einer Bauanleitung können auch „natürlich gewachsene“ Arbeitsumgebungen an den CiTAR-Dienst übergeben werden. Forschende müssen sicherstellen, dass ihre Entwicklungsumgebung „self-contained“ ist, d.h. keine externen Ressourcen-Abhängigkeiten aufweist. CiTAR gibt das Datenformat vor, in dem die Umgebungen vorliegen müssen, um importiert werden zu können. Derzeit werden die Containerlösungen Docker und Singularity unterstützt, eine Unterstützung für Virtuelle-Maschinen-Abbilder ist geplant. CiTAR unterstützt nur Linux-basierte Entwicklungsumgebungen.

Die Kooperation beider Projekte zielt darauf ab, dass archivierte Software-Artefakte auch zu einem späteren Zeitpunkt wieder in ihrer ursprünglichen Laufzeitumgebung nachnutzbar sind. Die Chance auf eine 1:1-Reproduzierbarkeit wird dadurch erheblich erhöht.

7. Fazit

Mit der Entwicklung des SARA-Dienstes sollen Software-Artefakte, die im Rahmen von Forschung entstehen, bereits prozessbegleitend archiviert und publiziert werden können. Dabei wird darauf geachtet, dass u.a. die eingangs genannten Anforderungen und Empfehlungen von FORCE 11 und des Arbeitskreises Open Science der Helmholtz-Gemeinschaft berücksichtigt werden. Wichtige Merkmale des SARA-Dienstes sind, dass Zeitpunkt und Umfang der Archivierung/Veröffentlichung selbst bestimmt werden können. Dabei ist auch eine Archivierung ohne Veröffentlichung möglich. Die Änderungshistorie kann zudem ebenfalls mitarchiviert werden und auf einfache Weise mitpubliziert und für Dritte online zur Exploration angeboten werden. Das modulare Konzept des geplanten SARA-Dienstes ermöglicht verschiedene Einsatzszenarien, beispielsweise als kooperativ betriebener Landesdienst und stellt eine sinnvolle Ergänzung zu bestehenden Services einer Einrichtung im Bereich des Forschungsdatenmanagements dar.

Acknowledgements

Das SARA-Projekt dankt dem Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg für die Förderung des Projektes.

Literaturverzeichnis

- Ackermann, Franziska, Petra Enderle, Matthias Fratz, Vladimir Kushnarenko, Daniel Scharon, Pia Schmücker, Marcel Waldvogel und Stefan Wesner. „SARA-Service: Langzeitverfügbarkeit und Publikation von Softwareartefakten“ (Poster auf den E-Science-Tagen in Heidelberg März 2017). https://doi.org/10.11588/heidok.00022887.

- Arbeitskreis Open Science. „Helmholtz Open Science: Zugang zu und Nachnutzung von wissenschaftlicher Software.“ 2017. Zuletzt geprüft am 12.02.2018. https://os.helmholtz.de/?id=2766.

- „bwIDM | Föderiertes Identitätsmanagement der baden-württembergischen Hochschulen.“ Zuletzt geprüft am 13.02.2018. https://www.bwidm.de/.

- Clarivate Analytics. „Data Citation Index.” Zuletzt geprüft am 15.02.2018. http://wokinfo.com/products_tools/multidisciplinary/dci/.

- Deutsche Forschungsgemeinschaft. *Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“*. Ergänzte Auflage. Weinheim: Wiley-VHC, 2013. https://doi.org/10.1002/9783527679188.oth1.

- Klaß, Christian. „Geocities geschlossen - und für die Nachwelt archiviert.“ *Golem.de*. Zuletzt geprüft am 12.02.2018. https://www.golem.de/0910/70706.html.

– Ministerium für Wissenschaft, Forschung und Kunst. „E-Science: Wissenschaft unter neuen Rahmenbedingungen.“ Zuletzt geprüft am 12.02.2018. https://mwk.baden-wuerttemberg.de/de/forschung/forschungslandschaft/e-science/.

– Smith, Arfon M., Daniel S. Katz, Kyle E. Niemeyer und FORCE11 Software Citation Working Group. „Software Citation Principles.“ *PeerJ Computer Science* 2 (2016): e86. https://doi.org/10.7717/peerj-cs.86.