Umgang der DELFI-Community mit Forschungsdaten und Softwareartefakten

Eine Erhebung auf Basis der Tagungsbände im Zeitraum 2018-2022

Wabi Melkamu Jate[1] und Michael Striewe[2]

Abstract: Um Forschungsergebnisse validieren und weiterverwenden zu können, ist ein möglichst umfassender Zugriff auf die zugrundeliegenden Forschungsdaten notwendig. Die FAIR-Prinzipien geben dazu Leitlinien, die für eine umfassende Veröffentlichung von Daten befolgt werden sollten. Der vorliegende Beitrag untersucht, in wie weit die Veröffentlichungen der DELFI-Tagungen von 2018 bis 2022 Forschungsdaten auffindbar und zugänglich machen. Das Ergebnis zeigt, dass bisher nur ein Bruchteil der Daten verfügbar ist, wobei Softwareartefakte tendenziell besser verfügbar sind.

Keywords: Open Science, Open Data, Open Source, FAIR-Prinzipien, DELFI-Community

1 Einleitung

Wissenschaftliche Tagungen dienen der Veröffentlichung von Forschungsergebnissen, um innerhalb der Fachcommunity inhaltliche und methodische Fortschritte zu erzielen, Erkenntnisse zu verbreiten und Resultate unabhängig zu validieren. Dafür, aber auch zum Aufgreifen und Fortführen existierender Ideen ist es notwendig, auf die bei Forschungsaktivitäten anfallenden Daten zuzugreifen. In diesem Kontext sind Daten in einem breiten Sinne zu verstehen und umfassen z. B. sowohl Messdaten als auch Gesprächsprotokolle und Softwareartefakte. Insbesondere Software und IT-Infrastruktur erfahren im Rahmen der Veröffentlichung von Forschungsergebnissen oft eine geringe Beachtung [KS22], obwohl gerade in einer interdisziplinären Fachcommunity wie die der DELFI-Tagung, in der u. a. Forschende aus Informatik, Fachdidaktik und Psychologie kooperieren, viele unterschiedliche Daten involviert sind. Um z. B. eine Technologie in einem veränderten Kontext zu erproben, muss die Software verfügbar sein; um Erprobungen aus verschiedenen Kontexten zu vergleichen, müssen die dabei entstandenen Log- und Erhebungsdaten verfügbar sein, und um eine Evaluationsmethode weiterverwenden zu können, müssen Befragungsbögen, Interviewfragen und ähnliches Material verfügbar sein.

Um die Veröffentlichung derartiger Daten zu fördern, wurden 2016 die sogenannten FAIR-Prinzipien veröffentlicht [Wi16], die konkrete Anleitung zur Veröffentlichung von

[1] Universität Duisburg-Essen, Universitätsstraße 2, 45141 Essen, wabi.melkamu-jate@stud.uni-due.de
[2] Universität Duisburg-Essen, paluno – The Ruhr Institut for Software Technology, Gerlingstraße 16, 45127 Essen, michael.striewe@paluno.uni-due.de, https://orcid.org/0000-0001-8866-6971

doi: 10.18420/delfi2023-27

Daten geben, um sie für eine weitere Nutzung verfügbar zu machen. Das Akronym FAIR steht dabei für die Auffindbarkeit (Findable), Zugänglichkeit (Accessible), Interoperabilität (Interoperable) und Wiederverwendbarkeit (Reusable) von Daten.

Der vorliegende Beitrag konzentriert sich als erster Versuch einer summarischen Analyse auf die ersten beiden Aspekte und untersucht, wie Forschungsdaten in Publikationen auf der DELFI-Tagung verfügbar gemacht werden. Im Vordergrund steht daher die Frage, ob Forschungsdaten zu DELFI-Beiträgen leicht auffindbar sind (indem Beiträge direkt auf ihren Speicherort verweisen) und ob die Daten dort auch tatsächlich zugänglich sind (indem enthaltene Verweise auf Ressourcen zeigen, die ohne Einschränkung zugreifbar sind). Die Untersuchung beschränkt sich dabei auf die fünf DELFI-Tagungen der Jahre 2018 bis 2022 und deckt damit einen Zeitraum ab, in dem die FAIR-Prinzipien der DELFI-Community bekannt gewesen sein können. Eine detaillierte Analyse der Vollständigkeit und Qualität der bereitgestellten Daten sowie eine Aufschlüsselung nach Beitragsformen erfordert weiteren Aufwand und muss nachfolgenden Publikationen vorbehalten bleiben.

2 Methode

Für die Erhebung wurden alle Beitragskategorien der DELFI-Hauptkonferenz berücksichtigt. Beiträge zu den Workshops wurden nicht berücksichtigt. Alle Beiträge wurden vom Erstautor des vorliegenden Beitrags daraufhin untersucht, ob ihrem Inhalt Forschungsdaten zugrunde liegen. Als Indikatoren dafür wurde die Forschungsmethode (sofern beschrieben) auf Tätigkeiten überprüft, bei der Forschungsdaten anfallen. Ferner wurden die textuell beschriebenen Ergebnisse sowie Abbildungen (sofern vorhanden) daraufhin untersucht, ob sie Ausschnitte aus Forschungsdaten oder aggregierte Daten enthalten. Im Ergebnis wurde für jeden Beitrag festgehalten, ob ein unmittelbarer Bezug zu Forschungsdaten vorliegt. Die so gewonnenen Erhebungsdaten wurden vom Zweitautor des vorliegenden Beitrags in einer zweiten Sichtung aller Beiträge überprüft.

Für alle Beiträge, die Forschungsdaten enthalten wurde geprüft, ob im Beitrag Verweise auf den Veröffentlichungsort der Daten in Form einer URL, einer DOI oder eines Verweises auf eine andere schriftliche Publikation enthalten sind. Im Fall von URLs und DOIs wurde ferner geprüft, ob diese Verweise tatsächlich zu einer zugreifbaren Ressource führen. Bei schriftlichen Publikationen wurde überprüft, ob es eine solche Veröffentlichung gibt. Eine inhaltliche Prüfung der Vollständigkeit oder Qualität der referenzierten Daten wurde in keinem der Fälle vorgenommen. Es wurde allerdings erhoben, ob im Beitrag selbst Informationen dazu gegeben wurde, wenn nur ein Teil der Forschungsdaten über die genannten Referenzen verfügbar war, beispielsweise lediglich ein verwendeter Fragebogen, nicht jedoch alle darüber erhobenen Rohdaten.

Alle genannten Daten wurden in Form einer Excel-Tabelle[3] erhoben und wurden anschließend manuell ausgewertet, um einen summarischen Überblick zu erhalten. Da in

[3] Verfügbar unter https://doi.org/10.5281/zenodo.7774014

der Bildungstechnologie Softwareartefakten als Informatik-spezifische Form von Forschungsdaten eine besondere Rolle spielen, wurden diese in der Erhebung gesondert erfasst. In den nachfolgenden Ergebnissen meint „Forschungsdaten“ daher alle Daten außer Softwareartefakte, die jeweils getrennt diskutiert werden.

3 Ergebnisse

Im Folgenden werden zunächst die Ergebnisse der Erhebung nach Jahren betrachtet und anschließend Beobachtungen über den Zeitverlauf zusammengefasst. Zuletzt werden weitere Beobachtungen aufgeführt, zu denen keine systematische Erhebung erfolgte.

3.1 Nach Jahren

Die Zahlen für Forschungsdaten ohne Softwareartefakte sind in Tab. 1 zusammengefasst. In gut 50% bis knapp 70% der Beiträge jedes Jahres konnten Forschungsdaten identifiziert werden. Insbesondere in Beiträgen der Kategorie „Demo“ sind dabei in der Regel keine Forschungsdaten enthalten, da sich diese Beiträge auf die Vorstellung eines Softwareartefakts konzentrieren. Ebenso gibt es in jedem Jahrgang Beiträge, die Forschungskonzepte und theoretische Überlegungen vorstellen, die nicht durch Forschungsdaten untermauert werden. Während Forschungsdaten in 14 Fällen vollständig auffindbar und zugänglich sind, sind in 17 Fällen nur Teile auffindbar und zugänglich.

Jahr	Beiträge	enthalten	voll zugänglich	teilweise zugänglich	nicht zugänglich	Anteil voll oder tlw. zugänglich
2018	44	30 ≅ 68%	2 (1)	3 (2)	25	10%
2019	59	33 ≅ 56%	4	3 (2)	26	18%
2020	60	39 ≅ 65%	5	2	32	18%
2021	63	43 ≅ 68%	1	3	39	9%
2022	49	25 ≅ 51%	2	6	17	32%

Tab. 1: Anzahl der Beiträge, die Forschungsdaten enthalten und deren Zugänglichkeit. Zahlen in Klammern geben an, wie viele der per URL zugänglich gemachten Daten tatsächlich verfügbar sind. Bei fehlender Angabe in Klammern sind alle URLs noch verfügbar.

Die separat erfassten Zahlen für Softwareartefakte sind in Tab. 2 dargestellt. Knapp 60% bis etwa 70% der Beiträge jedes Jahres beziehen sich auf konkrete Softwareartefakte als Forschungsgegenstand. Dies trifft wie erwartet insbesondere auf Beiträge der Kategorie „Demo“ zu. Regelmäßig anzutreffende Beiträge zu Erhebungen und Befragungen weisen dagegen zwar Forschungsdaten, aber keine Softwareartefakte auf. Anders als bei den Forschungsdaten ist die teilweise Zugänglichkeit von Softwareartefakten die deutliche Ausnahme und tritt nur in vier von insgesamt 59 Fällen zugänglicher Softwareartefakte auf.

Jahr	Beiträge	enthalten	voll zugänglich	teilweise zugänglich	nicht zugänglich	Anteil voll oder tlw. zugänglich
2018	44	31 ≅ 70%	7 (5)	2	22	23%
2019	59	34 ≅ 58%	9	0	25	27%
2020	60	36 ≅ 60%	10	1	25	31%
2021	63	40 ≅ 63%	19 (18)	0	21	45%
2022	49	34 ≅ 69%	10	1	23	32%

Tab. 2: Anzahl der Beiträge, die Softwareartefakte enthalten und deren Zugänglichkeit. Zahlen in Klammern geben an, wie viele der per URL zugänglich gemachten Artefakte tatsächlich verfügbar sind. Bei fehlender Angabe in Klammern sind alle URLs noch verfügbar.

3.2 Vergleich über die Jahre hinweg

Der Jahresvergleich zeigt keine starken Trends. Insbesondere bei den Forschungsdaten sind Schwankungen zu beobachten. Während 2019 und 2020 jeweils 18% der Beiträge und in 2022 sogar 32% der Beiträge Forschungsdaten zumindest teilweise auffindbar und zugänglich gemacht haben, trifft dies für 2018 und 2021 jeweils auf nur 10% der Beiträge zu. Bei den Softwareartefakten ist für die Jahre 2018 bis 2021 eine kontinuierliche Steigerung von 23% auf 45% der Beiträge zu erkennen, die Softwareartefakte auffindbar und zugänglich machen. Im Jahr 2022 sinkt dieser Anteil jedoch wieder auf 32% und liegt damit auf demselben Niveau wie bei den Forschungsdaten.

3.3 Weitere Beobachtungen

Die Nutzung dedizierter Dienste für die langfristige Verfügbarmachung von Forschungsdaten oder Softwareartefakten wurde nur selten beobachtet. Viermal wird eine DOI des Dienstes Zenodo[4] angegeben. Dreizehnmal wird auf GitHub[5] verwiesen, wobei dies nicht ausschließlich für Softwareartefakte der Fall ist. In allen anderen Fällen beziehen sich Verweise auf augenscheinlich projektspezifische URLs oder Webseiten von Hochschulen.

4 Diskussion

Zunächst muss festgestellt werden, dass eine eher geringe Auffindbarkeit und Verfügbarkeit von Forschungsdaten und Softwareartefakten gegeben ist. Mit Ausnahme von Software im Jahr 2021 liegt die Quote durchweg bei unter einem Drittel; bei den Forschungsdaten mit Ausnahme von 2022 sogar bei unter einem Fünftel. Als überraschend erwies sich, dass oft sogar bei Demo-Beiträgen, die explizit der Vorstellung eines Software-

[4] https://zenodo.org/
[5] https://github.com/

Artefakts dienen, keine Verweise zu finden waren, unter denen das Artefakt auffindbar und zugänglich gemacht wurde. In einigen dieser Fälle wurde zumindest der Name der Software genannt. Fraglich ist jedoch, ob alleine das (ggf. zuzüglich des Hinweises auf eine Verfügbarkeit im App-Store) als hinreichend für die Auffindbarkeit gelten sollte. Immerhin könnte auf diesem Weg inzwischen eine neuere oder auch gänzlich veränderte Version der Software verfügbar sein, als diejenige, die in dem Beitrag besprochen wird.

Ebenfalls kritisch zu beurteilen sind URLs, die auf interne Webseiten verweisen. Während die Notwendigkeit zur Registrierung der Zugänglichkeit nicht prinzipiell im Wege steht, ist dies anders, wenn Accounts nur an Angehörige einer bestimmten Organisation vergeben werden. Die betroffenen Daten sind damit über die URL zwar eindeutig auffindbar, faktisch aber für einen Großteil der Forschungscommunity nicht zugänglich.

Positiv zu bemerken ist die längerfristige Verfügbarkeit der Forschungsdaten. Nur in insgesamt sechs Fällen sind die Ziele von in Beiträgen angegeben URLs bereits nicht mehr erreichbar. Vier dieser Fälle liegen allerdings im Jahr 2018, so dass die Verfügbarkeit von Daten nach vier bis fünf Jahren möglicherweise spürbar abnimmt. Die Prüfung dieser Hypothese erfordert weitere Untersuchungen älterer Jahrgänge der Proceedings.

Bei der Auswertung erwies es sich in einigen Fällen als schwierig, Forschungsdaten und Softwareartefakte eindeutig zu klassifizieren bzw. eine klare Trennung zu ziehen, wann ein Softwareartefakt in einem so engen Bezug zum Beitrag steht, dass seine Auffindbarkeit und Verfügbarkeit sichergestellt werden sollte. Insbesondere zwei Sachverhalten scheinen im Kontext der DELFI-Community vermehrt aufzutreten:

- Die Forschungsdaten enthalten digitale Artefakte, die nur mit der zugehörigen Software nutzbar sind, bei denen die Software selber aber nicht primärer Gegenstand der Forschung ist. Ein Beispiel dafür sind Lerneinheiten in einem frei verfügbaren oder kommerziellen Lern-Management-System. In der vorliegenden Erhebung wurde in solchen Fällen jeweils die Auffindbarkeit und Zugänglichkeit einer lauffähigen Version im Sinne eines Softwareartefakts gewertet, da der Untersuchungsgegenstand ansonsten nicht nachvollziehbar ist.
- Es ist ein Softwareartefakt Gegenstand der Forschung, das selbst wiederum ein anderes Softwareartefakt als Grundsystem benötigt, um lauffähig zu sein. Beispiele dafür sind Plugins für Lern-Management-Systeme oder Erweiterungen für Spieleplattformen. In der vorliegenden Erhebung wurde in solchen Fällen jeweils nur das Plugin bzw. die Erweiterung als relevantes Softwareartefakt betrachtet, nicht jedoch das zugehörige Grundsystem. Dies geschah in der Annahme, dass der Forschungsgegenstand insgesamt nur für denjenigen Personenkreis relevant ist, dem das Grundsystem ohnehin zur Verfügung steht.

Beide Annahmen sollten für eine detailliertere Untersuchung ggf. überdacht oder die Kategorisierung von Forschungsdaten verfeinert werden.

Schließlich sollte bei der Beurteilung der Ergebnisse berücksichtigt werden, dass die DELFI-Tagung bisher keine explizite Policy hat, die Autorinnen und Autoren zur

Veröffentlichung von Forschungsdaten verpflichtet oder dafür formale Vorgaben macht. Beim Vergleich mit anderen Konferenzen muss daher auch berücksichtigt werden, ob es dort solche Policies gibt und ob diese ggf. sogar die Bereitstellung von Begleitmaterial zu einem Paper untersagen [KS23].

5 Fazit und Ausblick

Die Erhebung zeigt, dass in der DELFI-Community bereits Ansätze für die systematische Bereitstellung von Forschungsdaten existieren, indem in allen untersuchten Jahrgängen Beiträge aufgefunden wurden, die Forschungsdaten und Softwareartefakte vollständig auffindbar und zugänglich machen. Gleichzeitig kann festgestellt werden, dass derartige Beiträge in der Minderheit sind und dedizierte Dienste für die langfristige Bereitstellung von Forschungsdaten selten genutzt werden.

Die Untersuchung in diesem Beitrag ist keineswegs als abschließend zu betrachten. Es wurde insbesondere nicht untersucht, ob die verfügbaren Forschungsdaten vollständig und qualitativ geeignet sind, um die jeweilige Forschungsaktivität vollständig nachvollziehen zu können. Bei Softwareartefakten wurde nicht untersucht, ob die Artefakte lediglich zur Nutzung bereitstehen oder ob sie als Open Source Software verfügbar sind. Auch die Lauffähigkeit von Software sowie die Reproduzierbarkeit von aggregierten Ergebnissen auf Basis der veröffentlichten Rohdaten und Auswertungsskripte wurde nicht geprüft. Außerdem wurde auf eine Analyse von Metadaten für Forschungsdaten und Softwareartefakte gänzlich verzichtet. Alle diese Aspekte bieten Raum für zukünftige Erhebungen; letzterer insbesondere auch mit Blick auf den Entwurf eines Metadaten-Schemas, das für die Verfügbarmachung von Forschungsdaten im Fachgebiet der Bildungstechnologie dienlich ist. Dabei müssen auch die am Ende von Abschnitt 4 genannten Beobachtungen berücksichtigt werden, nach denen eine einfache Trennung zwischen reinen Softwareartefakten und sonstigen Forschungsdaten nicht ausreichend zu sein scheint oder gar nicht ohne weiteres möglich ist.

Literaturverzeichnis

[KS22] Kiesler, N., Schiffner, D.: On the Lack of Recognition of Software Artifacts and IT Infrastructure in Educational Technology Research. In: 20. Fachtagung Bildungstechnologien (DELFI). Bonn: Gesellschaft für Informatik e.V., S. 201-206, 2022. https://doi.org/10.18420/delfi2022-034

[KS23] Kiesler, N., Schiffner, D.: Why We Need Open Data in Computer Science Education Research. In: Proc. 28th annual ACM conference on Innovation and Technology in Computer Science Education (ITiCSE), 2023.

[Wi16] Wilkinson, M. et al.: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018, 2016. https://doi.org/10.1038/sdata.2016.18