HUBzero als open-source Science Gateway im Rahmen des Science Data Centers BioDATEN

Holger Gauza, Fabian Wannenmacher, Johannes Werner, Thomas Zajac und Jens Krüger

Eberhard Karls Universität Tübingen, High Performance und Cloud Computing Gruppe, Zentrum für Datenverarbeitung, Wächterstraße 76, 72074 Tübingen, Germany

Der offene Austausch von Forschungsdaten ist essentiell für die Kollaboration von Forscherinnen und Forschern und steigert den Erkenntnisgewinn. Im Unterschied zu Plattformen für Datenpublikation oder kollaboratives Arbeiten ermöglichen Science Gateways eine umfangreiche Integration von Storage- und Compute-Infrastrukturen sowie die Anbindung von Repositorien für fachspezifische Communities. Darüber hinaus bieten Science Gateways Module zur Interaktion mit anderen Forscherinnen und Forschern und zur Dokumentation von Prozessen, Know-how und Metadaten. Der Zugang zu Analysewerkzeugen, Storage und Modulen zur Erleichterung von Projekt- und Wissensmanagement erfolgt webbasiert. Durch die breite Integration von Infrastrukturen und Modulen unter anderem zur Analyse, Speicherung und Veröffentlichung von Forschungsdaten decken sie den gesamten Lebenszyklus von Forschungsdaten ab. Zusätzlich können Science Gateways dazu dienen, die öffentliche Sichtbarkeit von wissenschaftlichen Communities und Forschungsgebieten zu verbessern. Das Science Data Center BioDATEN baut ein solches Science

1 Einleitung

Science Gateways dienen einer wissenschaftlichen Community als zentraler und webbasierter Einstiegspunkt zu Ressourcen wie Schulungsmaterialien, Werkzeugen für die Datenanalyse, Datensätzen und erlauben den einfachen Zugang zu verteilten IT-Infrastrukturen für Storage und Computation [1]. Das Science Data Center *Bioinformatics DATa ENvironment* (SDC BioDATEN) baut ein solches Science Gateway auf Basis von HUBzero auf [2]. Der Anspruch eines Science Gateways ist es, Forscherinnen und Forscher über den gesamten Lebenszyklus von Forschungsdaten hinweg zu unterstützen. Darüber hinaus bieten Science Gateways Werkzeuge für die Außendarstellung einer wissenschaftlichen Community wie Newsletter, Blogartikel und Ankündigungen. Ein gutes Beispiel für die

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI: https://doi.org/10.11588/heidok.00029705 veröffentlicht.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13749 (CC BY-SA 4.0)

Umsetzung eines Science Gateways liefert das nanoHUB [3]. Der positive Einfluss von Science Gateways auf die Nutzung verteilter IT-Infrastrukturen konnte bereits nachgewiesen werden. So übertraf die Nutzung verteilter IT-Infrastrukturen über Science Gateways im Jahr 2014 die deren Nutzung via Kommandozeile [4]. Die internationalen Bemühungen um die Verbreitung von Science Gateways werden von dem *Science Gateway Community Institute* (SGCI) und dem *International Workshop on Science Gateways* (IWSG) vorangetrieben [5, 6, 7]. Der Katalog des SGCI listete im April 2021 über 620 Science Gateways aus unterschiedlichen Forschungsfeldern auf [8]. Neben den dort gelisteten Gateways existieren weitere Projekte und Initiativen die sich als Science Gateway beschreiben lassen, wie beispielsweise das CAMPOS Data Cockpit [9], MoSGrid [10] und NFDI4Chem [11]. Im Folgenden werden der gegenwärtige Stand und die weiteren Zielsetzungen für den Aufbau eines Science Gateways für BioDATEN dargestellt.

2 Ein Science Gateway für BioDATEN

BioDATEN setzt für den Aufbau eines Science Gateways für die Bioinformatik-Community in Baden-Württemberg auf HUBzero. HUBzero ist ein Open-Source-Framework und erlaubt durch das integrierte Joomla Content-Management-System (CMS) den Aufbau einer Website für die Außendarstellung des Projekts. Der Vorteil von HUBzero, im Vergleich zur Nutzung eines einfachen CMS, liegt in seinem weitreichenden Funktionsumfang für die wissenschaftliche Community. Durch die Integration von Komponenten, Modulen und Plugins können Funktionen erstellt und eingebunden werden, welche die Nutzerinnen und Nutzer bedarfsgerecht in ihrer Arbeit unterstützen. Darüber hinaus umfasst HUBzero ein vorkonfiguriertes Rechte- und Rollenmanagement, mit dem Berechtigungen für die Administration des Gateways und der Veröffentlichung von Inhalten gesteuert werden können. Ein weiteres, vorkonfiguriertes Modul erlaubt die Erstellung von News-Artikeln und steuert bei Bedarf Zeitpunkt und Dauer der Veröffentlichung. Durch die Kombination von Maßnahmen zur Außendarstellung mit der Integrierbarkeit weiterer Funktionen in einem Science Gateway wird die Community sowohl hinsichtlich ihrer Sichtbarkeit als auch ihrer wissenschaftlichen Arbeit unterstützt. Bei letzterem liegt der Fokus auf frühzeitige Berücksichtigung und Umsetzung des Lebenszyklus von Forschungsdaten.

2.1 Gegenwärtiger Stand

Die öffentliche Webseite des Projekts BioDATEN (*https://portal.biodaten.info*) basiert bereits auf HUBzero. Daneben steht den registrierten Nutzerinnen und Nutzern ein Mitgliederbereich zur Verfügung. Zum Login wird die etablierte *ELIXIR Authentication and Authorization Infrastructure* (AAI) in Kombination mit Keycloak genutzt [12, 13]. Die Anbindung an die ELIXIR AAI hat für die Nutzerinnen und Nutzer den Vorteil, dass diese sich mit den bereits vorhandenen Zugangsdaten der Heimatorganisation am Gateway anmelden können. Der Einsatz von Keycloak erlaubt die Anbindung weiterer Dienste mit gleicher Nutzerbasis. Durch die ELIXIR AAI wird der Login als Single Sign-on

zur Nutzung aller Dienste des Gateways umgesetzt (Abbildung 1). Nach der Anmeldung ermöglicht ein zentrales und konfigurierbares Dashboard eine Übersicht über Gruppen, Projekte und offene Tickets. Zusätzlich können Nutzerinnen und Nutzer über ihre Forschungsarbeit in Blogs berichten, an Kursen teilnehmen und sich in Gruppen austauschen. Besondere Bedeutung kommt der Verwaltung von Projekten zu: Nutzerinnen und Nutzer können eigene Projekte anlegen, Mitarbeiterinnen und Mitarbeiter einladen, Dateien austauschen, To-do-Listen pflegen und Aufgaben zuweisen. Dem jeweiligen Projekt steht eine Speicherplatz-Quota zur Verfügung, die durch die Betreiber angepasst werden kann. Das Gateway ermöglicht auf diese Weise eine zentrale Datenablage und den einfachen Datenaustausch unabhängig von kommerziellen Diensten, E-Mails oder USB-Sticks. In Kombination mit dem Gruppenmodul steht Projekten ein Wiki-System zur Verfügung, um das Wissen im Projekt für alle verfügbar und auf aktuellem Stand zu halten.

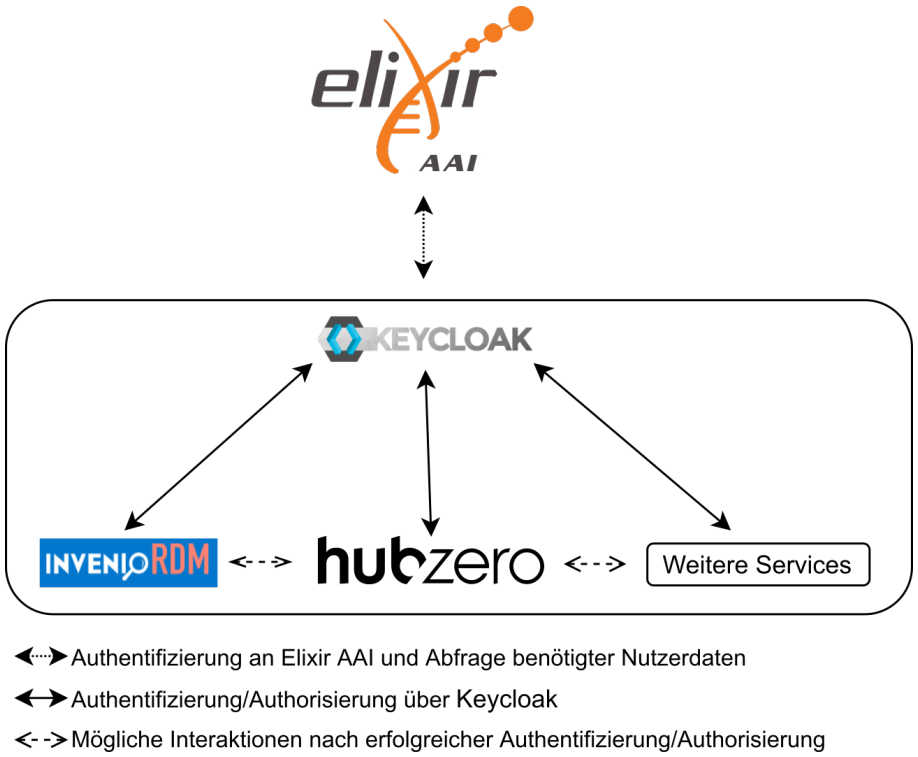

Abbildung 1: Anbindung verschiedener Services via Keycloak und ELIXIR AAI.

2.2 Zukünftige Entwicklung

Die Zielsetzung von BioDATEN liegt in der Unterstützung der Wissenschaftlerinnen und Wissenschaftlern über den gesamten Lebenszyklus der erhobenen Forschungsdaten, von der Datengenerierung über die Annotation mit Metadaten bis zur Datenpublikation. Das zentrale Science Gateway auf Basis von HUBzero dient als erste Anlaufstelle für die Community und bündelt bereits etablierte und neu zu implementierende Ressourcen und In-

frastrukturen. Die Integration etablierter Systeme und bestehender Infrastrukturen trägt zur Nachhaltigkeit des Science Data Centers BioDATEN bei.

Anbindung an die bwHPC-Infrastruktur BioDATEN entwickelt einen Workflow, um direkt bei Lifecycle-Beginn Daten und Metadaten aus der bwHPC-Infrastruktur zu übernehmen [14]. Hierbei dienen Job-Feedback-Skripte des Bioinformatics and Astrophysics Cluster (BinAC) als Ausgangspunkt für die automatische Generierung eines Rumpf-Metadatensatzes, der primär prozessuale Metadaten über die erzeugten Dateien und die verwendeten Ressourcen enthält. Dieser Metadatensatz wird anschließend von den Forscherinnen und Forschern über das Gateway mit Angaben zur Person und wissenschaftlichen Angaben erweitert. Soweit möglich wird dabei automatisch auf bereits vorhandene Metadaten zurückgegriffen, um den Aufwand für die Nutzerinnen und Nutzer so gering wie möglich zu halten. Die notwendigen Angaben richten sich nach DataCite-Schema und wissenschaftlich einschlägigen Schemata, wie beispielsweise dem *Minimum Information about any (x) Sequence* Schema [15, 16]. Durch die Integration des BinAC und vorhandener Metadatenschemata wird ein Forschungsdatensatz früh auf eine mögliche Publikation mit allen benötigten Metadaten vorbereitet.

Datenpublikation Zur Publikation von Forschungsdaten setzt das SDC BioDATEN auf InvenioRDM [17]. Die Publikation soll direkt aus dem Gateway ermöglicht werden, indem die Daten und Metadaten über eine Schnittstelle an InvenioRDM übertragen und dort veröffentlicht werden. Durch den Rückgriff auf das DataCite-Schema liegen alle notwendigen Angaben für die Veröffentlichung und Registrierung eines DOIs vor, so dass Nutzerinnen und Nutzer von der frühen Integration standardisierter Schemata in den Forschungsalltag profitieren können.

Suche Nutzerinnen und Nutzer können die Inhalte des Gateways mithilfe einer entsprechenden Funktion durchsuchen. Zur Realisierung der Suche wird Apache Solr in Kombination mit VuFind eingesetzt [18, 19]. Die Umsetzung der facettierten Suche basiert auf einem eigens erstellten Metadatenschema.

Storage Die Langzeit-Speicherung großer Datenmengen ist eine Herausforderung in datenintensiven Forschungsfeldernwie der Bioinformatik oder Astrophysik. Beim Umgang mit solchen (größtenteils heterogenen) Daten spielt die Wahl eines geeigneten Speichermodells, das den verschiedenen Anforderungen während der Laufzeit eines Projekts genügt, eine wichtige Rolle. In letzter Zeit gewinnt S3-Objektspeicher immer mehr an Bedeutung. Dabei werden Dateien als Objekte abgebildet, die in so genannten Buckets hinterlegt und ähnlich einer Ordnerstruktur mithilfe von Präfixen gruppiert werden können. Im Gegensatz zu traditionellen Dateisystemen müssen sich der Nutzerinn und Nutzer dabei keine Informationen merken, die seine Daten nicht direkt betreffen, wie Laufwerksbuchstabe und den vorausgehenden Pfad, was den Zugriff erleichtert. S3-Objektspeicher erlaubt zudem die Zugriffsbeschränkung auf hinterlegte Objekte und Freigabe dieser unter einer URL.

So können bestimmte Forschungsdaten vor der Publikation auf eine einfache Weise einem wachsenden Kreis an Interessenten freigegeben werden. Das baden-württembergische Landesprojekt *Storage for Science* (bwSFS) stellt S3-Objektspeicher ebenso bereit wie die Cloud des *Deutschen Netzwerks für Bioinformatik-Infrastruktur* (de.NBI) [20, 21]. Den Nutzerinnen und Nutzern wird mittels Gateway Objektspeicher über die de.NBI Cloud angeboten, dessen Nutzung durch die Hinterlegung von S3-Credentials im Gateway erfolgt. Durch diese Integration wird ein wichtiger Beitrag zur Nachhaltigkeit des SDC BioDATEN geleistet und der Nutzen für die Forscherinnen und Forscher erhöht.

Integration von Tools Die enge Anbindung an die bwHPC/BinAC-Infrastruktur, de.NBI-Cloud und die Integration von Galaxy Workflows [22] ermöglicht die Verwendung und Integration vorhandener Workflows und Tools zur Datenanalyse. Die web- und UI-basierte Bereitstellung dieser Tools und Workflows entlang des Lebenszyklus von Forschungsdaten über das Gateway ist ein weiteres Ziel des SDC BioDATEN.

3 Zusammenfassung

Science Gateways ermöglichen den Aufbau eines zentralen Einstiegspunkts für die wissenschaftliche Community zu Diensten, Daten und Materialien. Darüber hinaus erlauben sie den Aufbau einer Website für die Verbreitung von News, Blogs und Öffentlichkeitsarbeit. Ein Beispiel für die Integration von Diensten und verbesserte Sichtbarkeit einer Community ist das nanoHUB [3]. Im Rahmen des Science Data Centers BioDATEN wird ein solches Science Gateway auf Basis von HUBzero aufgebaut, das Wissenschaftlerinnen und Wissenschaftler über den gesamten Lebenszyklus von Forschungsdaten hinweg unterstützen soll. Die Integration in den wissenschaftlichen Alltag erfolgt durch den Rückgriff auf etablierte und verbreitete Infrastrukturen wie die ELIXIR AAI für Login und Accountgenerierung sowie auf de.NBI Cloud und bwHPC/BinAC für die Bereitstellung von Storage und Computation. Die Anbindung an den BinAC ermöglicht die automatische Generierung eines Rumpf-Metadatensatzes, welcher im Laufe des weiteren Lebenszykluses vervollständigt wird. Durch die frühzeitige Erhebung notwendiger Metadaten wird die Veröffentlichung in den späteren Phasen des Lebenszyklus vorbereitet. Die geplante Integration von Analysewerkzeugen und Workflows wird einen weiteren Beitrag zur Integration des Science Gateways in den wissenschaftlichen Alltag leisten und die Unterstützung und Umsetzung des Lebenszyklus von Forschungsdaten weiter ausbauen.

Förderung

Das Science Data Center BioDATEN wird vom Ministerium für Wissenschaft, Forschung und Kunst Baden -Württemberg aus Mitteln der Landesdigitalisierungsstrategie digital@bw gefördert.

Literaturverzeichnis

[1] Nancy Wilkins-Diehr, Michael Zentner, Marlon Pierce, Maytal Dahan, Katherine Lawrence, Linda Hayden, and Nayiri Mullinix. The science gateways community institute at two years. In *Proceedings of the Practice and Experience on Advanced Research Computing*, PEARC '18, New York, NY, USA, 2018. Association for Computing Machinery. https://doi.org/10.1145/3219104.3219142.

[2] Michael McLennan and Rick Kennell. HUBzero: A Platform for Dissemination and Collaboration in Computational Science and Engineering. *Computing in Science Engineering*, 12(2):48–53, 2010. https://doi.org/10.1109/MCSE.2010.41.

[3] nanoHUB. https://nanohub.org/. Zuletzt abgerufen am 29.07.2021.

[4] Katherine A Lawrence, Nancy Wilkins-Diehr, Julie A Wernert, Marlon Pierce, Michael Zentner, and Suresh Marru. Who cares about science gateways? a large-scale survey of community use and needs. In *2014 9th Gateway Computing Environments Workshop*, pages 1–4. IEEE, 2014.

[5] IWSG - International Workshop on Science Gateways. https://sites.google.com/site/iwsglife. Zuletzt abgerufen am 27.04.2021.

[6] Katherine A. Lawrence, Nayiri Mullinix, Maytal Dahan, Linda Hayden, Marlon Pierce, Nancy Wilkins-Diehr, and Michael Zentner. How the science gateways community institute supports those who are creating websites to access shared resources. In *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (Learning)*, PEARC '19, New York, NY, USA, 2019. Association for Computing Machinery. https://doi.org/10.1145/3332186.3333256.

[7] SGCI - Science Gateways Community Institute. https://sciencegateways.org/. Zuletzt abgerufen am 27.04.2021.

[8] Science Gateways Catalog. https://catalog.sciencegateways.org/#/home. Zuletzt abgerufen am 27.04.2021.

[9] M. Finkel, A. Baur, T. K. D. Weber, K. Osenbrück, H. Rügner, C. Leven, M. Schwientek, J. Schlögl, U. Hahn, T. Streck, O. A. Cirpka, T. Walter, and P. Grathwohl. Managing collaborative research data for integrated, interdisciplinary environmental research. *Earth Science Informatics*, 13(3):641–654, Sep 2020. https://doi.org/10.1007/s12145-020-00441-0.

[10] Jens Krüger, Richard Grunzke, Sandra Gesing, Sebastian Breuers, André Brinkmann, Luis de la Garza, Oliver Kohlbacher, Martin Kruse, Wolfgang E. Nagel, Lars Packschies, Ralph Müller-Pfefferkorn, Patrick Schäfer, Charlotta Schärfe, Thomas Steinke, Tobias Schlemmer, Klaus Dieter Warzecha, Andreas Zink, and Sonja Herres-Pawlis. The mosgrid science gateway – a complete solution for molecular simulations. *Journal of Chemical Theory and Computation*, 10(6):2232–2245, 2014. PMID: 26580747. https://doi.org/10.1021/ct500159h.

[11] Nicole Jung, Steffen Neumann, Oliver Koepler, Felix Bach, Christian Popp, Sonja Herres-Pawlis, Johannes Liermann, Matthias Razum, and Christoph Steinbeck. NFDI4Chem – Infrastruktur für den digitalen Wandel in der Chemischen Forschung. *Bunsen-Magazin*, 2021. URL: https://bunsen.de/bmo/nfdi4chem, https://doi.org/10.26125/r978-6f93.

[12] Mikael Linden, Michal Prochazka, Ilkka Lappalainen, Dominik Bucik, Pavel Vyskocil, Martin Kuba, Sami Silén, Peter Belmann, Alexander Sczyrba, Steven Newhouse, et al. Common elixir service for researcher authentication and authorisation. *F1000Research*, 7, 2018.

[13] Keycloak - Open Source Identity and Access Management. https://www.keycloak.org/. Zuletzt abgerufen am 27.04.2021.

[14] bwHPC. https://www.bwhpc.de/. Zuletzt abgerufen am 29.07.2021.

[15] DataCite Metadata Working Group. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.4. 2021. URL: https://schema.datacite.org/meta/kernel-4.4/, https://doi.org/10.14454/3W3Z-SA82.

[16] Pelin Yilmaz, Renzo Kottmann, Dawn Field, Rob Knight, James R Cole, Linda Amaral-Zettler, Jack A Gilbert, Ilene Karsch-Mizrachi, Anjanette Johnston, Guy Cochrane, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature biotechnology*, 29(5):415–420, 2011.

[17] InvenioRDM. https://inveniosoftware.org/products/rdm/. Zuletzt abgerufen am 27.04.2021.

[18] Apache Solr. https://solr.apache.org/. Zuletzt abgerufen am 27.04.2021.

[19] VuFind. https://vufind.org/vufind/. Zuletzt abgerufen am 27.04.2021.

[20] bwSFS - Storage for Science. https://www.alwr-bw.de/bwsfs/. Zuletzt abgerufen am 27.04.2021.

[21] Peter Belmann, Björn Fischer, Jan Krüger, Michal Procházka, Helena Rasche, Manuel Prinz, Maximilian Hanussek, Martin Lang, Felix Bartusch, Benjamin Gläßle, et al. de. NBI Cloud federation through ELIXIR AAI. *F1000Research*, 8, 2019.

[22] E. Afgan, D. Baker, B. Batut, M. van den Beek, D. Bouvier, M. Cech, J. Chilton, D. Clements, N. Coraor, B. A. Grüning, A. Guerler, J. Hillman-Jackson, S. Hiltemann, V. Jalili, H. Rasche, N. Soranzo, J. Goecks, J. Taylor, A. Nekrutenko, and D. Blankenberg. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res*, 46(W1):W537–W544, 07 2018.