Ein Werkzeug zur XSD-basierten Metadatenannotation

Olaf Brandt[1], Holger Gauza[1], Jan Kaltenbach[1], Maximilian E. Müller[2], Gabriel Schneider[2], Claus Zinn[1]

[1]Universität Tübingen;
[2]Universität Konstanz

Der Umgang mit Forschungsdaten entlang des Forschungsdatenlebenszyklus erlangt immer mehr Relevanz und erfordert nicht nur eine gewissenhafte Planung und eine sichergestellte Speicherung der Forschungsdaten, sondern auch die Auszeichnung der Forschungsdaten mit geeigneten Metadaten, um deren Auffindbarkeit und Nachnutzbarkeit im Sinne der FAIR-Prinzipien umzusetzen (Wilkinson u. a. 2016). Aus diesem Grund stehen viele Dienste im Bereich des Forschungsdatenmanagements vor der Herausforderung, ihren Nutzer:innen ein Werkzeug anzubieten, mit dem sie ihre Daten mit passenden Metadatenstandards beschreiben können. Die Metadatenstandards werden dabei aber nur in den wenigsten Fällen durch die Dienstanbieter:innen selbst entwickelt, sondern durch verschiedene Konsortien, wie z.B. in den nationalen Forschungsdateninfrastrukturen (NFDI)[1], entwickelt und gepflegt. Die Dienstanbieter:innen müssen entsprechend auf Änderungen im Upstream reagieren und ihre Dienste und Werkzeuge anpassen. Gleichzeitig muss die Verwendung eines Werkzeugs zur Annotation von Metadaten für die Nutzer:innen unter Aspekten der Usability dahingehend ausgelegt sein, dass diese sich auf die einschlägigen Metadaten konzentrieren können und z.B. durch Dropdownmenüs und Hilfetexte unterstützt werden. Zusätzlich zur Usability tragen Dropdownmenüs mit hinterlegten fachspezifischen Vokabularen und Ontologien zur Qualität der Metadaten bei, da nicht nur Freitext vermieden, sondern auch semantische Verbindungen durch Übernahme von Uniform Resource Identifiern (URI) etc. ermöglicht werden. Unter Aspekten der Nachhaltigkeit ist es außerdem wünschenswert, dass ein Werkzeug nicht spezifisch für nur eine wissenschaftliche Community maßgeschneidert wird, sondern dass es potentiell durch Austausch von Metadatenschemata zur Erfassung relevanter Inforamtionen auf andere wissenschaftliche Disziplinen übertragen und angewendet und somit auch von anderen Communities genutzt werden kann. Es ergeben sich vier grundlegende Anforderungen an ein Werkzeug zur Metadatenannotation:

1. Reduzierter Betreuungsaufwand für Dienstanbieter:innen durch die einfache Integration bzw. den Austausch von Schemata. Integration von Ontologien und Voka-

Publiziert in: Vincent Heuveline, Nina Bisheh und Philipp Kling (Hg.): E-Science-Tage 2023. Empower Your Research – Preserve Your Data. Heidelberg: heiBOOKS, 2023. DOI: https://doi.org/10.11588/heibooks.1288.c18088

1 Für weitere Informationen siehe https://www.nfdi.de.

bularen aus externen Quellen und eine einfache Integration in bestehende Systeme wie universitäre Idendity Provider (IDP) oder Publikationsplattformen.

2. Erhöhte Usability für die Nutzer:innen durch die Integration von Dropdownmenüs mit integrierter Autocomplete-Funktion und dynamische Reduktion der Schemakomplexität auf einschlägige Felder. Berücksichtigung von Anforderungen im Sinne der Barrierefreiheit.

3. Generischer Anspruch und Einsatzmöglichkeit durch den einfachen Austausch von Schemata zur Bereitstellung für weitere Communities. Entsprechend ein Verzicht auf Hardcoding und Nutzung von Standards.

4. Mehrwert für die Community durch die Generierung von maschienenlesbarem Output bei Erhalt semantischer Relationen sowie automatisierte Qualitätskontrolle durch automatische Validierung der Eingaben sowie Vorbereitung für eine Publikation der Forschungsdaten.

Im Folgenden wird ein Werkzeug zur XSD-basierten[2] Metadatenannotation vorgestellt, das einen generischen Ansatz bei der Bereitstellung von Formularen zur Metadatenbeschreibung verfolgt.[3] Anstatt ein spezifisches Metadatenschema zu hinterlegen, können jegliche XSD-basierten Schemata verwendet werden. Aus dieser formalen Beschreibung werden anschließend durch einen XSLT-Processor HTML-Formulare generiert, die Nutzer:innen webbasiert in einem Browser ausfüllen können. Hierdurch wird eine Trennung von Inhalt (XSD) und Form (HTML) erzeugt, die eine erhöhte Flexibilität durch den Austausch von XSD-Schemata bedeutet. Gleichzeitig lassen sich beispielsweise mehrere Schemata integrieren und nebeneinander befüllen, um so eine Metadatenbeschreibung zu erreichen, die exakt auf die Bedarfe der Nutzer:innen der Dienstanbieter:innen zugeschnitten sind. Eine automatisierte Validierung der Ergebnisse ist auf diese Weise ebenfalls möglich, wodurch Fehler direkt bei der Eingabe erkannt und korrigiert werden können. Der Aufwand bei der Implementierung von Aktualisierungen wird dadurch reduziert, dass diese sich direkt in die Formulare integrieren lassen, solange sie als XSD vorliegen. Eine Administrationsoberfläche ermöglicht Anpassungen an der Darstellung der Formulare hinsichtlich Reihenfolge und Sichtbarkeit von Elementen und den Eingabefeldern, um eine Fokussierung auf die einschlägigen Metadaten umzusetzen. Auf diese Weise werden Anpassungen in der XSD-Datei umgangen und gleichzeitig Flexibilität erzeugt. Das Annotationstool ist kompatibel mit allen IDPs, die OIDC[4] anbieten.

1 Reduzierter Betreuungsaufwand

Die Auszeichnung von Forschungsdaten mit Metadaten ist kein Selbstzweck, sondern zielt darauf ab, die Forschungsdaten im Sinne der FAIR-Prinzipien zu beschreiben, um so unter anderem Angaben für die Auffindbarkeit, die wissenschaftliche Nachnutzbarkeit und

2 XML Schema Definition, siehe `https://www.w3.org/TR/xmlschema11-1`.

3 Der verwendete XSLT-Processor basiert auf dem ursprünglichen Projekt XSD2HTML2XML von Meulendijk (2019).

4 Für weitere Informationen siehe `https://openid.net/connect`.

dauerhafte Referenzierbarkeit zu erfassen. Für die dauerhafte Referenzierbarkeit hat sich weitgehend die Vergabe von DOIs[5] durchgesetzt, woraus sich die Notwendigkeit der Erfassung von Metadaten nach dem DataCite-Schema ergibt.[6] Neben diesen deskriptiven Metadaten werden wissenschaftliche Metadaten benötigt, um Such- und Filterfunktion zu ermöglichen und Wissenschaftler:innen einen schnellen Überblick über die Forschungsdaten zu ermöglichen. Hierfür notwendige Schemata sind Anpassungen unterworfen, die entsprechend in einem Annotationswerkzeug nachgezogen werden müssen. Die Integration neuer Versionen oder gänzlich neuer Schemata erfordert bei dem hier vorgestelltem Werkzeug lediglich den Austausch oder die Bereitstellung einer neuen XSD, wodurch der Betreuungsaufwand stark reduziert wird. Im Rahmen des SDC BioDATEN wurde das DataCite-Schema 4.4 integriert und um wissenschaftliche Metadaten aus dem BioDATEN-Minimalschema ergänzt.[7] Diese Kombination erfüllt die projektinternen Anforderung hinsichtlich Publikation und Suchbarkeit von Forschungsdaten bei gleichzeitiger Anwendbarkeit auf mehrere Omics-Disziplinen. Um die Nutzer:innen beim Umgang mit dem Werkzeug zu unterstützen und die Qualität ihrer Eingaben zu erhöhen, wurde die Einbindung von Ontologien und Vokabularen über Dropdownmenüs mit Autocomplete-Funktion in das Werkzeug integriert. Hierfür wird die API von Bioportal[8] eingesetzt, um die Anforderungen an eine eigene Datenaufbereitung der hinterlegten Ontologien zu minimieren.[9] Über eine Adminstrationsoberfläche lässt sich konfigurieren, welche Metadatenfelder aus welchen Vokabularen befüllt werden sollen, siehe Abbildung 1. Hierdurch und durch die Integration der Bioportal API wird der Betreuungsaufwand minimiert und gleichzeitig die Datenqualität erhöht.

2 Mehrwert

Die Annotation von Metadaten ist ein wichtiger Baustein des Forschungsdatenlebenszyklus und bildet die Grundlage für die Auffindbarkeit und Nachnutzbarkeit der Forschungsdaten.[10] Entsprechend wurde bei der Entwicklung des Annotationswerkzeugs darauf Wert gelegt, dass die erfassten Metadaten eine Grundlage für anschließende Prozesse sind. Dies geschieht im Kontext von BioDATEN durch die Anbindung an eine Publikationsplattform auf Basis von InvenioRDM[11]. Die erfassten deskriptiven Metadaten erfüllen die Anforderungen des DataCite-Schemas 4.4[12] und bilden die Grundlage für die DOI-Registrierung. Für die Nutzer:innen bedeutet dies einen Verzicht auf die nochmalige Eingabe ihrer Daten. Während des Annotationsprozesses werden die Angaben validiert und wo immer möglich durch ein Vokabular bzw. eine Ontologie supplementiert. Somit kann durch den Verzicht

5 Digital Object Identifier, siehe `https://www.doi.org`.
6 Für die aktuelle Version des Schemas siehe `https://schema.datacite.org`.
7 Für weitere Informationen zum BioDATEN Minimalschema siehe `https://github.com/ubtue/BioDATEN-Minimalschema`.
8 Für weitere Informationen siehe `https://bioportal.bioontology.org`.
9 Einen alternativen Ansatz bietet der Dienst Semlookup der ZB MED, siehe hierzu Madan u. a. (2018).
10 Für weitere Informationen siehe `https://forschungsdaten.info/themen/informieren-und-planen/datenlebenszyklus`.
11 `https://inveniosoftware.org/products/rdm`
12 `https://datacite.org`

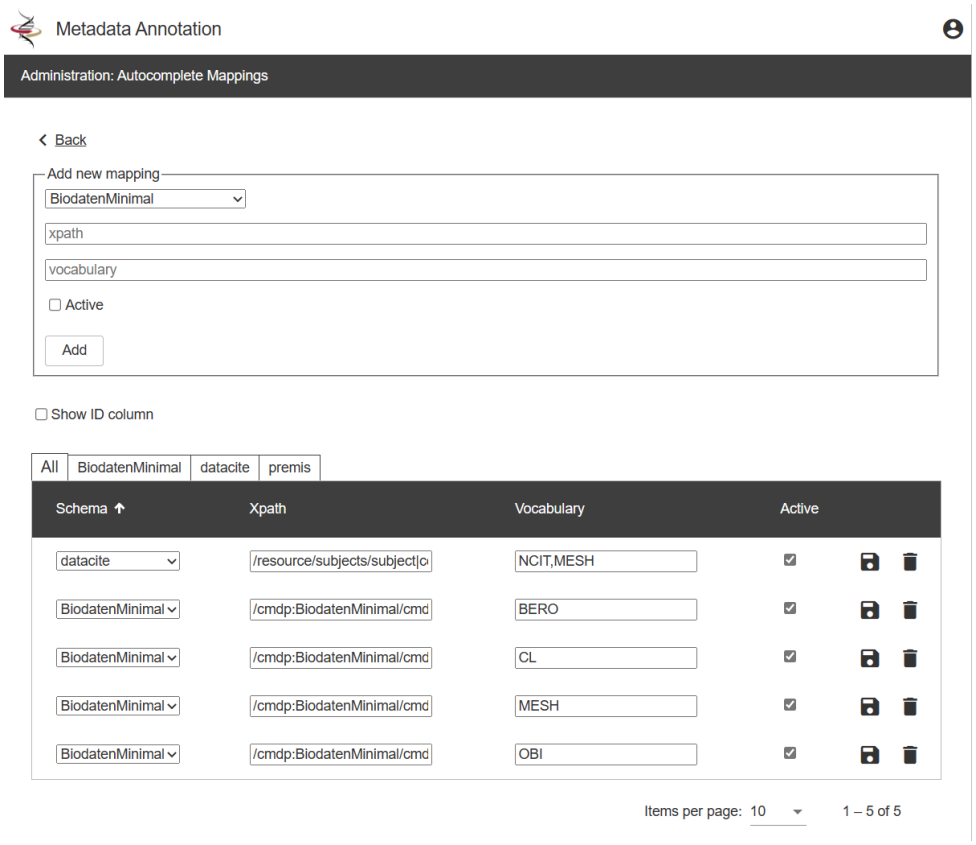

Abbildung 1: Screenshot der Administrationsoberfläche. Auswahl der Metadatenfelder mit zugeordnetem Vokabular.

auf Freitext die Qualität der Metadaten erhöht und die Erfassung von semantischen Relationen generiert werden.

3 Erhöhte Usability

Die Auszeichnung von Forschungsdaten bringt einen gewissen Mehraufwand für die Forscher:innen mit. Um diesen zu reduzieren, wurden Dropdownmenüs integriert und Vokabulare hinterlegt, siehe Abbildung 2. Eine weitere Maßnahme zur Verbesserung der Usability liegt in der fokussierten Integration einschlägiger Metadatenfelder unter Berücksichtigung des Anspruchs auf generische Einsetzbarkeit. Das Annotationswerkzeug unterstützt deshalb generell und speziell für das BioDATEN Minimalschema die Verwendung von konditional-obligatorischen Metadatenfeldern und entsprechende Abhängigkeiten, wie sie unter anderem durch das CMDI-Framework generiert und in XML abgebildet werden können (siehe Brandt u. a. 2021). Hierdurch werden Metadatenfelder nur dann angezeigt, sofern diese aufgrund definierter Abhängigkeiten von getätigten Eingaben relevant sind. Zusätzlich stand bei der Entwicklung des Annotationswerkzeugs die Berücksichtigung und Umsetzung digitaler Barrierefreiheit im Fokus.

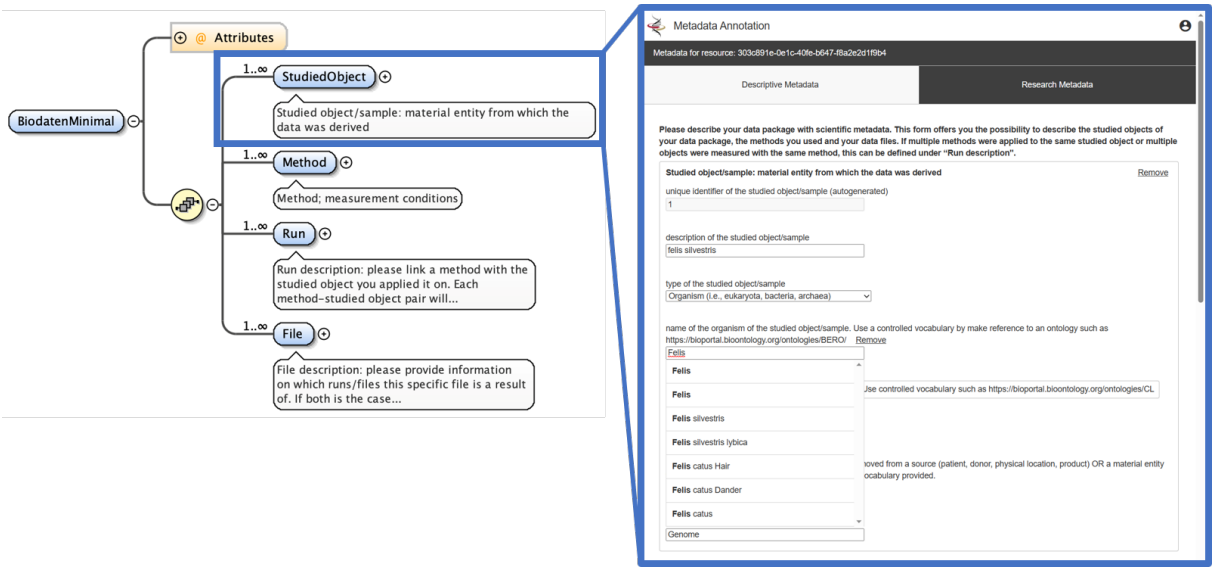

Abbildung 2: Screenshot der Annotationsoberfläche. Auswahl des untersuchten Organismus mit Dropdownmenü in der Beschreibung des „Studied Object“.

4 Generischer Anspruch

Das Annotationswerkzeug wurde im Rahmen des SDC BioDATEN mit Fokus auf Bioinformatik und mehrerer der sogenannten Omics-Disziplinen entwickelt.[13] Dennoch wurde bei der Entwicklung großer Wert darauf gelegt, das Annotationswerkzeug durch eine einfache Austauschbarkeit von Metadatenschemata breit und generisch anbieten zu können. Der Verzicht auf Hardcoding erlaubt in Kombination mit der Trennung von Funktion und Inhalt eine einfache Anpassung an andere Communities. Entsprechend kann das Annotationswerkzeug beispielsweise auch im Rahmen von bwHPC eingesetzt werden, wo ebenfalls der Bedarf nach einer generischen und anpassungsfähigen Lösung besteht. Die Grundlage für den Aufbau der webbasierten Annotationsoberfläche bilden vorhandene XSDs. XSDs haben den Vorteil, Metadaten-Standards abbilden und gleichzeitig die Validierung gegen diese Standards ermöglichen zu können.

5 Open Source Software

Das Tool wurde als Open-Source-Software entwickelt. Es steht unter der AGPL-3-Lizenz[14] zur Verfügung. Für das Frontend wurde das auf TypeScript basierte Webapplikationsframework Angular[15] verwendet. Der Backend-Service des Annotationstools wurde mit

13 Für weitere Informationen siehe `https://portal.biodaten.info`.
14 `https://www.gnu.org/licenses/agpl-3.0.de.html`
15 `https://angular.io`

dem Spring Boot Framework[16] erstellt. Die Einzelkomponenten liegen öffentlich zugänglich in GitHub-Repositorien: Backend[17], Frontend[18] und XSLT-Prozessor[19].

6 Fazit

Die Annotation von Forschungsdaten mit Metadaten erzeugt Aufwand bei den Forscher:innen und den Dienstanbieter:innen gleichermaßen. Erstere müssen Daten liefern und letztere müssen entsprechende Dienste bereitstellen. Das Ziel bei der Entwicklung des Tools liegt in der Entlastung beider Seiten gleichermaßen und der Schaffung von Mehrwert durch eine Annotation. Die Kombination aus XSD Input, menügeführter Administration, Integration von kontrollierten Vokabularen, Dropdownmenüs und der automatischen Validierung von Input resultiert in einem Werkzeug, das die eingangs genannten Anforderungen erfüllt. Der generische Ansatz bei der Erstellung des Werkzeugs und die konsequente Umsetzung von Open-Source sowie eine Nutzbarkeit ohne tiefes Expertenwissen verspricht generische Einsatzmöglichkeiten.

Literaturverzeichnis

Brandt, Olaf, Holger Gauza, Steve Kaminski, Mario Trojan, Thorsten Trippel und Johannes Werner. 2021. „Extending the CMDI Universe“. In *Linköping Electronic Conference Proceedings,* herausgegeben von Costanza Navarretta und Maria Eskevich, Bd. 180. DOI: https://doi.org/10.3384/ecp1806.

Madan, Sumit, Maksims Fiosins, Stefan Bonn und Juliane Fluck. 2018. „A Semantic Data Integration Methodology for Translational Neurodegenerative Disease Research“. https://doi.org/10.6084/m9.figshare.7339244.v1. See also https://semanticlookup.zbmed.de/, *Semantic Web Applications and Tools for Healthcare and Life Sciences.*

Meulendijk, Michiel. 2019. „XSD2HTML2XML“. Besucht am 6. September 2023. https://github.com/MichielCM/xsd2html2xml.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersbergand, Gabrielle Appleton, Myles Axtonand, Arie Baakand, Niklas Blombergand u. a. 2016. „The FAIR Guiding Principles for scientific data management and stewardship“. *Scientific data* 3 (1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.

16 https://spring.io
17 https://github.com/ubtue/BioDATEN-Metadata-Annotation-Backend
18 https://github.com/ubtue/BioDATEN-Metadata-Annotation-Tool
19 https://github.com/ubtue/BioDATEN-Metadata-XSLT-Processor