BioDATEN

Metadaten Harvesting und Annotation

Jan Kaltenbach

BioDATEN Workshop 12.06.2023

Übersicht

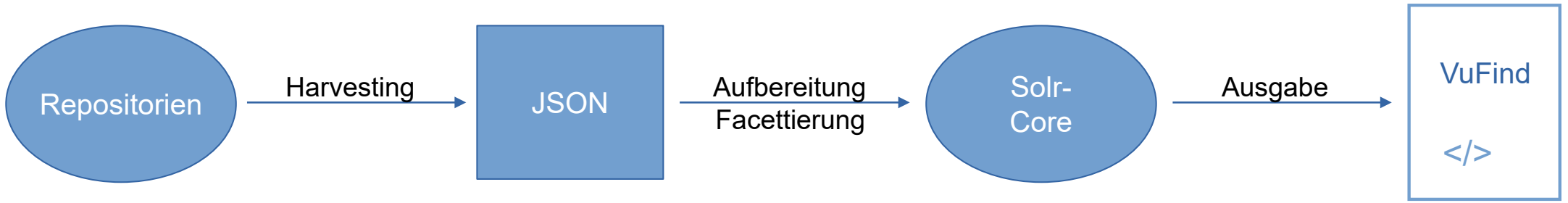
1. Metadaten Harvesting und Suche
2. Metadaten Annotation
3. Verwendung von Vokabularen
4. Administration

Metadaten Harvesting und Suche

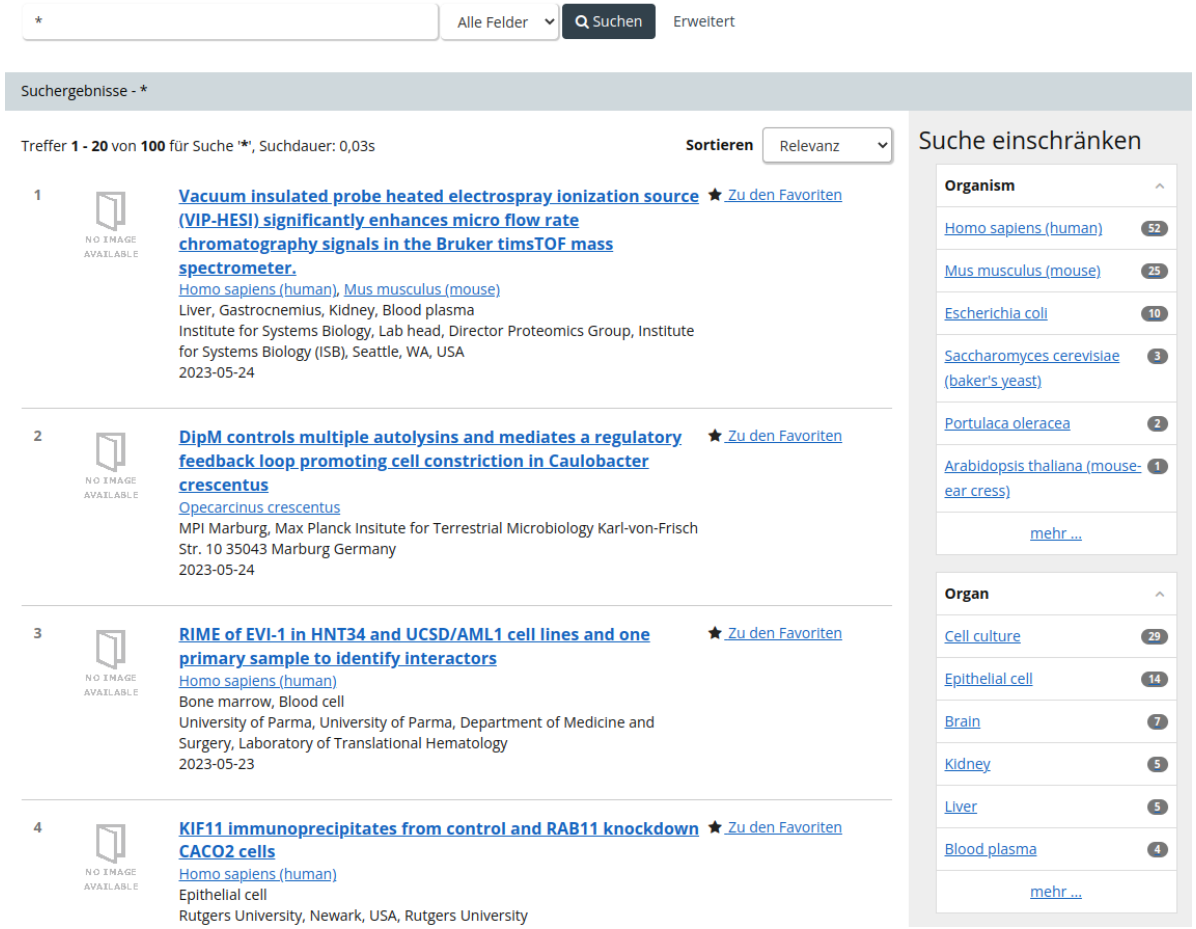
- Verschiedene Repositorien werden regelmäßig für das Harvesting der Daten durchsucht. Diese Daten werden zwischengespeichert.
- Die gespeicherten Daten werden in einen Solr-Core eingespielt und für die korrekte Facettierung umgewandelt.
- Die aufbereiteten Daten sind dann über eine VuFind-Suchmaschine (https://discover.biodaten.info/) durchsuchbar.

Metadaten Harvesting und Suche

- Harvesting von Daten aus Repositorien und Speichern als JSON-Datei.
- Import der JSON-Datei in Solr-Core inklusive Facettierung.
- Einbinden des Solr-Cores in VuFind-Suchmaschine.



Metadaten Harvesting und Suche

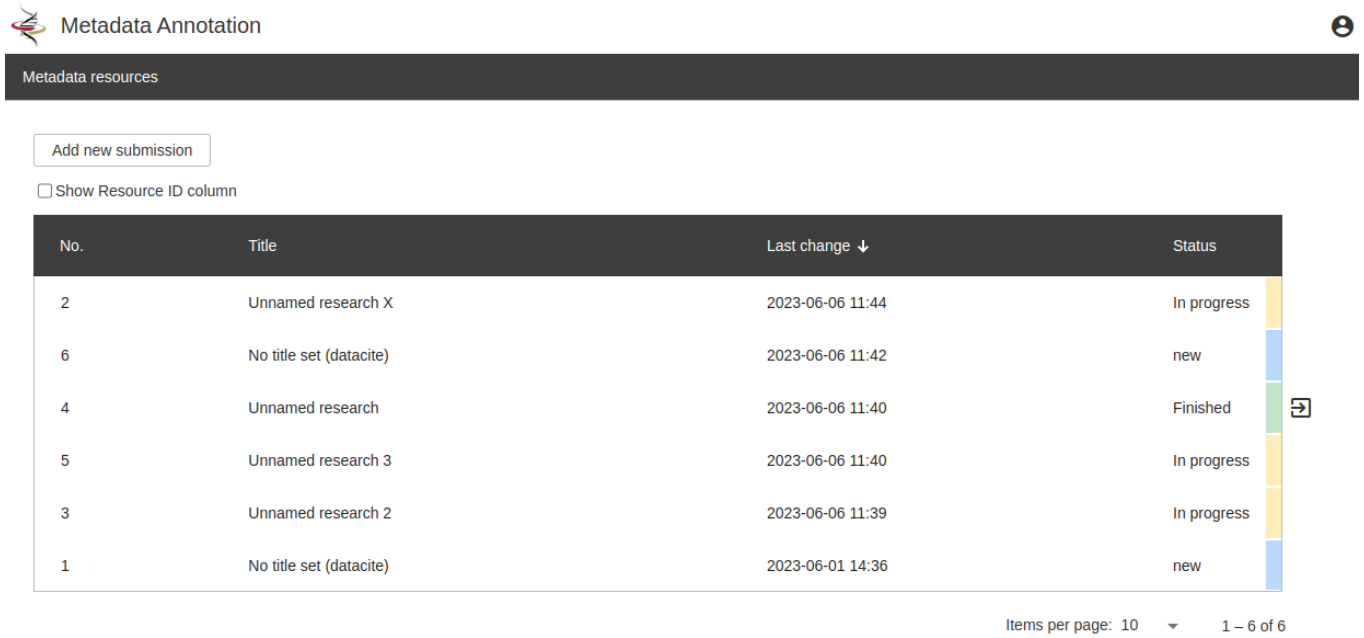


- VuFind-Suchmaschine zeigt Datensätze des Harvestings.
- Suche kann über die Facettierung eingeschränkt und verfeinert werden.
- Ergebnisse werden regelmäßig nach dem automatischen Harvesting aktualisiert.

Metadaten Annotation

- Metadaten Annotation findet über eigens entwickeltes webbasiertes Annotationstool statt (https://annotate.biodaten.info/).
- Anmeldevorgang wird via Life Science Login gesteuert.
- Übersicht der eigenen Datensätze.
- Über eine Profilseite können z.B. Authentifizierungsschlüssel, die für die Veröffentlichung notwendig sind, hinterlegt werden.

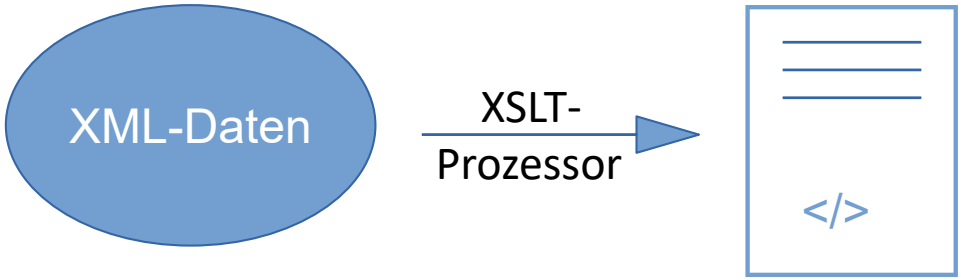
Metadaten Annotation



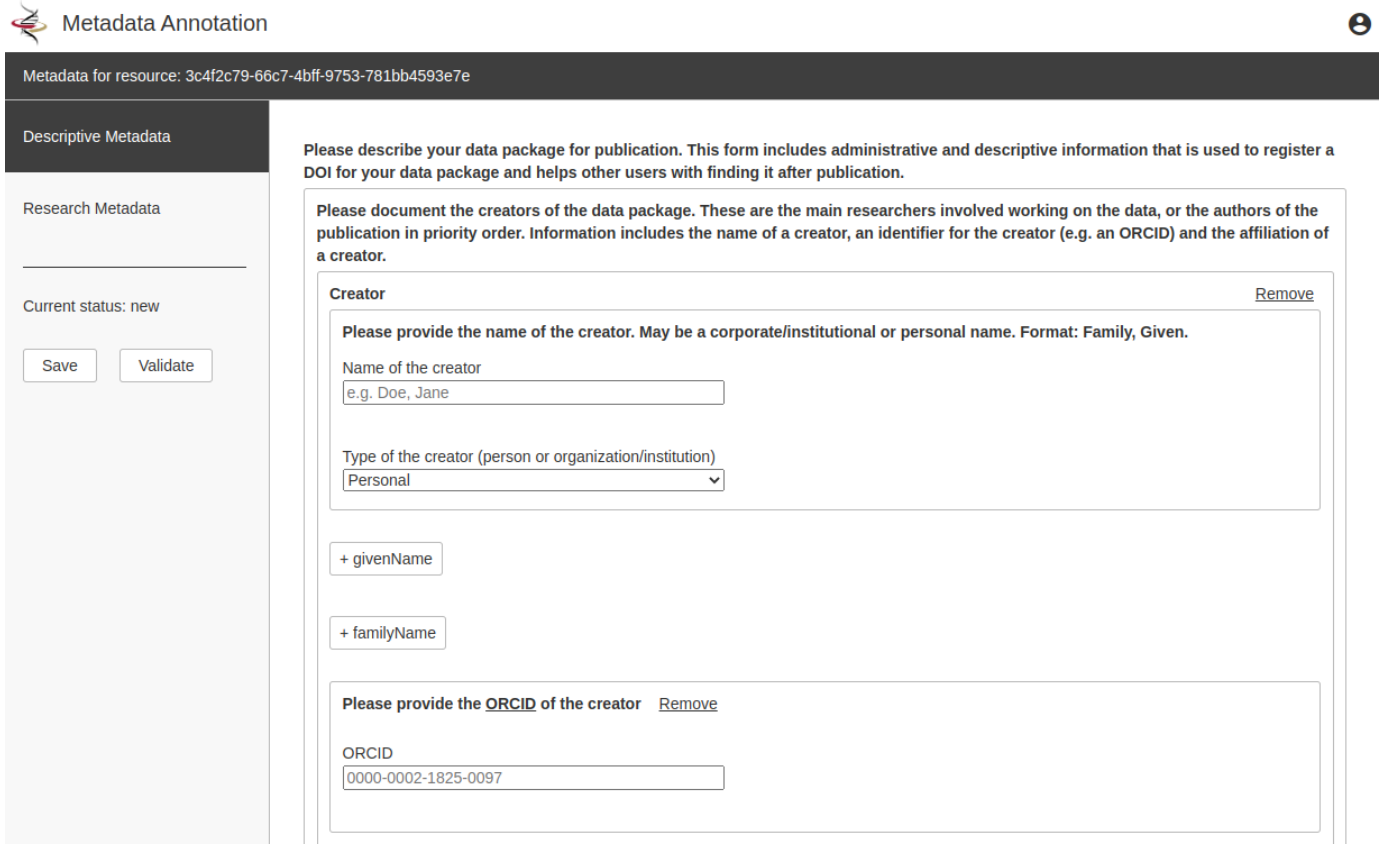
- Sortierbare Übersicht eigener Datensätze nach erfolgreichem Login.
- Anzeige des Titels (innerhalb des Datacite Schemas), der letzten Änderung und des Status.
- Status zeigt an, ob ein Datensatz neu, in Bearbeitung, fertig bearbeitet oder exportiert ist.
- Fertige Datensätze können zum Forschungsdatenmanagement (FDAT/Invenio) exportiert werden.

Metadaten Annotation

- XML-Daten werden gelesen und verwendete Schemas (Bsp.: datacite, premis, eigenes BioDATEN Minimalschema) ermittelt.
- Ein Webformular wird anhand der zuvor ermittelten Schema Dateien durch einen XSLT-Prozessor erstellt.
- Formularfelder werden mit den XML-Daten befüllt.

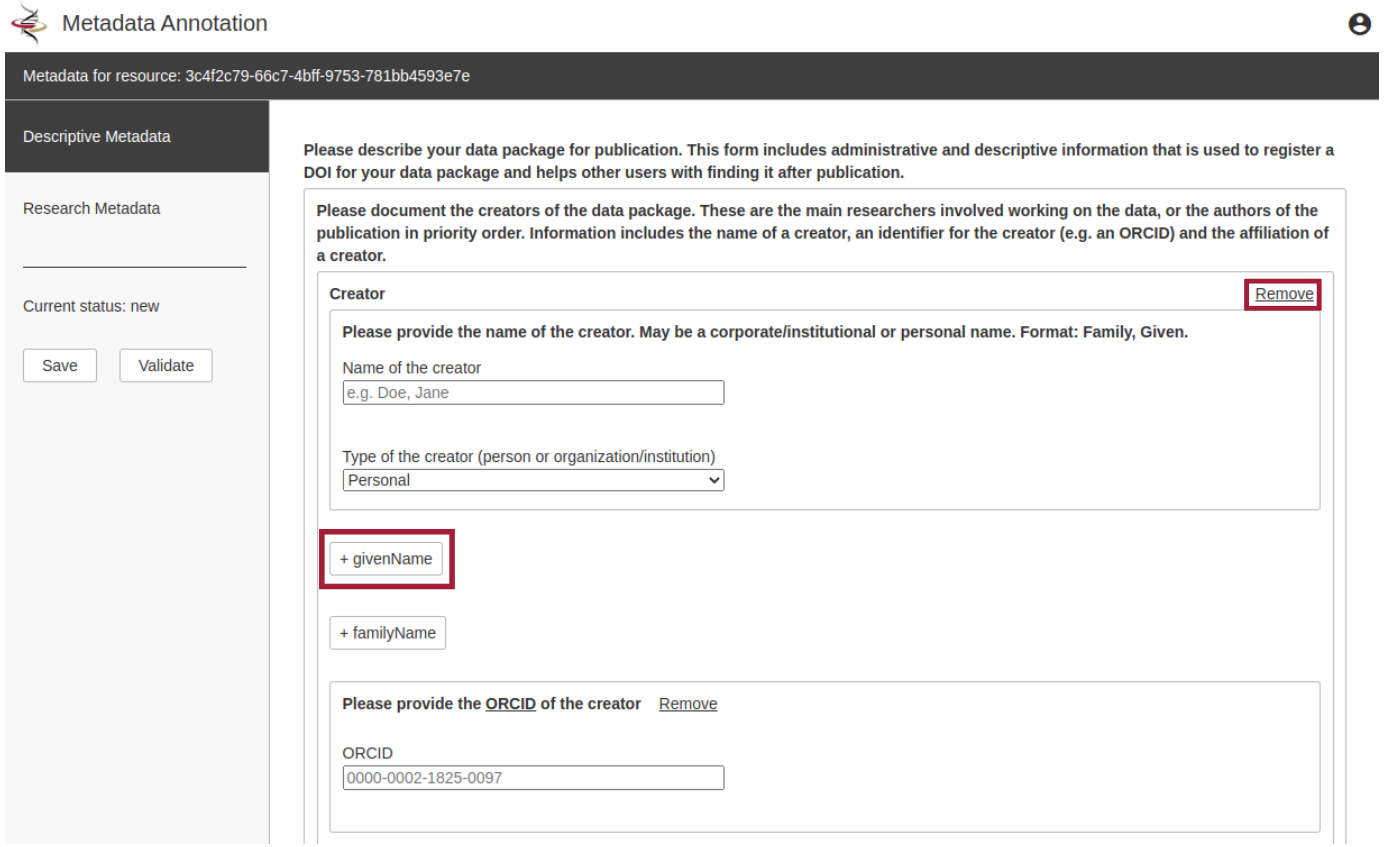


Metadaten Annotation



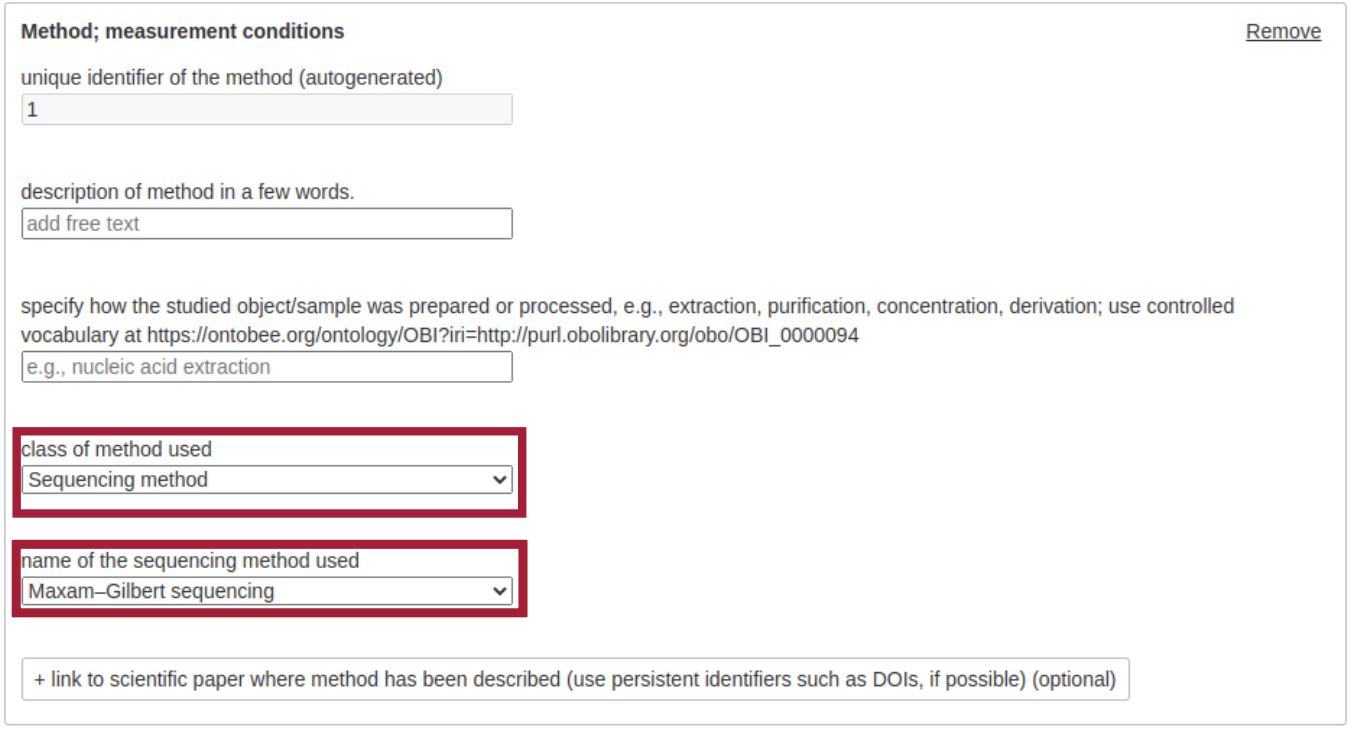
- Erstelltes Webformular aus den Schemas Datacite und BioDATEN Minimal.
- Felder der XSD-Schemadatei wurden in Eingabe- bzw. Drop-down-Felder umgewandelt.
- Verschiedene Labels, die entweder aus dem Schema oder der Administration kommen, helfen bei der Vervollständigung der Daten.

Metadaten Annotation



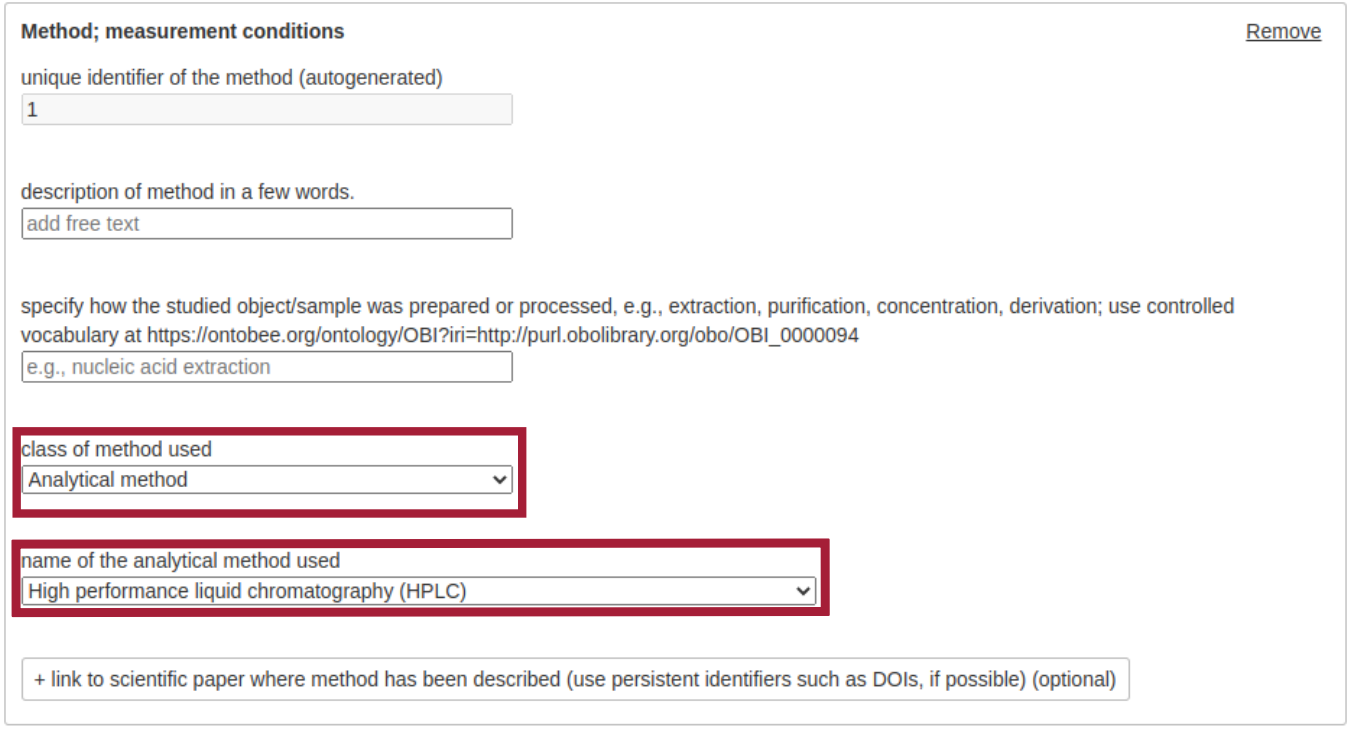
- Verschiedene Schaltflächen ermöglichen das Hinzufügen bzw. das Entfernen von optionalen oder mehrfach erstellten Inhaltsblöcken.
- Über die Schaltfläche “givenName” kann ein neues Eingabefeld eingeblendet werden.
- Über die Schaltfläche “Remove” kann der komplette Block entfernt bzw. geleert werden.

Metadaten Annotation



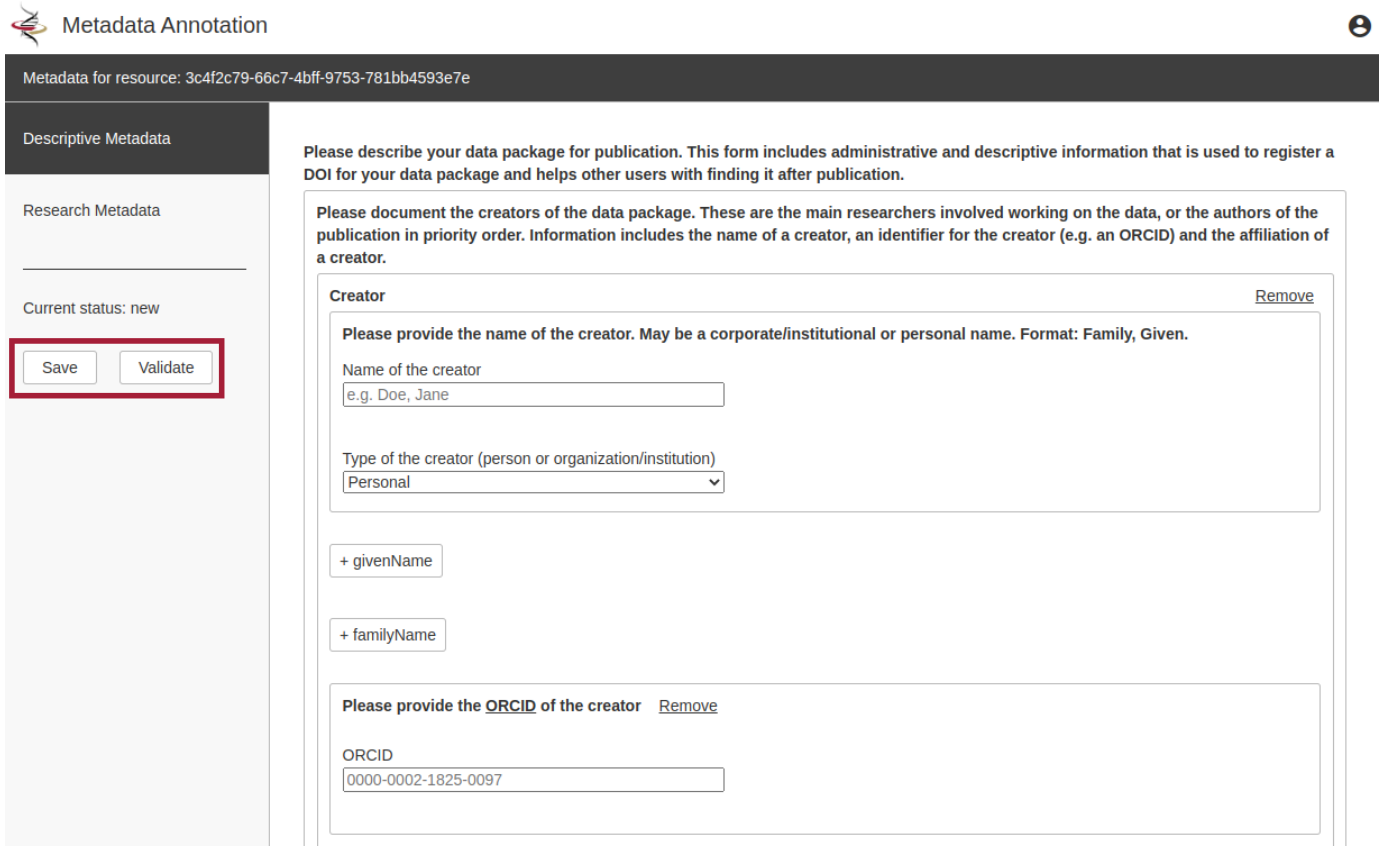
- Über spezifische Parameter in der XSD-Schemadatei besteht die Möglichkeit Elemente nur dann anzuzeigen, wenn eine andere Bedingung erfüllt wurde.
- Das Feld „name of sequencing method used“ wird nur angezeigt, wenn im Feld „class of method used“ der Wert „Sequencing method“ ausgewählt wurde.

Metadaten Annotation



- Wird der Wert „Analytical Method“ im Feld „class of method used“ ausgewählt, wird das zuvor sichtbare Feld „name of sequencing method used“ ausgeblendet und stattdessen das Feld für „name of the analytical method used“ eingeblendet.
- Dieses Verhalten kann für jedes beliebige Element über die Schemadatei gesteuert werden.

Metadaten Annotation

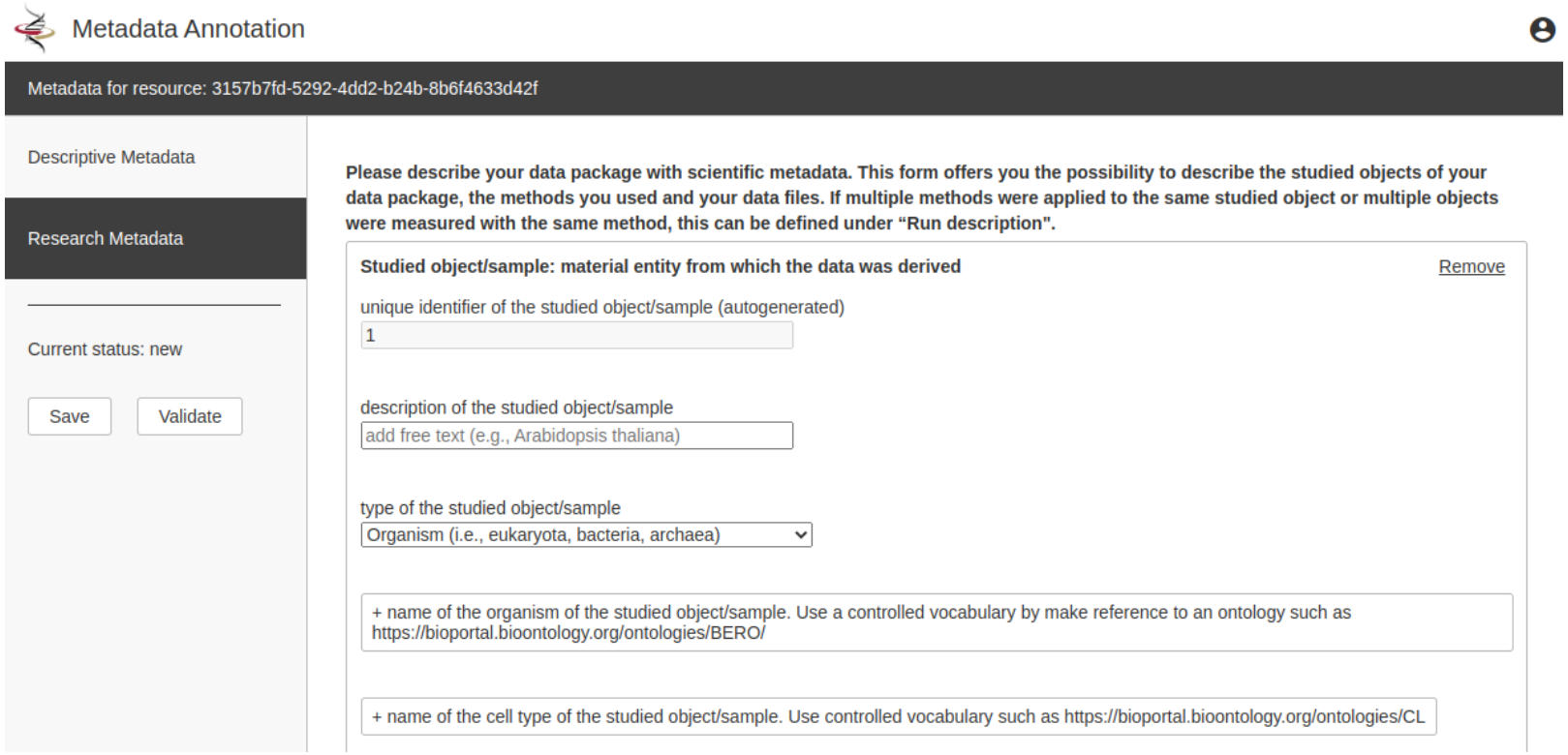


- Über die Schaltfläche “Save” kann der Datensatz gespeichert werden. Dieser kann dann zu einem späteren Zeitpunkt weiter bearbeitet werden.
- Über die Schaltfläche “Validate” kann das Formular auf Fehleingaben geprüft werden. Gefundene Fehler werden automatisch hervorgehoben.
  Werden keine Fehler gefunden, wird das Formular als “fertig” markiert.

Verwendung von Vokabularen

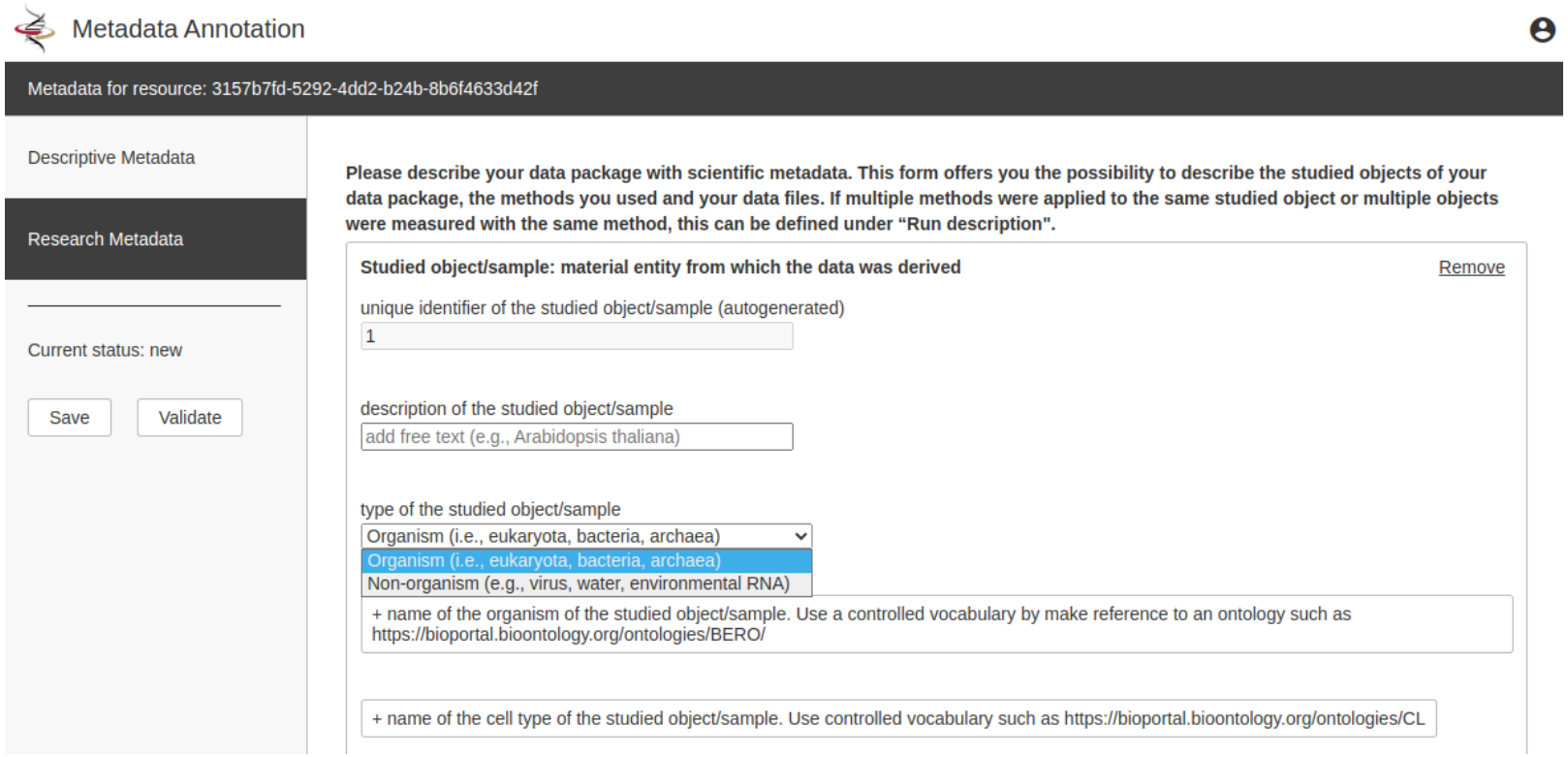
- Vokabulare werden verwendet, um eine hohe Datenqualität zu sichern und um die Eingabe der Daten zu vereinfachen.
- Die verschiedenen Vokabulare können den Eingabefeldern über die Administrationsoberfläche zugeordnet werden.
- Als Datenquelle kann entweder eine selbst exportierte JSON-Datei oder eine Referenz auf eine Bioportal Ontologie verwendet werden.
- Es werden jeweils Bezeichnung und eindeutige Identifier URL des ausgewählten Begriffs gespeichert. Diese URL beinhaltet alle Informationen innerhalb der Ontologie, zu der das Vokabular gehört.

Verwendung von Vokabularen



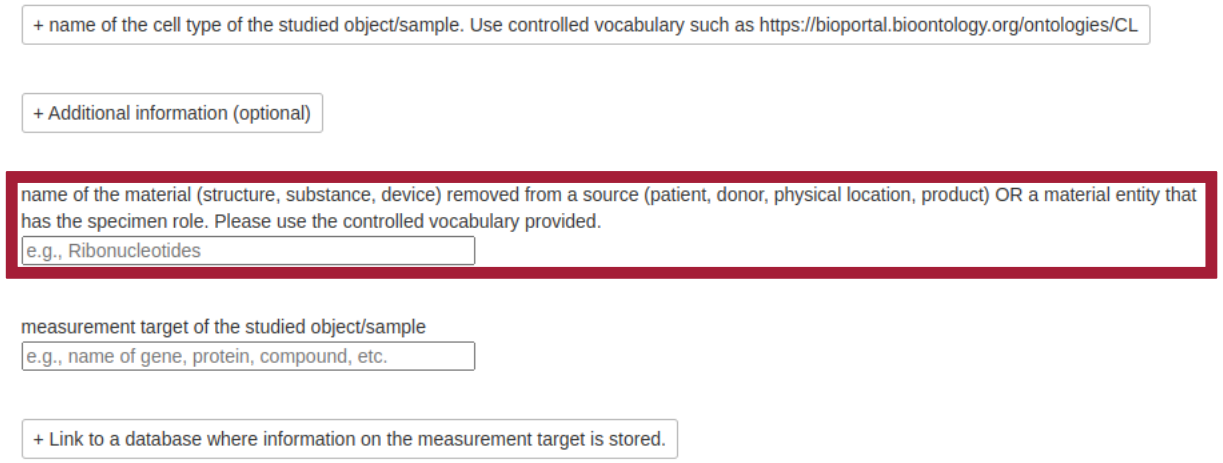
- Erstelltes Webformular aus den Schemas Datacite und BioDATEN Minimal.
- Felder mit begrenzten Optionen werden als Drop-Down dargestellt.
  Bsp.: type of the studied object/sample

Verwendung von Vokabularen



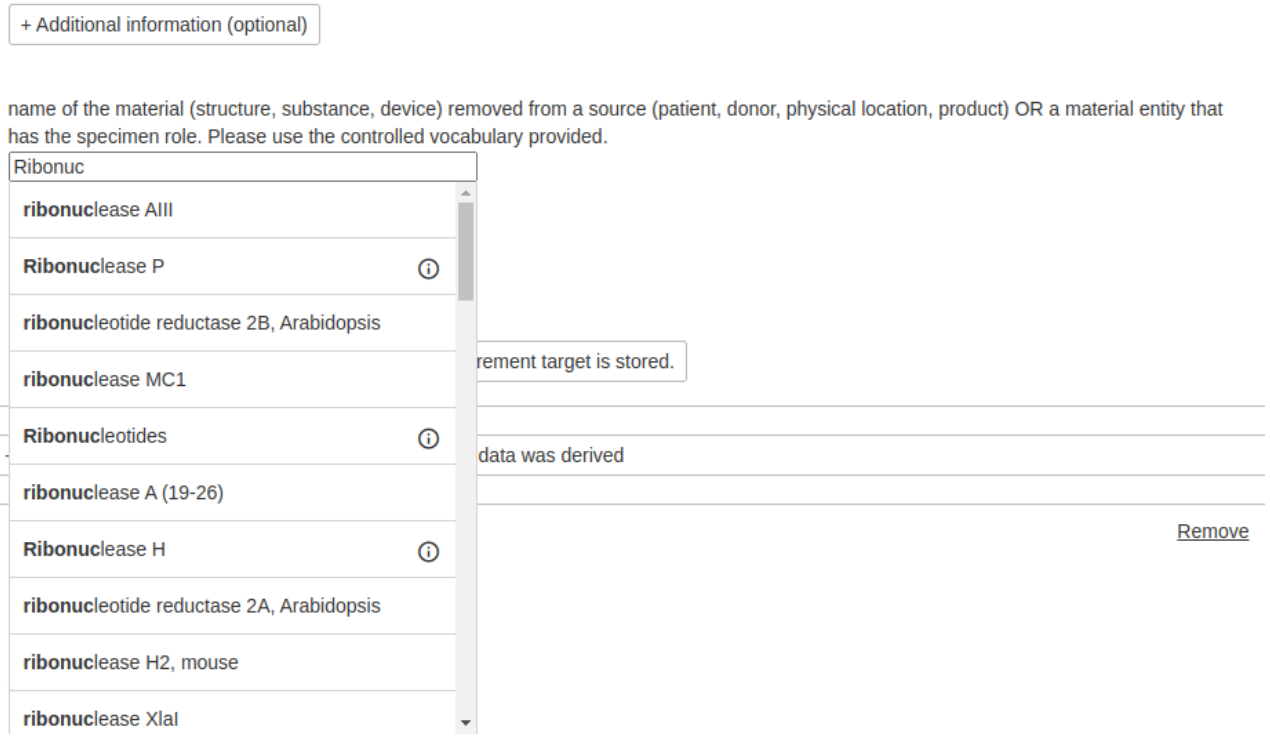
- Wählbare Optionen werden im Schema definiert.
- Drop-Down Anwendung durch die Anzahl der Optionen begrenzt. Wird bei zu viel Inhalt schnell unübersichtlich bzw. schlecht bedienbar.

Verwendung von Vokabularen



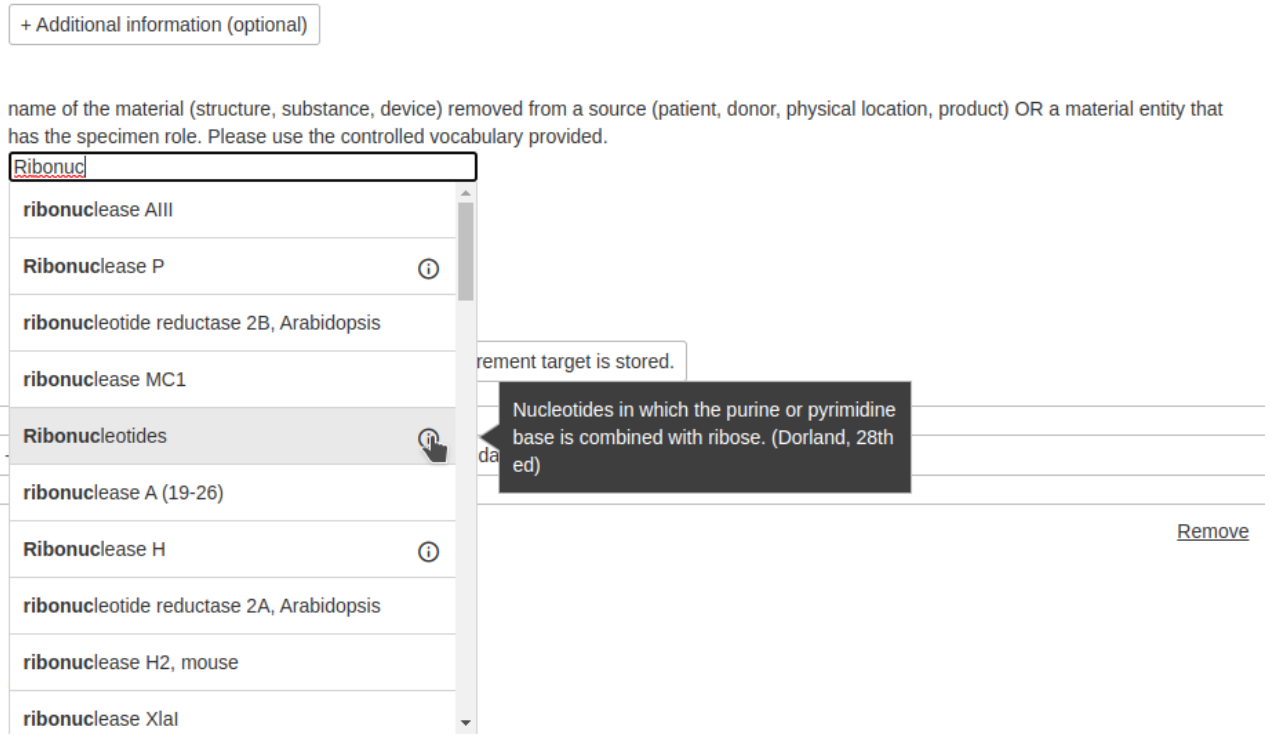
- Felder, die für ein Drop-Down zu viele Eingabemöglichkeiten haben, werden als normale Eingabefelder generiert.
- Problem: Durch eine freie Eingabe z.B. beim markierten Feld, führt dies zu verschiedenen Ausführungen der gleichen Werte (Statt einem – ein _ verwendet). Dies kann problematisch für die spätere Suche werden.

Verwendung von Vokabularen



- Vokabulare werden den Eingabefeldern zugeordnet.
- Zu jedem Begriff ist innerhalb des Vokabulars eine eindeutige Identifier URL hinterlegt, über die alle Informationen abgerufen werden können.
  Hierdurch wird sichergestellt, dass der Begriff eindeutig zuordenbar ist.
  Bsp.: Ribonucleotides: http://purl.bioontology.org/ontology/MESH/D012265

Verwendung von Vokabularen

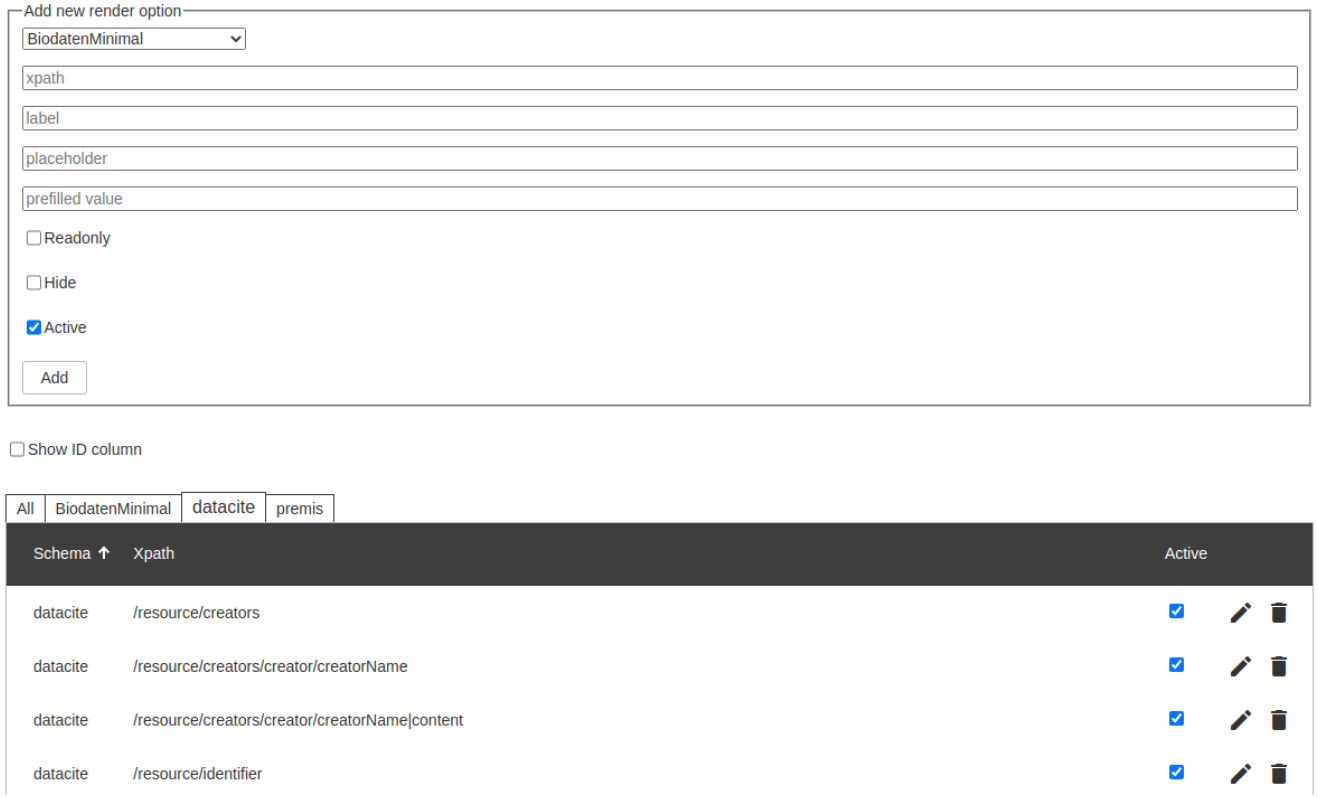


- Ist für den Begriff im Vokabular ein Beschreibungstext hinterlegt, kann dieser über ein Info Icon angezeigt werden
- Beim Speichervorgang werden die Eingaben zu den in dem Vokabular hinterlegten Identifier URLs konvertiert und in diesem Format innerhalb des XML-Datensatzes gespeichert. Die Anzeige im Formular ist davon nicht betroffen.

Administration

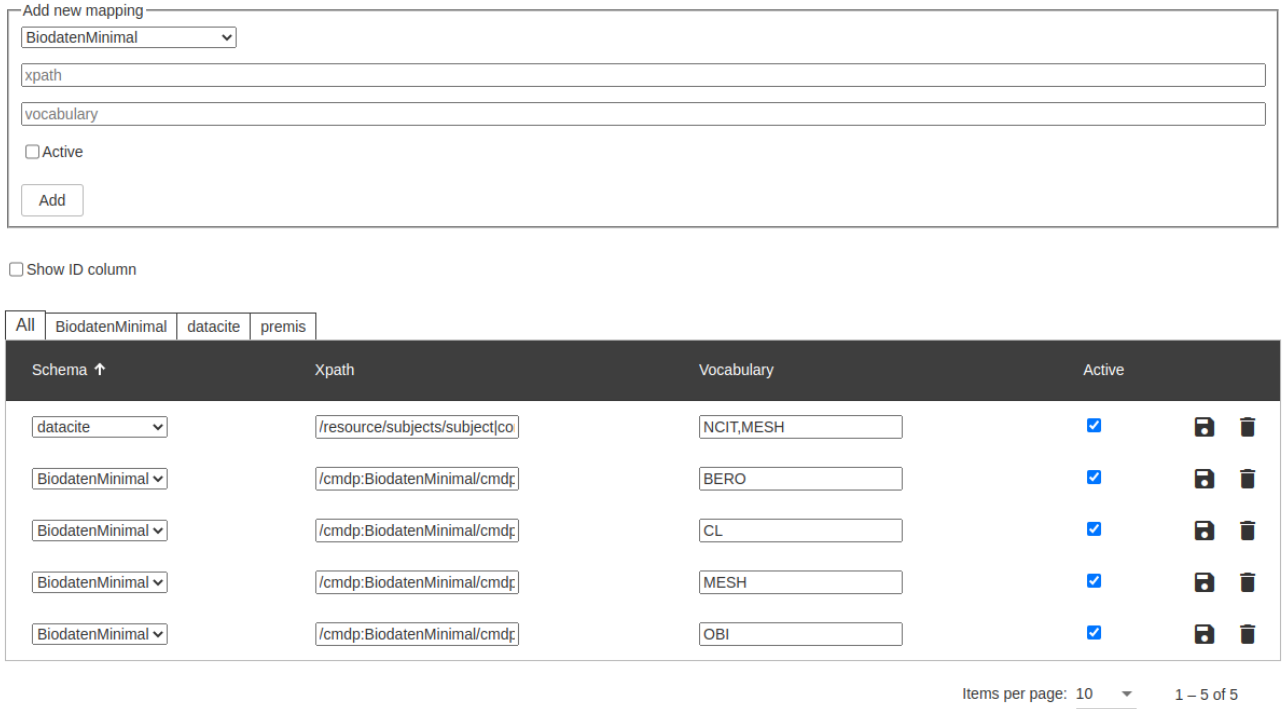
- Um die Anzeige des Webformulars einfach und schnell anpassen zu können, wurde eine Administrationsoberfläche in das Tool integriert.
- Die Administrationsoberfläche ermöglicht folgende Anpassungen:
  - Die Bearbeitung von Überschriften bzw. Anzeigelabels, das Hinzufügen von Platzhaltern, das Vorausfüllen von Eingabefeldern, sowie die volle Kontrolle über die Sichtbarkeit von einzelnen Anzeigeelementen. Hierfür sind keine Schemaanpassungen notwendig.
  - Die Zuweisung der Vokabulare.
  - Steuerung der verwendeten und anzuzeigenden Schema Dateien.

Administration



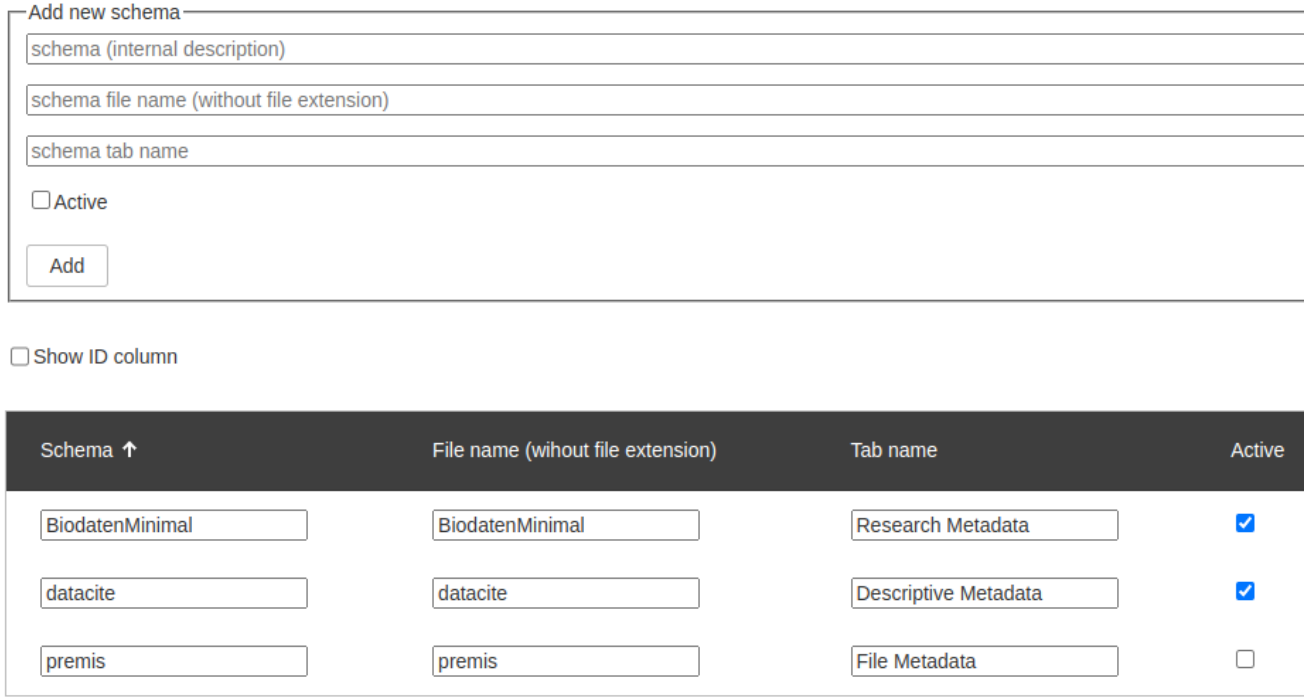
- Über die Render-Options kann die Anzeige der einzelnen Eingabefelder bzw. ganzer Inhaltsblöcke bearbeitet werden.
- Es können verschiedene vorausgefüllte Werte, Platzhalter oder Anzeigelabels angepasst werden.
- Die Sichtbarkeit der einzelnen Elemente kann komplett angepasst werden.

Administration



- In der Vokabularzuweisung können die einzelnen Eingabefeldern mit den gewünschten Vokabularen verknüpft werden.
- Es können auch mehrere Vokabulare zu einzelnen Eingabefeldern zugeordnet werden.

Administration



- Übersicht über aller hinzugefügten Schemas, inklusive Anzeigename des entsprechenden Reiters im Formular.
- Es können beliebig neue Schemas hinzugefügt werden. Diese können auch je nach Bedarf ein- und ausgeschaltet werden.
  Für jedes Schema muss eine XSD-Datei vorhanden sein.

Vielen Dank für Ihre Aufmerksamkeit

- Der Quellcode für das Annotationstool und die zugehörigen Services ist auf GitHub unter den folgenden Links verfügbar:
  - Annotationstool Frontend: https://github.com/ubtue/BioDATEN-Metadata-Annotation-Tool
  - Annotationstool Backend: https://github.com/ubtue/BioDATEN-Metadata-Annotation-Backend
  - XSLT-Prozessor: https://github.com/ubtue/BioDATEN-Metadata-XSLT-Processor