

BioDATEN

Metadaten Harvesting und Annotation

Jan Kaltenbach

BioDATEN Workshop 12.06.2023

Übersicht

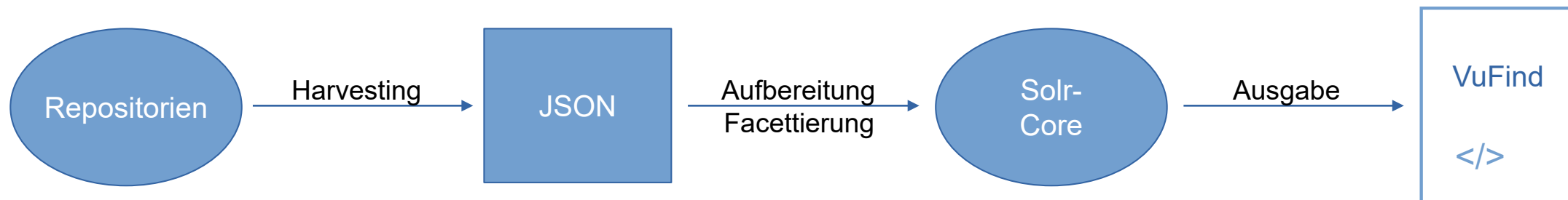
1. Metadaten Harvesting und Suche
2. Metadaten Annotation
3. Verwendung von Vokabularen
4. Administration

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


* Alle Felder Erweitert

Suchergebnisse - *


Treffer 1 - 20 von 100 für Suche '*', Suchdauer: 0,03s Sortieren Relevanz

1  [Vacuum insulated probe heated electrospray ionization source \(VIP-HESI\) significantly enhances micro flow rate chromatography signals in the Bruker timsTOF mass spectrometer.](#) ★ [Zu den Favoriten](#)


[Homo sapiens \(human\)](#), [Mus musculus \(mouse\)](#)
 Liver, Gastrocnemius, Kidney, Blood plasma
 Institute for Systems Biology, Lab head, Director Proteomics Group, Institute for Systems Biology (ISB), Seattle, WA, USA
 2023-05-24

2  [DipM controls multiple autolysins and mediates a regulatory feedback loop promoting cell constriction in Caulobacter crescentus](#) ★ [Zu den Favoriten](#)

[Opeparcinus crescentus](#)
 MPI Marburg, Max Planck Institute for Terrestrial Microbiology Karl-von-Frisch Str. 10 35043 Marburg Germany
 2023-05-24

3  [RIME of EVI-1 in HNT34 and UCSD/AML1 cell lines and one primary sample to identify interactors](#) ★ [Zu den Favoriten](#)

[Homo sapiens \(human\)](#)
 Bone marrow, Blood cell
 University of Parma, University of Parma, Department of Medicine and Surgery, Laboratory of Translational Hematology
 2023-05-23

4  [KIF11 immunoprecipitates from control and RAB11 knockdown CACO2 cells](#) ★ [Zu den Favoriten](#)

[Homo sapiens \(human\)](#)
 Epithelial cell
 Rutgers University, Newark, USA, Rutgers University

Suche einschränken

Organism ^

[Homo sapiens \(human\)](#) 52

[Mus musculus \(mouse\)](#) 25

[Escherichia coli](#) 10

[Saccharomyces cerevisiae \(baker's yeast\)](#) 3

[Portulaca oleracea](#) 2

[Arabidopsis thaliana \(mouse-ear cress\)](#) 1

[mehr ...](#)

Organ ^

[Cell culture](#) 29

[Epithelial cell](#) 14

[Brain](#) 7

[Kidney](#) 5

[Liver](#) 5

[Blood plasma](#) 4

[mehr ...](#)

- 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

Metadata Annotation ⊕

Metadata resources

Show Resource ID column

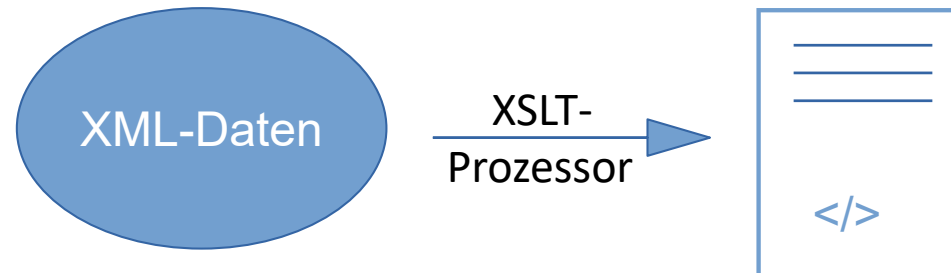
No.	Title	Last change ↓	Status
2	Unnamed research X	2023-06-06 11:44	In progress
6	No title set (datacite)	2023-06-06 11:42	new
4	Unnamed research	2023-06-06 11:40	Finished
5	Unnamed research 3	2023-06-06 11:40	In progress
3	Unnamed research 2	2023-06-06 11:39	In progress
1	No title set (datacite)	2023-06-01 14:36	new

Items per page: 10 ▾ 1 – 6 of 6


- 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

 Metadata Annotation ⊞

Metadata for resource: 3c4f2c79-66c7-4bff-9753-781bb4593e7e

Descriptive Metadata

Please describe your data package for publication. This form includes administrative and descriptive information that is used to register a DOI for your data package and helps other users with finding it after publication.

Research Metadata

Current status: new

Please document the creators of the data package. These are the main researchers involved working on the data, or the authors of the publication in priority order. Information includes the name of a creator, an identifier for the creator (e.g. an ORCID) and the affiliation of a creator.

Creator [Remove](#)

Please provide the name of the creator. May be a corporate/institutional or personal name. Format: Family, Given.

Name of the creator

Type of the creator (person or organization/institution)

Please provide the **ORCID** of the creator [Remove](#)

ORCID

- 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

Metadata Annotation

Metadata for resource: 3c4f2c79-66c7-4bff-9753-781bb4593e7e

Descriptive Metadata

Research Metadata

Current status: new

Save Validate

Please describe your data package for publication. This form includes administrative and descriptive information that is used to register a DOI for your data package and helps other users with finding it after publication.

Please document the creators of the data package. These are the main researchers involved working on the data, or the authors of the publication in priority order. Information includes the name of a creator, an identifier for the creator (e.g. an ORCID) and the affiliation of a creator.

Creator [Remove](#)

Please provide the name of the creator. May be a corporate/institutional or personal name. Format: Family, Given.

Name of the creator
e.g. Doe, Jane

Type of the creator (person or organization/institution)
Personal

+ givenName

+ familyName

Please provide the **ORCID** of the creator [Remove](#)

ORCID
0000-0002-1825-0097

- 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

Method; measurement conditions [Remove](#)

unique identifier of the method (autogenerated)

description of method in a few words.

specify how the studied object/sample was prepared or processed, e.g., extraction, purification, concentration, derivation; use controlled vocabulary at https://ontobee.org/ontology/OBI?iri=http://purl.obolibrary.org/obo/OBI_0000094

class of method used

name of the sequencing method used

- 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

Method; measurement conditions [Remove](#)

unique identifier of the method (autogenerated)

description of method in a few words.

specify how the studied object/sample was prepared or processed, e.g., extraction, purification, concentration, derivation; use controlled vocabulary at https://ontobee.org/ontology/OBI?iri=http://purl.obolibrary.org/obo/OBI_0000094

class of method used

name of the analytical method used

- 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

Metadata Annotation ⊙

Metadata for resource: 3c4f2c79-66c7-4bff-9753-781bb4593e7e

Descriptive Metadata

Research Metadata

Current status: new

Save **Validate**

Please describe your data package for publication. This form includes administrative and descriptive information that is used to register a DOI for your data package and helps other users with finding it after publication.

Please document the creators of the data package. These are the main researchers involved working on the data, or the authors of the publication in priority order. Information includes the name of a creator, an identifier for the creator (e.g. an ORCID) and the affiliation of a creator.

Creator [Remove](#)

Please provide the name of the creator. May be a corporate/institutional or personal name. Format: Family, Given.

Name of the creator

Type of the creator (person or organization/institution)

+ givenName

+ familyName

Please provide the **ORCID** of the creator [Remove](#)


ORCID

- 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

 Metadata Annotation e

Metadata for resource: 3157b7fd-5292-4dd2-b24b-8b6f4633d42f

Descriptive Metadata

Research Metadata

Current status: new

Please describe your data package with scientific metadata. This form offers you the possibility to describe the studied objects of your data package, the methods you used and your data files. If multiple methods were applied to the same studied object or multiple objects were measured with the same method, this can be defined under "Run description".

Studied object/sample: material entity from which the data was derived [Remove](#)

unique identifier of the studied object/sample (autogenerated)

description of the studied object/sample


type of the studied object/sample

+ name of the organism of the studied object/sample. Use a controlled vocabulary by make reference to an ontology such as <https://bioportal.bioontology.org/ontologies/BERO/>

+ name of the cell type of the studied object/sample. Use controlled vocabulary such as <https://bioportal.bioontology.org/ontologies/CL>

- 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

 Metadata Annotation e

Metadata for resource: 3157b7fd-5292-4dd2-b24b-8b6f4633d42f

Descriptive Metadata

Research Metadata

Current status: new

Please describe your data package with scientific metadata. This form offers you the possibility to describe the studied objects of your data package, the methods you used and your data files. If multiple methods were applied to the same studied object or multiple objects were measured with the same method, this can be defined under "Run description".

Studied object/sample: material entity from which the data was derived [Remove](#)

unique identifier of the studied object/sample (autogenerated)

description of the studied object/sample

type of the studied object/sample

+ name of the organism of the studied object/sample. Use a controlled vocabulary by make reference to an ontology such as <https://bioportal.bioontology.org/ontologies/BERO/>

+ name of the cell type of the studied object/sample. Use controlled vocabulary such as <https://bioportal.bioontology.org/ontologies/CL>

- 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

+ name of the cell type of the studied object/sample. Use controlled vocabulary such as <https://bioportal.bioontology.org/ontologies/CL>

+ Additional information (optional)

name of the material (structure, substance, device) removed from a source (patient, donor, physical location, product) OR a material entity that has the specimen role. Please use the controlled vocabulary provided.

e.g., Ribonucleotides

measurement target of the studied object/sample

e.g., name of gene, protein, compound, etc.

+ Link to a database where information on the measurement target is stored.




- 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

+ Additional information (optional)

name of the material (structure, substance, device) removed from a source (patient, donor, physical location, product) OR a material entity that has the specimen role. Please use the controlled vocabulary provided.

Ribonuc

ribonuclease AIII	
Ribonuclease P	
ribonucleotide reductase 2B, Arabidopsis	
ribonuclease MC1	
Ribonucleotides	
ribonuclease A (19-26)	
Ribonuclease H	
ribonucleotide reductase 2A, Arabidopsis	
ribonuclease H2, mouse	
ribonuclease XlaI	

ment target is stored.

Nucleotides in which the purine or pyrimidine base is combined with ribose. (Dorland, 28th ed)

[Remove](#)

- 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 Vorfüllen 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

Add new render option

BiodatenMinimal

xpath

label

placeholder

prefilled value

Readonly

Hide

Active

Add

Show ID column

All	BiodatenMinimal	datacite	premis
Schema ↑	Xpath	Active	
datacite	/resource/creators	<input checked="" type="checkbox"/>	 
datacite	/resource/creators/creator/creatorName	<input checked="" type="checkbox"/>	 
datacite	/resource/creators/creator/creatorName content	<input checked="" type="checkbox"/>	 
datacite	/resource/identifier	<input checked="" type="checkbox"/>	 

- 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

Add new mapping

BiodatenMinimal

xpath

vocabulary

Active

Add

Show ID column

All BiodatenMinimal datacite premis

Schema ↑	Xpath	Vocabulary	Active		
datacite	/resource/subjects/subject{co	NCIT,MESH	<input checked="" type="checkbox"/>		
BiodatenMinimal	/cmdp:BiodatenMinimal/cmdp	BERO	<input checked="" type="checkbox"/>		
BiodatenMinimal	/cmdp:BiodatenMinimal/cmdp	CL	<input checked="" type="checkbox"/>		
BiodatenMinimal	/cmdp:BiodatenMinimal/cmdp	MESH	<input checked="" type="checkbox"/>		
BiodatenMinimal	/cmdp:BiodatenMinimal/cmdp	OBI	<input checked="" type="checkbox"/>		

Items per page: 10 1 – 5 of 5

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

Administration

Add new schema

schema (internal description)

schema file name (without file extension)

schema tab name

Active

Add

Show ID column

Schema ↑	File name (without file extension)	Tab name	Active
BiodatenMinimal	BiodatenMinimal	Research Metadata	<input checked="" type="checkbox"/>
datacite	datacite	Descriptive Metadata	<input checked="" type="checkbox"/>
premis	premis	File Metadata	<input type="checkbox"/>

- 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>