BIO
DIVERSITY
NEXT

OpenBiodiv-O Ontology: Bridging the Gap Between Biodiversity Data and Biodiversity Publishing

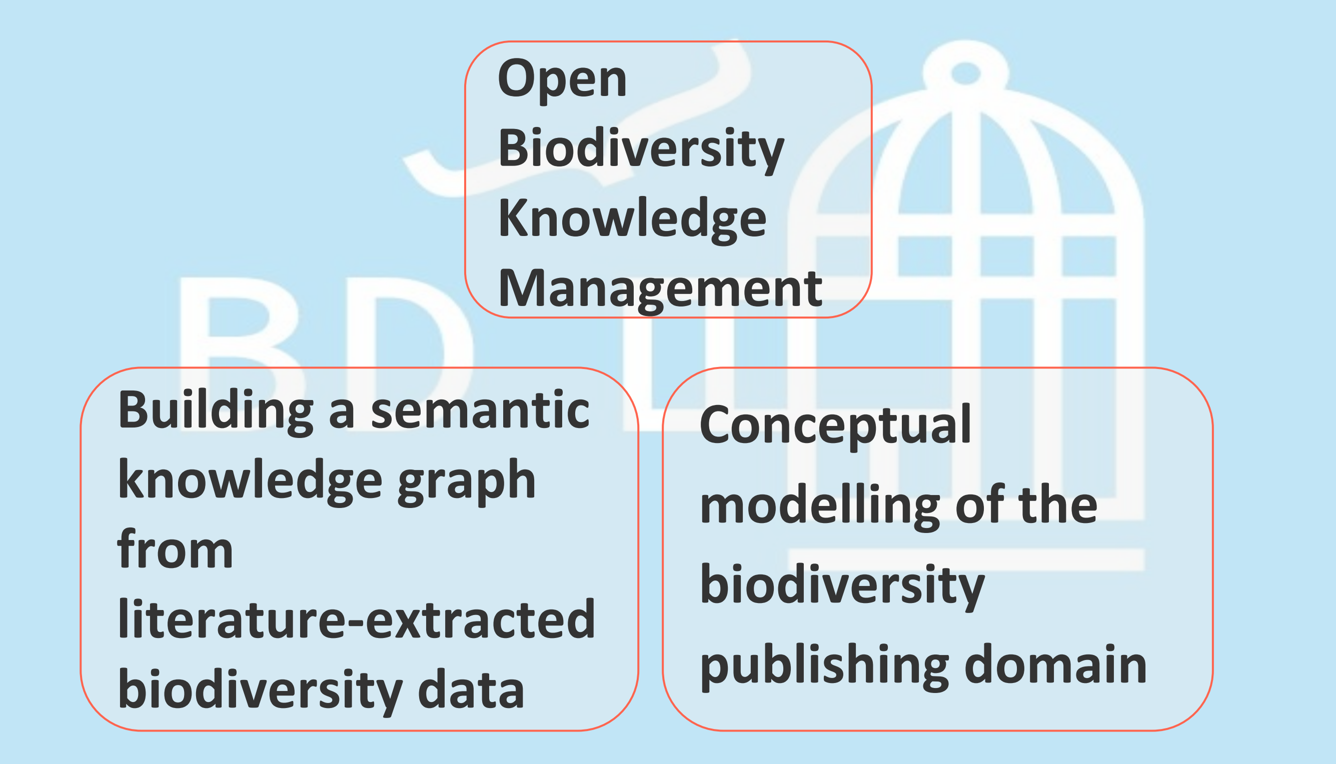
Mariya Dimitrova, Viktor Senderov, Kiril Simov, Teodor Georgiev, Lyubomir Penev | Pensoft Publishers & Bulgarian Academy of Sciences

PENSOFT®

ignite

150
1869
BULGARIAN ACADEMY OF SCIENCES

Motivation & Aims

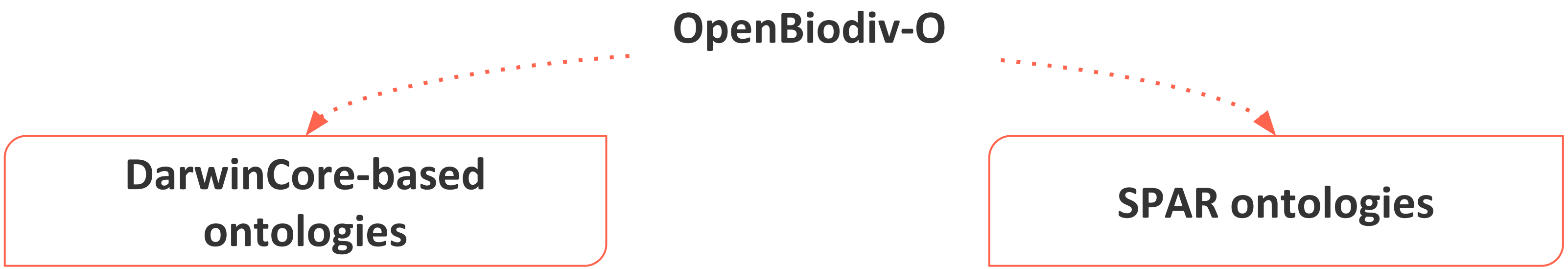


Open Biodiversity Knowledge Management

Building a semantic knowledge graph from literature-extracted biodiversity data

Conceptual modelling of the biodiversity publishing domain

Introducing OpenBiodiv-O



OpenBiodiv-O

DarwinCore-based ontologies

SPAR ontologies

Senderov *et al. Journal of Biomedical Semantics* (2018) 9:5
DOI 10.1186/s13326-017-0174-5

Journal of
Biomedical Semantics

RESEARCH Open Access

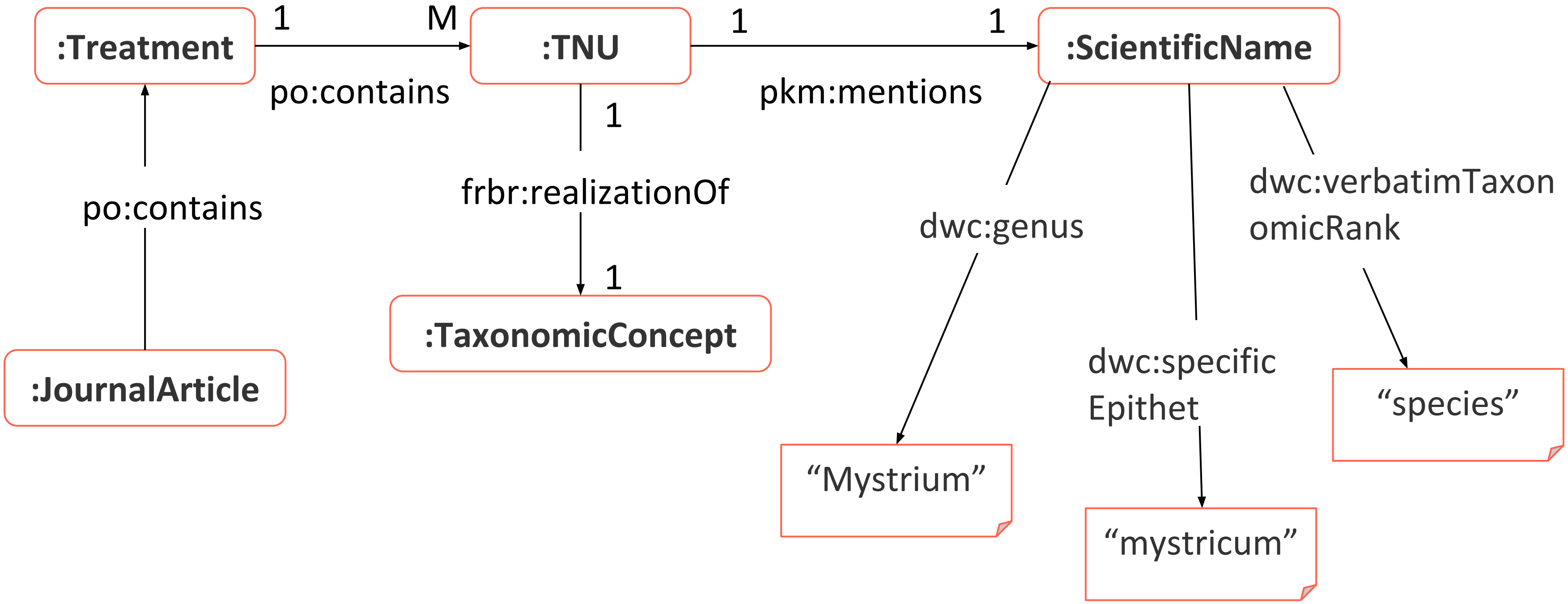
OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system

Viktor Senderov[1,2*], Kiril Simov[3], Nico Franz[4], Pavel Stoev[1,7], Terry Catapano[5], Donat Agosti[5], Guido Sautter[5], Robert A. Morris[6] and Lyubomir Penev[1,2]

Biodiversity Publishing

- Taxonomic articles:
- Abstract, Introduction, Materials and Methods, Results, Conclusions
- Taxonomic treatment:
    - Nomenclature
    - Type material
    - Etymology
    - Diagnosis
    - Description
    - Distribution
- Metadata and various identifiers (DOI, ORCID, Zoobank ID)
- **Taxonomic names**

Modelling Taxonomic Names

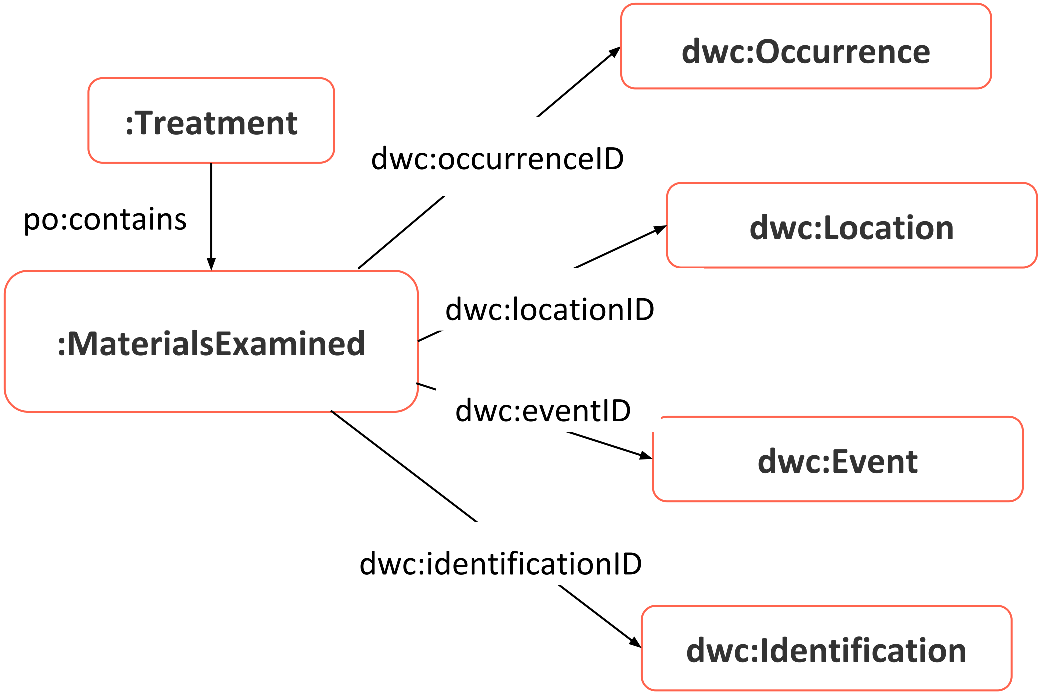


BIO DIVERSITY NEXT

Bottom-up approach to modifying OpenBiodiv-O

BIO
DIVERSITY
NEXT

1. Materials Examined



- catalogNumber
- recordNumber
- recordedBy
- individualCount
- sex
- lifeStage

- verbatimCoordinates
- country
- locality
- elevation
- depth
- waterBody

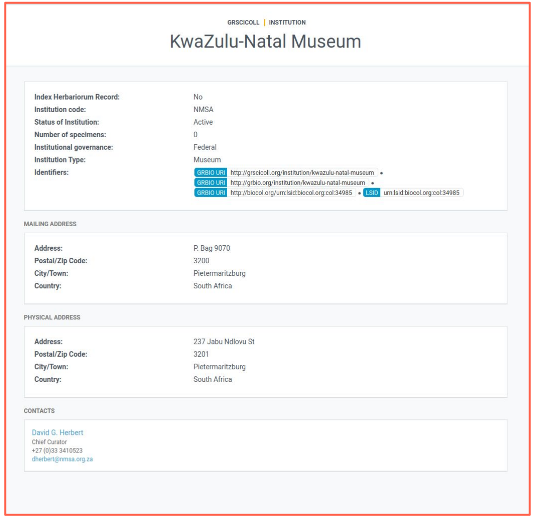
- collectionYear
- collectionMonth
- collectionDay
- eventDate
- habitat
- samplingProtocol

- identifiedBy
- typeStatus

2. Institutional identifiers

GRSciColl: The Global Registry of Scientific Collections

http://biocol.org/urn:lsid:biocol.org:col:34985



GRSCICOLL | INSTITUTION

KwaZulu-Natal Museum

Index Herbariorum Record:	No
Institution code:	NMSA
Status of Institution:	Active
Number of specimens:	0
Institutional governance:	Federal
Institution Type:	Museum
Identifiers:	GRBIO URI http://grscicoll.org/institution/kwazulu-natal-museum • GRBIO URI http://grbio.org/institution/kwazulu-natal-museum • GRBIO URI http://biocol.org/urn:lsid:biocol.org:col:34985 • LSID urn:lsid:biocol.org:col:34985

MAILING ADDRESS

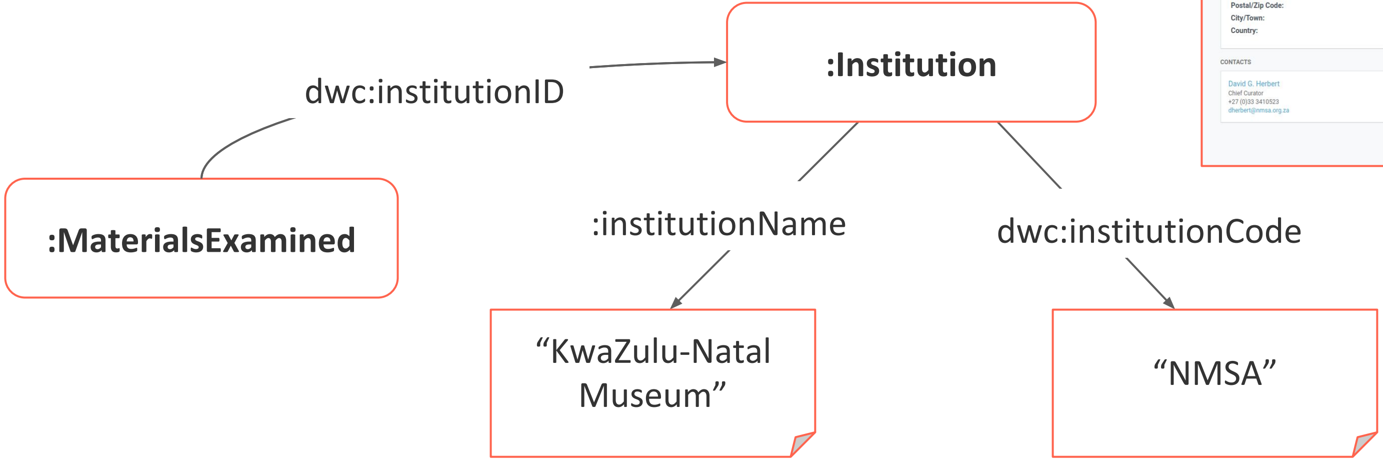
Address:	P. Bag 9070
Postal/Zip Code:	3200
City/Town:	Pietermaritzburg
Country:	South Africa

PHYSICAL ADDRESS

Address:	237 Jabu Ndlovu St
Postal/Zip Code:	3201
City/Town:	Pietermaritzburg
Country:	South Africa

CONTACTS

David G. Herbert
Chief Curator
+27 (0)33 3410523
dherbert@nmsa.org.za



:Institution

dwc:institutionID

:MaterialsExamined

:institutionName

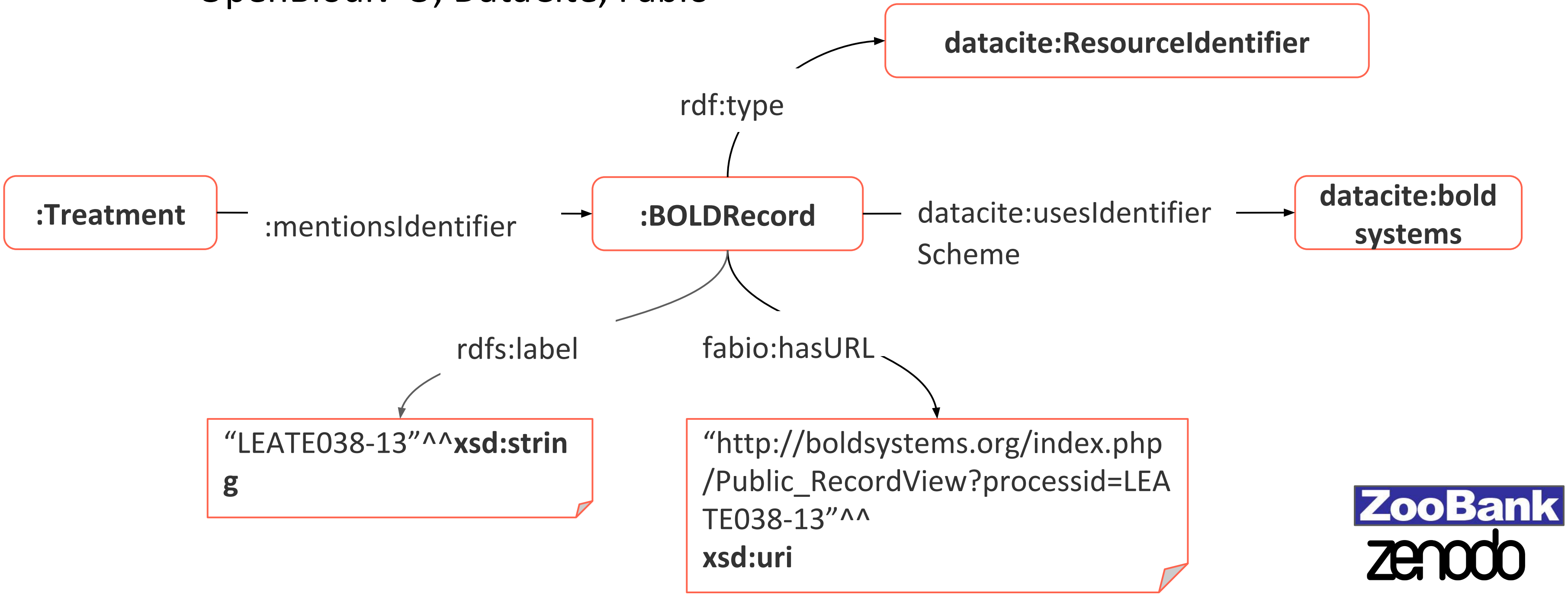
dwc:institutionCode

"KwaZulu-Natal Museum"

"NMSA"

3. Molecular resource identifiers

- OpenBiodiv-O, DataCite, Fabio



:Treatment

:mentionsIdentifier

:BOLDRecord

rdf:type

datacite:ResourceIdentifier

datacite:usesIdentifier Scheme

datacite:bold systems

rdfs:label

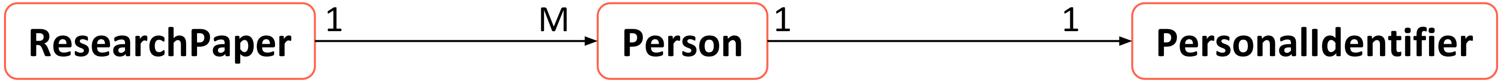
fabio:hasURL

"LEATE038-13"^^**xsd:string**

"http://boldsystems.org/index.php/Public_RecordView?processid=LEATE038-13"^^ **xsd:uri**

4. Personal identifiers

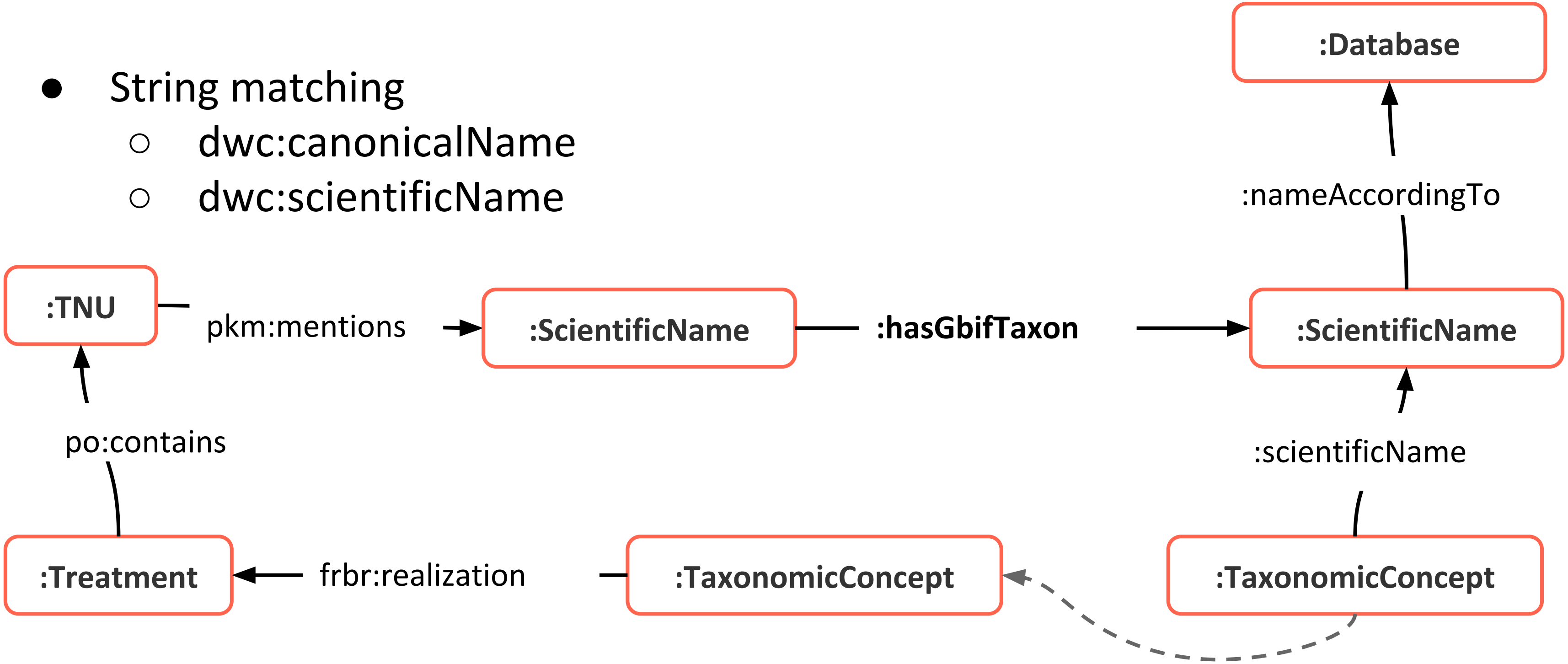
```
openbiodiv:8B0A6890-3094-4431-8262-23748A86B071   rdf:type   fabio:ResearchPaper ;
        dcterms:creator   openbiodiv:7F19D49E-B4CA-4E7A-808D-A57DAA7E02A3.

openbiodiv:7F19D49E-B4CA-4E7A-808D-A57DAA7E02A3   rdf:type   foaf:Person ;
        rdfs:label   "Mengmeng Liu" ;
        openbiodiv:affiliation   "College of Ecology, Lishui University, Lishui, Zhejiang, China" ;
        datacite:hasIdentifier   orcid:0000-0002-0985-5852 .

orcid:0000-0002-0985-5852   rdf:type   datacite:PersonalIdentifier ;
        datacite:usesIdentifierScheme   datacite:orcid ;
        rdfs:label   "0000-0002-0985-5852" .
```



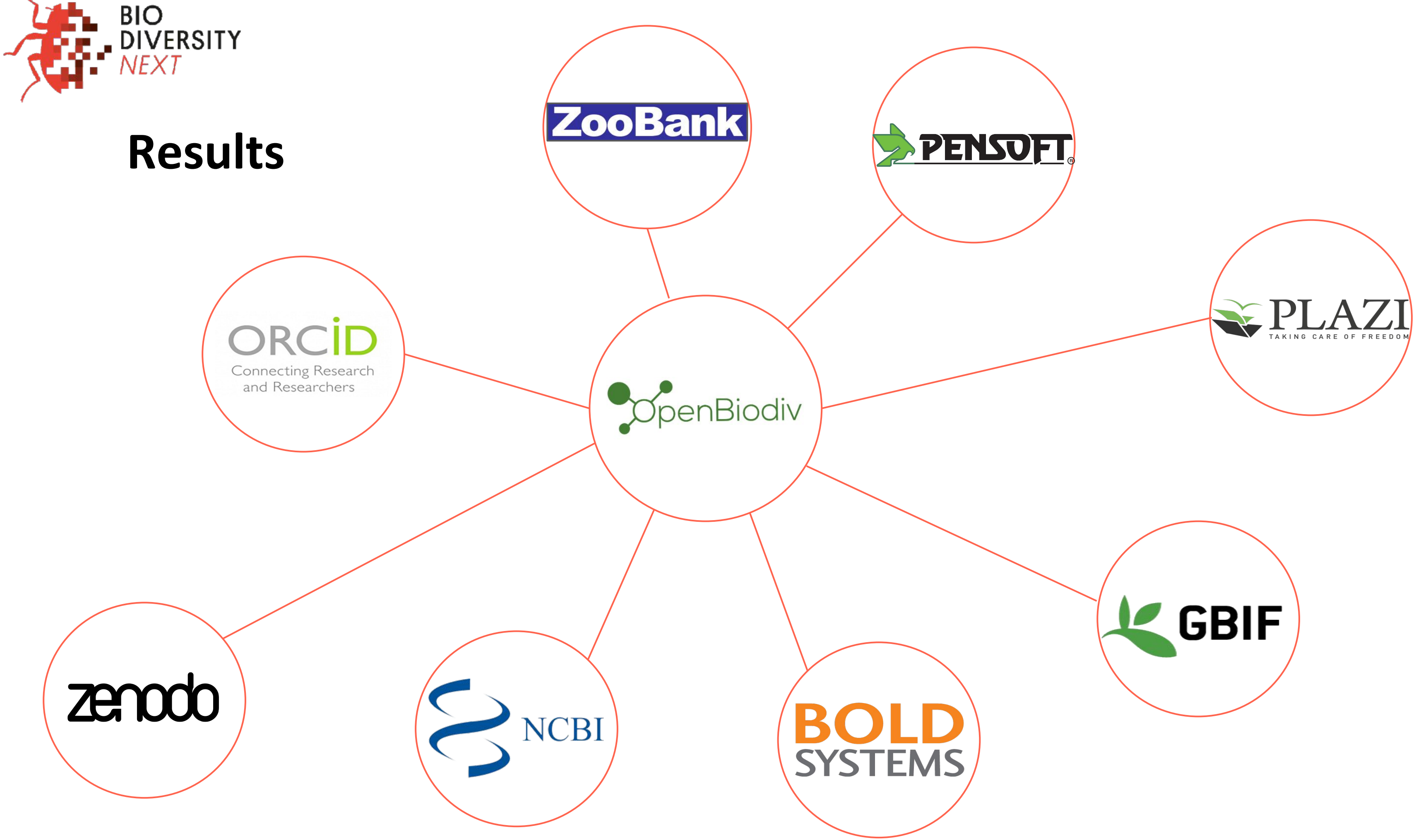
ResearchPaper 1 → M Person 1 → 1 PersonalIdentifier

5. Explicit mapping to GBIF’s taxonomy

- String matching
  - dwc:canonicalName
  - dwc:scientificName



Results

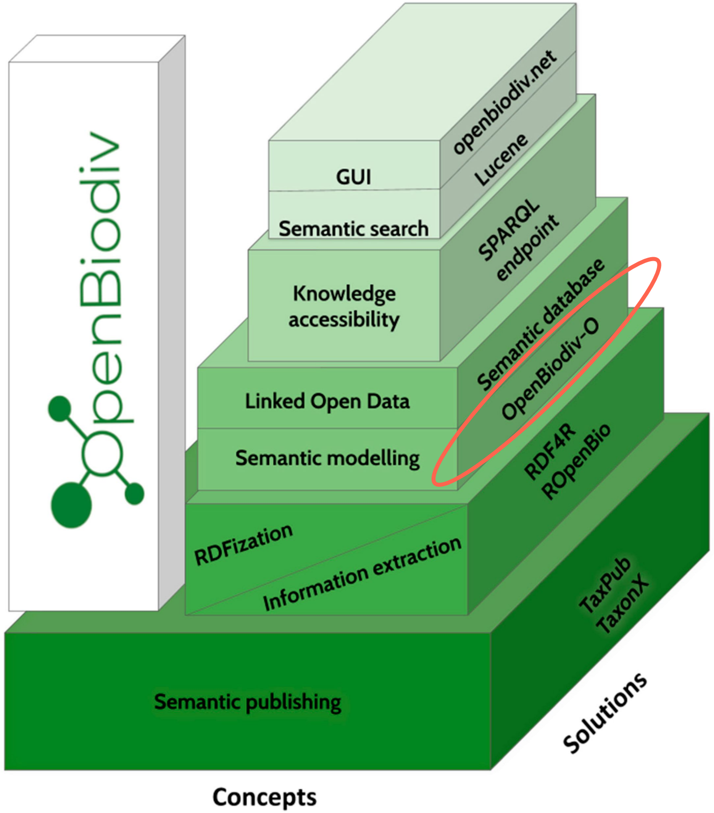

ZooBank
PENSOFT
PLAZI
TAKING CARE OF FREEDOM
ORCID
Connecting Research and Researchers
OpenBiodiv
zenodo
NCBI
BOLD SYSTEMS
GBIF

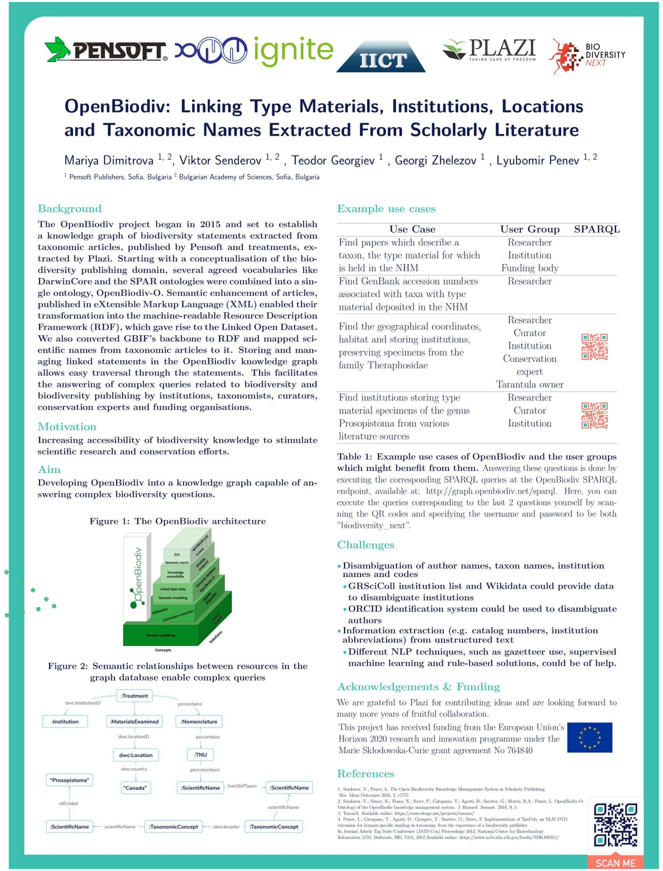
Use cases

- Modelling and linking resources across domains
  - Publishing
  - Taxonomy
  - Genomics
- Serving users from different groups
  - Taxonomists
  - Ecologists
  - Curators
  - Institutions

Application

PS 0101





OpenBiodiv: Linking Type Materials, Institutions, Locations and Taxonomic Names Extracted From Scholarly Literature

Mariya Dimitrova [1, 2], Viktor Senderov [1, 2] , Teodor Georgiev [1] , Georgi Zhelezov [1] , Lyubomir Penev [1, 2]

[1] Pensoft Publishers, Sofia, Bulgaria [2] Bulgarian Academy of Sciences, Sofia, Bulgaria

Background

The OpenBiodiv project began in 2015 and set to establish a knowledge graph of biodiversity statements extracted from taxonomic articles, published by Pensoft and treatments, extracted by Plazi. Starting with a conceptualisation of the biodiversity publishing domain, several agreed vocabularies like DarwinCore and the SPAR ontologies were combined into a single ontology, OpenBiodiv-O. Semantic enhancement of articles, published in eXtensible Markup Language (XML) enabled their transformation into the machine-readable Resource Description Framework (RDF), which gave rise to the Linked Open Dataset. We also converted GBIF's backbone to RDF and mapped scientific names from taxonomic articles to it. Storing and managing linked statements in the OpenBiodiv knowledge graph allows easy traversal through the statements. This facilitates the answering of complex queries related to biodiversity and biodiversity publishing by institutions, taxonomists, curators, conservation experts and funding organisations.

Motivation

Increasing accessibility of biodiversity knowledge to stimulate scientific research and conservation efforts.

Aim

Developing OpenBiodiv into a knowledge graph capable of answering complex biodiversity questions.

Figure 1: The OpenBiodiv architecture

Figure 2: Semantic relationships between resources in the graph database enable complex queries

Example use cases

Use Case	User Group	SPARQL
Find papers which describe a taxon, the type material for which is held in the NHM	Researcher Institution Funding body	
Find GenBank accession numbers associated with taxa with type material deposited in the NHM	Researcher	
Find the geographical coordinates, habitat and storing institutions, preserving specimens from the family Theraphosidae	Researcher Curator Institution Conservation expert Tarantula owner	
Find institutions storing type material specimens of the genus Prosopistoma from various literature sources	Researcher Curator Institution	

Table 1: Example use cases of OpenBiodiv and the user groups which might benefit from them. Answering these questions is done by executing the corresponding SPARQL queries at the OpenBiodiv SPARQL endpoint, available at: http://graph.openbiodiv.net/sparql. Here, you can execute the queries corresponding to the last 2 questions yourself by scanning the QR codes and specifying the username and password to be both "biodiversity_next".

Challenges

- **Disambiguation of author names, taxon names, institution names and codes**
  - **GRSciColl institution list and Wikidata could provide data to disambiguate institutions**
  - **ORCID identification system could be used to disambiguate authors**
- **Information extraction (e.g. catalog numbers, institution abbreviations) from unstructured text**
  - **Different NLP techniques, such as gazetteer use, supervised machine learning and rule-based solutions, could be of help.**

Acknowledgements & Funding

We are grateful to Plazi for contributing ideas and are looking forward to many more years of fruitful collaboration.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 764840

References

1. Senderov, V.; Penev, L. The Open Biodiversity Knowledge Management System in Scholarly Publishing. Res. Ideas Outcomes 2016, 2, e7757.
2. Senderov, V.; Simov, K.; Franz, N.; Stoev, P.; Catapano, T.; Agosti, D.; Sautter, G.; Morris, R.A.; Penev, L. OpenBiodiv-O: Ontology of the OpenBiodiv knowledge management system. J. Biomed. Semant. 2018, 9, 5.
3. TaxonX. Available online: https://sourceforge.net/projects/taxonx/
4. Penev, L.; Catapano, T.; Agosti, D.; Georgiev, T.; Sautter, G.; Stoev, P. Implementation of TaxPub, an NLM DTD extension for domain-specific markup in taxonomy, from the experience of a biodiversity publisher. In Journal Article Tag Suite Conference (JATS-Con) Proceedings 2012; National Center for Biotechnology Information (US): Bethesda, MD, USA, 2012.Available online: https://www.ncbi.nlm.nih.gov/books/NBK100351/

SCAN ME

BIO DIVERSITY NEXT

Next steps

- Facilitate federated SPARQL queries by mapping external ontologies to OpenBiodiv-O.
  - Wikidata
- Expanding the ontology as the number of extracted entity types increases

BIO
DIVERSITY
NEXT

Thank you!

Questions?

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 764840 & No 642241