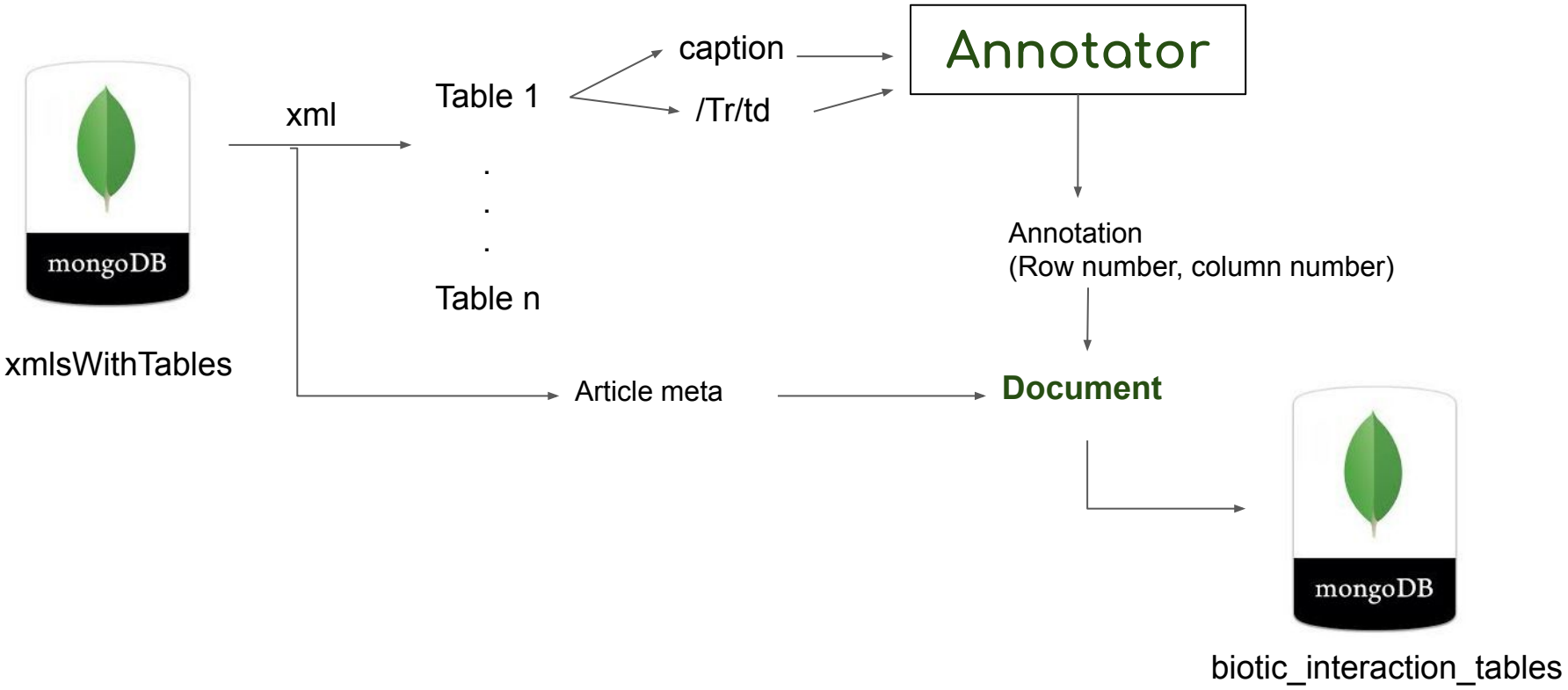
Detecting Pensoft-published tables containing biotic interactions

2020-06-25

Mariya Dimitrova

Identifying tables (possibly) containing biotic interactions



Representation in MongoDB

Article and table meta


```
{
    "_id" : ObjectId("5ef1b4d297119b779a4e7f3a"),
    "table_id" : "<http://openbiodiv.net/FE06B7F0-DBC2-4419-92DF-39FC80F2BAD8>",
    "table_content" : "<table id=\"TID0EC6AE\" [...] </table>",
    "caption" : "Species overview. Scientific and vernacular names of insects and host plants according to local Kikongo dialect [...]",
    "table_number" : "TID0EC6AE",
    "article_doi" : "10.3897/afrinvertebr.58.21083",
    "annotations" : [
        {
            "id" : "http://purl.obolibrary.org/obo/RO_0002453",
            "lbl" : "host",
            "length" : 4,
            "position" : 65,
            "ontology" : "custom",
            "type" : "PROPERTY",
            "context" : "species overview. scientific and vernacular names of insects and host plants according to local kikongo dialect, except one
name, which is marked with (kim.) according to kimbundu language; plant names",
            "is_synonym" :  true,
            "is_word" :    true,
        },
        {
            [...]
```

Article and table meta


```
{
   "_id" : ObjectId("5ef1b4d297119b779a4e7f3a"),
   "table_id" : "<http://openbiodiv.net/FE06B7F0-DBC2-4419-92DF-39FC80F2BAD8>",
   "table_content" : "<table id=\"TID0EC6AE\" [...] </table>",
   "caption" : "Species overview. Scientific and vernacular names of insects and host plants according to local Kikongo dialect [...]",
   "table_number" : "TID0EC6AE",
   "article_doi" : "10.3897/afrinvertebr.58.21083",
```


- table_id - The OpenBiodiv identifier of the table (statements will be added to the graph database)
- table_content - the full xml of the table (String type)
- caption - the table caption
- table_number - identifier to find the exact table in the xml of the article
- article_doi - the doi

-> You can obtain the table in 2 ways:

1. Directly from the MongoDB document as a xml
2. By resolving the article doi and then finding the table via its table number

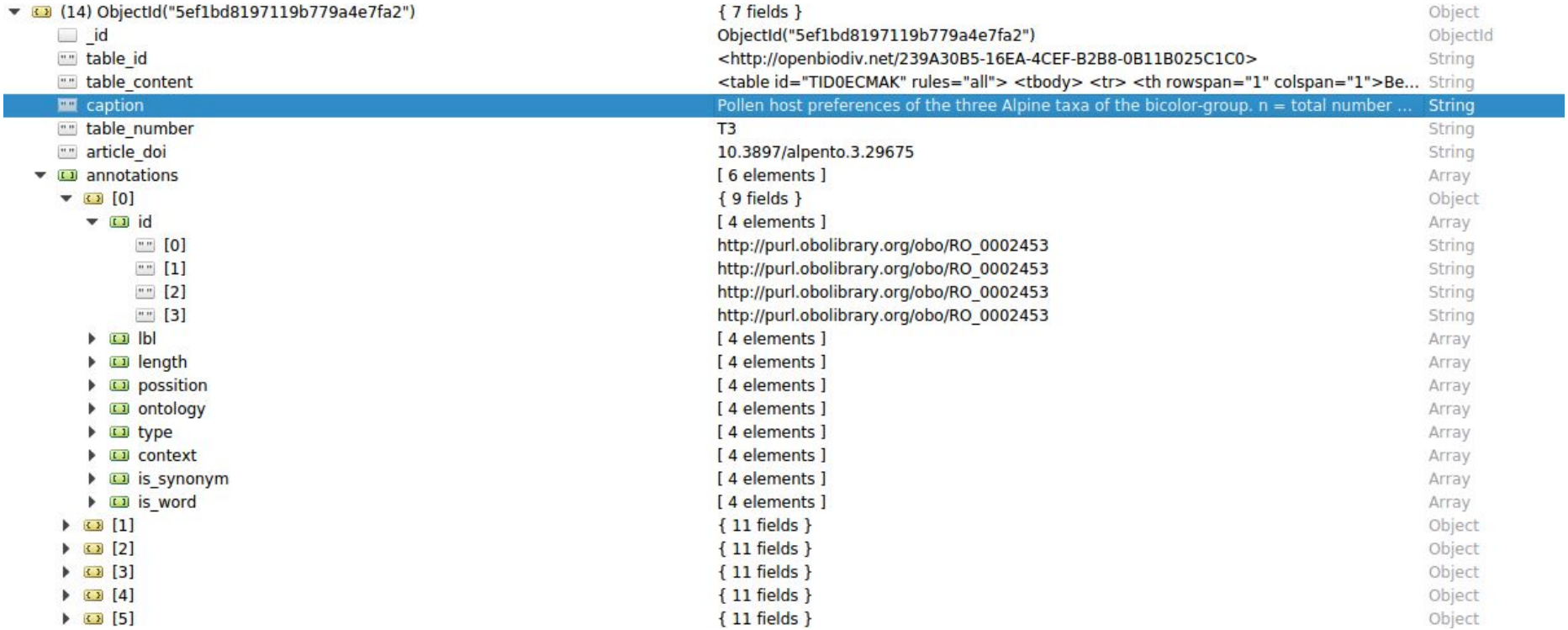
The annotation

- id - Term id
- lbl - the label of the ontology term
- length, position - the length of the matched term and its position in the text
- ontology - which ontology we annotated with
- type - Class or Property
- context - 10 words
- is_synonym - the matched text can be a synonym of a term
- is_word - the term can be a word or a phrase


```
{
   "id" : "http://purl.obolibrary.org/obo/RO_0002453",
   "lbl" : "host",
   "length" : 4,
   "position" : 65,
   "ontology" : "custom",
   "type" : "PROPERTY",
   "context" : "species overview. scientific and vernacular names of insects and host plants according to local kikongo dialect, except one name, which is marked with (kim.) according to kimbundu language; plant names",
   "is_synonym" :  true,
   "is_word" :    true,
},
```

Ontology - custom

- We call it ontology but it is essentially a vocabulary
- Modified RO ontology to include only subProperties of term labeled 'biotically interacts with' , removed all other terms
- Added different word forms and spellings to each term as exact synonyms to the term (our annotator filters out any broad, narrow and related synonyms)
    - **host of**: **host, hostof** (table headings may be formatted in camelCase)
    - **is killed by**: **killed, killedby, iskilledby**
- We don't need complete accuracy because we only use the 'ontology' for detection of tables and do not use it any further


(14) ObjectId("5ef1bd8197119b779a4e7fa2") { 7 fields } Object
_id ObjectId("5ef1bd8197119b779a4e7fa2") ObjectId
table_id <http://openbiodiv.net/239A30B5-16EA-4CEF-B2B8-0B11B025C1C0> String
table_content <table id="TID0ECMAK" rules="all"> <tbody> <tr> <th rowspan="1" colspan="1">Be... String
caption Pollen host preferences of the three Alpine taxa of the bicolor-group. n = total number ... String
table_number T3 String
article_doi 10.3897/alpento.3.29675 String
annotations [6 elements] Array
[0] { 9 fields } Object
id [4 elements] Array
[0] http://purl.obolibrary.org/obo/RO_0002453 String
[1] http://purl.obolibrary.org/obo/RO_0002453 String
[2] http://purl.obolibrary.org/obo/RO_0002453 String
[3] http://purl.obolibrary.org/obo/RO_0002453 String
lbl [4 elements] Array
length [4 elements] Array
possition [4 elements] Array
ontology [4 elements] Array
type [4 elements] Array
context [4 elements] Array
is_synonym [4 elements] Array
is_word [4 elements] Array
[1] { 11 fields } Object
[2] { 11 fields } Object
[3] { 11 fields } Object
[4] { 11 fields } Object
[5] { 11 fields } Object

Questions

- How can this workflow contribute to GLoBI harvesting?
- Should we aim for a workflow to help generate a GLoBI spreadsheet (sourceTaxonName - interactionTypeName) ?
  - XMLs are tagged with taxonomic names so we can extract them but the table structure can be ambiguous
- Can we improve the custom ontology/vocabulary to include more interactions?
- Taxonomic names

Federated queries between GLoBI and OpenBiodiv - opportunities

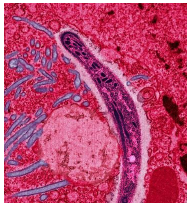
Passer domesticus

openbiodiv:ScientificName

parasiteOf

Plasmodium relictum

openbiodiv:ScientificName



contains

openbiodiv:Materials Examined

dwc:institutionID

Ornithology Collection at Academy of Natural Sciences of Drexel University

openbiodiv:Institution
openbiodiv:GRSciCollInstitution

openbiodiv:Treatment

mentions

openbiodiv:TNU

contains

contains

openbiodiv:Nomenclature Section