# AnnoSys2: Reaching out to the Semantic Web

Okka Tschöpe[1], Lutz Suhrbier[2], Anton Güntsch[3] and Walter G. Berendsohn[4]

BGBM, Freie Universität Berlin, Germany
[1] o.tschoepe@bgbm.org
[2] l.suhrbier@bgbm.org
[3] a.guentsch@bgbm.org
[4] w.berendsohn@bgbm.org

**Abstract.** AnnoSys is a web-based open-source system for correcting and enriching specimen data in publicly available data portals, thereby bringing traditional annotation workflows for biodiversity data to the Internet. During its first phase, the project developed a fully functional prototype of an annotation data repository for complex and cross-linked XML-standardized data, including back-end server functionality, web services and an on-line user interface. Annotation data are stored using the Open Annotation Data Model and an RDF-database. The current project phase aims at extending the generic qualities of AnnoSys to further structured data formats including RDF data with machine readable semantic concepts, thus opening up the data gathered through AnnoSys for the Semantic Web. We developed a semantic concept-driven annotation management, including the specification of a selector concept for RDF data and a repository for original records extended to RDF and other formats. Since many of the biodiversity data standards in use are still not defined in a semantic-web compliant way, mechanisms for referencing elements in such data sets need to be developed. We therefore developed an AnnoSys ontology based on DwC RDF terms and the ABCD ontology, which deconstructs the ABCD XML-schema into individually addressable RDF-resources published via the TDWG Terms Wiki. We mapped the terms from these standards into annotation types we defined, based on semantic concepts.

**Keywords:** AnnoSys, Ontology, Annotation.

## 1 Introduction

Biodiversity data are aggregated, linked and made globally accessible via a range of Internet portals and services. Globally, natural history collections contain 2–3 billion specimens [1]. These provide materials and primary data for a wide range of research questions and form the basis for the classification of organisms into species and other “taxa”. Traditionally, specimens are annotated by researchers with written annotation labels which are applied directly to the physical object, thus becoming accessible to succeeding observers of the specimen. These annotations improve the data quality of the collection and document research developments over time (e.g. the understanding of taxon concepts).

To ensure the continuance of the traditional data sharing and incremental documentation of specimens in the on-line environment, the AnnoSys project developed an annotation data repository [2] for complex XML data following the ABCD [3] and DwC [4] standards. This includes back-end server functionality, web services and an on-line user interface [5]. Annotation data are stored using the Web Annotation Data Model [6] and an RDF-database [7].

In a second step, AnnoSys2 aims at extending the generic qualities of AnnoSys to further structured data formats including RDF data with machine readable explicit semantic concepts.

## 2 Motivation/State of the art

Since many of the biodiversity data standards in use are still not defined in a semantic-web compliant way, mechanisms for referencing elements in such data sets need to be developed. We therefore compiled an AnnoSys ontology based on DwC RDF terms and the ABCD ontology, which deconstructs the ABCD XML-schema into individually addressable RDF-resources published via the TDWG Terms Wiki [8].

One of our motivations for the new ontology was to harmonize annotatable elements to allow unambiguous comparability between different versions of a record. For example, depending on the data publishing portal, a record can be displayed either in the DwC or the ABCD standard. In AnnoSys 1 we were facing the problem that those records were not directly comparable, because not all ABCD elements are part of the DwC standard and vice versa. We therefore needed different versions histories for different data standards (DwC, ABCD 2.06, ABCD 2.1 etc.). The AnnoSys ontology defines matching rules describing how these different elements are transformed into annotatable elements, resulting in harmonized records with only one, unambiguously comparable version history.

Additionally, via the different SKOS-relations equivalence levels for matches of elements can be specified, which potentially allows restricting the use of elements to those with a minimum level of equivalence. This may be important for data formats that need to be integrated in the future

## 3 Model construction

We used Protégé [9] to build an AnnoSys ontology based on DwC terms [4] and the ABCD ontology [8], which uses ABCD property terms as RDF predicates. We created a subclass “RecordConcept” comprising all ontology concepts as a subclass of *skos:concept* (Fig. 1). We also defined nine different “annotation types” as instances of the SubClass “annotation type” of *oa:Motivation* (Fig.1, Fig. 2). Individual concepts were related to the different annotation types via the skos:related relation. We then mapped the elements of the two standards to semantic concepts using the skos:Concepts exactMatch, broadMatch, narrowMatch, or closeMatch, respectively, to represent the different levels of matches (Table 1).

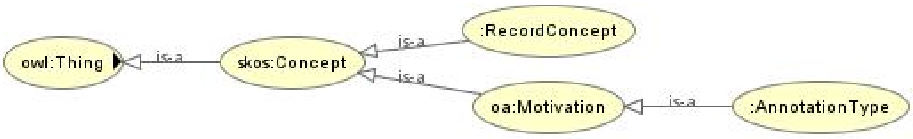

**Fig. 1.** Subclasses of skos:Concept in the AnnoSys Ontology.

Concepts that refer to identifiers of the institution, the collection or the unit, are not related to an annotation type but are also instances of the subclass "Record Concept". These concepts are not annotatable, but are important in their function as identifiers (e.g. to query for records related to a given triple id – the identifier originally used in schemas describing specimens, composed of three ids designating the holding institution, a collection within the institution, and the catalogue number within that collection).

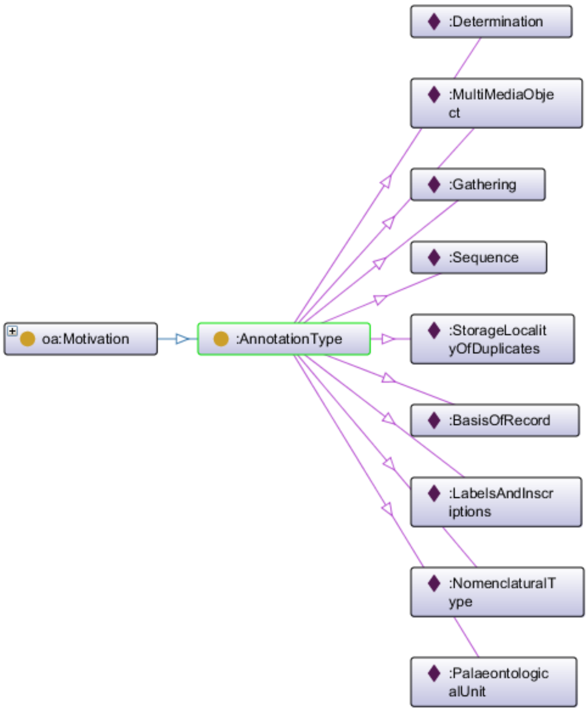

**Fig. 2.** "Annotation type" is a subclass of *oa:Motivation*, which is a subclass of *skos:concept*.

**Table 1.** Example concepts of the AnnoSys Ontology and their mappings for annotation type "Determination"

| Concept in AnnoSys ontology | Skos exact match | Skos close match | Narrow match | Skos related |
|---|---|---|---|---|
| Full scientific name | Dwc:Scientific Name | abcd2:TaxonIdentified-FullScientific-NameString | | Annotation type: Determination |
| Scientific Name Authorship | dwc:Scientific NameAuthorship | | abcd2:TaxonIdentified-AuthorTeam<br>abcd2:TaxonIdentified-AuthorTeamAndYear<br>abcd2:TaxonIdentified-AuthorTeamOriginalAndYear | Annotation type: Determination |
| Scientific Name Authorship Parenthetical | | | abcd2:TaxonIdentified-ParentheticalAuthor-TeamAndYear<br>abcd2:TaxonIdentified-AuthorTeamParenthesis<br>abcd2:TaxonIdentified-AuthorTeamParenthesisAndYear | Annotation type: Determination |

A prototype of the system is available under https://dev-annosys.bgbm.fu-berlin.de/AnnoSys/AnnoSys.

## 4 Evaluation

The ontology is composed of around 150 data properties that are related to nine annotation types. Since concepts are now defined in a semantic-web compliant way, they can be stored together with the record in the same triple store (whereas in AnnoSys 1, records have been stored in an XML database). This allows more complex searches and significantly improves the performance of the system. AnnoSys data properties cover the classic annotation workflows in the biodiversity collection data domain. However, the ontology is potentially expandable for other workflows and other domains.

When aiming to integrate annotations for specimens from different data portals, it is essential to be able to identify annotated specimens universally. Therefore, AnnoSys 2 builds persistent identifiers for all objects (records, specimens and annotations) from UUIDs, making the system independent of the previously used tripleIds.

## 5 Conclusion

Our work tackles the development of an extensible and format-independent system for virtual annotation of biological specimen label data. To this end, we compiled an "AnnoSys-Ontology" mapping essential concepts defined by the widely accepted community standards DarwinCore and ABCD. Annotations are entered via an open browser interface and stored centrally in an RDF triple store following the W3C Web Annotation Data Model.

The system is currently in the testing phase and will be released in 2018. In future research, we will examine the use of AnnoSys for taxon-level data as well as its integration with image annotation systems.

## References

1. Duckworth, W.D., Genoways, H.H., Rose, C.L. et al.: Preserving Natural Science Collections: Chronicle of Our Environmental Heritage. National Institute for the Conservation of Cultural Property, Washington, DC. (1993)
2. AnnoSys portal, https://annosys.bgbm.fu-berlin.de/AnnoSys/AnnoSys, last accessed 2017/07/18
3. Berendsohn W.G. (ed.). Access to biological collection data. ABCD Schema 2.06 – ratified TDWG Standard. Berlin: Botanischer Garten und Botanisches Museum Berlin-Dahlem (BGBM), Freie Universität Berlin. (2007) http://www.bgbm.org/TDWG/CODATA/Schema/default.htm.
4. DwC terms homepage, http://rs.tdwg.org/dwc/terms/index.htm, last accessed 2017/07/18
5. Tschöpe, O., Macklin, J.A., Morris, R.A. et al. Annotating biodiversity data via the Internet. Taxon, 62, 1248–1258 (2013)
6. Web Annotation Data Model homepage, https://www.w3.org/TR/annotation-model/, last accessed 2017/07/18
7. Suhrbier, L., Kusber, W.-H., Tschöpe, O., Güntsch, A. & Berendsohn, W. G.: AnnoSys - implementation of a generic annotation system for schema-based data using the example of biodiversity collection data. Database (2017). doi:10.1093/database/bax018
8. ABCD2 homepage, https://terms.tdwg.org/wiki/ABCD_2, last accessed 2017/09/07
9. Protege homepage, http://protege.stanford.edu/products.php, last accessed 2017/07/18