OAC Demo Report: "Adapting the OAC data model for automated annotation of the biomedical literature"

Karin Verspoor and Kevin Livingston
University of Colorado Denver
February 3, 2013

Executive summary

We introduce our use case of representation of annotations over the biomedical literature produced by automated natural language processing tools. We propose an extension to the OAC data model for structured bodies and representation of fine-grained provenance. We further investigated the relationship of the OAC data model to the Linguistic Annotation Framework, and lay the foundation for representation of detailed linguistic annotations over text using the OAC data model. We did not get as far as we had hoped in producing resources or tools using the OAC data model, due to the significant investment in the investigation of representational matters.

Use Case Context and Description

The context of our use case is annotation of the scientific literature using automated natural language processing tools. We are developing tools for information extraction of concepts and events in the biomedical literature. Our tools typically have output annotations over the texts in an ad hoc representation formalism derived from the text analysis framework that we have adopted (UIMA, the Unstructured Information Management Architecture); these annotations are not interoperable with any other frameworks or tools. Therefore we were interested in exploring the applicability of the OAC model in this context.
We have identified several key characteristics that an annotation representation must address for our use case:

1. Annotations can identify arbitrary segments of text as annotation targets, including discontiguous spans.
2. Structured bodies. We have a direct requirement for the representation of richly structured bodies involving sets of assertions.
3. Provenance. The sources (whether manually derived or system-generated) of assertions used in interpretation and analysis of text must be tracked. Specifically, we have explored the representational requirements of compositional analysis of a text. That is, if an annotation is based directly on other annotations, we wish to record that provenance.

Description of Annotation Classes Associated with Use Case

Our use case addresses the very top of the Open Annotation (OA) model, specifically `oa:Annotation`.
The following is an RDFS alignment of our model with the OA model in N3 notation.

```
@prefix kiao: <http://kabob.ucdenver.edu/iao/>
@prefix oa:   <http://www.w3.org/ns/openannotation/core/>
@prefix rdfs: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
```

```
@prefix rdfg: <http://www.w3.org/2004/03/trix/rdfg-1/>
```

We introduce two subclasses of `oa:Annotation` which are also subclasses of our own `kiao:Annotation`.

```
kiao:RdfResourceAnnotation rdfs:subClassOf oa:Annotation.
kiao:RdfGraphAnnotation    rdfs:subClassOf oa:Annotation.
```

These subclasses capture our requirement to represent structured bodies. `RdfResourceAnnotation` is for a standard single resource body. `RdfGraphAnnotation` is a RDF graph, composed of a set of one or more RDF statements, that is being used to annotate another information content entity and is of `rdf:type kiao:RdfGraphAnnotation`. A graph annotation is connected to a named graph of RDF statements using the property `iao:denotes`. While a graph annotation is directly linked to a named graph, it actually denotes the *content* of the named graph (*i.e.*, the RDF graph that the named graph encodes or represents) and not the named graph itself; this is consistent with the semantics of named graphs proposed by Carroll *et al.* [1], which states that any assertion in RDF about the graph structure of a named graph is understood to refer to the underlying RDF graph.

Illustrative Annotations

We introduce here a few examples of the annotations we have targeted, focusing on the representation of provenance.

The simplest way to record provenance is to make coarse-grained `basedOn` assertions between annotations. A `basedOn` statement can be made between two annotations either when there is a direct relationship between the annotations, such as one directly using one or more elements of another, or when there is an indirect relationship.

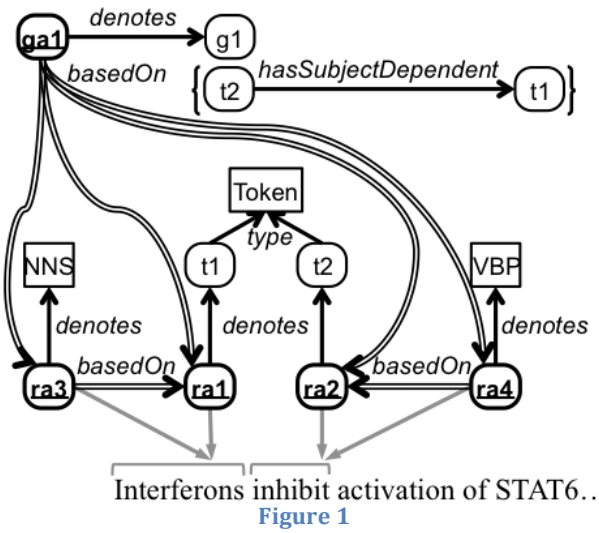

Figure 1

Most syntactic dependency parsers use tokenization and part-of-speech tags produced by other annotation systems as input. Figure 1 depicts six different annotation-level `basedOn` assertions between syntactic resource annotations. Those from `ra3` to `ra1` and from `ra4` to `ra2` have been asserted because `ra3` and `ra4`, denoting parts of speech, were created based on `ra1` and `ra2`, denoting tokenization, respectively. Provenance relations are analogously depicted among the semantic annotations in Figure 2. Graph annotation `ga2` was built using information from resource annotations `ra6` and `ra7`. Similarly, the larger graph annotation `ga3` records that it was built using information from resource annotation `ra5` and from graph annotation `ga2`. The provenance information can be traced from annotation to annotation, and in this case one can see that `ga3` is (partly) based on `ga2`, which in turn is based on `ra6` and `ra7`.

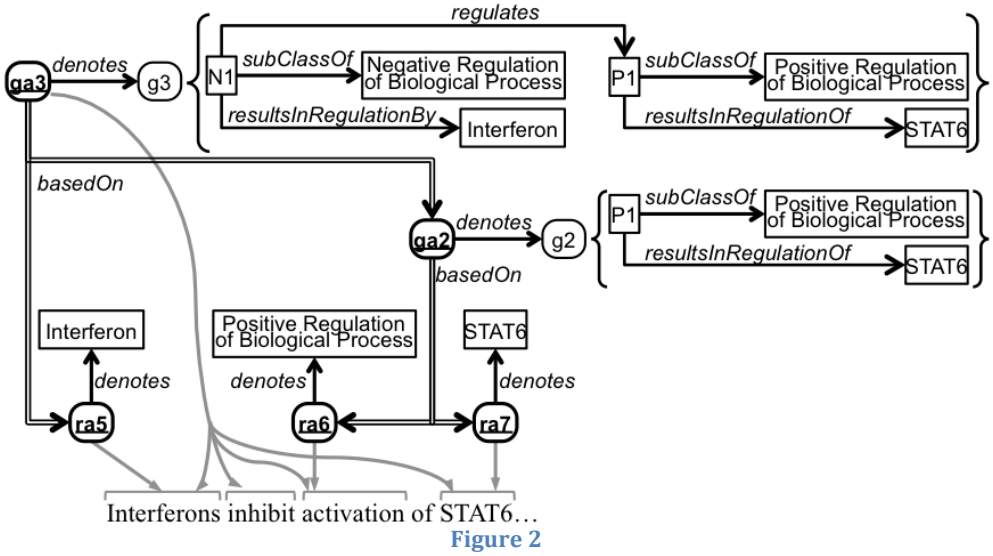

Figure 2

The second type of provenance represented in our model records detail at a more fine-grained level. Referring back to Figure 2, while the assertion `ga2 basedOn ra7` is sufficient to model that at least some part of `ga2` was based on `ra7`, it doesn't capture *which* elements of `ga2` are based on `ra7`. In RDF, the typical way to make statements about statements is to reify the statement itself as an instance of `rdf:Statement`. An RDF statement identifies its subject, property, and object via the relations `rdf:subject`, `rdf:property`, and `rdf:object`, respectively. However, RDF statements and their elements are conceptual representations; for example, in Figure 1, the RDF statement `t2 hasSubjectDependent t1` represents the assertion that token `t2` has as its subject token `t1`. To explicitly represent RDF statements as information content entities, we introduce the class `kiao:RdfStatement`, which is `rdfs:subClassOf iao:information content entity`. An example of a reified `kiao:RdfStatement` is the instance `s1` in Figure 3. A graph annotation can then be connected to each reified statement of the graph annotation using the property `ro:has_part`.

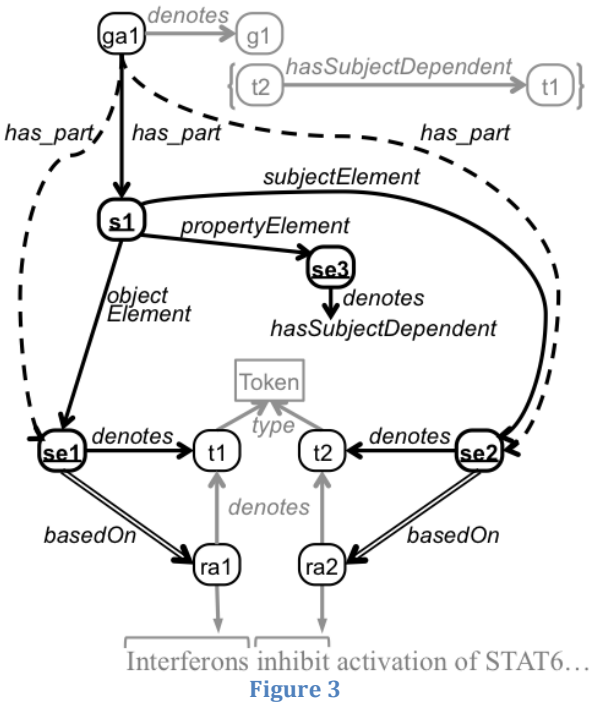

Figure 3

In order to record the provenance about individual parts of a statement, these parts must also be reified as instances of `kiao:RdfStatementElement`. A reified statement is linked to its component instances of `kiao:RdfStatementElement` using three properties that mirror the properties used to reify the `rdf:Statement` itself (*i.e.*, `rdf:subject`, `rdf:property`, and `rdf:object`): `kiao:subjectElement`, `kiao:propertyElement`, and `kiao:objectElement`, each of which is `rdfs:subPropertyOf ro:has_part`; that is, a reified statement has these subject, property, and object elements as parts. While relations are typically

named as verbs or verb phrases, we modeled these relation names to be analogous to the core RDF statement model. Two reified instances of `kiao:RdfStatementElement`, `se1` and `se2`, can be seen in Figure 3. The corresponding `iao:denotes` assertions from these statement elements to their denoted concepts (*i.e.,* tokens `t1` and `t2`, respectively) are also depicted.

Summary of obstacles

Our use case is perhaps different from other OAC demo contexts, in that we did not have a clear scholar user group to target, but rather our own automated tools. We were also more concerned with aligning the representational requirements of the results of our natural language processing pipelines with the OA model, rather than building tools per se. Therefore we focused our effort on the proposals to extend OA to handle structured bodies and provenance. Those proposals have been fleshed out in a paper which is currently under review for publication.

While we hoped to produce a fully OA-compliant release of the Colorado Richly Annotated Full Text (CRAFT) corpus we did not finally achieve this goal, primarily because we got side-tracked performing an in-depth analysis of the ISO standard Linguistic Annotation Framework (LAF; ISO 24612:2012), as the only existing formal standard for linguistic annotation, and how it aligns to the OA proposals and the requirements we have identified above. We found that LAF does not immediately accommodate the requirements, and is generally not adequate to cleanly capture the semantics of our annotations over text. We presented this analysis in the Lingusitic Annotation Workshop at the Association for Computational Linguistics annual meeting [2]; this represents a notable product of our project.

We also have a significant portion of a UIMA-to-OA conversion tool completed, but not yet ready for public release. One of the main challenges we have faced in building this tool is in tracking the ongoing discussion of the open annotation model, and trying to determine the implications of changes to the proposed model for our implementation. While we could, per the experiment instructions, freeze the OA model we were targeting, we did not want our tool to be obsolete before it was even released. However we did not anticipate some of the fundamental changes in the OA proposals that would appear over the course of the project. Even now we feel that we have not had adequate bandwidth to really absorb the discussion on the OA mailing list, let alone to participate actively or fully appreciate the implications of decisions, given that this is not a full-time activity for us and there are many details of the OA model discussed there that are not directly relevant to our use case.

We reuse or extend existing community-curated ontologies where possible, and we have developed our proposal in terms of the Information Artifact Ontology (IAO), which is a member of the Open Biomedical Ontologies library of ontologies [3] (though not all of the concepts of these ontologies are specific to the biomedical realm). The IAO focuses on the representation of types of *information content entities*, which are defined to stand “in relation of aboutness” to other entities; that is, an information content entity is in some way “about” some other concept(s). For example, within the biomedical domain, data, images, and text are all in some way about sets of biomedical concepts. The IAO provides a hierarchy of types of information content entities as well as types of aboutness, including *denotation*, in which the information content entity specifically refers to some other concept (*e.g.,* the word “apple” denotes either a specific apple or the more general concept of an apple). We hold that an annotation is a type of information content entity, as it is in some way about the entity it is annotating. We are engaged in the ongoing process of submitting our model to the IAO for inclusion and feel that it is an appropriate effort to relate the OA model to.

Technical Lessons Learned

We can make some observations about the formal semantics of the Annotation class in relation to our `kiao` classes, which provide some insight into the semantics of the OA model. The class `oa:Annotation` is more specific than `kiao:Annotation` and more general than `kiao:RdfResourceAnnotation` and `kiao:RdfGraphAnnotation`. Thus the following relations hold:

```
oa:Annotation rdfs:subClassOf kiao:Annotation.
```

If OA annotations are being converted to KIAO annotations, there are several ambiguities. Primarily the OA definitions place no cardinality constraints on `oa:hasBody` or `oax:hasSemanticTag` (relations used to map an annotation to its denoted knowledge representation); one annotation can contain multiple of assertions using each relation. In such cases, each assertion in the OA model should likely be converted into an independent annotation in the KIAO model. The `oax:hasSemanticTag` property is a more specific type of `iao:denotes`, and annotations using this property should be converted into instances of `kiao:RdfResourceAnnotations`.

```
oa:hasSemanticTag rdfs:subPropertyOf iao:denotes.
```

It is unknown if the object of `oa:hasBody` is an `rdfs:Resource` (which should translate an annotation instance of the class `kiao:RdfResourceAnnotation`) or an `rdfg:Graph` (which should translate to an annotation instance of the class `kiao:RdfGraphAnnotation`). We can assert that `oa:hasBody` is also a subproperty of `iao:denotes`:

```
oa:hasBody rdfs:subPropertyOf iao:denotes.
```

Thus, each `oa:hasBody` assertion translates to an `iao:denotes` assertion.

For the conversion of KIAO annotations to OA annotations, one may think that `iao:denotes` could straightforwardly be made subproperty of `oa:hasBody`:

```
iao:denotes rdfs:subPropertyOf oa:hasBody.
```

Thus, each `iao:denotes` assertion would translate to an `oa:hasBody` assertion. However, `iao:denotes` is defined to hold not only among annotations but more generally among information content entities, while `oa:hasBody` only pertains to annotations. If this translation was broadly accepted, it is possible that `iao:denotes` assertions pertaining to information content entities other than annotations would be erroneously converted to `oa:hasBody` assertions, which would necessarily pertain to annotations. The only generally correct translation is to convert all `iao:denotes` assertions *for annotations only* (*i.e.*, with annotations as the subjects of the assertions) to `oa:hasBody` assertions. Using the OA extension model, it would also be acceptable to convert the `iao:denotes` assertions from instances of `kiao:RdfResourceAnnotation` to assertions using the property `oax:hasSemanticTag`.

Generalizable Results and Conclusions

We have proposed a basic distinction between Annotations that are related to a single web resource, and those that relate to a named graph, where a more complex set of information appears in the body of the annotation. We have also proposed a representation of both high-level and fine-grained representation of Provenance for tracking compositional or modular construction of Annotations from existing Annotations. While our proposals result in representations that are quite “heavy” in that they track many details, there is no explicit requirement to use them where they are not useful. Furthermore, we anticipate that they will be primarily useful for

computational processing of Annotations, where the availability of specific processing details in the provenance representation will enable more sophisticated reasoning about the validity of particular inferences.
We are very interested in producing an OA-based representation of the syntactic structures in the CRAFT Treebank (syntactic parse trees for each sentence in the CRAFT corpus) but this required learning what other representational proposals existed, LAF and its XML-based implementation GrAF, and comparing them to the OA model. Our investigation of LAF has also led us to the Linguistic Linked Open Data community (http://linguistics.okfn.org/resources/llod/) [4] which is an active community of linguists interested in RDF-based representation of linguistic data; further investigation is required to really understand whether and how their efforts, particularly POWLA [5] and OliA [6], can be harnessed for our needs; an initial survey suggests they do not address provenance tracking. Neither LAF nor POWLA appear to make a primary distinction between an Annotation and the Body of an Annotation, meaning that meta-data about the Annotation cannot be represented separately from meta-data about the Body. We feel that this is an important distinguishing feature of the OAC model.

References

1. Carroll JJ, Bizer C, Hayes P, Stickler P: **Named graphs, provenance and trust**. In*: 2005*. ACM: 613-622.
2. Verspoor K, Livingston K: **Towards adaptation of linguistic annotations to scholarly annotation formalisms on the Semantic Web**. In: *Proceedings of the Sixth Linguistic Annotation Workshop (LAWVI)*. Association for Computational Linguistics 2012.
3. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ *et al*: **The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration**. *Nat Biotech* 2007, **25**(11):1251-1255.
4. Chiarcos C, Hellmann S, Nordhoff S: **Linking linguistic resources: Examples from the Open Linguistics Working Group**. In: *Linked Data in Linguistics Representing Language Data and Metadata.* Edited by Chiarcos C, Nordhoff S, Hellmann S. Heidelberg: Springer; 2012: 201-216.
5. Chiarcos C: **Interoperability of corpora and annotations**. In: *Linked Data in Linguistics Representing Language Data and Metadata.* Edited by Chiarcos C, Nordhoff S, Hellmann S. Heidelberg: Springer; 2012: 161-179.
6. Chiarcos C: **An ontology of linguistic annotations**. In: *LDV Forum*. 2012: 1-16.