Towards Adaptation of Linguistic Annotations to Scholarly Annotation Formalisms on the Semantic Web

Karin Verspoor[*†]
[*]National ICT Australia
Victoria Research Lab
Melbourne VIC 3010 Australia
karin.verspoor@nicta.com.au

Kevin Livingston[†]
[†]University of Colorado Denver
12801 E 17th Ave, MS 8303
Aurora, CO 80045 USA
kevin.livingston@ucdenver.edu

Abstract

This paper explores how and why the Linguistic Annotation Framework might be adapted for compatibility with recent more general proposals for the representation of annotations in the Semantic Web, referred to here as the Open Annotation models. We argue that the adapted model, in addition to being interoperable with other annotations and annotation tools, also resolves some representational limitations and semantic ambiguity of the original data model.

1 Introduction

Formal annotation of language data is an activity that dates back at least to the classic work of Kucera and Francis on the Brown Corpus (Kucera 1967). Many annotation representations have been developed; some proposals are specific to a given corpus, e.g., the Penn Treebank (Marcus et al. 1993)) or type of annotation, e.g., CONLL dependency parse representation[1]), while others aim towards standardization and interoperability, most recently the Linguistic Annotation Framework [2] (LAF) (ISO 2008). All such proposals, however, are closely tied to the requirements of linguistic annotation.

Annotation, however, is not an activity limited to language data but rather is a general scholarly activity used both by the humanist and the scientist. It is a method by which scholars organize existing knowledge and facilitate the creation and sharing of new knowledge. Museum artifacts are annotated with meta-data relating to artist or date of creation, or semantic descriptors for portions of the artifacts (e.g. an eye of a statue) (Hunter & Yu 2011). Medieval manuscripts or ancient maps are annotated with details resulting from careful study (Sanderson et al. in press). Beyond scholarship, annotation is becoming increasingly pervasive in the context of social media, such as Flickr tags on images or FaceBook comments on news articles. Recognition of the widespread importance of annotation has resulted in recent efforts to develop standard data models for annotation (Ciccarese et al. 2011; Hunter et al. 2011), specifically targeting Web formalisms in order to take advantage of increasing efforts to expose information on the Web, such as through Linked Data initiatives[3].

In this paper, we will explore the adoption of the more general scholarly annotation proposals for linguistic annotation, and specifically look at LAF in relation to those proposals. We will show that with a few adaptations, LAF could move into use within the Semantic Web context, and, importantly, achieve compatibility with data models under development in the broader scholarly annotation community.

This generalization of the model is particularly pertinent to collaborative annotation scenarios; exposing linguistic annotations in the *de facto* language of the Semantic Web, the W3C's Resource Description Framework (RDF), provides several advantages that we will outline below.

[1] http://conll.cemantix.org/2012/data.html
[2] http://www.cs.vassar.edu/~ide/papers/LAF.pdf
[3] http://linkeddata.org/

Proceedings of the 6th Linguistic Annotation Workshop, pages 75–84,
Jeju, Republic of Korea, 12-13 July 2012.

2 Characteristics of the Semantic Web

There are two converging cultures within the Semantic Web community (Ankolekar et al. 2008) – one of providing structured data, and one of promoting community sharing of data. Sharing is supported by four principles of linked data (Bizer et al. 2009):

1. Use URIs (Uniform Resource Identifiers) as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using standards.
4. Include links to other URIs, so that they can discover more things.

These principles are built on top of the basic technology of the Web, HTTP and URIs, and represent best practices for making structured data available on the Web. They are the foundation for any Semantic Web model.

RDF provides a generic graph-based data model for structuring and relating information, through simple assertions. The RDF model encodes data in the form of subject, predicate, object triples. The predicate specifies how the subject and object are related. The linked data principles mean that the subject and predicates of an RDF triple are typically dereferenceable URIs representing concepts or entities.

3 The relevance of the Semantic Web for linguistic annotation

There are several clear reasons to explore a linguistic annotation formalism that is compatible with general Semantic Web annotation efforts. Some are not unique to the Web, but there do exist some special opportunities in the Web context.

3.1 Interoperability

Interoperability refers to enabling different entities (agents, services) to exchange information. Interoperability is impeded by both the syntax and format of data representations, and also by the ability to accurately represent the semantics of one data source in another.

Data can be exchanged in an *ad hoc* manner, for instance by having an individual system understand the syntax and semantics of the information produced by a given source and translating or mapping that information to an internal representation. However, this leads to significant duplication of effort, with each system having to manage data import and conversion from a given source independently.

Data compatibility problems also exist when attempting to use multiple data sources simultaneously. If two independent sources refer to "annotation 1" do they mean the same annotation or different annotations? And if these annotations are different are the tools processing them equally aware of the distinction?

The Semantic Web overcomes syntax and format issues through the use of RDF. While agreeing on semantics will continue to be challenging, the use of unique and resolvable URIs goes a long way toward formalizing meaning, or at least agreeing on references. Additionally as the use of more formal subsets of RDF, such as OWL, grows, more precise definitions of concepts will also become available.

3.2 Information Sharing and Reuse

Interoperability in turn enables reuse of information. The results of any annotation effort are generally intended to be shared. Agreement on a standard representation of annotations, with a consistent semantics, facilitates integration.

With interoperability, tools can directly build on annotations made by others. For the natural language processing community, this has several potentially significant advantages. Individual research groups need not build an end-to-end processing pipeline, but can reuse existing annotations over a common resource. For domains where there are commonly used shared document sets, such as standard annotated corpora used for training or testing, or document repositories that are the primary target of a body of text-related work – e.g. the Medline repository of biomedical journal abstracts – annotations can be made available for incorporation into downstream processing, without the need for re-computation and to ensure consistency. Tokens, parts of speech, even syntactic structures and basic named entities, can all be computed once and made available as a starting point for subsequent processing.

Where there is considerable investment in linked data, such as the biomedical domain, it also opens the possibility of taking advantage of external resources in language processing algorithms: if a

document has been semantically annotated by a domain expert, or semantically connected to external information, those annotations can be used to enable more sophisticated analysis of that document. For instance, (Livingston et al. 2010) demonstrated that incorporating existing background knowledge about proteins when extracting biological activation events from biological texts allows some inherent ambiguities in recognizing those events to be resolved.

3.3 Web-scale collaboration and analysis

Targeting the semantic web provides new opportunities in terms of collecting, analyzing and summarizing data both within and across annotation sets on the web. The methods on the Semantic Web for creating and providing data are fundamentally "open-world" and allow for data to be added at any time.

The Web is the natural place for collaborative annotation activities, which is by necessity a distributed activity. Whether a collaborative annotation project is undertaken by a focused community of interest or by crowd sourcing, using semantic models that can represent and document contradiction or multiple competing views allows data to be collected and aggregated from multiple sources.

Collaboration is also about coordinating and cooperating with the consumers of annotation. The Semantic Web has defined ways in which data can be shared and distributed to others. This includes the preference for resolvable URIs, such that automated tools can seek out data and definitions as needed. Additionally data is being provided through access points, such as SPARQL end points. Vocabularies exist for documenting what is in a dataset, such as VoID (Alexander & Hausenblas 2009), and there is work underway to standardize data sharing within domains, for example health care and life science.[4]

The availability of Linked Open Data also enables unforeseen novel use of the data. This is evident in the large number of popular "mash-ups" connecting existing tools and data in new ways to provide additional value. Tools even exist for end-users to create mash-ups, such as Yahoo! Pipes[5].

3.4 Availability of tools

Adoption of Semantic Web standards for annotation makes available mature and sophisticated technologies for annotation storage (e.g. triple-stores) and to query, retrieve, and reason over the annotations (e.g. SPARQL).

Perhaps of particular interest to the computational linguistics community are tools under development to visualize and manipulate annotation information in the dynamic context of the web. For instance, the DOMEO tool (Ciccarese et al. in press) provides support for display of annotation over the text of biomedical journal publications in situ, by adopting strategies for managing dynamic HTML. The Utopia Documents tool (Attwood et al. 2010) is oriented towards annotation of PDF documents and provides visualization of annotations that dynamically link to web content. The Utopia tool has been recently updated to consume Annotation Ontology content[6].

Finally, enabling compatibility of linguistic annotation tools with Semantic Web standards opens up the possibility of making those tools useful to a much broader community of annotators.

4 RDF data models for annotation

Beyond fundamental Semantic Web compatibility, we believe that linguistic annotation formalisms can benefit from compatibility with the Web-based scholarly annotation models. We are aware of two such models, namely, the Annotation Ontology (Ciccarese et al. 2011) and the Open Annotation Collaboration (OAC) (Hunter et al. 2011) models. Each of these models incorporates elements from the earlier Annotea model (Kahan et al. 2002). These two groups have now joined together to bring their existing proposals together, through the Open Annotation W3C community group[7]. As a result, we will focus on their commonalities, and use the OAC model and terminology for the purposes of our discussion. We refer to the models collectively as the Open Annotation models.

4.1 High-level model for scholarly annotation

The basic high-level data model of the two primary Open Annotation models defines an *Annotation* as

[4] http://www.w3.org/blog/hcls/
[5] http://pipes.yahoo.com/pipes/
[6] http://www.scivee.tv/node/26720
[7] http://www.w3.org/community/openannotation/

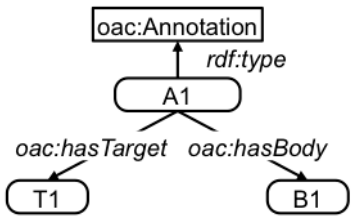

Figure 1: Base model for OAC[8].

an association created between two elements, a *Body* or content resource and (one or more) *Target* resources. The annotation provides some information about the target through the connection to the body. For instance, an annotation may relate the token "apple" in a text (the target of the annotation) to the concept of an apple, perhaps represented as WordNet (Fellbaum 1998a) synset "apple#1" (the body of the annotation).

Figure 1 shows the base model defined in the OAC model. The model, following linked data principles, assumes that each element of an annotation is a web-addressable entity that can be referenced with a URI.

Annotations can be augmented with meta-data, e.g. the author or creation time of the annotation. The model allows for each element of the annotation – the annotation itself, the target, and the body – to have different associated meta-data, such as different authors. Other features of the OAC model are that it can accommodate annotations over not only textual documents, but any media type including images or videos (for details, see the OAC model[8]). Text fragments are typically referred to using character positions.

4.2 Graph Annotations

The initial use cases for Open Annotation focused on single target-concept relationships, formalized as an expectation that the body of an Annotation be a single web resource. Recently, an extension that supports representation of collections of statements as the body of an annotation has been proposed (Livingston et al. 2011). In a revision of that extension (Livingston, personal communication), a *GraphAnnotation* is connected to a Body which is not a single web resource, but a set of RDF statements captured in a construct known as a *named graph* (Carroll et al. 2005). The named graph as a whole has a URI.

[8] http://www.openannotation.org/spec/beta/

This extension enables complex semantics to be associated with a resource, as well as supporting fine-grained tracking of the provenance of compositional annotations. These developments make possible the integration of linguistic annotation with the scholarly annotation models.

5 Adapting LAF to Open Annotation

The Linguistic Annotation Framework, or LAF, (ISO 2008) defines an abstract data model for annotations which consists of nodes and edges. Both nodes and edges can be elaborated with arbitrary feature structures, consisting of feature-value pairs. Nodes can link via edges to other nodes, or directly to regions in the primary data being annotated. An example of a LAF annotation is shown in Figure 2.

While LAF has made significant progress towards unified, unambiguous annotation representations, adopting some representation decisions of the Open Annotation models will not only facilitate interoperability with those models, but also resolve some ambiguities and limitations inherent to the LAF model.

5.1 High-level representation compatibility

At a high level, the LAF model aligns well with the Open Annotation RDF models. Fundamentally, the LAF model is based on directed graphs, as is RDF. The abstract data model in LAF consists of a referential structure for associating annotations with primary data, and a feature structure for the annotation content. These are similar to the Open Annotation notions of target and body.

Importantly, these models agree that the source material being annotated is separate from the annotations. In other words, stand-off annotation is assumed. In a web context, this is particularly significant as it is often not possible to directly manipulate the underlying resource. It also facilitates collaborations and distribution, as

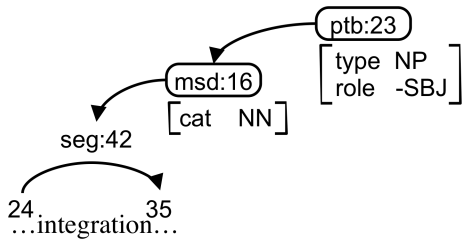

Figure 2: A sample LAF annotation, based on (Ide & Suderman 2007)

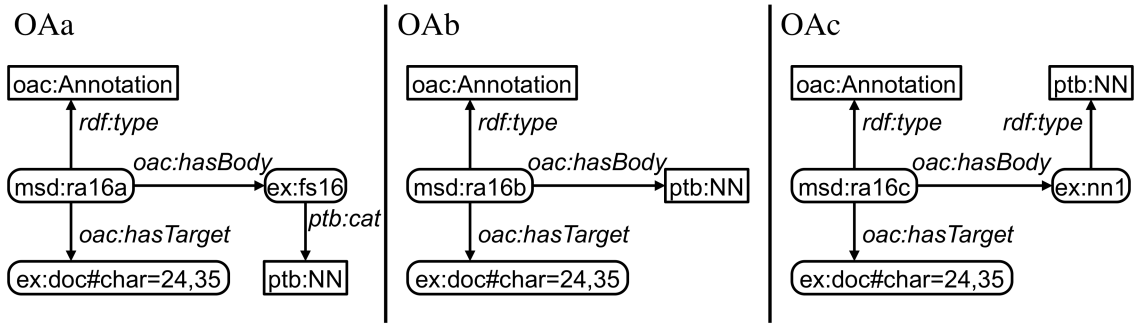

Figure 3: Options for an Open Annotation-compatible representation of the annotation msd:16 of Figure 2. Ovals represent instances, classes are boxed, and relations are italic labels on directed edges from subject to object.

annotations can be individually distributed and sets of annotations from different sources can be collected and used simultaneously.

5.2 Changes to LAF for Open Annotation

In order to facilitate integration of LAF with the Open Annotation models currently under development, a few changes would be required. A key difference is the separation in the Open Annotation models of three distinct elements: a target, a body, and the annotation itself, relating the previous two. These distinctions allow relations between any two elements to be made explicit and unambiguous, and further allow more detailed provenance tracking (Livingston et al. 2011).

5.2.1 Annotation content

In the LAF model, feature structures can be added to any node in the annotation graph. It has been shown that feature structures can be losslessly represented in RDF (Denecke 2002; Krieger & Schäfer 2010). In the XML serialization of LAF, GrAF (Ide & Suderman 2007), feature structures are represented within an annotation. An example of a LAF annotation from that paper is in Figure 2.

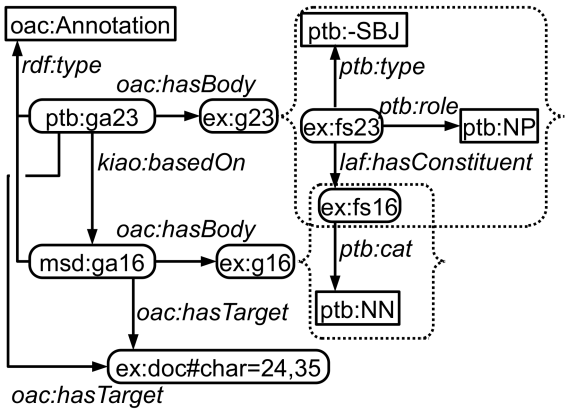

Figure 4: Open Annotation compatible representation of Figure 2 using GraphAnnotations. Graph contents are surrounded by dotted lines connected to their name.

In an Open Annotation model, the LAF feature structure corresponds to the body of the annotation. Figures 3 and 4 show several possibilities for representing the information in Figure 2 in a model compatible with the Open Annotation proposals. The most literal transformation for the part of speech annotation msd:16, Figure 3:OAa, utilizes an explicit feature structure representation in the body, consistent with automated feature structure transformations (Denecke 2002; Krieger & Schäfer 2010). Since RDF prefers URIs, concepts in the Open Annotation model are made explicit (pointing to an external definition for the Penn Treebank category of "NN", ptb:NN), in contrast to the LAF string representation of the feature and value. A named feature value pair is not necessarily needed and the concept could be annotated to directly, as is shown in Figure 3:OAb. This example, although much simpler, does lose the ability to refer to the specific instance. An instance could therefore be reified so that it could be referred to later, as is shown in Figure 3:OAc.

5.2.2 Named graphs

A GraphAnnotation explicitly separates the annotation from its content and provides a handle for the content as a whole, separate from the handle for the annotation, through reification of the content graph. The content of Figure 2 is represented as GraphAnnotations in Figure 4. The graph encapsulation clearly delineates which assertions are part of which annotation. For example, the hasConstituent relation from fs23 to fs16 in Figure 4 is part of the g23 graph, which is the body of the ga23 annotation, even though it shares concepts with the g16 graph.

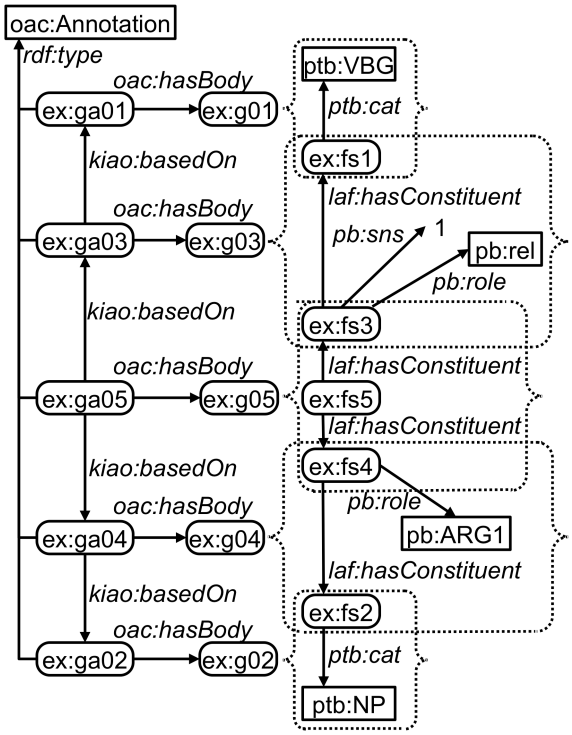

Figure 5: Literal RDF translation of a GrAF Propbank annotation representation from (Ide & Suderman 2007)

The separation of annotation and content also allows explicit provenance relations to be represented. For example, the relationship between the annotation for the NN part of speech (msd:ga16) and the annotation for the NP (ptb:ga23) as a kiao:basedOn relation (Livingston et al. 2011), indicating that the phrasal annotation is based on the part of speech annotation. This allows us to identify how analyses build on one another, and perform error attribution.

LAF annotations consist of feature structures, which have functional properties (restricted to only one object value per key), and a set of edges that connect nodes, which may have an unclear or ambiguous interpretation (see section 5.2.4). RDF-based graph annotations avoid these issues as they can directly contain any set of assertions in the annotation body that an annotator wishes to express. This includes capturing relations that are not functional, and information that might only be implicit in a LAF edge. This body representation is both more expressive and more explicit.

The greater expressivity and simpler structure of RDF based annotations can be clearly seen in contrasting Figure 5 with Figure 6. Both figures depict the same subset of information from a PropBank example in Section 3 of (Ide & Suderman 2007). Figure 5 represents a verbatim translation of the LAF following the feature structure in RDF conventions. In this figure, as in the original LAF figure, the proposition elements are distributed across 3 feature structures, for the relation (rel), arg1, and the proposition itself. In contrast, Figure 6 uses individual RDF triples in the annotation bodies; the representation is not only more succinct, it more naturally expresses the semantics of the information, with the relation and its argument within the same content graph. The pb:arg1 relation in Figure 6 alleviates the need for the entire ga04 annotation in Figure 5. Arguably it was an intentional choice by Ide and Suderman (2007) to use a LAF node/annotation instead of a LAF edge. However, this and other examples point to arbitrary selection of nodes and edges in LAF, with little surrounding semantics to ground them. While it is true that users must understand the semantics of any model to use it, the framework of RDF and the linked data best practices provide a structure for explicitly and formally defining the concepts and links, facilitating interoperability.

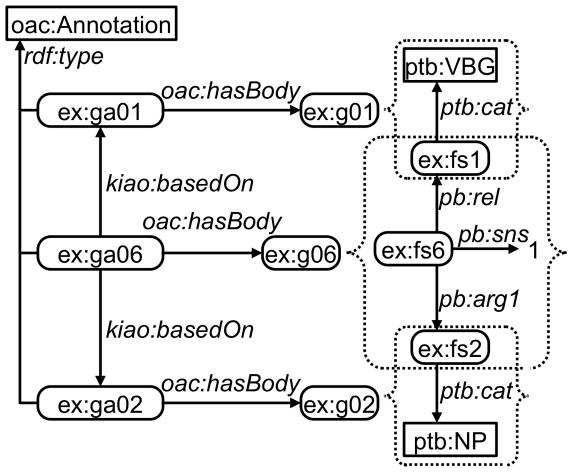

Figure 6: Streamlined representation of Figure 5, using a single feature structure for the core proposition (fs6).

5.2.3 Target objects

There are differences in how these models refer to specific region of a resource. LAF reifies structures to represent text spans but necessitates the use of a separate document enumerating (character-based) source text segmentation; subsequent annotations refer to those segments. The Open Annotation models have in common that they introduce a separate object (node in the graph) to point to the appropriate segment of the resource. OAC uses

fragment URIs or *ConstrainedTargets*. The Annotation Ontology uses a construct called a *Selector*. While the details vary slightly, these constructs are encoding essentially equivalent information and attaching it to a reified entity.

LAF further encourages only creating non-overlapping spans at the segmentation level. This appears to be due to properties of the particular XML-based segmentation language chosen by LAF influencing the model. This characteristic impedes representation of annotations over other linguistic modalities, such as speech streams, as noted by Cassidy (2010). An additional segmentation document is unnecessary in the Open Annotation approaches; the models do not restrict the organization of different aspects of the annotations across documents or web resources.

The use of separate reified entities as the target of annotations also allows locations to be specified in any number of ways. As discussed above, the models employ various strategies for this and therefore can flexibly accommodate different requirements for different media sources.

In Figure 4, we show a proposed treatment of targets in the case of embedded linguistic objects, i.e. linguistic constructs that build on other constructs. We suggest that the target of a higher-order constituent such as a noun phrase consists of the target(s) of its constituent parts. In our example, it is a single target that is shared between the part of speech annotation and the NP annotation. For a more complex set of constituents, such as the elements of a dependency relation, the targets may refer to a collection of non-contiguous spans of the source document. For example, the annotation ga06 in Figure 6 would have multiple targets (not shown), one for each constituent piece.

5.2.4 Graph Edges

Edges between nodes in LAF do not always have a clear interpretation. Edges are often left untyped; in this case an unordered constituency relationship is assumed. For transparency, an edge type that specifically defines the semantics of the relationship would be preferable to avoid any potential ambiguity.

Furthermore, the LAF model allows feature structures to be added to edges, as well as nodes. We agree with Cassidy (Cassidy 2010) that the intended use of this is likely to produce typed edges, and not to produce unique instance data for each edge. However, this is another source of ambiguity in the LAF representation. For example, annotations are sometimes directly connected to edges in the segmentation document (Ide & Romary 2006).

In the LAF model, the body and the annotation itself can at times appear conflated. When an edge connects two nodes it is unclear if that edge contains information that relates to the body of the annotation or metadata about the annotation itself. In LAF it sometimes appears to be both. There is a single link in the LAF representation in Figure 2 from ptb:23 to msd:16. This link simultaneously encodes information about the target of the annotation, the representation of the body of the annotation, and the provenance of the annotation. The Open Annotation models provide for more explicit and detailed representations. This single ambiguous arc in LAF can be represented accurately as three triples. In Figure 4, these are the hasTarget link from ptb:ga23, the hasConstituent link relating parts of the annotation body, and the basedOn link recording provenance.

5.3 Web Linguistic Category representation

A challenge that must be addressed in moving LAF to the Web context is the need for resolvable and meaningful URIs as names for resources, per the Linked Data principles. LAF intentionally avoids defining or requiring the use of standard or semantically typed identifiers in its feature structures. However, to enable true interoperability as an exchange formalism, semantic standardization is important.

While there are many standard names and tagsets that are used in the NLP community, for instance the Penn Treebank tags (Marcus et al. 1993), and there are recent efforts to formally specify and standardize linguistic categories (e.g. ISOcat (Kemps-Snijders et al. 2008)) the use of URIs to capture such names is not widespread. Recent efforts (Windhouwer & Wright 2012) show the use of the ISOcat data category registry terms as URIs, e.g. the category of *verb* is represented as `http://www.isocat.org/datcat/DC-1424`. The OLiA reference model explicitly tackles mapping among existing terminology resources for linguistic annotation (Chiarcos 2010), e.g. ISOcat and GOLD (Farrar & Langendoen 2003). A specific example of mapping part of speech tags from an existing category system can be found in

(Schuurman & Windhouwer 2011). Such mappings will be necessary for any tag set used by annotations on the Semantic Web; while the work is not complete there is clear movement towards Linked Data compatibility for linguistic data.

Recent efforts to standardize of lexical representation in RDF, e.g. the W3C Ontology-Lexica Community Group [9] and the Working Group on Open Data in Linguistics[10], also will contribute to improved reuse and systematicity of annotations, and may in fact greatly simplify annotations at the lexical level. The *lemon* model (Buitelaar et al. 2011), for instance, provides for an ontology-based (RDF) representation of lexical information. Such lexical entries could be used directly as the content of an annotation, associating a word with its word form information, including all of the elements currently captured in, e.g., a LAF feature structure for a token.

5.4 DADA: LAF in RDF

The DADA annotation store (Cassidy 2010) provides an adaptation of LAF to RDF. We review it here for completeness; it is the only other work we are aware of that addresses the representation of LAF in RDF. However, this implementation does not conform entirely to the structure of the current scholarly annotation proposals.

Although the DADA model explicitly reifies anchors in a document, each anchor refers to only a single location in the document. A span of text that is the target of an annotation is captured by two or more such anchors and the span as a whole is not explicitly reified. Additional properties must be used to associate that structure with the annotation, in essence conflating the annotation with its target.

In some uses, the annotation in DADA appears conflated with its body. For instance, in Figure 3 of (Cassidy 2010) a type-specific relation (*biber*) is used to connect the annotation (*s1*) to the body, making it necessary to understand the annotation's content before that content can be located. That is, a system cannot know generically which relation to follow to access annotation content. Additionally, the model treats relations that could best be interpreted as existing between annotation content (e.g. a temporal relationship between two events) as a direct relationship between two annotations, instead of between their denoted content (the events). The proposed DADA representation of LAF is similar to the OAa subfigure of Figure 3. It therefore suffers from the same limitations with respect to attribution and provenance as the original LAF model.

6 Conclusions and Future Work

In this paper, we have examined linguistic annotation efforts from the perspective of the Semantic Web. We have identified several reasons to bring linguistic annotation practices in line with more general web-based standards for scholarly annotation, and specifically examined what would be required to make Linguistic Annotation Framework representations compatible with the Open Annotation model.

While the required changes are not trivial due to some variation in how LAF has been applied, they will result in several key benefits: (1) explicit, semantically typed concepts and relations for the content of annotations; (2) the opportunity for more expressivity in the content of annotations; (3) a representation which formally separates the construct of an annotation itself from both the content and the document targets of the annotation, enabling significantly richer source attribution and tracking; and (4) increased clarity and specificity – and hence, reusability – of the annotations produced based on the model.

In future work, we will refine our proposals for the representation of linguistic annotations in an Open Annotation-compatible model through discussion with the broader linguistic annotation community. We plan to release a version of the CRAFT Treebank (Verspoor et al. in press) in Open Annotation RDF based on those proposals.

Acknowledgments
We would like to thank Mike Bada for critiquing our proposals. We also thank Steve Cassidy for helpful discussions about DADA and LAF, and for providing examples to explore.

This work was supported by Award No. 2011-02048-05 from the Andrew W. Mellon Foundation to KV. Additional support came from NIH grant 3T15 LM00945103S1 to KL. KV receives funding through NICTA, which is supported by the Australian Government as represented by the

[9] http://www.w3.org/community/ontolex/
[10] http://wiki.okfn.org/Working_Groups/Linguistics

Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

Alexander, K. & M. Hausenblas. 2009. Describing linked datasets-on the design and usage of void, the vocabulary of interlinked datasets, 2009.

Ankolekar, A., M. Krötzsch, T. Tran & D. Vrandecic. 2008. The two cultures: Mashing up Web 2.0 and the Semantic Web. Web Semantics: Science, Services and Agents on the World Wide Web 6.70-75.

Attwood, T. K., D. B. Kell, P. McDermott, J. Marsh, S. R. Pettifer & D. Thorne. 2010. Utopia documents: linking scholarly literature with research data. Bioinformatics 26.i568-i74.

Bizer, Christian, Tom Heath & Tim Berners-Lee. 2009. Linked Data—The Story So Far. International Journal on Semantic Web and Information Systems 5.1-22.

Buitelaar, Paul, Philipp Cimiano, John McCrae, Elena Montiel-Ponsoda & Thierry Declerck. 2011. Ontology Lexicalisation: The lemon Perspective. Paper presented at the 9th International Conference on Terminology and Artificial Intelligence, Paris.

Carroll, J.J., C. Bizer, P. Hayes & P. Stickler. 2005. Named graphs, provenance and trust, 2005.

Cassidy, Steve. 2010. Realisation of LAF in the DADA Annotation Server. Paper presented at the Fifth Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation (ISA-5), Hong Kong.

Chiarcos, Christian. 2010. Grounding an Ontology of Linguistic Annotations in the Data Category Registry. Paper presented at the Language Resource and Language Technology Standards workshop at LREC 2010, Malta.

Ciccarese, Paolo, Marco Ocana & Tim Clark. in press. Domeo: a web-based tool for semantic annotation of online documents. J Biomed Semantics.

Ciccarese, Paolo, Marco Ocana, Leyla Garcia Castro, Sudeshna Das & Tim Clark. 2011. An open annotation ontology for science on web 3.0. Journal of Biomedical Semantics 2.S4.

Denecke, Matthias. 2002. Signatures, Typed Feature Structures and RDFS. Paper presented at the Language, Resources and Evaluation Conference.

Farrar, Scott & Terry Langendoen. 2003. A linguistic ontology for the semantic web. GLOT International 7.1-4.

Fellbaum, C. 1998a. WordNet: An Electronic Lexical Database (Language, Speech, and Communication) Cambridge, Massachusetts: The MIT Press.

Hunter, J., T. Cole, R. Sanderson & H. Van de Sompel. 2011. The open annotation collaboration: A data model to support sharing and interoperability of scholarly annotations, 2011.

Hunter, J. & C. Yu. 2011. Assessing the Value of Semantic Annotation Services for 3D Museum Artefacts. Paper presented at the Sustainable Data from Digital Research Conference, Melbourne.

Ide, Nancy & Laurent Romary. 2006. Representing Linguistic Corpora and Their Annotations. Paper presented at the Proceedings of the Fifth Language Resources and Evaluation Conference.

Ide, Nancy & Keith Suderman. 2007. GrAF: A Graph-based Format for Linguistic Annotations. Paper presented at the Linguistic Annotation Workshop at ACL 2007, Prague.

ISO. 2008. ISO TC37 SC4 WG1. In *Language resource management -- Linguistic Annotation Framework*.

Kahan, J., M.R. Koivunen, E. Prud'Hommeaux & R.R. Swick. 2002. Annotea: An Open RDF Infrastructure for Shared Web Annotations. Computer Networks 39.589-608.

Kemps-Snijders, Marc, Menzo Windhouwer, Peter Wittenburg & Sue Ellen Wright. 2008. Corralling Data Categories in the Wild. Paper presented at the Sixth International Conference on Language Resources and Evaluation (LREC'08).

Krieger, HU & U Schäfer. 2010. DL Meet FL: A Bidirectional Mapping between Ontologies and Linguistic Knowledge. Paper presented at the 23rd International Conference on Computational Linguistics.

Kucera, H., and W. N. Francis. 1967. Computational analysis of present-day American English: Brown University Press.

Livingston, Kevin, Michael Bada, Lawrence Hunter & Karin M Verspoor. 2011. An Ontology of Annotation Content Structure and Provenance. Paper presented at the Proc Intelligent Systems in Molecular Biology: Bio-ontologies SIG.

Livingston, Kevin, Helen L. Johnson, Karin Verspoor & Lawrence E. Hunter. 2010. Leveraging Gene Ontology Annotations to Improve a Memory-Based Language Understanding System. Paper presented at the Fourth IEEE International

Conference on Semantic Computing (IEEE ICSC2010).

Marcus, Mitchell, Beatrice Santorini & Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics 19.313-30.

Sanderson, R., B. Albritton, R. Schwemmer & H. Van de Sompel. in press. Shared Canvas: A Collaborative Model for Medieval Manuscript Layout. International Journal of Digital Libraries.

Schuurman, Ineke & Menzo Windhouwer. 2011. Explicit Semantics for Enriched Documents. What Do ISOcat, RELcat and SCHEMAcat Have To Offer? Paper presented at the 2nd Supporting Digital Humanities conference (SDH 2011), Copenhagen.

Verspoor, Karin, K. Bretonnel Cohen, Arrick Lanfranchi, Colin Warner, Helen L. Johnson, Christophe Roeder, Jinho D. Choi, Christopher Funk, Yuriy Malenkiy, Miriam Eckert, Nianwen Xue, Jr. William A. Baumgartner, Michael Bada, Martha Palmer & Lawrence E. Hunter. in press. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. BMC Bioinformatics.

Windhouwer, Menzo & Sue Ellen Wright. 2012. Linking to Linguistic Data Categories in ISOcat. Linked Data in Linguistics, ed. by C. Chiarcos, S. Nordhoff & S. Hellmann, 99-107: Springer Berlin Heidelberg.