NICTA

computational bioscience program
university of colorado school of medicine

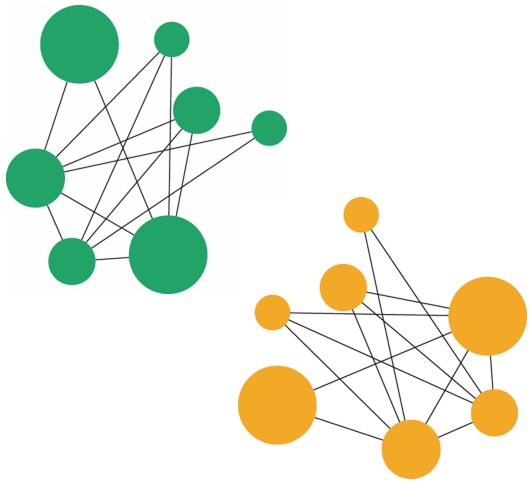
Towards Adaptation of Linguistic Annotations to Scholarly Annotation Formalisms on the Semantic Web

Karin Verspoor, Ph.D.
Senior Researcher, Health and Life Sciences Business Area
National ICT Australia

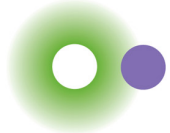
-and-

Kevin Livingston, Ph.D.
Computational Bioscience Program
School of Medicine
University of Colorado Denver

Karin.Verspoor@nicta.com.au
http://textminingscience.com



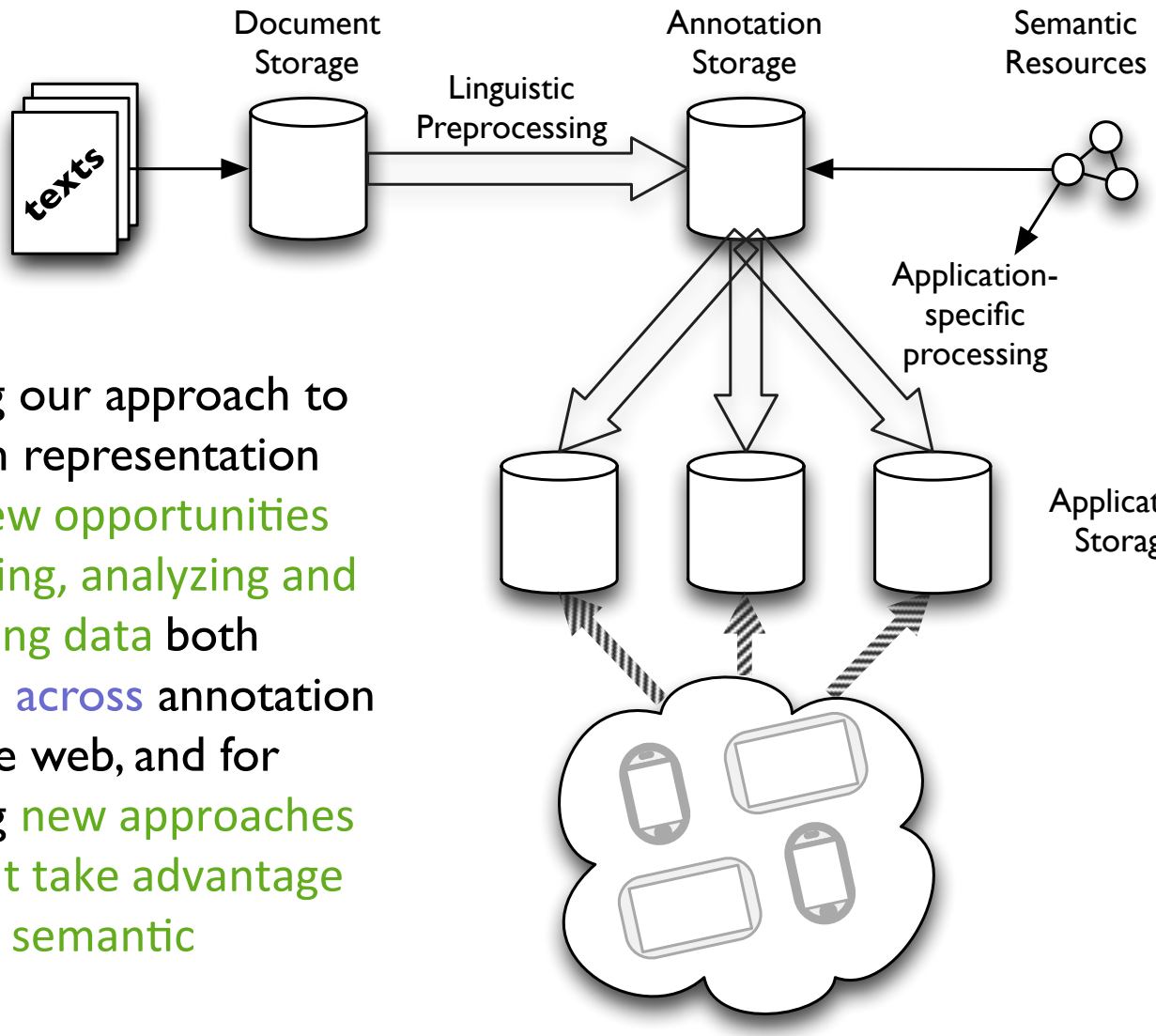
Linguistic Annotations: Who are the consumers?



- Creation of linguistically annotated resources is a hugely expensive enterprise (cf. LDC, ELRA, NIST)
- The community has typically viewed such resources primarily as training data for specific tasks
  - so there is an abundance of tools that can *produce* annotations of the specific types represented in those resources, and in the specific format of those resources
  - we are really good at evaluating tools on their ability to *reproduce* manual annotations over the same data
- But what about re-using those tools and annotations in other contexts and for other purposes?

Reuse

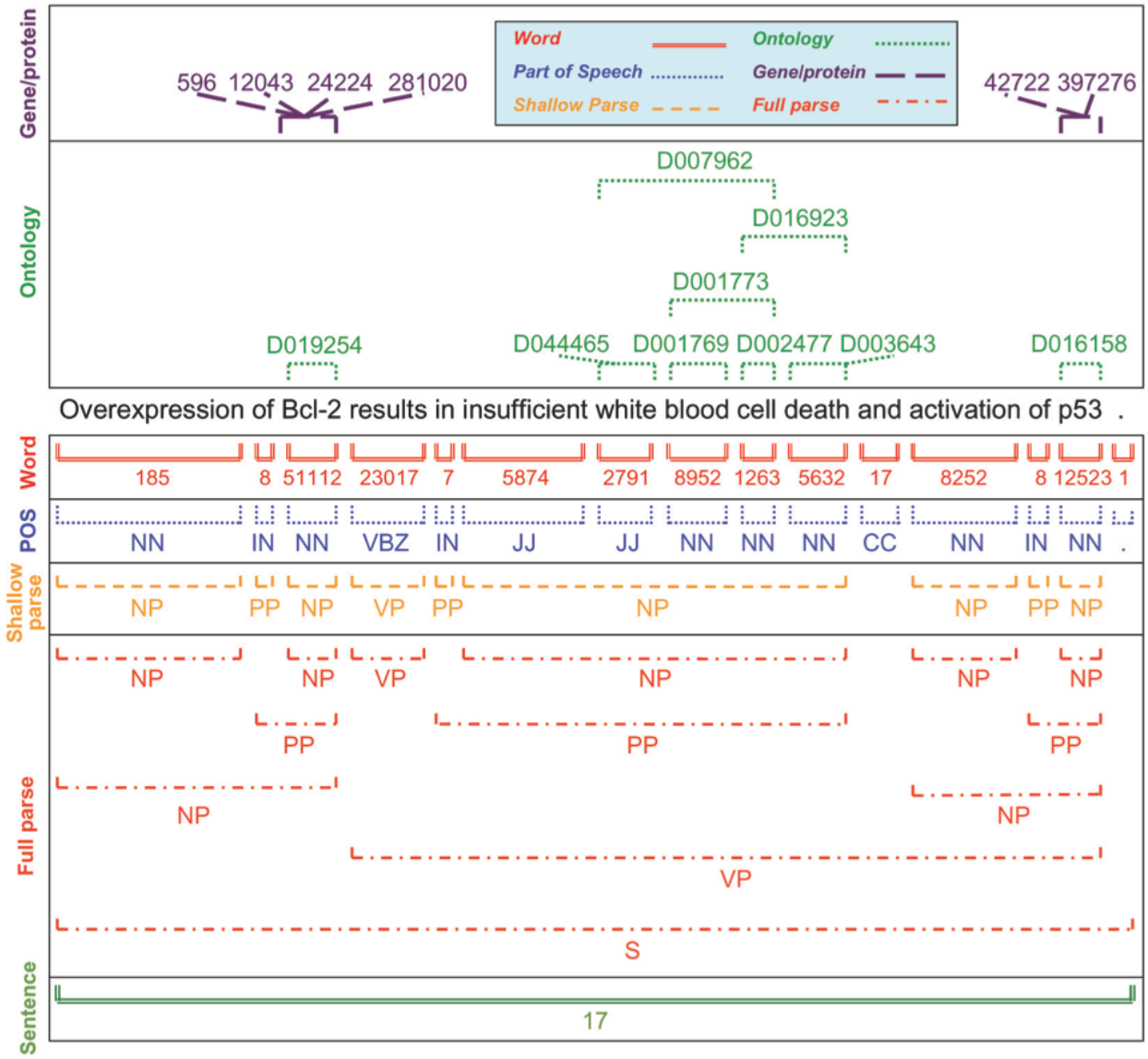
NICTA



Rethinking our approach to annotation representation enables new opportunities for collecting, analyzing and summarizing data both within and across annotation sets on the web, and for developing new approaches to NLP that take advantage of existing semantic resources.

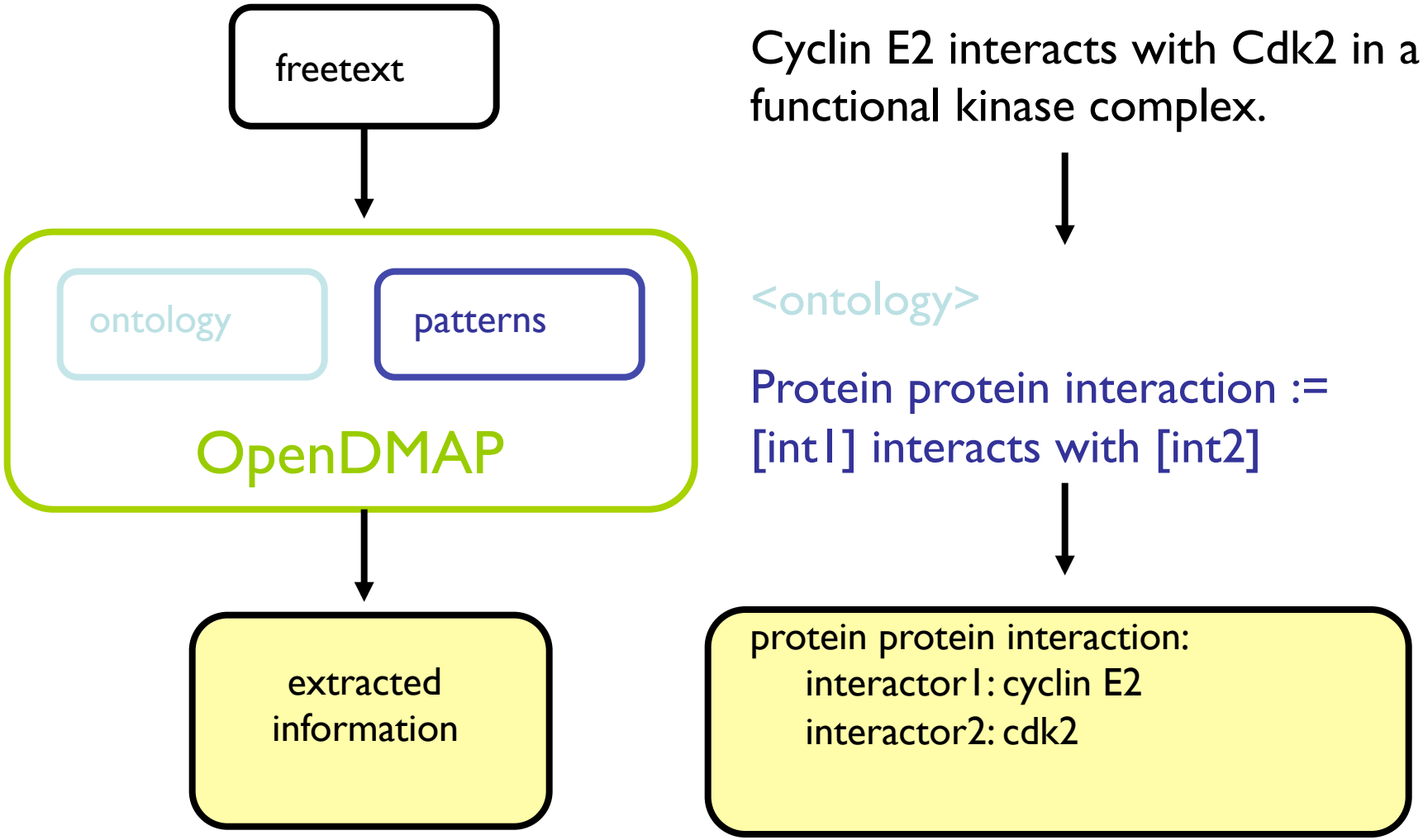
Interoperability

NICTA



Information Extraction via OpenDMAP

NICTA



Information Extraction patterns can cross linguistic levels


```
//
// SAMPLE PATTERNS FOR DETECTING PROTEIN BINDING EVENTS
//
{binding_trig_word} := r'bind.*', r'bound.*', r'crosslink.*', r'cross-link.*';
{binding_trig_word} := r'interact.*', r'ligat.*', r'co-ligat.*', r'coligat.*';

// SINGLE PROT BINDING
{binding} := [event_action binding_trig_word] {prep} {det}? [Theme];

// MULTI PROT BINDING
{binding} := [Theme] [Site]? [Theme] [Site]? [event_action binding_trig_word];
```

Linguistic Annotation formalisms

NICTA

- Many ad hoc representation formalisms have been developed for linguistic annotations
  - While some may be *de facto* standards, they are usually specific to one particular kind of linguistic construct

    (e.g. Penn Treebank syntax trees)
  - These formalisms do not enable interoperability with other representations
- These formalisms are used to share manually produced annotations
- They are also a target representation for automated annotation systems

Towards interoperability of Linguistic Annotations

- More recently, there have been efforts to address generalizability and interoperability in linguistic annotation formalisms
- LAF, the Linguistic Annotation Framework, is the leading solution
- GrAF is the XML serialization of LAF

The Semantic Web

- Web of data
  - with semantic metadata
- Data made available in a standard formalism (RDF)
  - Using URIs to identify resources
  - Dereferencing those URIs should lead to something useful
  - the use of unique and resolvable URIs helps to formalize meaning, or at least to improve consistency of references
- To establish links among data sets
  - Using ontologies as a semantic backbone
- Promoting community sharing
- Supported by a rich ecosystem of infrastructure tools

Open Annotation model (W3C Community Group draft)

NICTA

- Aim: a common, RDF-based, specification for annotating digital resources
- Plus tools supporting annotation of digital content
- Interoperable annotation environments

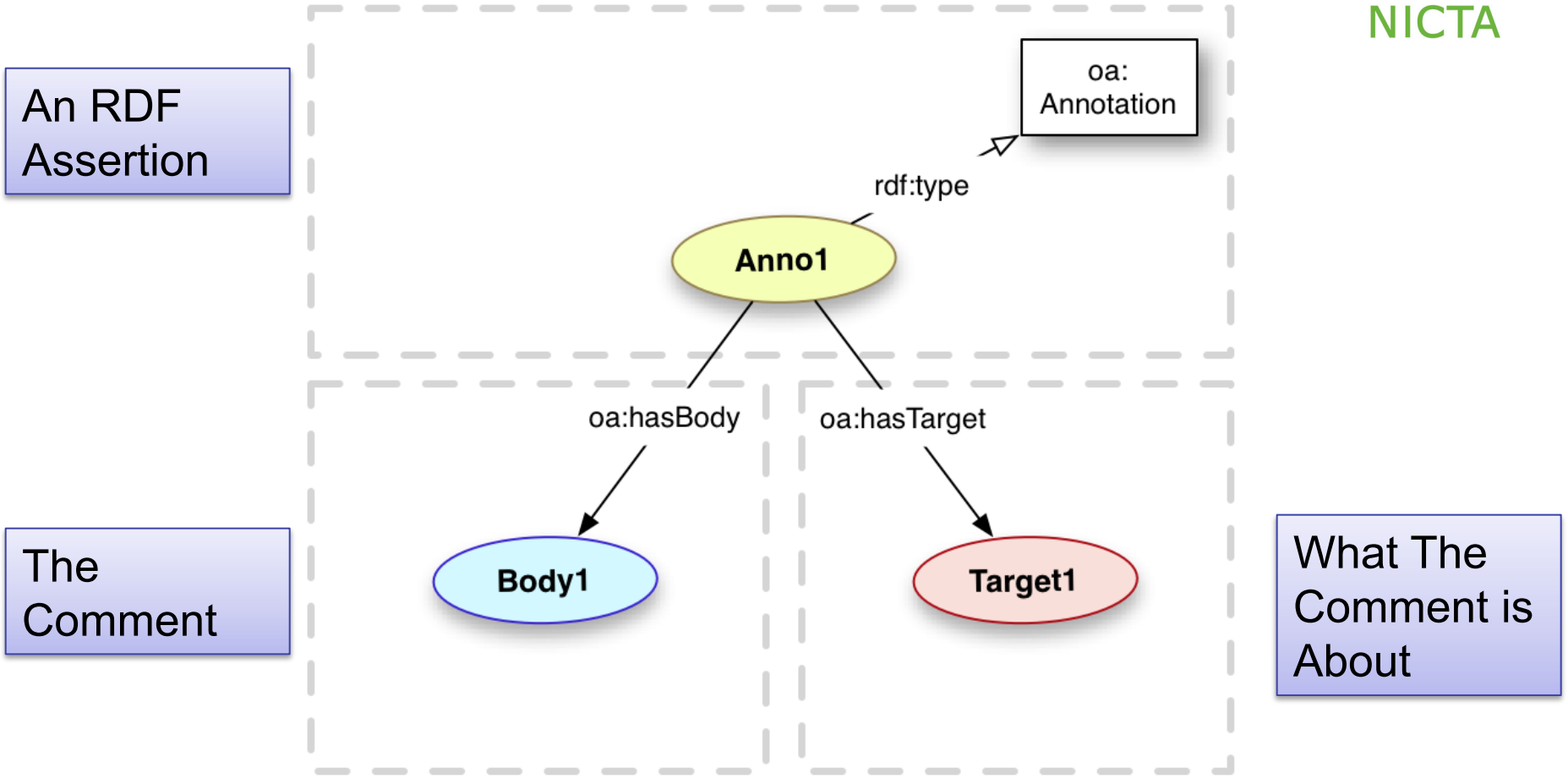
http://www.w3.org/community/openannotation/

Open Annotation model use cases

- Many scholarly annotation applications
- Meta-data about web resources
  - Tags (think Flickr)
  - Comments (think Facebook)
- Connecting web resources
  - An article discussing (part of) an image/video/map

Open Annotation Basic Model

NICTA



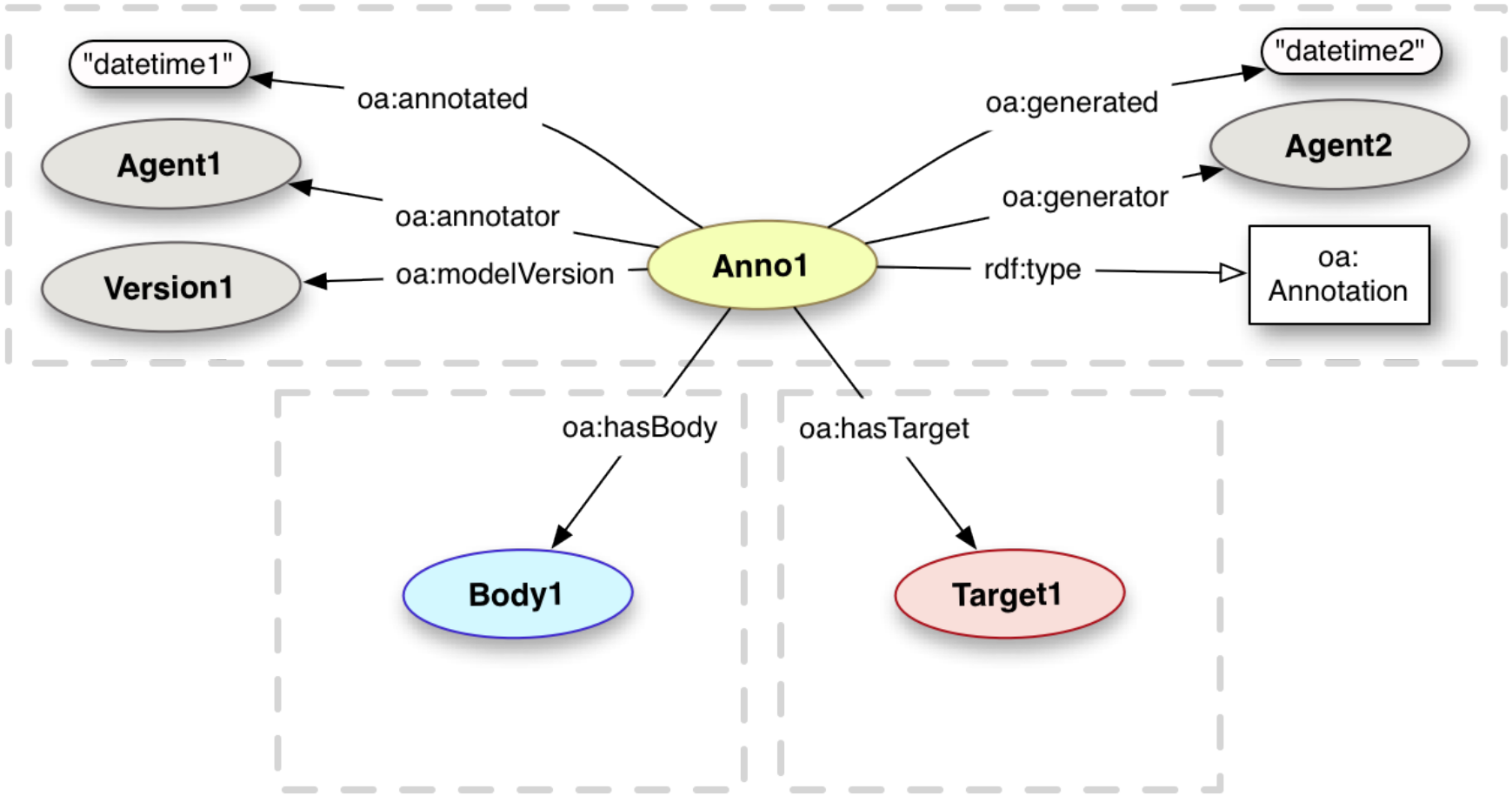
http://www.openannotation.org/spec/core/

Slide courtesy of Robert Sanderson, Los Alamos National Laboratory

Annotation Metadata

NICTA

Additional information can be associated with the Annotation

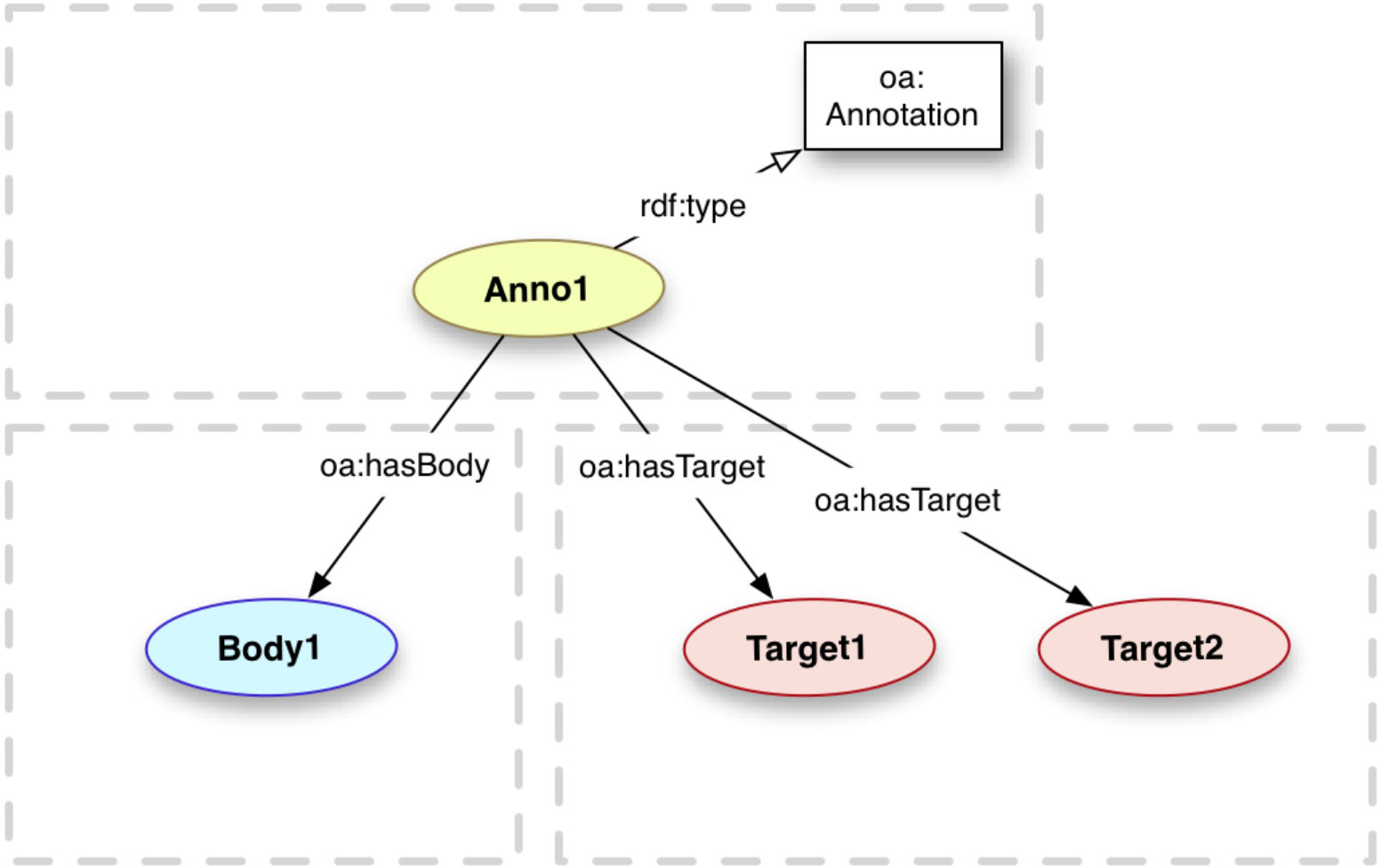


Slide courtesy of Robert Sanderson, Los Alamos National Laboratory

Multiple Targets

NICTA

There can be more than one Target, e.g. for compare/contrast, or discontinuous constituents.



Slide courtesy of Robert Sanderson, Los Alamos National Laboratory

Open Annotation
key characteristics

NICTA

- Annotations, including meta-data about that annotation, are kept separate from both the content of the annotation (body) and the resource being annotated (target)
  - This allows meta-data about the annotation itself (e.g. the system which captured the annotation and the time the relationship was created) to be distinguished from meta-data about the target (e.g. the author of the text) and about the body (e.g. the author of a comment about the text)
- Use RDF principles to refer directly to web artifacts and their segments

Extensions for “structured bodies”

NICTA

- Typical Open Annotation use case is a direct relationship between two web resources
- For linguistic annotations, we need to capture complex content that is not necessarily best represented via a single URI
- We create a GraphAnnotation that denotes a full RDF named graph, which captures a subgraph
- We also add kiao:basedOn to enable Annotation provenance tracking

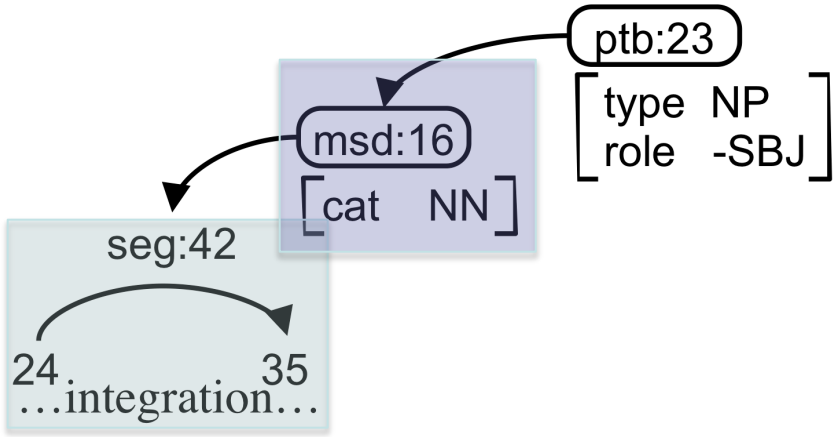
```
kiao:GraphAnnotation rdfs:subClassOf oa:Annotation
kiao:GraphAnnotation kiao:denotesGraph rdfg:Graph
```

Aligning LAF to Open Annotation

NICTA

- High-level data model compatibility
  - Graph-based representation
  - Stand-off annotation
- LAF
  - Links can exist between any two nodes; no formal distinction between text segment and content
  - Edges are often implicitly typed
  - Requires a separate segmentation document from the annotations themselves

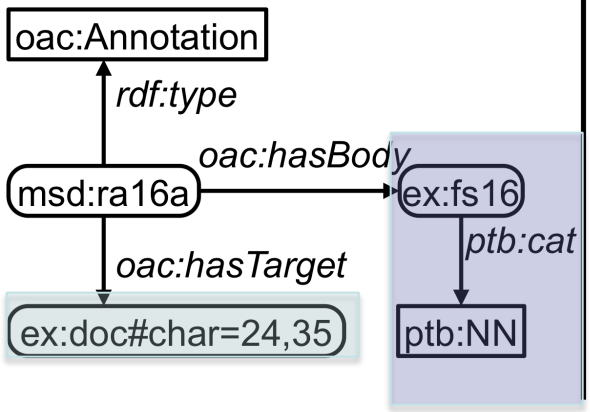
A LAF snippet based on (Ide and Suderman 2007)



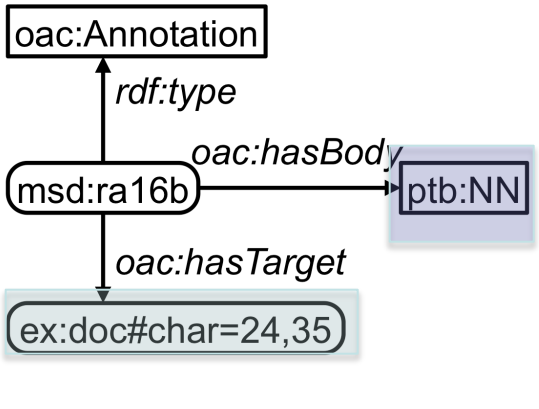
Target in text

Body of annotation

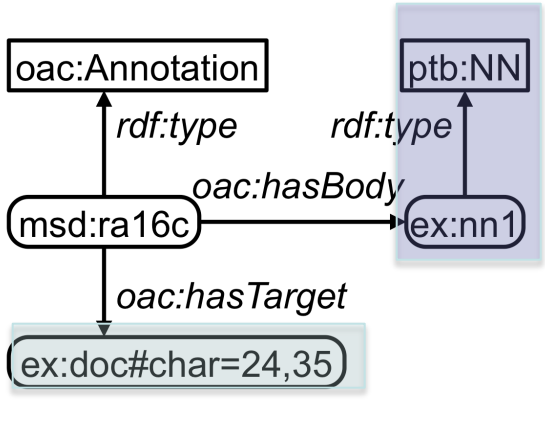
OAa



OAb

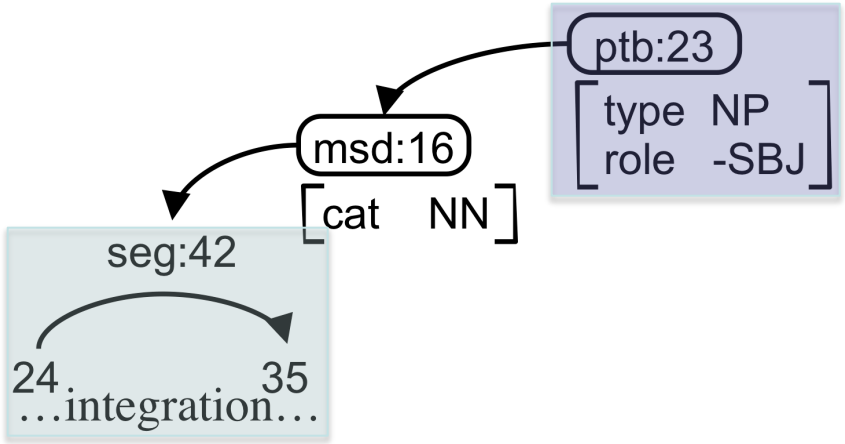


OAc



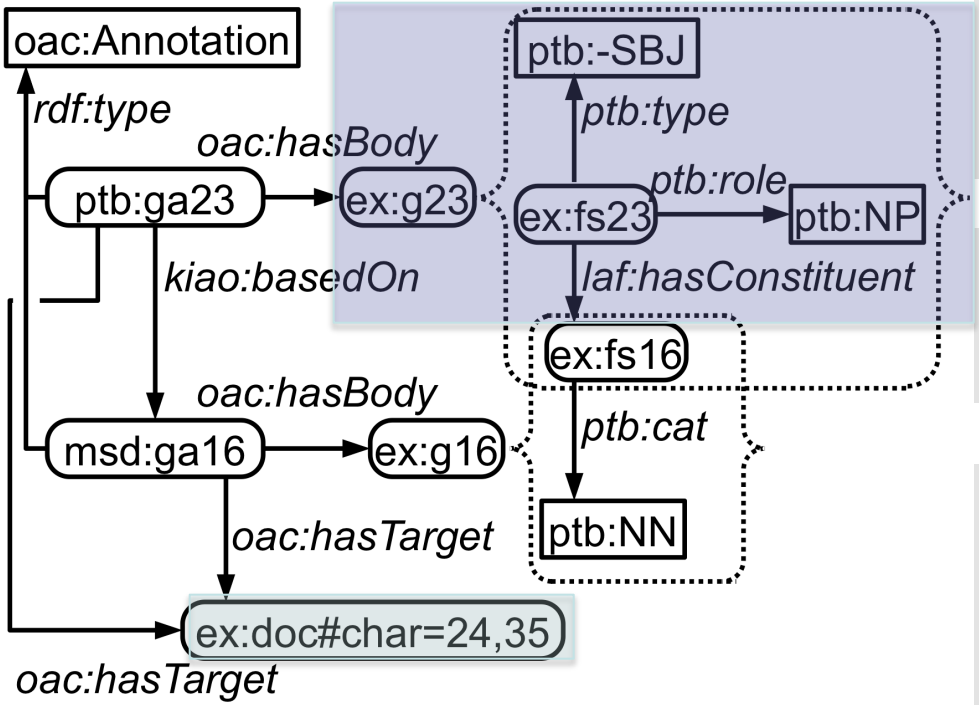
Target in text

Body of annotation



NICTA

Annotation and Body correspond to separate constructs; metadata associated with one vs. the other will have different interpretation

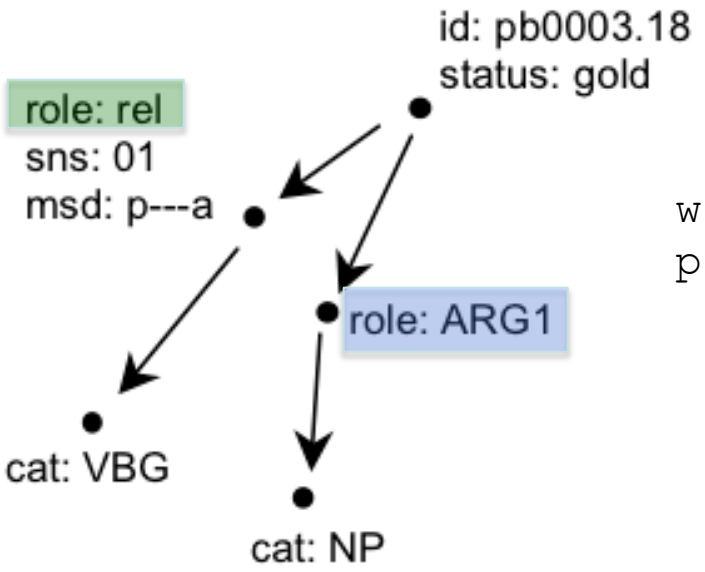


Standardize semantics of feature structure elements in terms of externally defined categories (ptb:-SBJ) and relations (ptb:type)

Make relationship between elements NP and NN explicit (laf:hasConstituent)

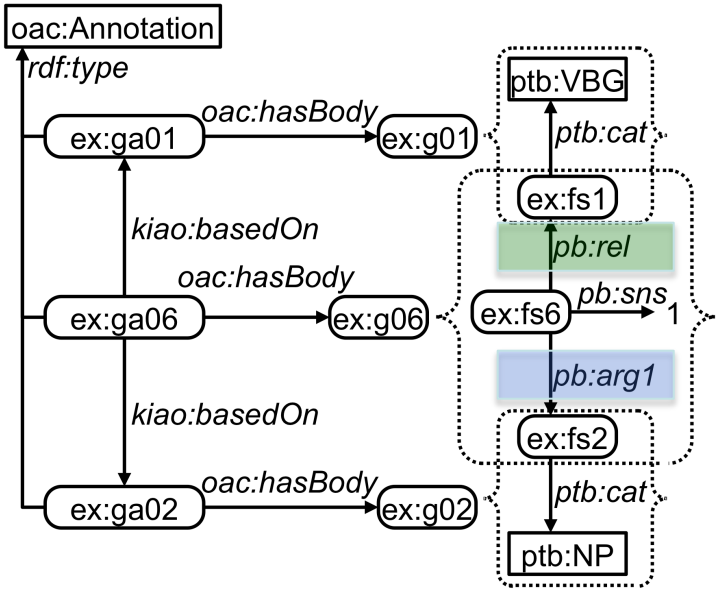
The elements of the body (ex:fs23 and ex:fs16) are connected; the annotations themselves are as well (kiao:basedOn)

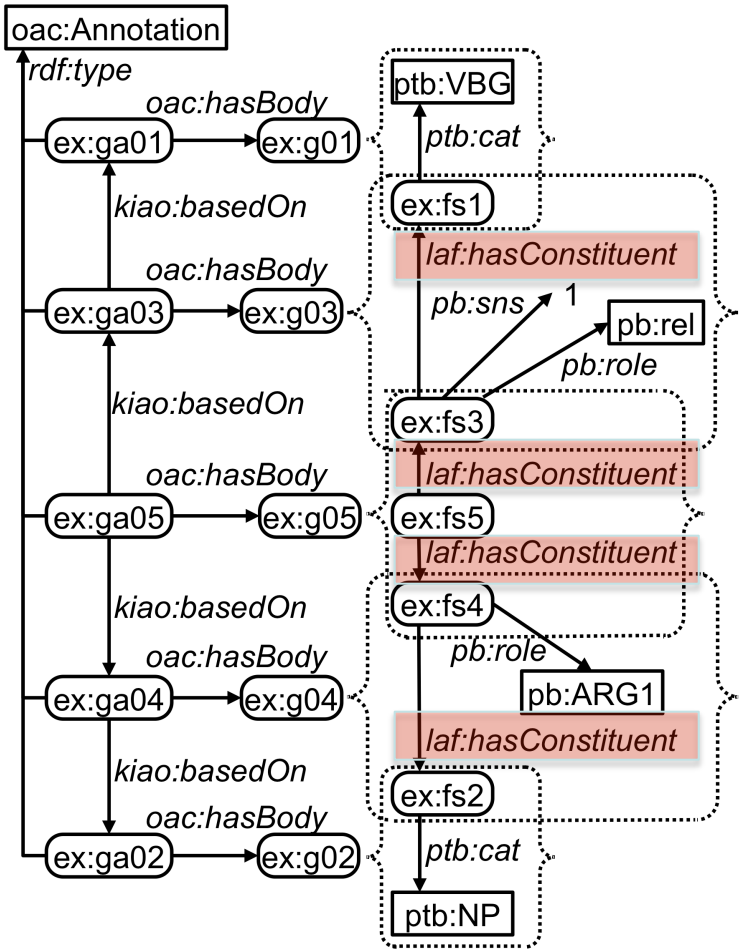
Each annotation independently connected to the target text



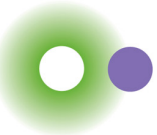

```
wsj/00/wsj_0003.mrg 18 18 gold include.01
p---a 14:1,16:1-ARG2 18:0-rel 19:1-ARG1
```

A LAF snippet of a PropBank annotation from (Ide and Suderman 2007)





So why is LAF/GrAF not enough to achieve our goals?



- XML is document-centric, not annotation-centric
  - what if only a subset of the XML is relevant to you?
  - can't refer to annotations from outside of the document
  - creating, querying, and manipulating the representations requires processing a complete XML document
- "Inward" focus of semantics
  - representation of arbitrary entity and relationship types without concern for consistency and reuse of those types
  - Lack of consideration of other, potentially relevant annotation types

Conclusions

NICTA

- A solution which enables interoperability of linguistic data with other, *possibly non-linguistic data*, about texts is preferable
- Open Annotation provides a model which is a good candidate for achieving this
- There is high-level compatibility between existing models
  - The devil is in the details, of course
  - Where things need to change, the changes will help with clarity and consistency

Acknowledgements

- Funding
  - Andrew W. Mellon Foundation
  - National ICT Australia
  - NIH grant 3T15 LM00945103S1 to Kevin Livingston
- Colleagues
  - University of Colorado Center for Computational Pharmacology
    - Kevin Livingston, Michael Bada, Kevin B. Cohen, Larry Hunter
  - Open Annotation Community Group
    - Robert Sanderson, Herbert van de Sompel, Jacob Jett, Tim Cole, Paolo Ciccarese, Tim Clark, Hennie Brugman
  - Linguistic Annotations in RDF
    - Steve Cassidy, Christian Chiarcos

Thank you!