A Model for Linguistic Resource Description

Nancy Ide and Keith Suderman

Department of Computer Science

Vassar College

Poughkeepsie, New York, USA

LAW VI

July 12-13, 2012

Jeju, Korea

Background

- Lots of effort has gone into defining standardized representation formats for linguistically annotated language resources
- Little effort towards standardizing documentation best practices for these resources
- Detailed information often sparse concerning
  - Provenance of data *and* annotations
  - Annotation schemes
  - Methodology
- **Need this information to use, assess quality, replicate processes and results, deal with idiosyncrasies/documented errors, etc.**

Background

- Virtually no effort to develop standardized strategies for formally describing the structure and organization of a resource
  - directory structure and relations among files typically provided in accompanying README files
  - provide no means to ensure that requisite components are in place or perform systematic processing without developing customized scripts

- **Formalized description of resource organization would enable automatic validation as well as enhanced processing capabilities**

Existing practices

- Multiple techniques proposed to specify resource provenance
  - W3C Working Group recently convened to define standards for exchange of provenance information
    - Provenance only
    - Primarily web data
- Some standard practices for resource publication/documentation through established data distribution centers (LDC, ELRA)
  - Not consistent, not comprehensive

Resource Description Standard

- ISO TC37 SC4 Linguistic Annotation Framework
  - *Now an official ISO standard!!!*
  - Specifies a comprehensive standard for resource description/documentation
  - Serialized in Graph Annotation Format (GrAF) XML headers
- Provides mechanisms for
  - describing organization of the resource
  - Documenting conventions used
  - associating data and annotation documents
  - defining and selecting defined portions of the resource and its annotations

Resource Description Standard

- Designed to accommodate the use of XML technologies for processing
  - XPath, XSLT, RDF/OWL
- Designed to accommodate linkage to web-based ontologies and data category registries
  - OLiA ontologies, ISOCat, etc.
- Designed to enable automatic validation of the resource
  - Check consistency, completeness
- Designed to enable automatic processing
  - E.g., select certain data and annotations, have information about which files are required /dependent

GrAF Overview

- Developed within ISO TC37 SC4 to provide a general framework for representing linguistically annotated resources

- Informed by previous and current approaches and tools, including but not limited to
  - UIMA CAS
  - GATE (annotation graphs)
  - ANVIL
  - ELAN
  - NLP Interchange Format (NIF)

- **Data model designed to capture the relevant structural generalization underlying best practices for linguistic annotation: directed (acyclic) graph**

GrAF as a “pivot” format

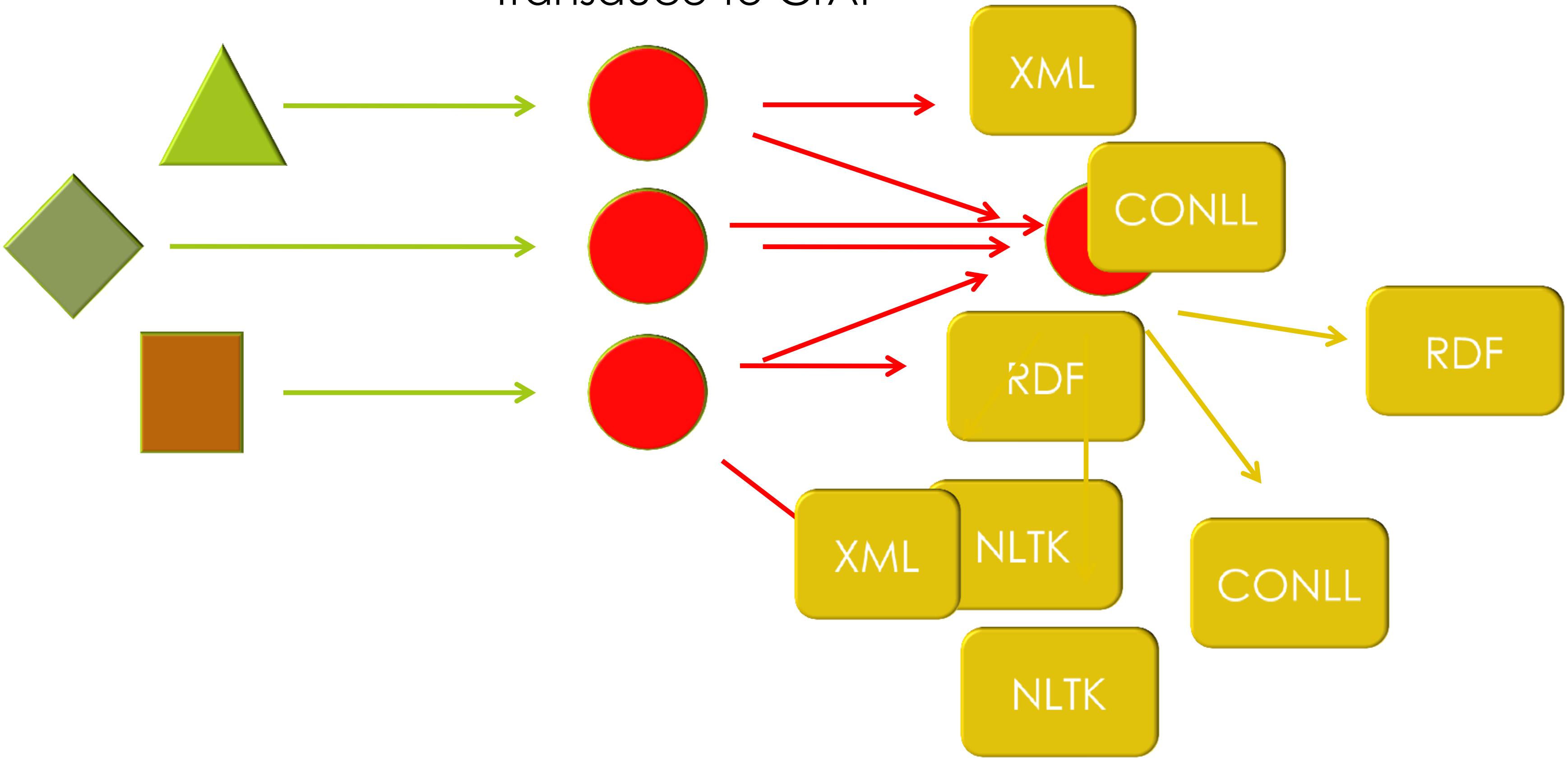
Different formats

Transduce merged graph

Transduce to other formats

Merge

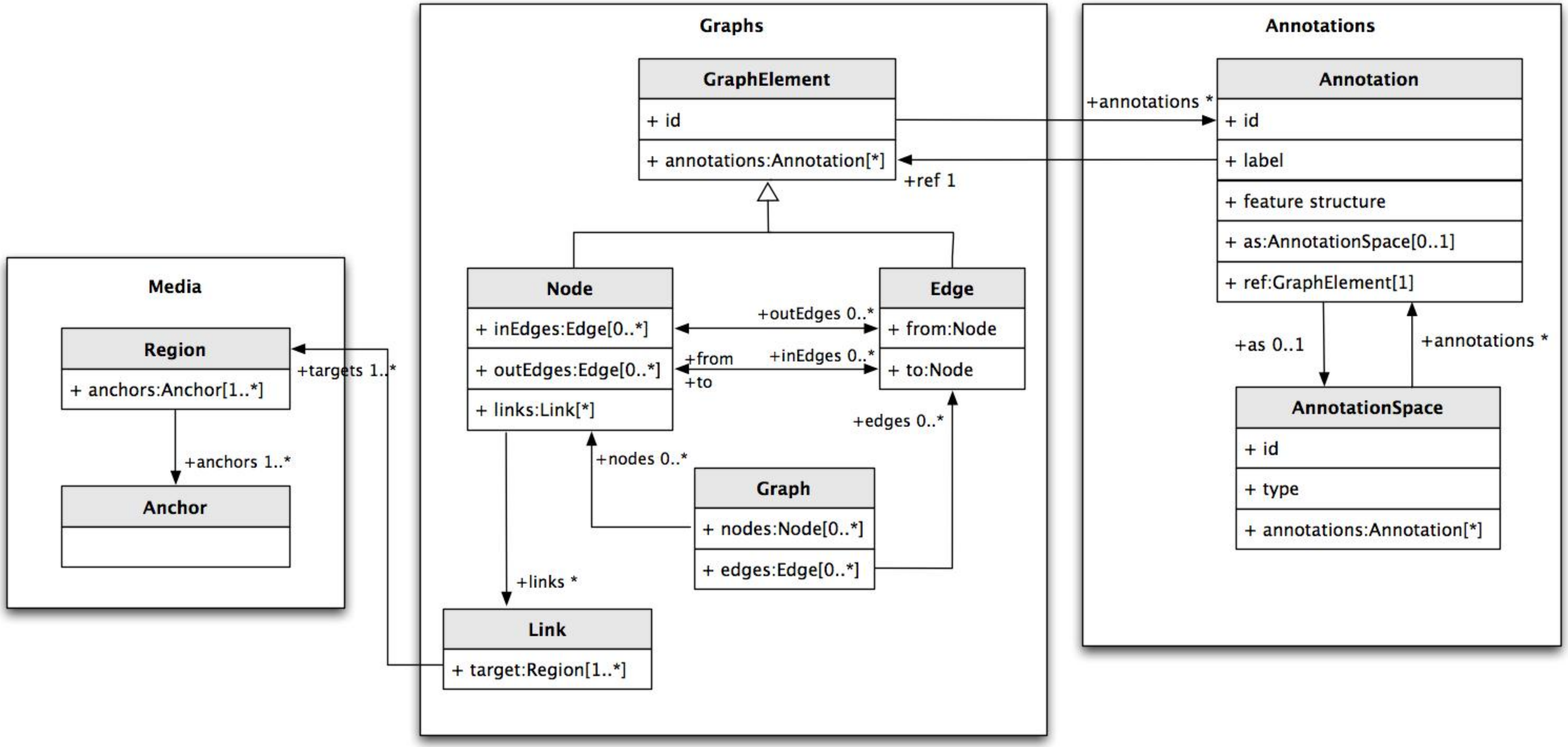
Transduce to GrAF



Overall GrAF resource architecture

- One or more **primary data documents**, in any medium
- One or more **base segmentation documents** defining a set of regions over a primary data document
- Any number of **annotation documents** containing feature structures associated with nodes and/or edges in a directed graph
- **Header documents** associated with each primary data document and annotation document, and a resource header that provides information about the resource as whole

GrAF Data Model



Media

Region
+ anchors:Anchor[1..*]

+anchors 1..*

Anchor

+targets 1..*

Graphs

GraphElement
+ id
+ annotations:Annotation[*]

+ref 1

Node
+ inEdges:Edge[0..*]
+ outEdges:Edge[0..*]
+ links:Link[*]

Edge
+ from:Node
+ to:Node

+outEdges 0..*

+inEdges 0..*

+from

+to

+edges 0..*

+nodes 0..*

Graph
+ nodes:Node[0..*]
+ edges:Edge[0..*]

+links *

Link
+ target:Region[1..*]

Annotations

+annotations *

Annotation
+ id
+ label
+ feature structure
+ as:AnnotationSpace[0..1]
+ ref:GraphElement[1]

+as 0..1

+annotations *

AnnotationSpace
+ id
+ type
+ annotations:Annotation[*]

GrAF Headers

- **Resource header**
  - Contains all the formal specifications for whole resource
    - Creator, project information, etc.; domain/genre category definitions, media definitions, annotation set definitions, layers/tiers, file structure definition, annotation types, pointer to annotation scheme documentation…
- **Primary data document header**
  - Contains information about the primary data
    - Provenance, medium (point to resource header), language, writing system, genre/domain information (point to resource header), associated annotations…
- **Annotation header**
  - Information about a particular annotation
    - Format, creator, location of original, dependencies on other annotations, medium, anchor types (references to text or other annotations), annotation set…

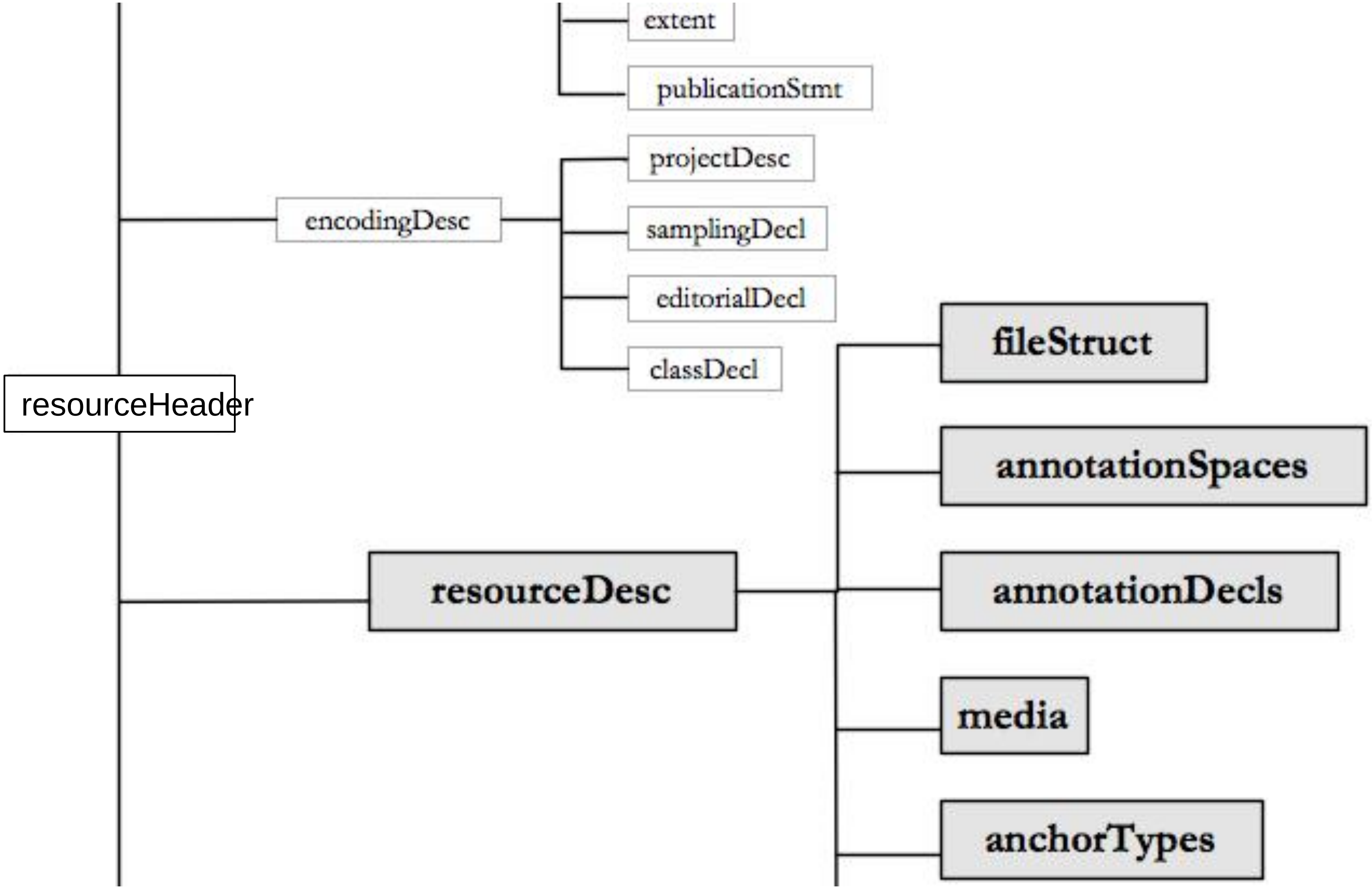
GrAF Resource Header

- Specifications formal enough for machine processing
  - Validation, selection of sub-parts of the corpus
- E.g. define domain / genre categories

```
<classDecl>
<!-- Category codes are referenced in the header of each primary data document -->
   <taxonomy id="MASC">
       <category id="WR">
                   <catDesc>Written</catDesc>
       </category>
       <category id="JO">
                   <catDesc>journal</catDesc>
            </category>
```

Would like to have a URI here

Overview of Resource Header



GrAF Resource Header

File structure

```
<fileType
    xml:id="f.entities"
    suffix="ne"
    a.ids="ne"
    medium="xml"
    requires="f.ptbtok"/>
```

Suffix in filenames

Id for reference

Id of medium type

Filetypes required to process this filetype

```
<r
      <fileType xml:id="f.seg" suffix="seg" a.ids="seg" medium="xml" requires="pr
      <fileType xml:id="f.logical" suffix="logical" a.ids="logical" medium="xml"
         requires="primary"/>
      <fileType xml:id="f.ptbtok" suffix="ptbtok" a.ids="ptbtok" medium="xml" requires="f.seg"/>
      <fileType xml:id="f.fntok" suffix="fntok" a.ids="fntok" medium="xml" requires="f.seg"/>
      <fileType xml:id="f.penn-pos" suffix="penn" a.ids="penn" medium="xml" requires="f.seg"/>
      <fileType xml:id="f.nounchunks" suffix="nc" a.ids="nc" medium="xml" requires="f.ptbtok"/>
      <fileType xml:id="f.verbchunks" suffix="vc" a.ids="vc" medium="xml" requires="f.ptbtok"/>
      <fileType xml:id="f.sentence" suffix="s" a.ids="s" medium="xml" requires="f.primary"/>
      <fileType xml:id="f.entities" suffix="ne" a.ids="ne" medium="xml" requires="f.ptbtok"/>
      <fileType xml:id="f.ptb" suffix="ptb" a.ids="ptb" medium="xml" requires="f.ptbtok"/>
      <fileType xml:id="f.framenet" suffix="fn" a.ids="fn" medium="xml" requires="f.fntok"/>
      <fileType xml:id="f.wordnet" suffix="wn" a.ids="wn" medium="xml" requires="f.sentence"/>
   </fileTypes>
</fileStruct>
```

GrAF Resource Header

Annotation spaces

```
<annotationSpaces>
    <annotationSpace xml:id="ptb" pid="http://www.cis.upenn.edu/~treebank/"/>
</annotationSpaces>
```

```
<annotationSpace
    xml:id="fn"
    pid="http://framenet.icsi.berkeley.edu/"/>
```

Reference to persistent identifier for the annotation type

GrAF Resource Header

Annotation Declaration

```
<annotationDecl xml:id="a.ne" as="xces">
  <a.desc>named entities</a.desc>
  <a.resp lnk:href="http://www.anc.org">ANC proj </a.resp>
  <a.method type="automatic-validated"/>
  <a.doc
   lnk:href="https://www.anc.org/wiki/wiki/Named
</annotationDecl>
```

Reference to previously declared annotation space

GrAF Resource Header

```
<medium xml:id = "text"
        type = "text/plain”
        encoding = "utf-8"
        extension = "txt"/>
```

```
<r
<r
<medium xml:id = "video" type = "video" encoding = "Cinepak" extens
<medium xml:id = "image" type = "image" encoding = "jpeg" extensio
...
<anchorType xml:id="text-anchor" medium = "text" default = "true"
      lnk:href = "http://www.xces.org/ns/GrAF/1.0/#character-anchor"/>
```

Extension on files containing this medium type

```
<anchorType
     xml:id="text-anchor"
     medium = "text"
  default = "true”
     lnk:href = "http://www.xces.org/ns/GrAF/1.0/#char-anchor"/>
```

Location of formal specification of anchor type

GrAF Resource Header

Associating files, annotations, media, anchors, etc.

```
<fileType xml:id = "f.entities" suffix = "ne" a.ids = ”a.ne"
          medium = "xml" requires = "f.ptbtok"/>
...
<annotationSpace xml:id = "xces" pid = "http://www.xces.org/schema/2003"/>
...
<annotationDecl xml:id="a.ne" as="xces">
     <a.desc>named entities</a.desc>
     <a.resp lnk:href="http://www.anc.org">ANC project</a.resp>
     <a.method type="automatic-validated"/>
     <a.doc lnk:href="https://www.anc.org/wiki/wiki/NamedEntities"/>
</annotationDecl>
...
 <medium xml:id = "text" type = "text/plain"
      encoding = "utf-8" extension = "txt"/>
 <medium xml:id = "xml" type = "text/xml"
      encoding = "utf-8" extension = "xml"/>
...
 <anchorType medium = "text" default = "true"
     lnk:href = "http://www.xces.org/ns/GrAF/1.0/#character-anchor"/>
```

GrAF Resource Header

Groups (layers, tiers, etc.)

```
<groups>
  <group xml:id = "g.token">
    <!-- all annotations in any annotation space with label "tok" -->
    <g.member value = "*:tok" type = "annotation"/>
  </group>
  <group xml:id = "g.example">
          <!-- all annotations of type logical -->
          <g.member value = "a.logical" type = "type"/>
          <!-- all files containing entity annotations -->
          <g.member value = "f.entities" type = "file"/>
          <!-- all annotations with a feature "speaker" with value "Alice" -->
          <g.member value = "@speaker = 'alice'" type = "expression"/>
          <!-- annotations with ids "id_1" to "id_n" in file "myfile.xml"-->
          <g.member xml:base = "myfile.xml" value = "id1 id2 ... idN"
                 type = "enumeration"/>
          <!-- the annotations included in group g.token, as defined earlier -->
          <g.member value = "g.token" type = "group"/>
  </group>
</groups>
```

GrAF Primary Document Header

```
<documentHeader xmlns="http://www.xces.org/ns/GrAF/1.0/" creator="KBS" date.created="2005-08-29"
           version="1.0.4">
   <fileDesc>
      <titleStmt>
         <title>Day3PMSession</title>
      </titleStmt>
      <extent wordCount="20817"/>
      <sourceDesc>
         <title>TRANSCRIPT OF PROCEEDINGS OF BENCH TRIAL - Day 3, Afternoon Session</title>
         <publisher>National Center for Science Education</publisher>
         <eAddress type="web">http://ncse.com/</eAddress>
         <pubPlace>http://ncse.com/files/pub/legal/kitzmiller/trial_transcripts/
                   2005_0928_day3_pm.pdf</pubPlace>
      </sourceDesc>
   </fileDesc>
   <profileDesc>
      <textClass catRef="SP TR">
         <domain>Government</domain>
         <subdomain>Court Transcript</subdomain>
         <subject>Darwin vs. Creationism</subject>
         <audience>Adult</audience>
         <medium>web</medium>
      </textClass>
      ...
```

Reference to category definition in resource header

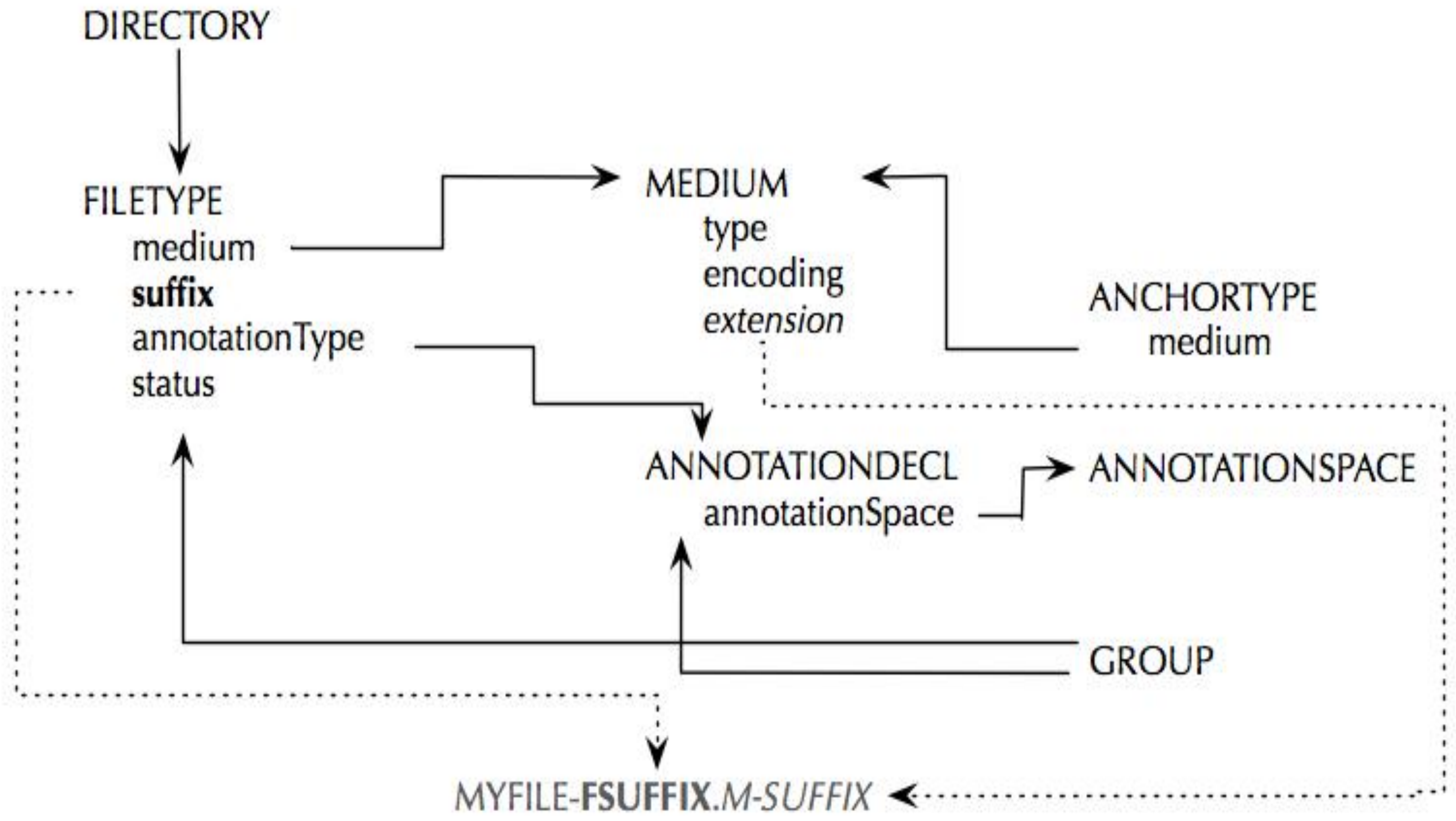
GrAF Primary Document Header

```
  <primaryData loc="Day3PMSession.txt" medium="text"/>
  <annotations>
    <annotation ann.loc="Day3PMSession-logical.xml" type="logical">Logical
                  structure</annotation>
    <annotation ann.loc="Day3PMSession-s.xml" type="s">Sentence boundaries
                  </annotation>
    <annotation ann.loc="Day3PMSession-nc.xml" type="nc">Noun chunks</annotation>
    <annotation ann.loc="Day3PMSession-penn.xml" type="penn">Penn part of speech
                  tags</annotation>
    <annotation ann.loc="Day3PMSession-ptb.xml" type="ptb">Penn Tree Bank</annotation>
    <annotation ann.loc="Day3PMSession-ptbtok.xml" type="ptbtok">Penn Tree Bank tokens
                  and part of speech tags</annotation>
    <annotation ann.loc="Day3PMSession-seg.xml" type="seg">Base segmentation
                  (quarks)</annotation>
    <annotation ann.loc="Day3PMSession-ne.xml" type="ne">Named Entities</annotation>
    <annotation ann.loc="Day3PMSession.txt" type="content">Document
                  content</annotation>
  </annotations>
</profileDesc>
<revisionDesc>
  ...
```

GrAF Annotation Documents

- Annotation documents contain both a header and the graph of feature structures comprising the annotation
- Header contains:
  - a list of the annotation labels used in the document and their frequencies;
  - a list of documents required to process the annotations, which will include a segmentation document and/or any annotation documents directly referenced in the document;
  - a list of annotation Spaces referenced in the document, one of which may be designated as a default for annotations in the document;
  - optional) The root node(s) in the graph, when the graph contains one or more graphs that comprise a well-formed tree.

Dependencies among headers



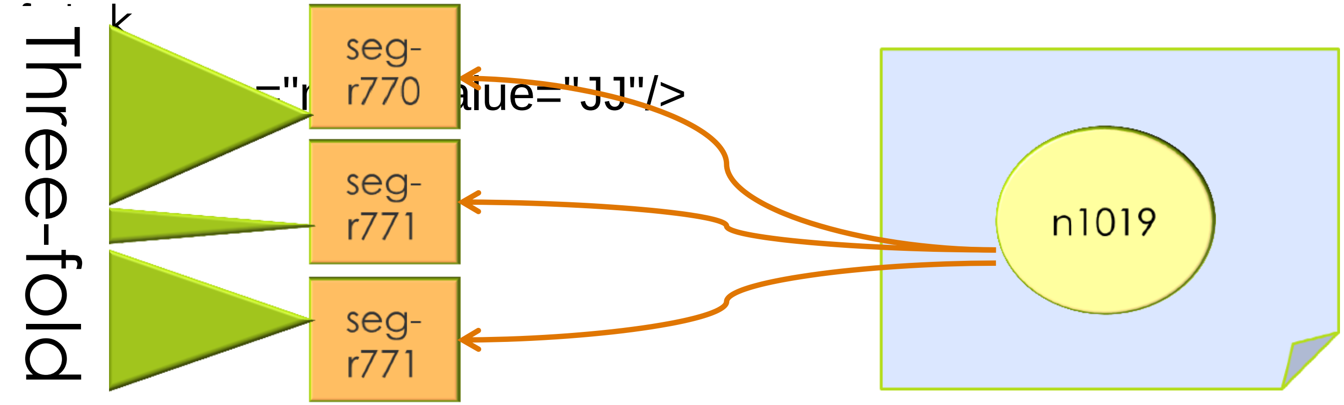
Anchors and Regions

```
<medium xml:id = "text" type = "text/plain" encoding = "utf-8" extension = "txt"/>
<medium xml:id = "audio" type = "audio" encoding = "MP4" extension = "mpg"/>
<medium xml:id = "video" type = "video" encoding = "Cinepak" extension = "mov"/>
<medium xml:id = "video" type = "image" encoding = "jpeg" extension = "jpg"/>
...
<anchorType xml:id="text-anchor" medium = "text" default = "true"
      lnk:href = "http://www.xces.org/ns/GrAF/1.0/#character-anchor"/>
<anchorType xml:id="time-slot" medium = "audio"
      lnk:href = "http://www.xces.org/ns/GrAF/1.0/#audio-anchor"/>
<anchorType xml:id="video-anchor" medium = "video"
      lnk:href = "http://www.xces.org/ns/GrAF/1.0/#video-anchor"/>
<anchorType xml:id="image-point" medium = "image"
      lnk:href = "http://www.xces.org/ns/GrAF/1.0/#image-point"/>
```

```
<!-- Regions in the segmentation document -->
<region xml:id="r1" anchor_type="time-slot" anchors="980 983"/>
<region xml:id="r2" anchor_type="image-point"
      anchors="10,59 10,173 149,173 149,59"/>
<region xml:id="r3" anchor_type="video-anchor"
      anchors="frame1(10,59) frame2(59,85) frame3(85,102)"/>
<region xml:id="r4" anchor_type="text-anchor"
      anchors="34 42"/>
```

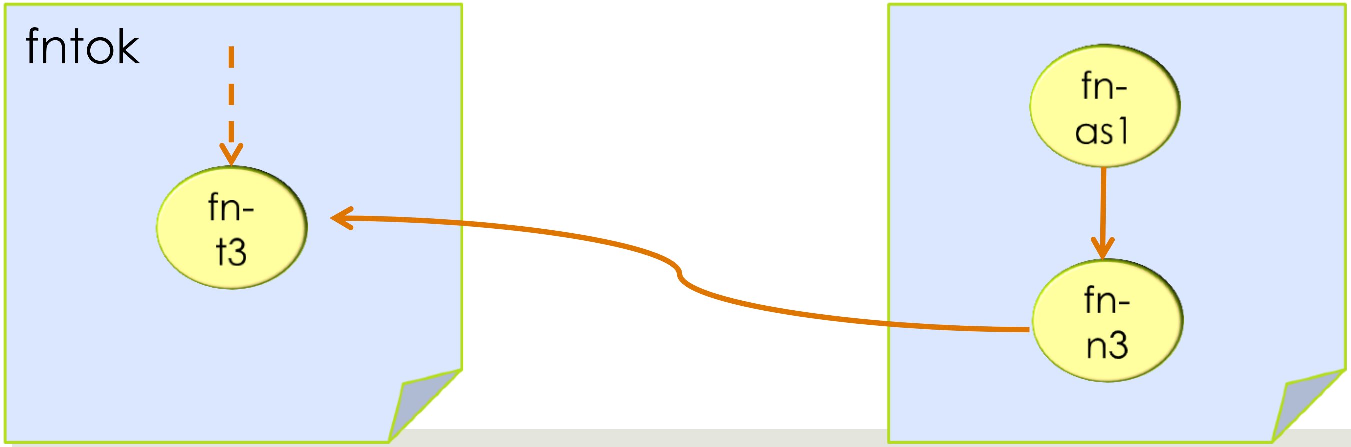
Nodes and Regions

```
<region xml:id="seg-r770" anchors="2211 2216"/>
<region xml:id="seg-r771" anchors="2216 2217"/>
<region xml:id="seg-r772" anchors="2217 2221"/>


<node xml:id="n1019">
        <link targets="seg-r770 seg-r771 seg-r772"/>
</node>
<a label="tok" ref="n1019" as="xces">
                    ="r        alue="JJ"/>

</a
```



Nodes and Edges

```
<node xml:id = "fn-n3"/>
<a label = "FE" ref = "fn-n3"  as = "FrameNet">
   <fs>
      <f name = "name" value = "Supplier"/>
      <f name = "GF" value = "Ext"/>
      <f name = "PT" value = "NP"/>
   </fs>
</a>
<edge xml:id = "e46" from = "fn-as1"  to = "fn-n3"/>
<edge xml:id = "e92" from = "fn-n3"  to = "fntok:fn-t3"/>
```



Summary

- GrAF headers provide
  - Guidance for what should be included in documentation of resources as a whole, annotations, and data
    - As with LAF approach overall, does not require specific information about an annotation scheme
      - At present requires reference (URI) to documentation
      - Awaits development of detailed “best practice standards”
  - Mechanisms that allow for automatic validation of
    - Overall resource structure
    - Presence of required files
    - Conformance of file contents in terms of type, medium, encoding, etc.
    - Conformance of data, anchors with specifications for the medium associated with it (encoding, anchor form)

Summary

- GrAF headers provide
    - Mechanisms for linking
        - Annotation documents to full description of annotation scheme, method of creation, etc.
        - Data to pre-defined information (e.g., genre)
        - Anchors in the resource to formal specification
    - Mechanisms for automatically selecting portions of a resource based on several different criteria
        - Annotations in a given annotation space or spaces
        - Annotations of a specific type, with specific features, etc.
        - Data belonging to a pre-defined type (e.g. genre), of a certain medium, etc.

Information

- http://www.anc.org/graf
- http://sourceforge.net/projects/graf/
- Paper to appear in *Language Resources and Evaluation*

  The Linguistic Annotation Framework: A Standard for Annotation Interchange and Merging

  Nancy Ide and Keith Suderman

Thank you