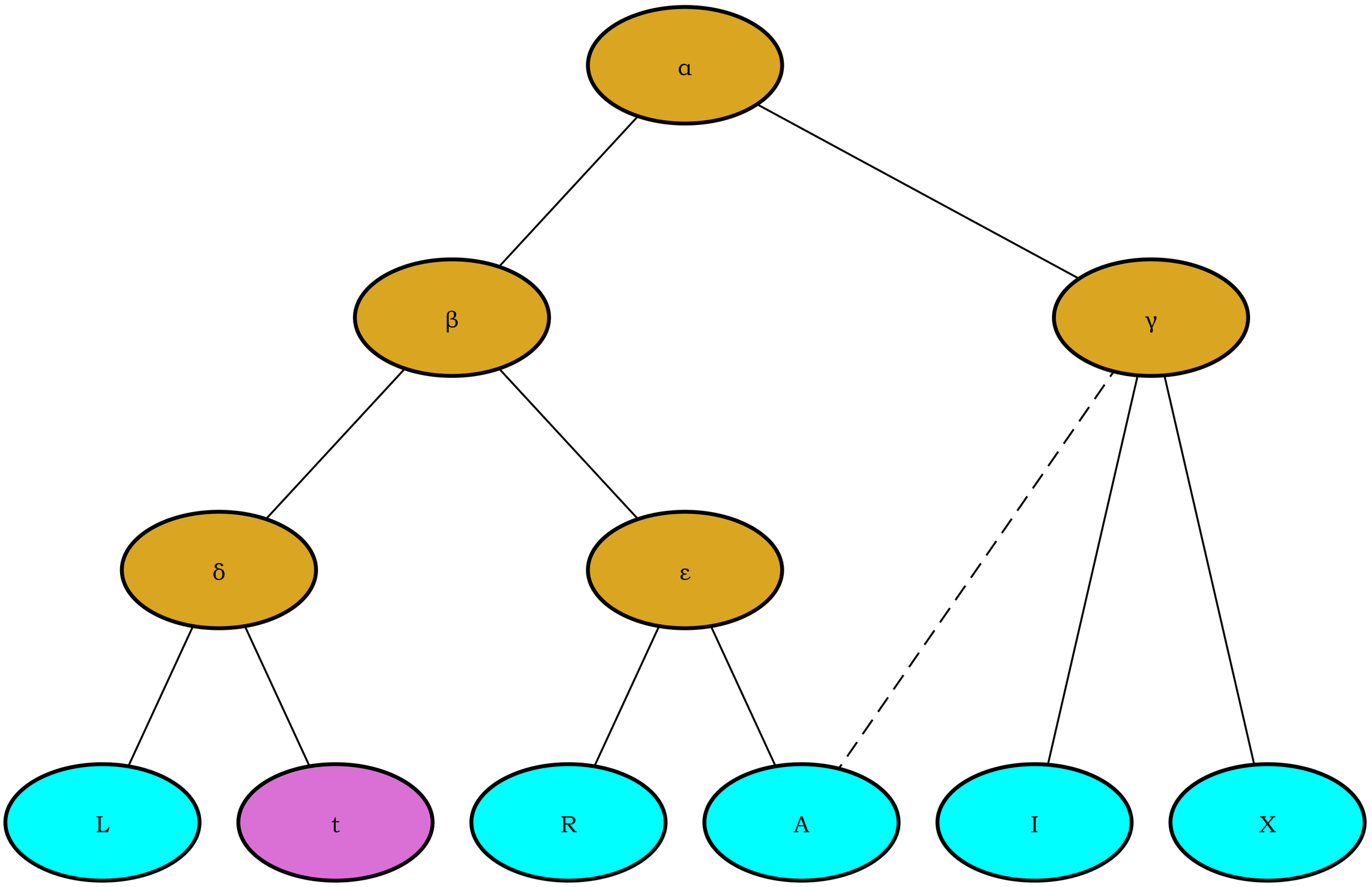
A proposal for encoding stemmata codicum in XML

David J. Birnbaum, University of Pittsburgh
djbpitt@pitt.edu
Full report: http://clover.slavic.pitt.edu/~djb/2007_xml-stemma/2007_xml-stemma.html

A *stemma codicum* (family tree of manuscripts) is an acyclic directed graph that consists of a *tree* of *typed nodes* with the addition of *contamination* (alternative parentage)



- A stemma models *textual transmission*.
- *Nodes* represent *extant* manuscripts (aqua), *lost* manuscripts (magenta), or *hypothetical stages* in the transmission (gold).
- There is an *original* version (α).
- Manuscripts are copied from other manuscripts or hypothetical stages (arc represented by a solid line). The nodes and solid lines constitute a *tree*.
- In addition to their primary source (parent), manuscripts may show the influence of other sources (arc represented by a dotted line), called *contamination.*

Stemmata codicum, p. 2

What's the point?

- A stemma is graphically demanding, which makes it tempting to encode the presentation (e.g., SVG, or a flat collection of `<node>` and `<arc>` elements, as in the main TEI graph model).
- A more proper XML approach would be descriptive, rather than presentational, but even if presentation isn't the only goal, one must be able to render the stemma.
- A descriptive (non-presentational) encoding is easily queried and traversed, unlike a presentational one.

Here's a model

- Parentage in the stemma is modeled by containment in the XML
- Nodes are labeled with the `@n` attribute
- Nodes are typed with the `@type` attribute
- Contamination is encoded with a `<contaminates>` element, whose parent in the XML is the contaminator and whose `@target` attribute identifies the contaminee

```
start = node
node = element node  { n, type, (node | contaminates)* }
n = attribute n { text }
type = attribute type { "hypothetical" | "extant" | "lost" }
contaminates = element contaminates { target, empty }
target = attribute target { text }
```

```
<node n="α" type="hypothetical">
  <node n="β" type="hypothetical">
    <node n="δ" type="hypothetical">
      <node n="L" type="extant"/>
      <node n="t" type="lost"/>
    </node>
    <node n="ε" type="hypothetical">
      <node n="R" type="extant"/>
      <node n="A" type="extant"/>
    </node>
  </node>
  <node n="γ" type="hypothetical">
    <contaminates target="A"/>
    <node n="I" type="extant"/>
    <node n="X" type="extant"/>
  </node>
</node>
```

Optional: Validate the node names

```
start = node
node = element node { atts_node, (node | contaminates)* }
atts_node =
   (attribute type { "hypothetical" },
    attribute n {
       xsd:token { pattern="\p{IsGreek}" pattern="\p{Ll}"} # lower-case Greek
    })
   | (attribute type { "extant" },
      attribute n {
         xsd:token { pattern="\p{IsBasicLatin}" pattern="\p{Lu}"} # upper-case Latin
      })
   | (attribute type { "lost" },
      attribute n {
         xsd:token { pattern="\p{IsBasicLatin}" pattern="\p{Ll}"} # lower-case Latin
      })
contaminates = element contaminates { target, empty }
target = attribute target { text }
```

Optional: Use schematron to fake ID/IDREF validation

```
<s:pattern name="Value of @target attribute must point to node/@n in document">
 <s:rule context="contaminates">
  <s:assert test="@target=//node/@n">The value of a @target attribute of an element of
   type <s:name/> must point to the @n attribute of an element of type node in the
   document.</s:assert>
 </s:rule>
</s:pattern>
```

Stemmata codicum, p. 3

Render in at least two ways

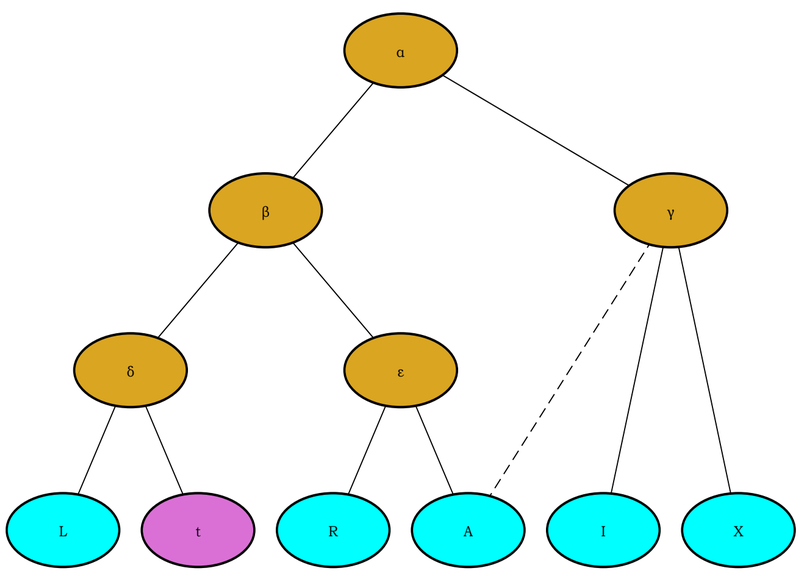
- Convert to Graphviz *dot* format with XSLT (http://www.graphviz.org)
- Transform directly to SVG with XSLT
- Sample XSLT for both transformations available at http://clover.slavic.pitt.edu/~djb/2007_xml-stemma/2007_xml-stemma.html

What's it good for other than rendering?

- A stemma is a hypothesis about how variation arises during textual transmission.
- For any pattern of variation, the application of parsimony to a stemma yields one of three results:
  - A particular reading is more likely than others to have stood in the original.
  - Different readings are equally likely to have stood in the original (a *crux*).
  - The stemma cannot help decide which reading is most likely to have stood in the original.
- Stemmatic evaluation is purely mechanical, i.e., well suited to automation (in this case, with the help of XSLT).

Why you should care

- XML engineers may recognize the virtues of descriptive (vs presentational) markup but find themselves unable to motivate or perform thorough and effective document analysis on elaborate graphical materials.
- A stemma, like a table (see David J. Birnbaum's presentation at Extreme Markup 2007), is amenable to descriptive markup, but its distinctive presentational features may encourage document analysts to encode presentation, instead of semantics.
- Presentational markup prioritizes presentation, and often presentation in only one way, at that. Descriptive markup, when implemented thoughtfully, may support not only presentation in different ways, but also querying and analysis.
- XSLT is able to convert from descriptive markup to presentational structures (HTML tables, SVG graphs, etc.), and can also support querying and analysis.
- Descriptive markup can model what the object is (the metaphysical argument) and can do so elegantly (the aesthetic argument). Ultimately, though, descriptive markup in hard, presentationally demanding cases like tables and stemmata is desirable because it is also often more capable, flexible, and powerful than presentational markup.



Number	Reading		Where introduced		Stemmatic reading
	Chocolate	Peanut butter	Chocolate	Peanut butter	
1	L	t R A I X	L	α (mixed)	Peanut butter
2	t	L R A I X	t	α (mixed)	Peanut butter
3	R	L t A I X	R	α (mixed)	Peanut butter
4	A	L t R I X	A	α	Peanut butter
5	I	L t R A X	I	α (mixed)	Peanut butter
6	X	L t R A I	X	α (mixed)	Peanut butter
7	L t	R A I X	δ	α (mixed)	Peanut butter
8	L R	t A I X	β (mixed)	α (mixed)	(Non-stemmatic pattern)
9	L A	t R I X	β (mixed)	α (mixed)	(Non-stemmatic pattern)
10	L I	t R A X	α (mixed)	α (mixed)	(Non-stemmatic pattern)
11	L X	t R A I	α (mixed)	α (mixed)	(Non-stemmatic pattern)
12	t R	L A I X	β (mixed)	α (mixed)	(Non-stemmatic pattern)
13	t A	L R I X	β (mixed)	α (mixed)	(Non-stemmatic pattern)
14	t I	L R A X	α (mixed)	α (mixed)	(Non-stemmatic pattern)
15	t X	L R A I	α (mixed)	α (mixed)	(Non-stemmatic pattern)
16	R A	L t I X	ε	α (mixed)	Peanut butter
17	R I	L t A X	α (mixed)	α (mixed)	(Non-stemmatic pattern)
18	R X	L t A I	α (mixed)	α (mixed)	(Non-stemmatic pattern)
19	A I	L t R X	γ (mixed)	α (mixed)	(Non-stemmatic pattern)
20	A X	L t R I	γ (mixed)	α (mixed)	(Non-stemmatic pattern)
21	I X	L t R A	γ	β	Chocolate or Peanut butter (Crux)
22	L t R	A I X	β	γ	Chocolate or Peanut butter (Crux)
23	L t A	R I X	β (mixed)	α (mixed)	(Non-stemmatic pattern)
24	L t I	R A X	α (mixed)	α (mixed)	(Non-stemmatic pattern)
25	L t X	R A I	α (mixed)	α (mixed)	(Non-stemmatic pattern)
26	L R A	t I X	β (mixed)	α (mixed)	(Non-stemmatic pattern)
27	L R I	t A X	α (mixed)	α (mixed)	(Non-stemmatic pattern)
28	L R X	t A I	α (mixed)	α (mixed)	(Non-stemmatic pattern)
29	L A I	t R X	α (mixed)	α (mixed)	(Non-stemmatic pattern)
30	L A X	t R I	α (mixed)	α (mixed)	(Non-stemmatic pattern)
31	L I X	t R A	α (mixed)	β (mixed)	(Non-stemmatic pattern)