Representation, Data format and Standards

Main objectives

- Sharing resources
  - Are we able to reuse annotations from others?
- Sharing tools
  - transcription, annotation, data query and visualization
- Sharing semantics
  - How do we ensure consistancy of the meaning of annotations accross:
    - Data collection
    - Annotation manuals
    - Evaluation methods

What is an annotated corpus?

- Some primary data…
  - A text , speech signal (possibly with a transcription), movie, etc.
- …associated with some information about its linguistic properties
  - Morpho-syntactic category for each word, syntactic categories and structure, discourse structure, co-references etc.

Where are the annotations?

- Often, in the same document/file as the primary data

MUC-7

```
<TEXT>
<p>
  <s>
    <lex pos=DT>The</lex>
    <lex pos=NNP>Federal</lex>
    <lex pos=NNP>Aviation</lex>
    <lex pos=NNP>Administration</lex>
    <lex pos=VBD>underestimated</lex>
    <lex pos=DT>the</lex>
    <lex pos=NN>number</lex>
    <lex pos=IN>of</lex> …
```

Penn Treebank

```
((S   (NP-SBJ-1 Jones)
      (VP followed)
         (NP him)
         (PP-DIR into
               (NP the front room))
               ,
         (S-ADV   (NP-SBJ *-1)
                  (VP  closing
                  (NP the door)
                  (PP  behind
                       (NP him))))).))
```

Main difficulties

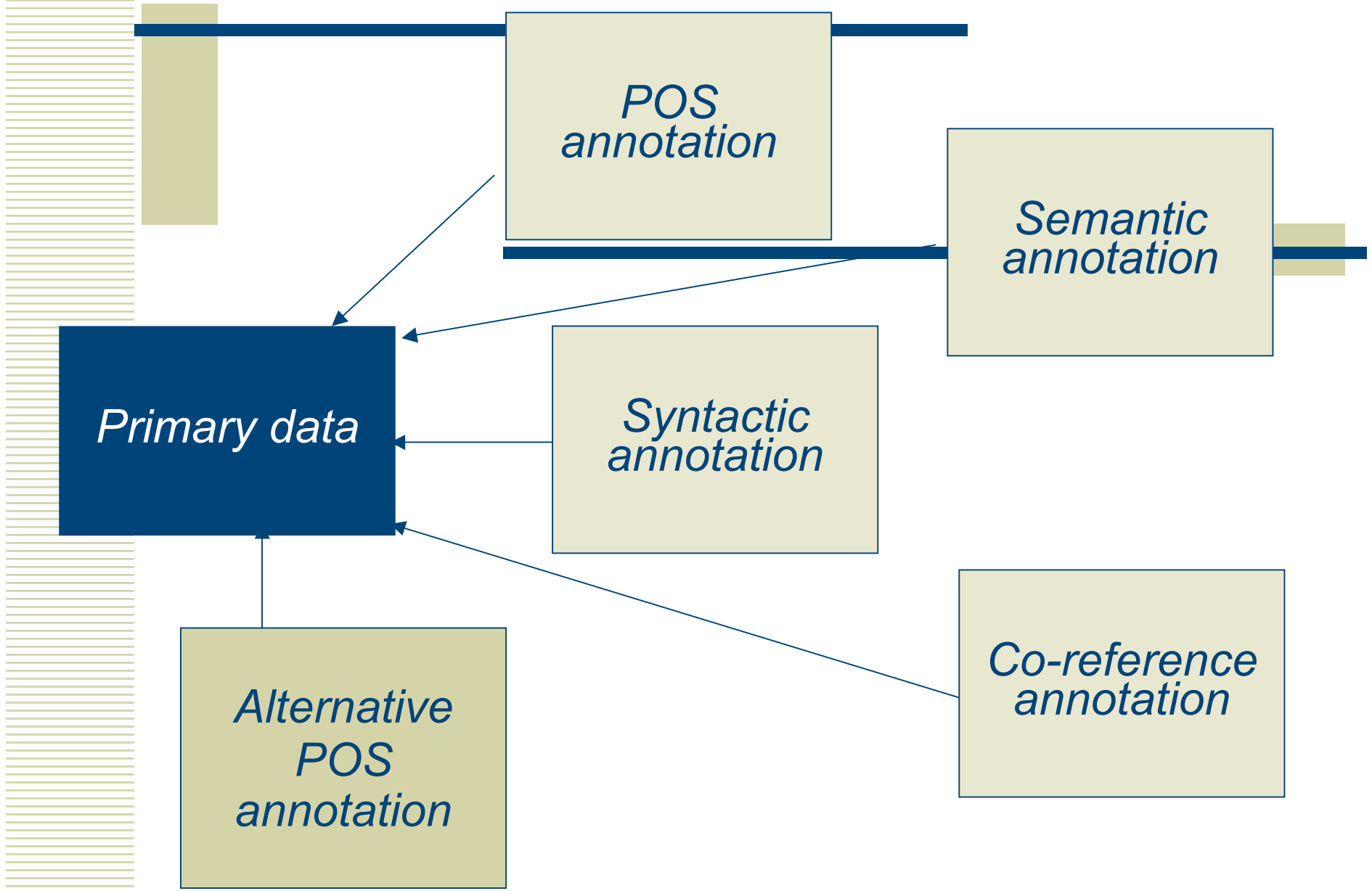
- Variety of underlying formats for representing annotations
  - Lack of interoperability between similar annotations
  - Hence lack of software for editing, accessing, viewing the annotated data
- Hard to have several different types of annotation for the same data
- Hard to have alternative annotations of same type
- Hard to maintain…

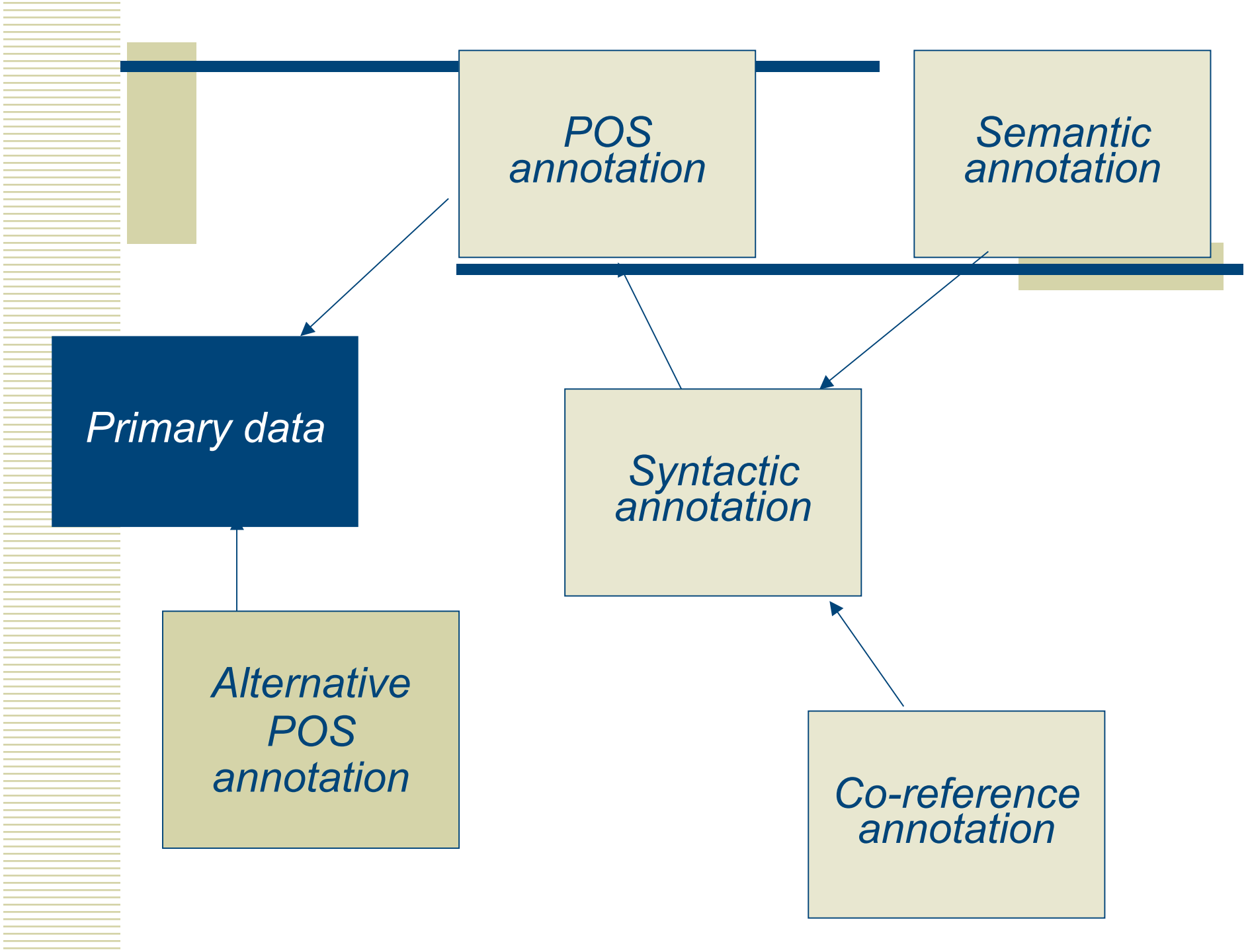
Preliminary answers

- XML as an underlying syntax
  - Adopted by a wide range of communities
  - Adapted for the web
  - Software availability
- But,
  - Need to go beyond the syntax
    - ⇨ Sharing structures within a community (Standards)

Preliminary answers (cont.)

- "Stand-off markup"
  - Annotations in separate XML documents, linked to original
- Leads to muti-level markup


POS annotation
Semantic annotation
Primary data
Syntactic annotation
Alternative POS annotation
Co-reference annotation


POS annotation
Semantic annotation
Primary data
Syntactic annotation
Alternative POS annotation
Co-reference annotation

Outline

- General context: past and present initiatives
  - newly formed ISO committee: TC 37/SC 4 (*Language Resource Management*)
- Present an abstract data model for linguistic annotations
  - Quick overview of XML
  - Main aspects of annotation scheme design

Background on past and present standardizing activities

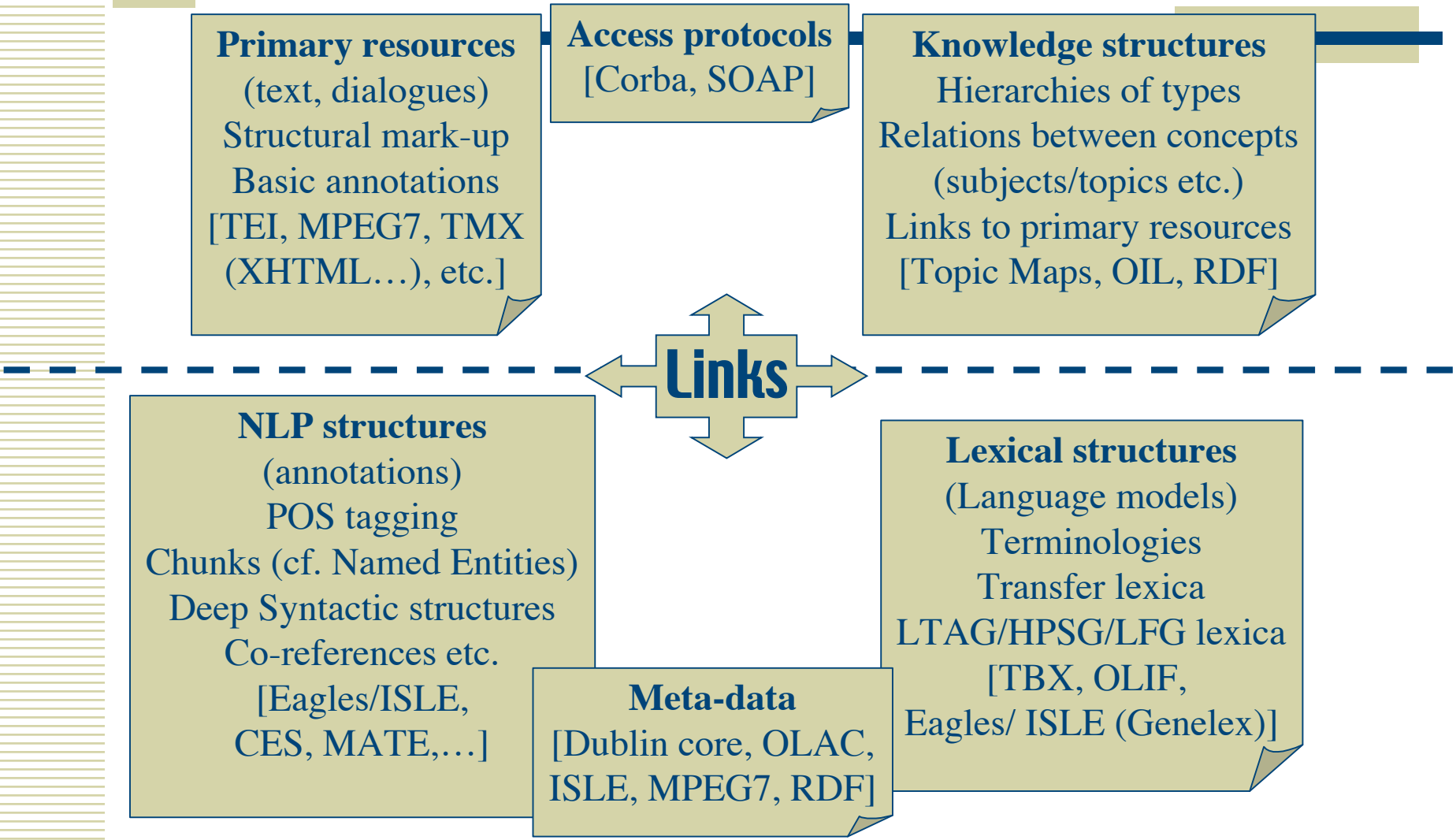
Standardization is Tricky

- Skepticism within the community
- Arguments against language resource standardization:
  1. diversity of languages and of theoretical approaches makes standardization impractical or impossible
  2. vast amounts of existing data and processing software will be rendered obsolete by the acceptance of new standards

Standardization is Tricky (cont.)

- Arguments…
  1. I don't want someone to impose a format that will not fit my needs
  2. Standards: they are too many of them for me to choose
     - ⇨ Do already existing initiatives fit my needs? (usability)
     - ⇨ Which are stable and which are not? (perenity)
     - ⇨ How do they interact with one another? (compatibility)
     - ⇨ Which one would I want to act upon? (involvement)

The variety of linguistic information sources



So…

- Work remains to be done to achieve interoperability in the domain of language resources
- One should not attempt to provide yet a new format
  - Tradeoff between flexibility and compatibility
- There is a need for an effort integrating existing initiatives (less “standards”)
  - Community specific standards vs. internationally recognized ones
- Hence the new ISO/TC37/SC4 committee: *Language Resource Management*

ISO in short

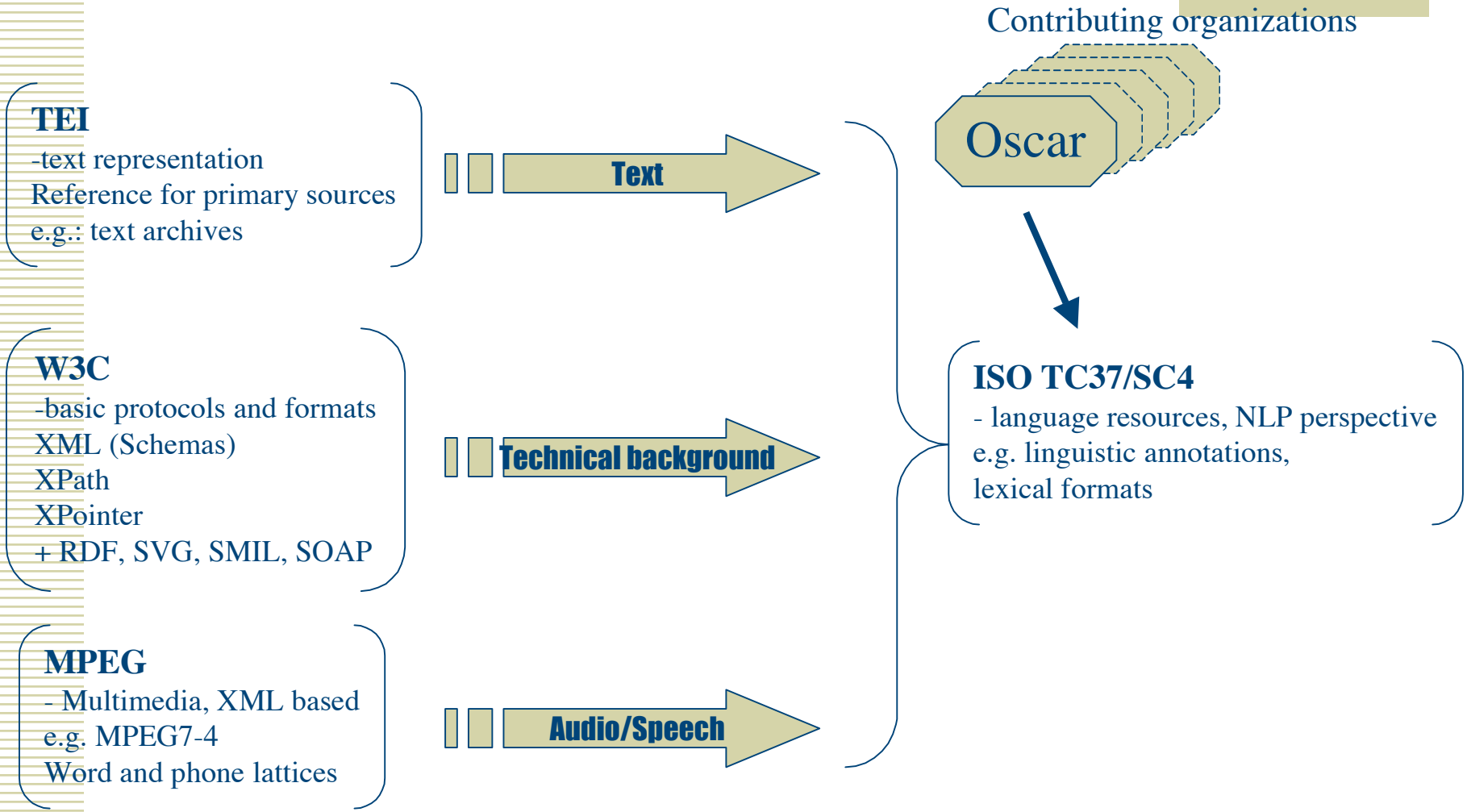
- ISO: International Standards Organization (www.iso.org)
  - Acts for its member bodies:
    - BSI, AFNOR, DIN, ANSI…
  - Is organized in Technical Committees (TC) and Sub-Committees (SC)
    - E.g. ISO/IEC JTC1
  - Produces international standards according to an established review process (WD, CD, DIS, F-DIS, IS)

Context

- **ISO TC37 -** Terminology and other language resources
  - **SC3 -** Computer applications in terminology
    - ISO 12200 - Martif
      - Latest version of TEI Terminology chapter
    - ISO 12620 - Data categories (under revision)
    - ISO 16642 - TMF (Terminological Markup Framework)
  - **SC4 -** Language Resource Management

www.tc37sc4.org

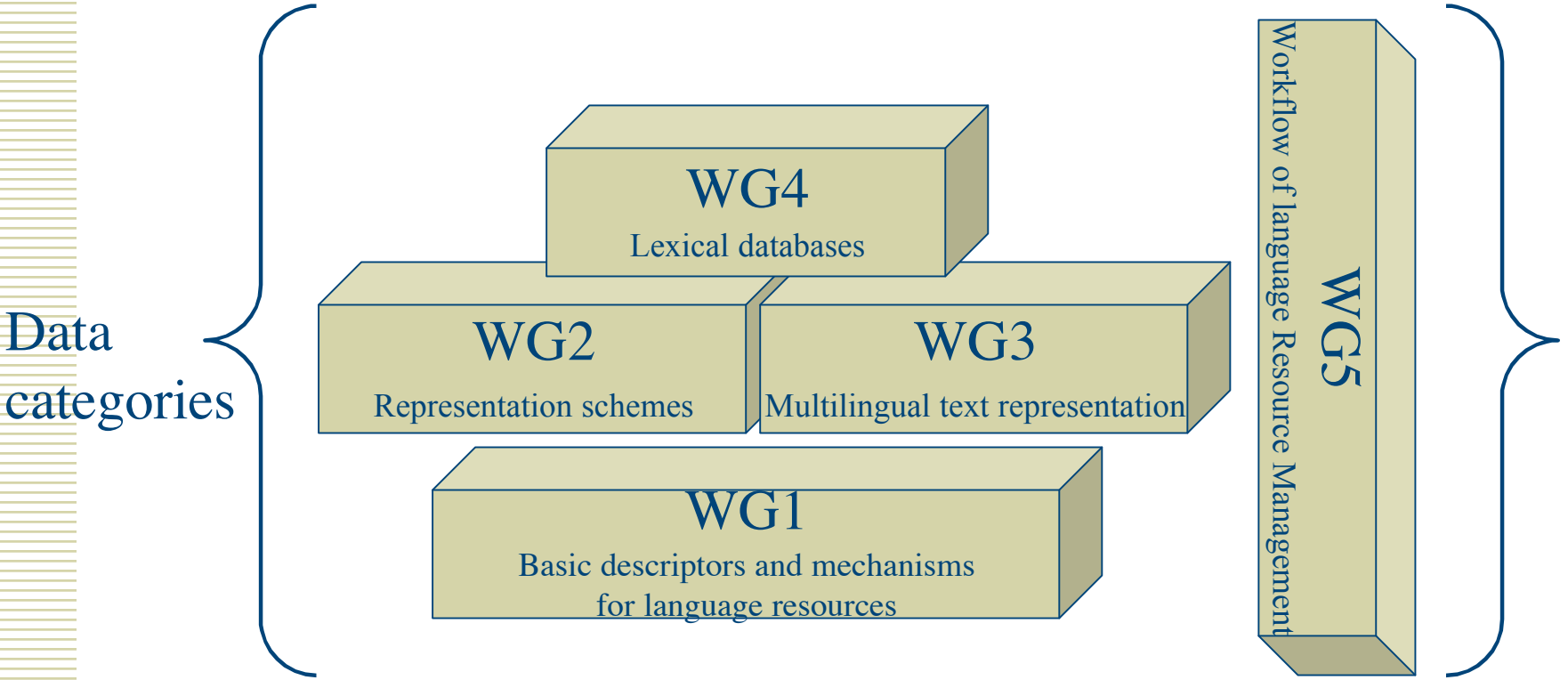
SC4 and other standardizing bodies



SC4 Approach

- Efforts geared toward defining abstract models and general frameworks for the creation and representation of language resources
  - In principle, abstract enough to accommodate diverse theoretical approaches
- Situate development squarely in the framework of XML and related standards
  - Ensure compatibility with established and widely accepted web-based technologies
  - Ensure feasibility of transduction from legacy formats into newly defined formats

ISO TC 37/SC 4 structure



On-going activies

- Feature structure representation (collaboration with the TEI)
- Morpho-syntactic annotation
- Lexical markup framework
- Meta-data for language resources (OLAC+IMDI)
  - E.g. communication setting and conditions
- Sigsem working group on multimodal content representation
- … Data category registry

Going into technical aspects

A (very) quick overview of XML

A quick historical overview

- 1986
  - SGML (*Standard Generalized Markup Language*)
  - ISO standard: *ISO:8879:1986*
- 1987
  - TEI (*Text Encoding Initiative*)
- 1990
  - HTML 1.0 (*HyperText Markup Language*)
- 1997/1998
  - XML 1.0 (*eXtensible Markup Language*)

What XML is:

- XML: eXtended Markup Language
- A W3C (World Wide Web Consortium) Recommendation
- A meta-language: it allows one to define his own markup language
- A simplified version of the SGML standard
  - SGML was intended to represent the “logical” structure of a document
  - HTML was conceived as an application of SGML

From HTML to XML - 1

- A simple HTML document:

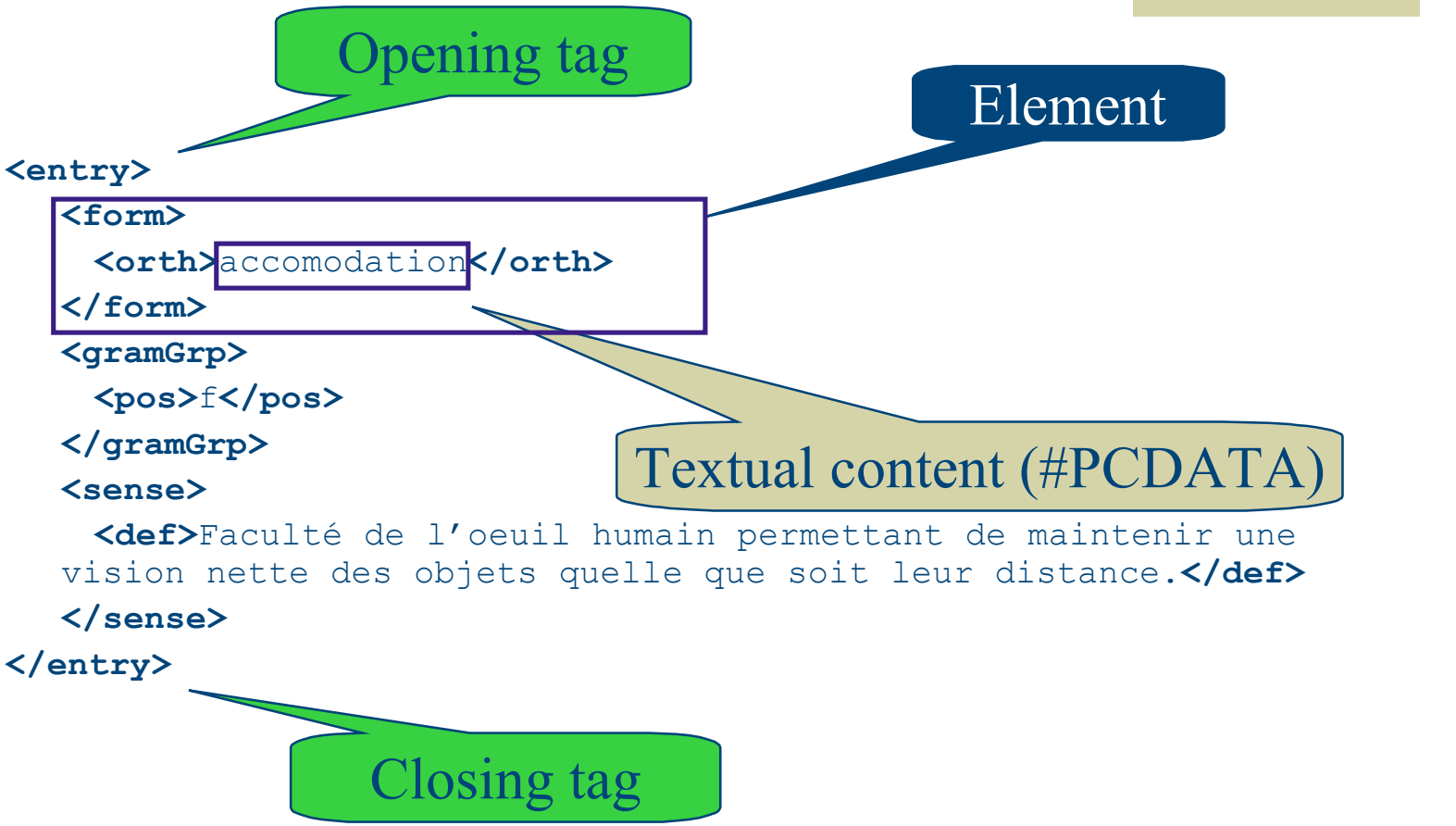
```
<p><b>Accomodation</b>,<i>f.</i>, faculté de
l'oeuil humain permettant de maintenir une
vision nette des objets quelle que soit leur
distance. <BR>
<b><I>En stéréoscopie</b></i>, faculté des
yeux humains d'obtenir la vision
stéréoscopique par superposition de deux
images.</p>
```

Accomodation,*f.*, faculté de l'oeuil humain permettant de maintenir une vision nette des objets quelle que soit leur distance.
En stéréoscopie, faculté des yeux humains d'obtenir la vision stéréoscopique par superposition de deux images.

From HTML to XML - 2

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE entry SYSTEM "http://…/entry.dtd">
<!-- A dictionary entry -->
<entry>
   <form>
     <orth>accomodation</orth>
   </form>
   <gramGrp>
     <pos>f</pos>
   </gramGrp>
   <sense>
     <def>Faculté de l'oeuil humain permettant de maintenir une
   vision nette des objets quelle que soit leur distance.</def>
   </sense>
   <sense>
     <usg type="dom">En STEREOSCOPIE</usg>,
     <def>faculté des yeux humains d'obtenir la vision stéréoscopique
   par superposition de deux images.</def>
   </sense>
</entry>
```

Elements and their content



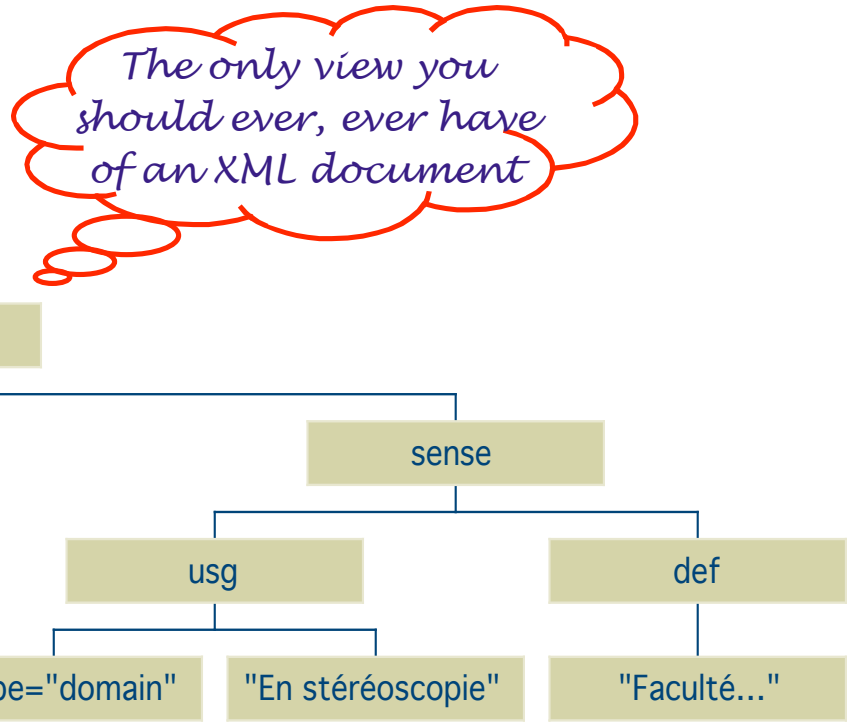
Elements and their attributes

```
<entry id="e25">
  <form>
    <orth lang="fr">accomodation</orth>
  </form>
  <gramGrp>
    <pos>f</pos>
  </gramGrp>
  <sense n="1">
    <def>…</def>
  </sense>
  <sense n="2">
    <usg type="dom">En STEREOSCOPIE</usg>,
    <def>Faculté des yeux humains…/def>
  </sense>
</entry>
```

Attribute value

Attribute name

XML to represent trees



Some properties of XML

- Underlying model: tree structure
  - Emphasis should be put on the “semantics” of a document
  - Possibility to imagine a script language to access any part of an XML document

    `e.g.: dictionary/entry[28]//orth/text()`
- XML supports Unicode/ISO 10646
  - Forbidden characters, expressed by means of XML *entities*:

    < → <

    & → &

Modeling linguistic annotation structures

Basic requirements - 1

- Expressive adequacy
  - represent all varieties of linguistic information
- Semantic adequacy
  - representation structures must have a formal semantics
- Incrementality
  - support for various stages of input interpretation and output generation
- Uniformity
  - representations utilize same “building blocks” and the same methods for combining them

Basic requirements - 2

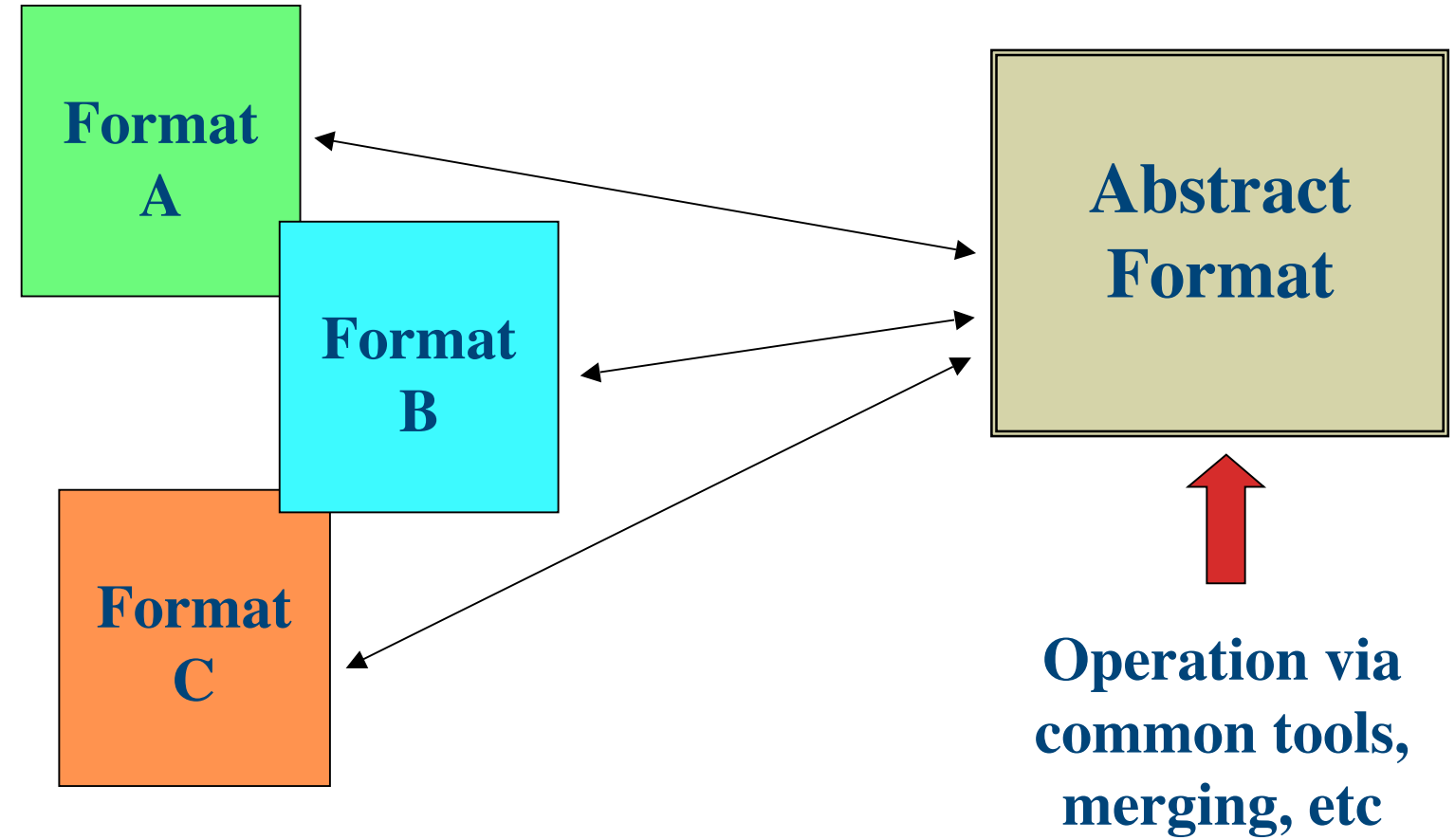
- Support for under-specification
  - representation of partial and intermediate results and ambiguities
- Openness
  - not dependent on a single linguistic theory
- Extensibility
  - compatible with alternative methods for designing representation schemas

The General Framework

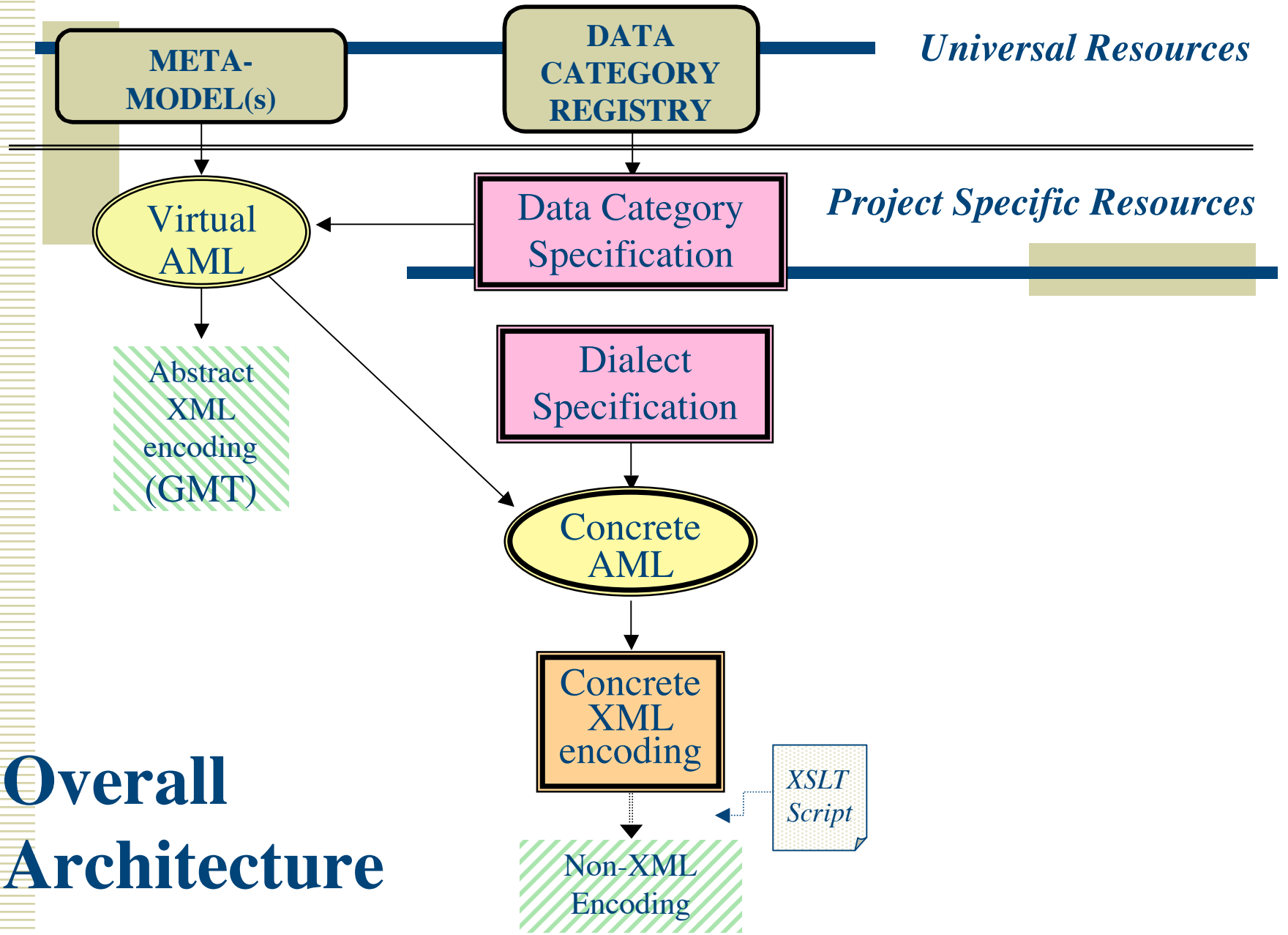
- Model for linguistic annotation that can
  - be instantiated in a standard representational format
    - GMT: Generic Mapping Tool
  - serve as a *pivot format* into and out of which proprietary formats may be transduced to enable
    - comparison
    - merging
    - manipulation via common tools

Overall Plan

Annotation Format Tower of Babel

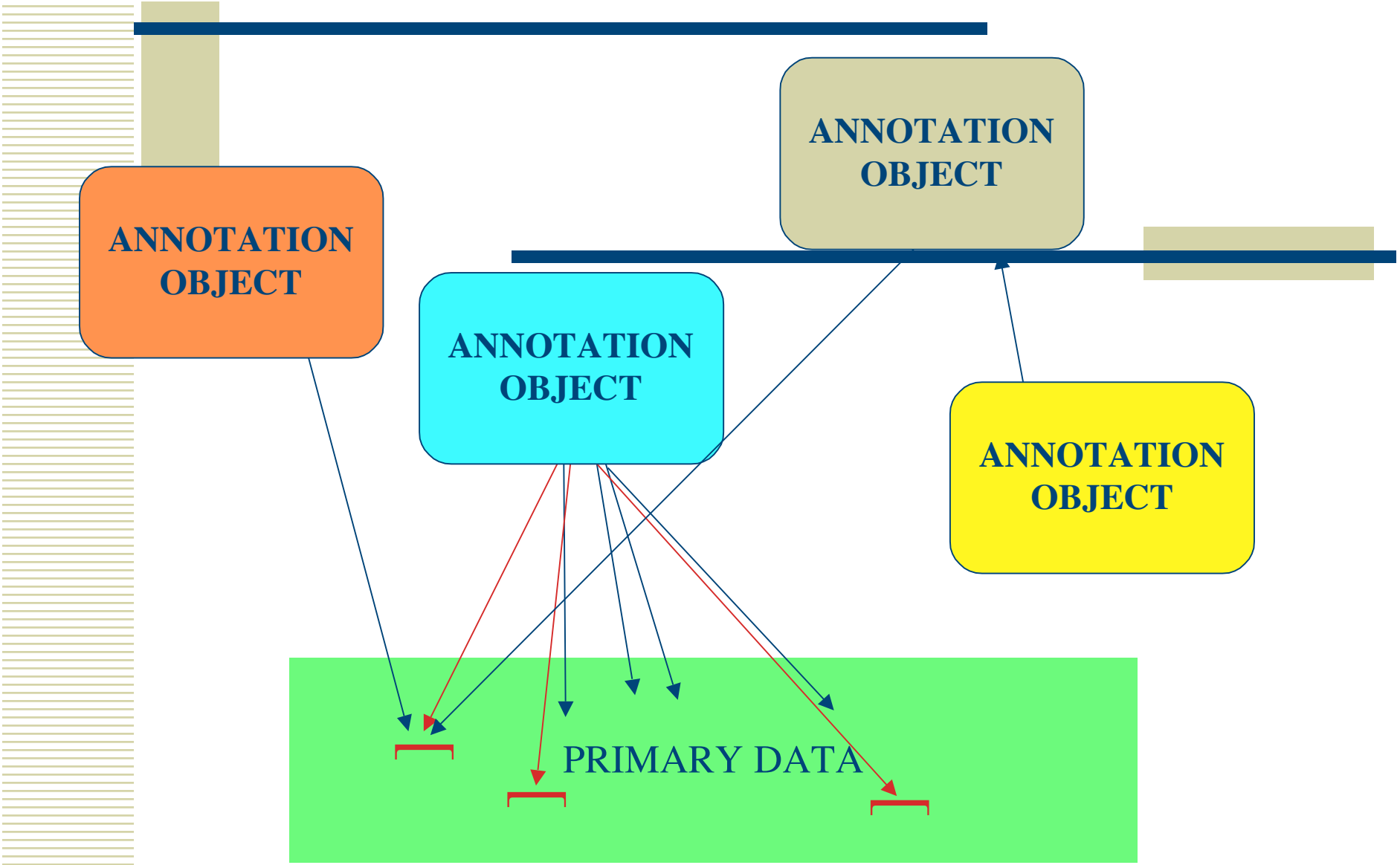


Overall Architecture



(Most) Abstract Model

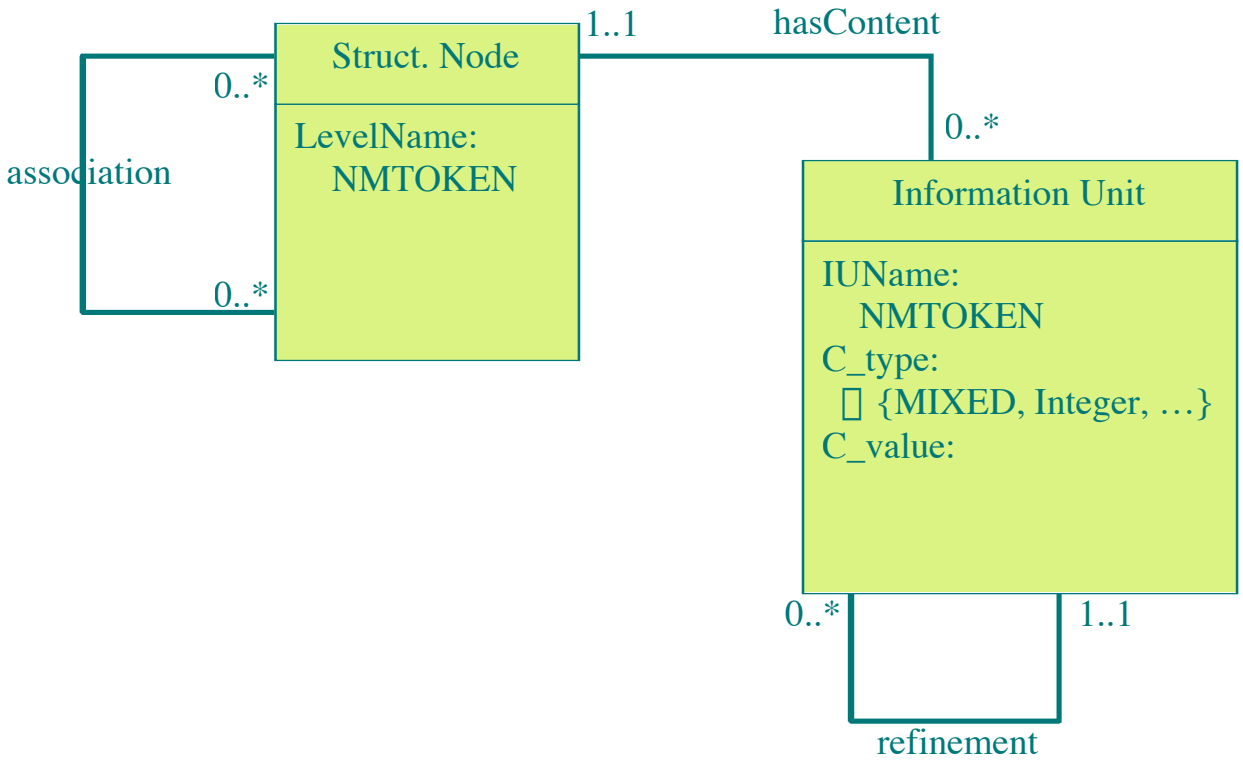
- An annotation is a set of data or information associated with some other data
- More precise: an annotation is a one- or two-way link between
  - an *annotation object*, and
  - a *point* or *span* (or a list/set of points or spans) within a "base" data set
- Links may or may not have a semantics
- Points and spans may be objects, or sets/lists of objects


ANNOTATION OBJECT
ANNOTATION OBJECT
ANNOTATION OBJECT
ANNOTATION OBJECT
PRIMARY DATA

Representing Annotation Objects

- Annotation objects may be relatively complex
- Abstract representation
  - graph of elementary *structural nodes* to which one or more *information units* are attached
  - distinction between structure and information units is critical to the design of a truly general model
- Annotations may be structured in several ways
  - Most common: hierarchical
    - phrase structure analyses of syntax
    - lexical and terminological information
    - etc.

UML representation



Main elements to consider

- A meta-model
  - A general, underlying model that informs current practice
- A set of data-categories
  - Provides to precise semantics of the format
  - Obtained:
    - By sub-setting a Data Category Registry
    - By providing application specific categories

Data category: example

- /Addressee/
  - Definition: The entity that is the intended hearer of a speech event. The scope of this data category is extended to deal with any multimodal communication event (e.g. haptics and tactile)
  - Source: (implicit) an event, whose /Event Type/ should be /Speak/
  - Target: a participant (user or system)
- ISO 12620-1: Background for representing data categories and deploying a data category registry
  - Will comprise language descriptions (ISO 639 series)

Additional mechanisms

Relations Among Annotations

Parallelism

two or more annotations refer to the same data object

Alternatives

two or more annotations comprise a set of mutually exclusive alternatives

Aggregation

two or more annotations comprise a list or set that should be taken as a unit

Dialect Specification

- Defines project-specific XML format
  - Use XML schemas, XSLT scripts, XSL style sheets
- Includes
  - *Instantiation styles*

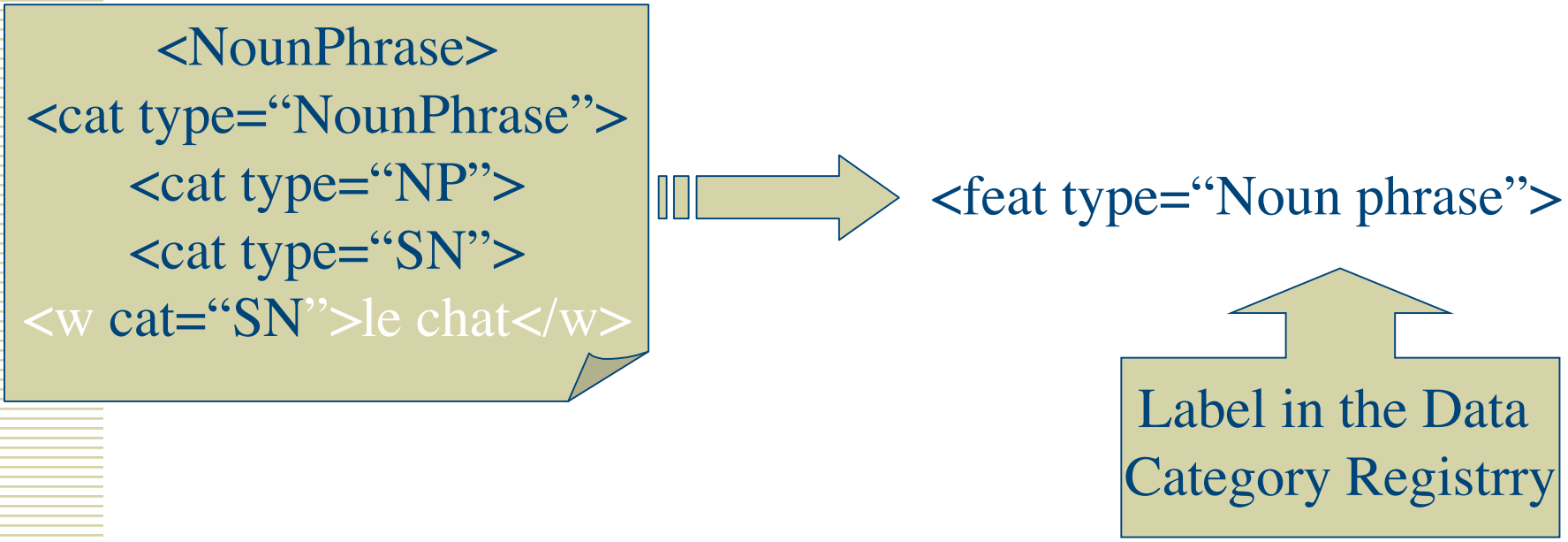
```
<NounPhrase>
<cat type="NounPhrase">
```

  - *Instantiation vocabulary*

```
<cat type="NounPhrase">
<cat type="NP">
<cat type="SN">
```

  - *Expansion structures*
    - ⇨ alter the structure expressed using the meta-model
    - ⇨ e.g. create two sub-nodes under a given node to group different types of information

A possible dump format: GMT (Generic mapping tool)



GMT in short (cont.)

- A simplified DTD
  - <struct>
    - represents a structural node in the annotation
  - <feat>
    - provides information attached to the node represented by the enclosing <struct>
  - <alt>
    - brackets alternative annotations
  - <brack>
    - groups information to be regarded as a single unit

Relating Annotation Levels

- Three ways:
  1. Temporal anchoring
     - ⇨ associates positional information with each structural level
  2. Event-based anchoring
     - ⇨ introduces a structural node to represent a location in the text to which all annotations can refer
  3. Object-based anchoring
     - ⇨ enables pointing from a given level to one or several structural nodes at another level

Temporal Anchoring

- Positional information
  - Usually, a pair of numbers expressing the starting and ending point of segment
- Example:

```
<struct type="phonetic">
  <feat
        type="phone"
        startsAt="2300"
        endsAt="3200">iy</feat>
</struct>
```

Event-based Anchoring

- Structural node (*landmark*) referred to by annotations for the defined span

```
<struct type="landmark">
    <seg startsAt="2300" endsAt="3200"/>
</struct>
```

- Annotation graph formalism explicitly designed for this

Object-based Anchoring

- Useful to make dependencies between two or more annotation levels explicit
  - Example: syntactic annotation can refer directly to the relevant nodes in a morpho-syntactically annotated corpus

Representation for “le chat” - 1

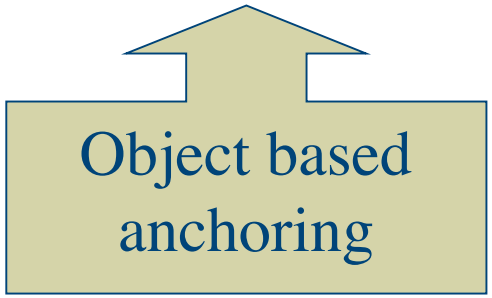
```
<!-- Syntactic level -->
<struct>
   <feat type=”synCat”>NP</feat>
   <seg targets=”#w1 #w2”/>
</struct>
```

```
<!-- Morphosyntactic level -->
<struct type=”W-level”>
   <seg target=”#w1”>
   <feat type=”lemma”>le</feat>
   <feat type=”pos”>DET</feat>
   <feat type=”gender”>masc</feat>
</struct>
<struct type=”W-level”>
   <seg target=”#w2”>
   <feat type=”lemma>chat</feat>
   <feat type=”pos”>NOUN</feat>
</struct>
```

Representation for “le chat” - 2

```
<!-- Syntactic level -->
<struct>
   <feat type="synCat">NP</feat>
   <seg targets="#s1 #s2"/>
</struct>
```

```
<!-- Morphosyntactic level -->
<struct type="W-level" id="s1">
   <seg target="#w1">
   <feat type="lemma">le</feat>
   <feat type="pos">DET</feat>
   <feat type="gender">masc</feat>
</struct>
<struct type="W-level" id="s2">
   <seg target="#w2">
   <feat type="lemma>chat</feat>
   <feat type="pos">NOUN</feat>
</struct>
```



Application 1: morpho-syntactic annotation

Linguistic Annotation (MuchMore)

Balint syndrom is a combination of symptoms including simultanagnosia, a disorder of spatial and object-based attention, disturbed spatial perception and representation, and optic ataxia resulting from bilateral parieto-occipital lesions.

```
<text>
 <token id="w1" pos="NN">Balint</token>
 <token id="w2" pos="NN">syndrom</token>
 <token id="w3" pos="VBZ" lemma="be">is</token>
 <token id="w4" pos="DT" lemma="a">a</token>
 <token id="w5" pos="NN" lemma="combination">combination</token>
 <token id="w6" pos="IN" lemma="of">of</token>
 <token id="w7" pos="NNS" lemma="symptom">symptoms</token>
 ...
 <token id="w20" pos="JJ" lemma="spatial">spatial</token>
 <token id="w21" pos="NN" lemma="perception">perception</token>
 ...

<chunks>
 <chunk id="c1" from="w1" to="w2" type="NP"/>
 <chunk id="c7" from="w20" to="w23" type="NP"/>
</chunks>
```

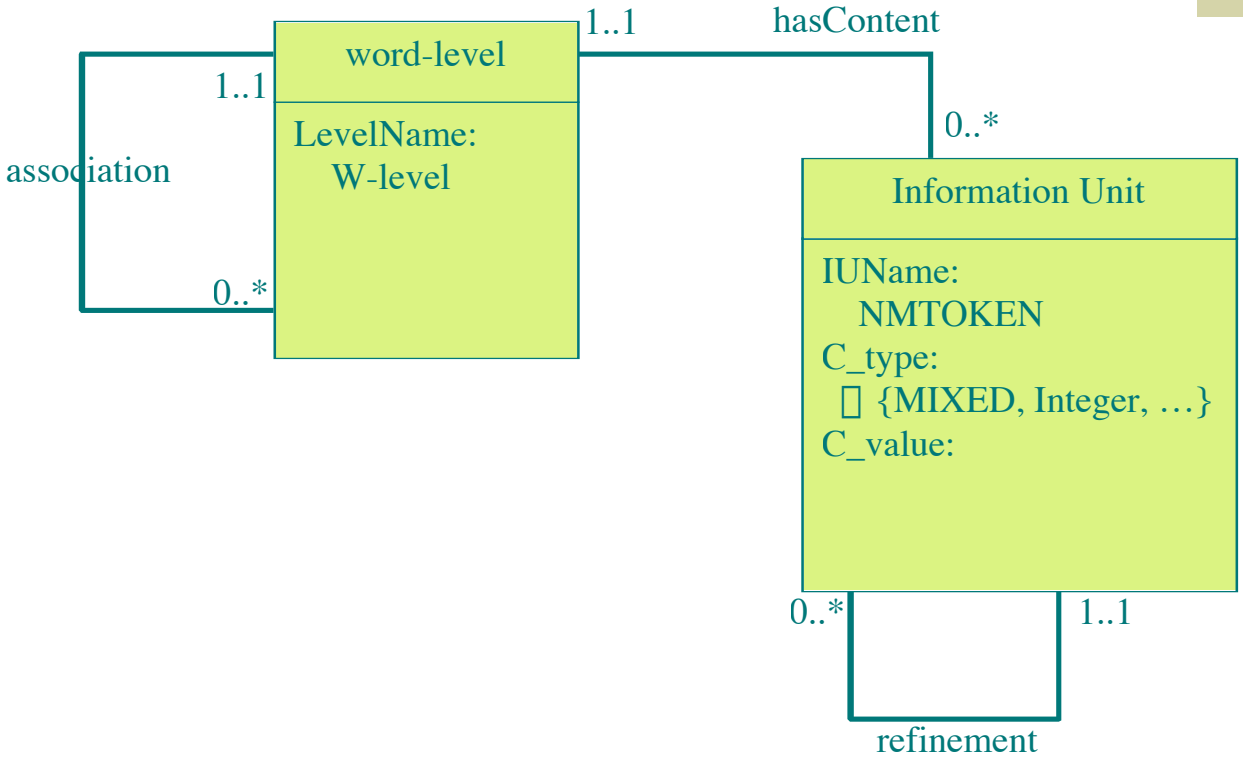
Principles

- Morpho-syntactic annotation
  - involves the identification of word classes over a continuous stream of word tokens
  - may refer to the segmentation of the input stream into word tokens or morphemes
  - may also involve grouping together sequences of tokens or identifying sub-token units (or morphemes)
  - description of word classes may include one or several features
    - ⇨ syntactic category, lemma, gender, number,…

Data model of morpho-syntactic annotation

- Single type of structural node
  - Represents a word-level structure unit
  - Organized as a hierarchy
- One or several information units associated with each structural node
- POS structures can be embedded into higher level structures
  - In general, sentences

UML representation for morpho-syntactic annotation



Rem.: sentence level is missing here.

Simple Case

"Paul aime les croissants"

```
<struct type="S-level">
   <struct type="W-level">
          <feat type="lemma">Paul</feat>
          <feat type="pos">PNOUN</feat>
          <seg target="#w1"/>
   </struct>
   <struct type="W-level">
          <feat type="lemma">aimer</feat>
          <feat type="pos">VERB</feat>
          <feat type="tense">present</feat>
          <feat type="person">3</feat>
          <seg target="#w2"/>
   </struct>
   <struct type="W-level">
          <feat type="lemma">le</feat>
          <feat type="pos">DET</feat>
          <feat type="number">plural</feat>
          <seg target="#w3"/>
   </struct>
   <struct type="W-level">
          <feat type="lemma">croissant</feat>
          <feat type="pos">NOUN</feat>
          <feat type="number">plural</feat>
          <seg target="#w4"/>
   </struct>
</struct>
```

Pointers to data in primary document

Representing More Complex Cases

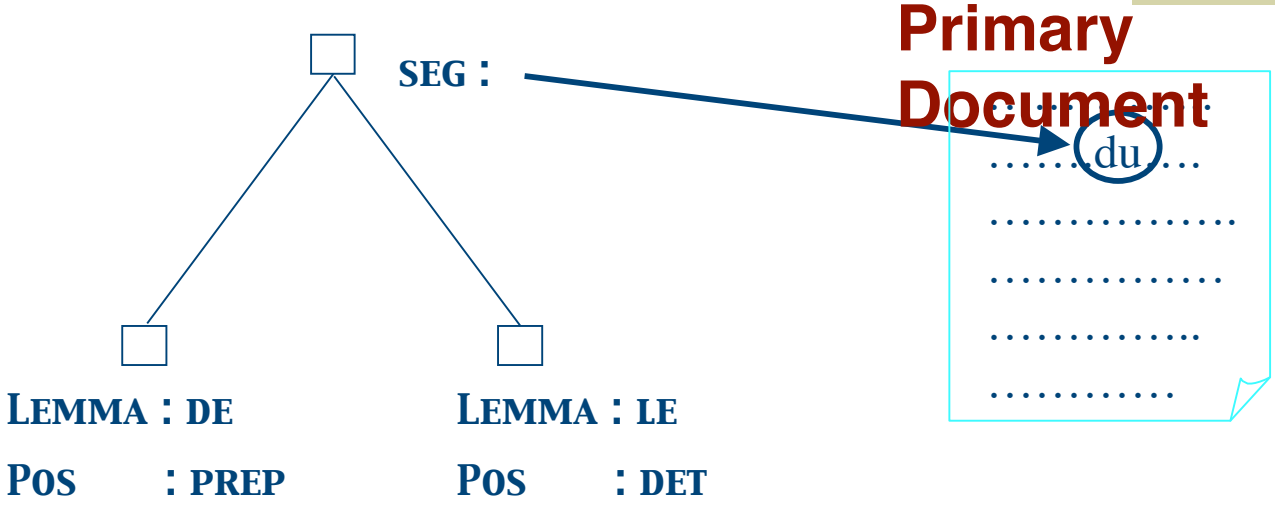
Example: "du" = "de" + "le" in French

Points to "du" in text

```
<struct type="W-level">
    <seg target="#w1"/>
    <struct type="W-level">
        <feat type="lemma">de</feat>
        <feat type="pos">PREP</feat>
    </struct>
    <struct type="W-level">
        <feat type="lemma">le</feat>
        <feat type="pos">DET</feat>
    </struct>
</struct>
```

Gives the structure of the "words" underlying the word

GMT as a Tree Structure



Compound Words

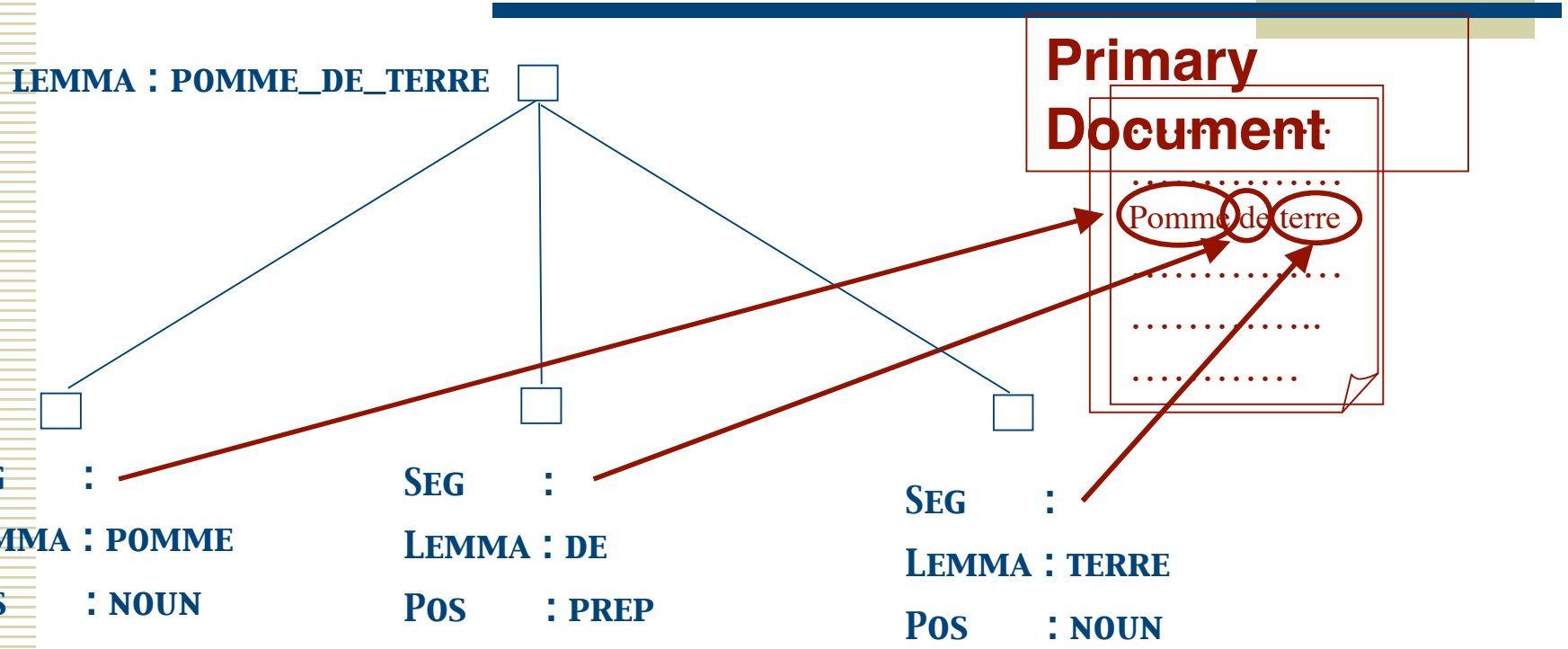
Example: “pomme de terre”

Primary lemma

Component lemmas

```
<struct type="W-level">
   <feat type="lemma">pomme_de_terre</feat>
   <feat type="pos">NOUN</feat>
   <struct type="W-level">
      <seg target="#w1"/>
      <feat type="lemma">pomme</feat>
      <feat type="pos">NOUN</feat>
   </struct>
   <struct type="W-level">
      <seg target="#w2"/>
      <feat type="lemma">de</feat>
      <feat type="pos">PREP</feat>
   </struct>
   <struct type="W-level">
      <seg target="#w3"/>
      <feat type="lemma">terre</feat>
      <feat type="pos">NOUN</feat>
   </struct>
</struct>
```

Tree

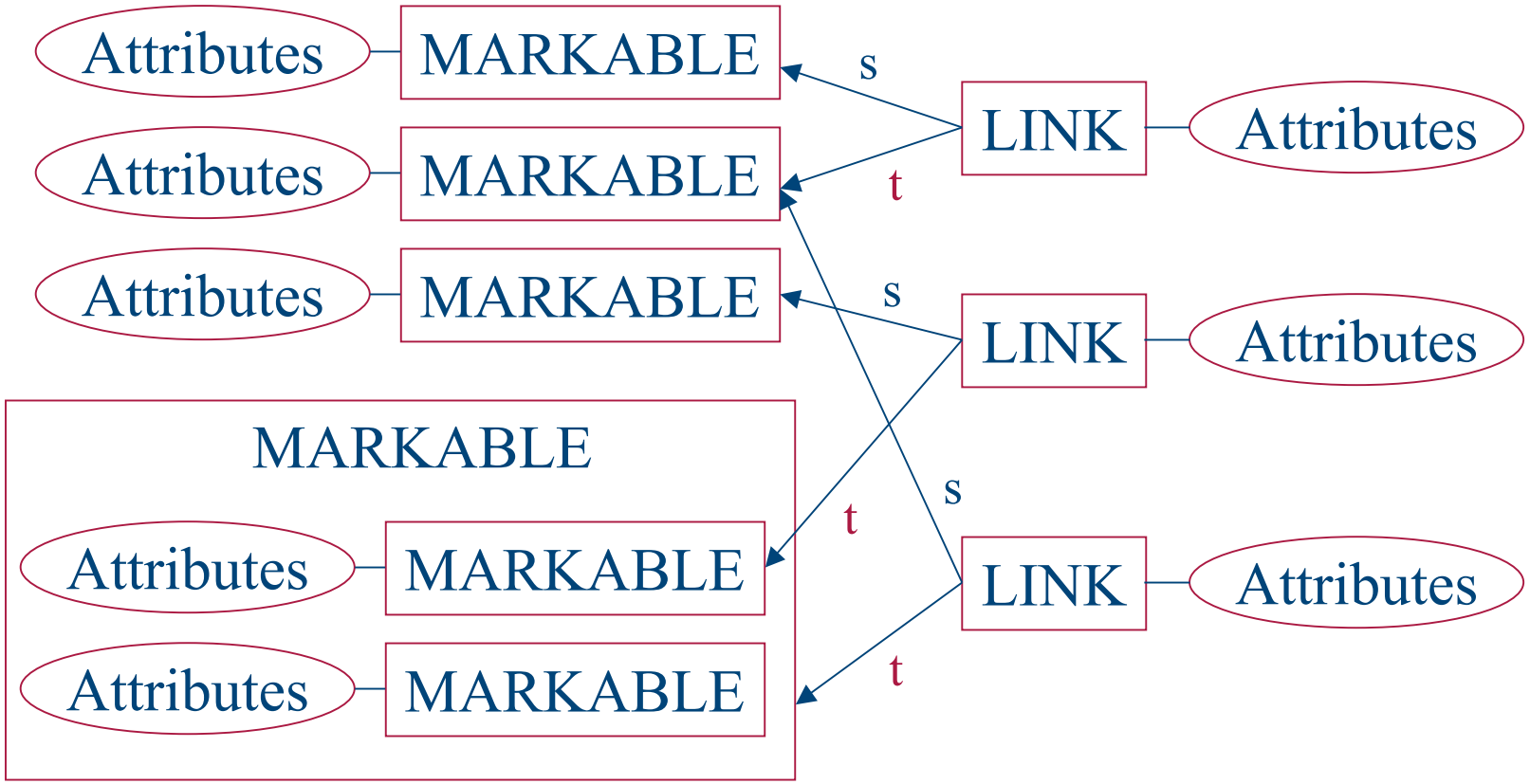

LEMMA : POMME_DE_TERRE
Primary Document
Pomme de terre
SEG :
LEMMA : POMME
POS : NOUN
SEG :
LEMMA : DE
POS : PREP
SEG :
LEMMA : TERRE
POS : NOUN

Application 2: reference annotation

Requirements

- keep the stand-off annotation principle
  - separate different annotation levels
- provide an explicit markable level
  - allow for any kind of markables
    - surface strings, morphological units, syntactic chunks, ...
    - grouping items into discontinued markables
      - disjoint antecedents (*cats... dogs... the animals*)
  - allow for adding relevant attributes to markables
- expressing links
  - typed relation between two markables

Basic structure



Simple case

a dog ... the animal

```
<struct id="m_1" type="markable" >
      <feat type="source text" > a dog </feat>
      <feat type= "Category">indefinite_NP</feat>
      ...
</struct>
<struct id="m_2" type="markable"
       <feat type= "source_text"> the animal </feat >
      <feat type= "Category">definite_NP</feat>
</struct>
```

MARKABLES

```
<struct id="l_1" type="ref_link" >
      <feat type="ref_link_type">coreference </feat>
      <feat type="source" target ="m_2" >
      <feat type="target" target ="m_1" >
</struct>
```

LINK

Recursive markables

a dog ... a cat... the animals

```
<struct id="m_3" type="markable" >
    <struct id="m_1" type="markable">
        <feat type ="source_text" >a dog </feat>
        <feat type= "Category">indefinite_NP</feat>
    </struct>
    <struct id="m_2" type="markable" >
        <feat type ="source_text" >a cat</feat>
        <feat type= "Category">indefinite_NP</feat>
    </struct>
</struct>

<struct id="m_4" type="markable" >
    <feat type ="source_text" >the animals</feat>
    <feat type= "Category">definite_NP</feat>
</struct>
                    <struct id="l_1" type="ref_link" >
                        <feat type="ref_link_type">coreference </feat>
                        <feat type="source" target ="m_4" />
                        <feat type="target" target ="m_3" />
                    </struct>
```

Ambiguity for antecedents

a dog ... a cat... the animal

```
<struct id="m_1" type="markable"  >
      <feat type ="source_text" > a dog </feat>
      <feat type= "Category">indefinite_NP</feat>
</struct>
<struct id="m_2" type="markable" >
      <feat type ="source_text" > a cat </feat>
      <feat type= "Category">indefinite_NP</feat>
</struct>
<struct id="m_3" type="markable" >
      <feat type ="source_text" > the animal </feat>
      <feat type= "Category">definite_NP</feat>
</struct>
```

```
<struct id="l_1" type="ref_link" >
    <feat type="ref_link_type">coreference </feat>
    <feat type="source" target ="m_3" >
    <alt>
            <feat type="target" target ="m_1" />
            <feat type="target" target ="m_2" />
    </alt>
</struct>
```

Conclusion

- You should not bother about standards
  - A good methodology should ease a progressive switch to future international standards
    - But you may not avoid XML
- You should bother about standardization
  - Make sure that your annotation scheme is precisely defined and documented so that it can be shared with other researchers and groups
    - You may even want to contribute to international standardization initiatives