AN ONTOLOGY FOR THE TEI (SIMPLE): ONE STEP BEYOND

F. Ciotti, F. Tomasi, S. Peroni, F. Vitali

TEI Conference 2015

IN PREVIOUS EPISODES ...

- This paper is directly connected with Ciotti and Tomasi presentation at the TEI Meeting 2014
- There we envisioned the adoption of ontological modeling as a mean to define formally the semantics of TEI markup constructs and content.
- In this paper we present the first outcomes of the work we have done in the last months, after the research team has been extended to include Fabio Vitali and Silvio Peroni.

TOPICS OF THIS TALK

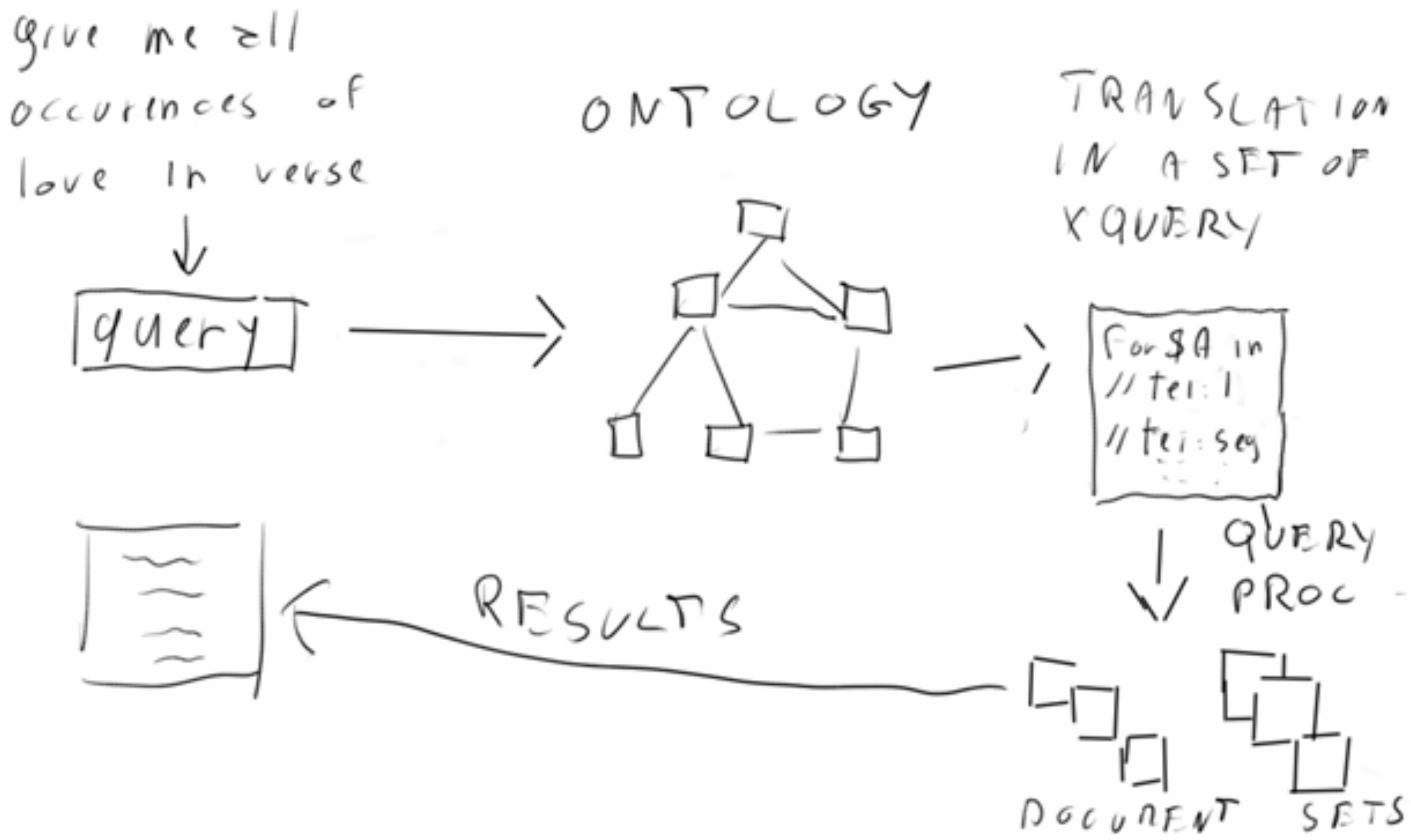
- Why: why do we think that the idea of giving TEI a formal semantic is a good idea, how it could enhance its usefulness for the community and expressive power?
- What: what in the TEI do we really think is conceivable to formalize in the form of an ontology? How far can we imagine to go in this direction?
- How: which are the better formal strategies to build a formal model of the TEI subset we have identified in the step before?

WHY ONTOLOGIZING

- Pragmatic and technical reasons
  - enabling parsers to perform both syntactic and semantic validation of document markup;
  - inferring facts from documents automatically by means of inference systems and reasoners;
  - simplifying the federation, conversion and translation of documents marked up with different markup vocabularies;
  - allowing users to query upon the structure of the document considering its semantics;
  - creating visualizations of documents by considering the semantics of their structure rather than the specific vocabulary in which they are marked up;
  - increasing the accessibility of documents' content, even in the case of tag abuse, i.e., "using markup languages construction in ways other than intended by the language designer";
  - promoting a more flexible software design for those applications that use markup languages, guaranteeing a better maintainability even when markup language schemas evolve.

WHY ONTOLOGIZING

- The advantages in this list are not TEI specific
- Some of the issues have special relevance for TEI
- Take for instance the query issue
- Having a set of ontological definitions of the conceptual level behind markup, that is a set of shared formal definitions of the textual features to which any single encoding project could bind idiosyncratic markup usage (keeping safe the need and the right to fine tune the encoding at local level), could solve this problem


give me all occurences of love in verse
query
ONTOLOGY
TRANSLATION IN A SET OF XQUERY
QUERY PROC
DOCUMENT SETS
RESULTS

```
for $A in
//tei:l
//tei:seg
....
```

WHY ONTOLOGIZING

- This result could be in principle attained adopting different technologies, but a semantic approach has some exclusive pros:
    - it is implementation independent
    - it is more expressive than the average ad hoc solutions
    - can take advantage of inference engines capabilities to extend or refine the query without previous knowledge of the details of encoding practices
- An extension of this argument is the possibility to define relationship between TEI data sets and other data and metadata models and languages, working directly at the abstraction level of ontology and not at the level of the document

WHY ONTOLOGIZING

- Linked Data extraction
    - an ontology to assign one and the same semantics to (quasi)-homonymic markup constructs like <rs type=”person”> or <persName> or even <seg type=”persName”>
    - and a bunch of owl:equivalentClass properties that bind our TEI:personal-name concept with foaf:name or lawd:PersonalName concepts
- seems to make the living easy: we can happily jump into the LOD cloud extracting data sets from our richly encoded documents

WHY ONTOLOGIZING

- Deeper theoretical and foundational advantage
  - The very core of digital methods application in humanities research is the notion of model/modeling
  - As far as we are using Turing machine like device for computation, the only workable notion of modeling is a formal one: model we should be interested in are formal models
  - Formalization is a set of semiotic and representational processes that generates a representation of a (or a set of) phenomenon/object algorithmically accessible and computable

WHY ONTOLOGIZING

- TEI is not only a markup facility but first and foremost a conceptual **model** of textuality.
- In fact, in the *Guidelines* we can find an explicit statement asserting this, when in chapter 23 we find the the concept of "TEI abstract model":
  - The *TEI Abstract Model* is the conceptual schema instantiated by the TEI Guidelines. These Guidelines define, both formally and informally, a set of abstract concepts such as 'paragraph' or 'heading', and their structural relationships, for example stating that 'paragraph's do not contain 'heading's. These Guidelines also define classes of elements, which have both semantic and structural properties in common. Those semantic and structural properties are also a part of the TEI Abstract Model
  - It is an important condition of TEI conformance that elements defined in the TEI Guidelines as having one specific meaning should not be used with another… The semantics of elements defined in the TEI Guidelines are conveyed in a number of ways, ranging from formally verifiable datatypes to informal descriptive prose

WHY ONTOLOGIZING

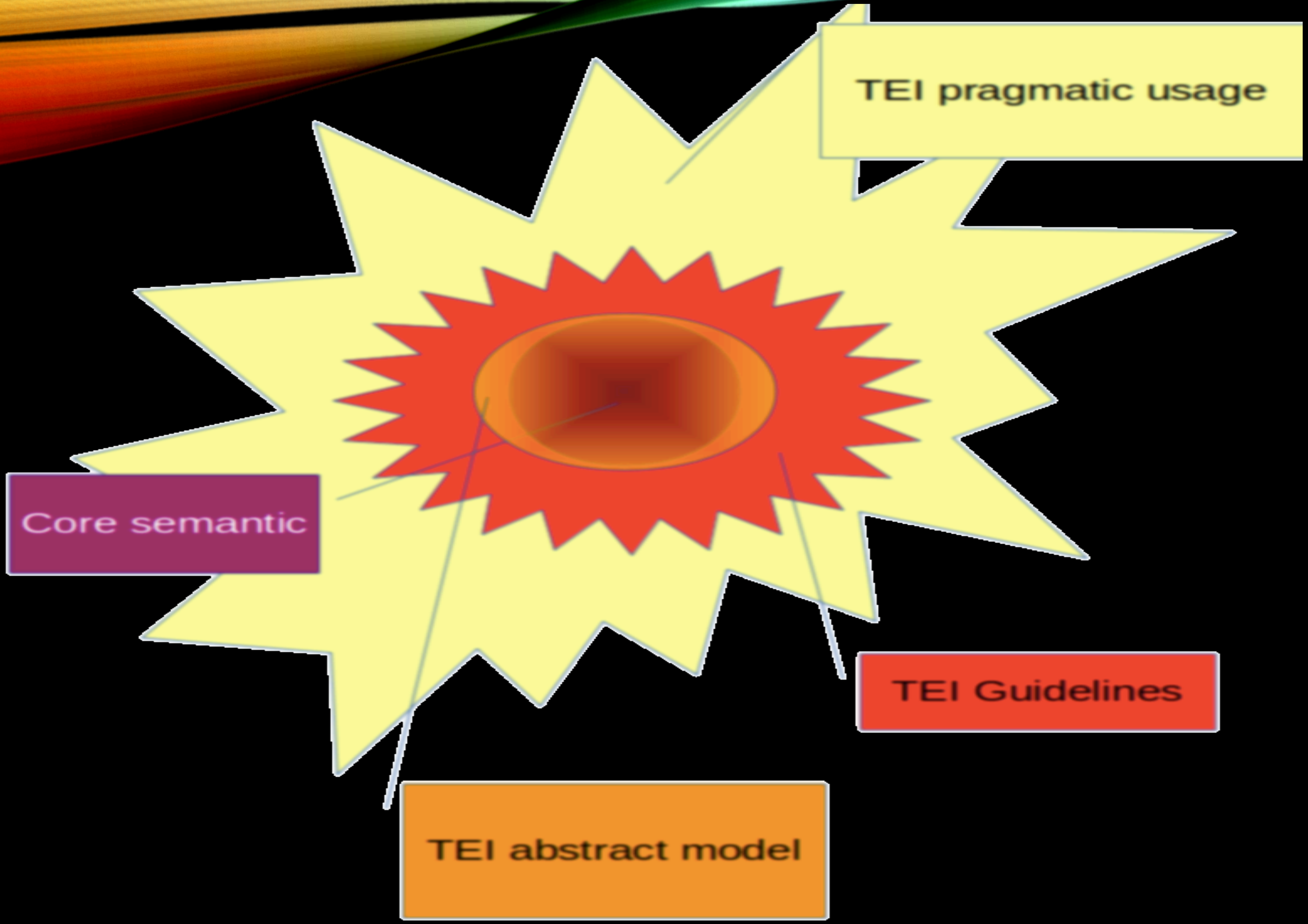
- We have the notion of an abstract model and this notion is used in many formal procedures but this very notion is *not* formally defined
- I suggest that we need to move the quasi-formal notion of TEI abstract model to a formal ontology, if it has to be of any use other than a sort of regulatory principle
- our suggestion to adopt contemporary Semantic Web formalisms to build this abstract conceptual model give us the possibility to have a “foundation” of TEI in a well-defined data model that is not dependent on the notion of a single hierarchical structure (OHCO), and that can accommodate at least to some extent the “pluralities” of textuality

WHY ONTOLOGIZING

- A corollary is that the approach we are suggesting can represent an **operational** solution for some long standing debates and controversies, like the good ole XML/non XML issue and the problem of interoperability/ interchange
- The definition of TEI abstract model as formal ontology, independent from serialization formats, could become the real TEI
- A set of serialization can be derived algorithmically in any language of choice

WHAT ONTOLOGIZING

- In our last year paper, and in what I have been saying until now, there is an optimistic stance: we seem to believe that the TEI as whole can be reduced to a formal ontology without residue…
- We were far too optimistic!!
  - TEI is huge and diverse, the result of decades of work, refinements, extensions, additions, it covers in details many different areas of application
  - Second, TEI real usage in the community is largely oriented by **pragmatic** factors. That is the intended meaning of the markup in concrete markup act is determined by the circumstances of usage, the context in which the markup happens, the presuppositions of the encoder himself



TEI pragmatic usage

Core semantic

TEI Guidelines

TEI abstract model

WHAT ONTOLOGIZING

- I think that is impossible to reduce to a unique formal semantic definition this cloud
- We can identify a subset of shared assumptions, a common ground of notions about the meaning of TEI markup and the nature of documents like object: this level is what in philosophy of language is referred to as **semantics**
- This subset can be the object of an ontological formalization

WHAT ONTOLOGIZING

- We can “prima facie” take the TEI Simple customization element set as an approximation of this common ontology.
- This is not an opportunistic *ad hoc* choice, as it may seem. TEI Simple in fact has been defined by a group of domain expert that
  - have analyzed the actual usage of markup in some big textual repositories
  - have selected and partially organized a set of one hundred or so elements that can describe all the textual features represented by TEI markup in those documents
- This process fits perfectly in the definition of formal ontology

WHAT ONTOLOGIZING

castlist

<actor> <castGroup> <castItem> <castList> <role> <roleDesc>

character

<g>

editorial

<abbr> <add> <addSpan> <am> <choice> <corr> <del> <desc> <ex> <expan> <gap> <handShift> <orig> <reg> <sic> <space> <subst> <supplied> <unclear>

header

<att> <biblFull> <biblScope> <biblStruct> <change> <charDecl> <charProp> <editor> <editorialDecl> <email> <encodingDesc> <extent> <fileDesc> <gi> <glyph> <glyphName> <idno> <imprint> <keywords> <licence> <listChange> <listPerson> <localName> <monogr> <msDesc> <msIdentifier> <person> <physDesc> <profileDesc> <publicationStmt> <relatedItem> <repository> <resp> <respStmt> <sourceDesc> <tag> <teiHeader> <term> <textClass> <textDesc> <titleStmt> <typeDesc> <val> <value>

interpretation

<author> <date> <foreign> <hi> <measure> <name> <num> <q> <quote> <ref> <rhyme> <rs> <seg> <time>

linguistic

<c> <pc> <s> <w>

pictures

<figDesc> <figure> <graphic>

structure

<ab> <address> <addrLine> <anchor> <back> <bibl> <body> <cb> <cit> <div> <floatingText> <formula> <front> <fw> <group> <head> <item> <l> <label> <lb> <lg> <list> <listBibl> <milestone> <note> <p> <pb> <sp> <speaker> <spGrp> <stage> <TEI> <teiCorpus> <text> <title>

table

<cell> <row> <table>

titlepage

<docAuthor> <docDate> <docEdition> <docImprint> <docTitle> <imprimatur> <publisher> <pubPlace> <titlePage> <titlePart>

wrapper

<argument> <byline> <closer> <dateline> <epigraph> <opener> <postscript> <salute> <signed> <trailer>

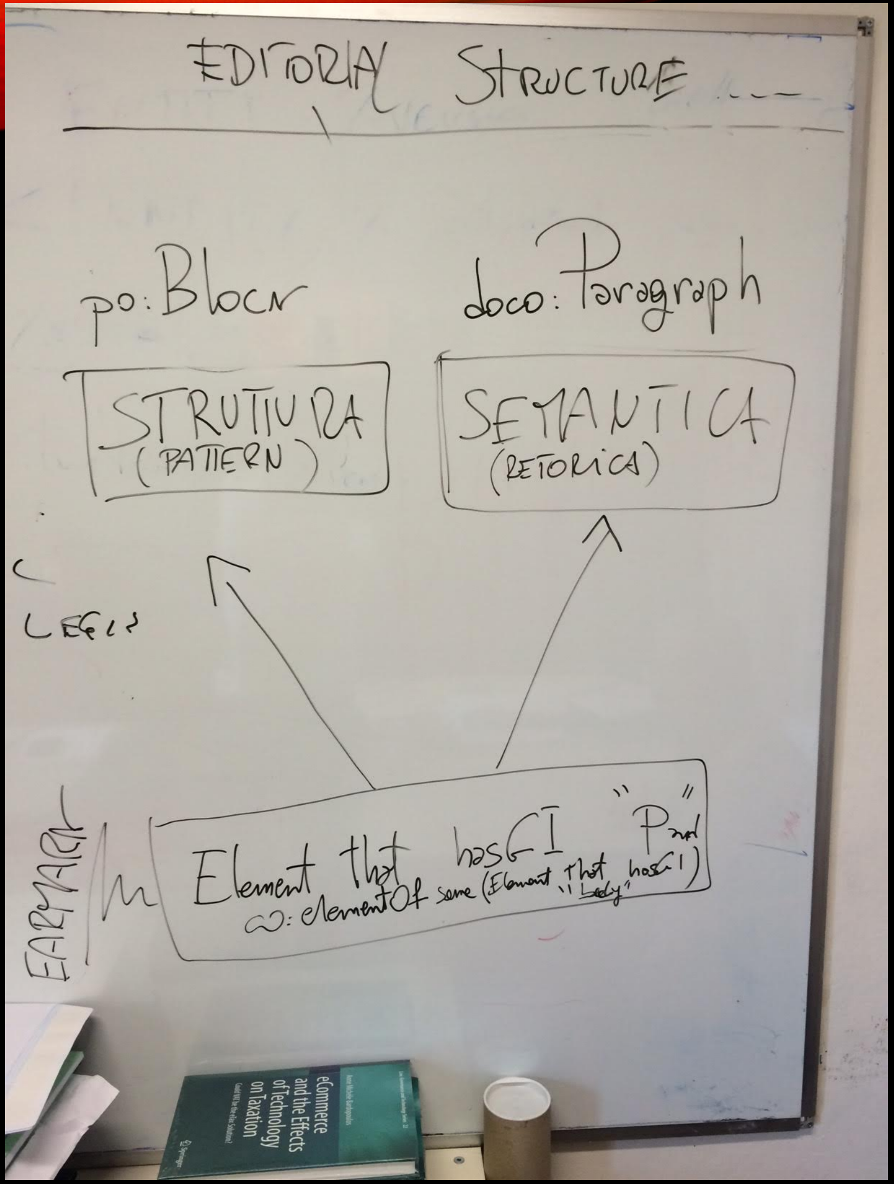
HOW ONTOLOGIZING: REQUIREMENTS

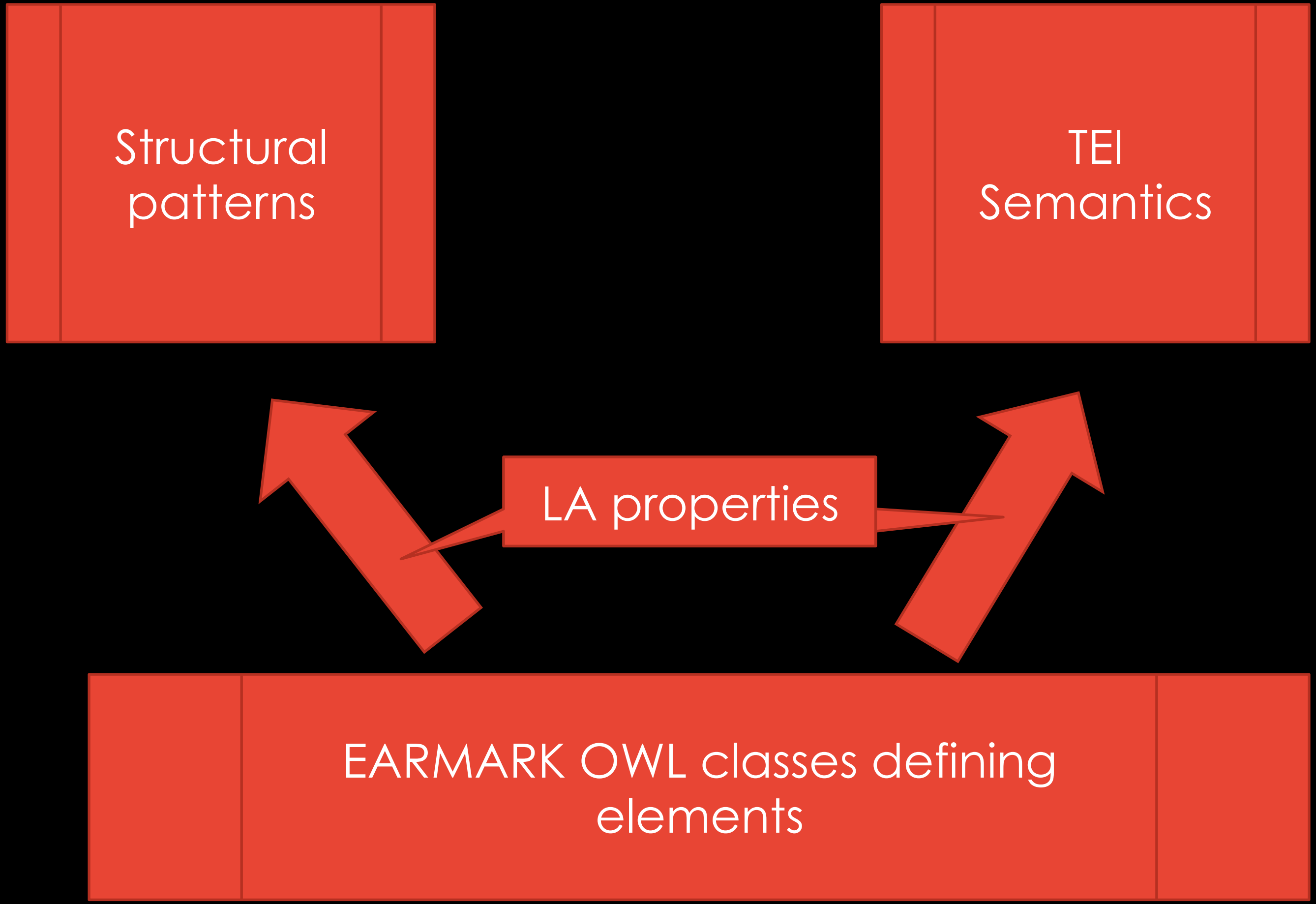
- the ontology should express at the same time an abstract characterization of TEI (Simple) elements' semantics and an ontological definition of their structural role
- the ontology should define a precise semantics of the elements having a clear characterization in the official TEI documentation (e.g., the element <p>), while it should relax the semantical constraints if the elements in consideration can be used with different semantic connotations depending on the context (e.g., the element <seg>)
- it should be possible to extend the ontology, reuse it and define alternative characterizations of elements semantics without compromising the consistency of the ontology itself
- where possible existing ontologies or meta-ontologies should be reused

HOW ONTOLOGIZING

- In accordance with these overall principles we have implemented a complex architecture using some pre-existing meta-ontology frameworks to express the meaning of TEI elements by the way of the relations and properties they define:
  - **LA-Earmark**, a markup metalanguage, that can express both the syntax and the semantics of markup as OWL assertions, and an ontology of markup that make explicit the implicit assumptions of markup languages. LA-EARMARK is an extension of EARMARK with the Linguistic Act module of the Linguistic Meta-Model that allows one to express and assess facts, constraints and rules about the markup structure as well as about the inherent semantics of the markup elements themselves.
  - **Structural Pattern Ontology**, whose goal is to identify a small number of patterns that are sufficient to express how the structure of digital documents can be segmented into atomic components.

HOW ONTOLOGIZING




Structural patterns

TEI Semantics

LA properties

EARMARK OWL classes defining elements

HOW ONTOLOGIZING

- The specification of markup semantics for TEI Simple elements is done by means of LA-EARMARK class and properties.
- The general Earmark class for any markup element is earmark:Element. The <abbr> element is defined as follows (in Manchester Syntax):

```
Prefix earmark: <http://www.essepuntato.it/2008/12/earmark#>
Prefix co: <http://purl.org/co/>
Prefix tei: <http://www.tei-c.org/ns/1.0/>

Class: tei:abbr a
       earmark:Element that
             earmark:hasGeneralIdentifier "abbr" and
             earmark:hasNamespace "http://www.tei-c.org/ns/1.0"
```

HOW ONTOLOGIZING

- We need to create restrictions in order to identify and characterize possible subsets of elements described by the schema
- For instance, the class of all the element <tei:p> that occurs inside <tei:text> and not inside <tei:teiHeader>

```
Class: tei:pInBody
   EquivalentTo:
    earmark:Element that
     earmark:hasGeneralIdentifier "p" and
         earmark:hasNamespace "http://www.tei-c.org/ns/1.0" and
         co:elementOf some (
                earmark:Element that
                earmark:hasGeneralIdentifier "body" and
                earmark:hasNamespace "http://www.tei-c.org/ns/1.0")
```

HOW ONTOLOGIZING

- LA-EARMARK allows us to link particular class of elements with the actual semantics they express
- There are two semantic levels that we explicitly define:
  - one concerning the structural behavior of markup (e.g., the fact that an element is a block rather than an inline, a container rather than a field), that can be described by means of Pattern Ontology
  - the other regarding the intended semantics of an element (e.g., the fact that an element is a paragraph rather than a section, a personal name reference rather than a geographical reference), that can be described by TEI Semantics Ontology or by a combination (and/or an extension) of already existing ontologies

HOW ONTOLOGIZING

- TEI Semantics Ontology is the core component that gives the actual semantics of TEI elements.
- Its definition is based on the categorization of elements in the TEI Simple we have seen before, that constitutes the upper level classes.
- The lower level classes are the concepts expressed by TEI construct. There is not a one to one relation between elements and lower level semantic classes, since we have identified at least three different markup “crystals” that can have a different ontological meaning:
  - one XML element: ex. <abbr> means teiOnt:Abbreviation
  - an XML element/attribute couple: ex. <corr resp=>
  - one XML element in a given context: ex. <p> in <text> vs <p> in <teiHeader>
- The middle level classes are derived from the TEI model Classes of so called “like” type

HOW ONTOLOGIZING

- The link between the classes describing kinds of elements and their related semantic characterization is expressed by the LA property "semiotics:expresses“ and OWL 2 class punning

```
Prefix teiOnt: <http://purl.org/spar/teiOnt/>
Prefix semiotics: <http://www.ontologydesignpatterns.org/cp/owl/
semiotics.owl#>
Prefix tei: <http://www.tei-c.org/ns/1.0/>
Individual: tei:pInBody
   Facts:
         semiotics:expresses teiOnt:Paragraph
```

FURTHER STEPS…

- Refine and factorize TEI Simple Semantics Ontology component of our model, that is related to the TEI model class structure
- Extend to some other areas of TEI that are suitable for formalization. Simple is not all, and with appropriate time and work force the ontology can be extended to some other areas of TEI
- In the long term this formalism could evolve to became the real formalization of TEI, independent of any serialization
- XML is still the better strategy to encode digital texts in real word projects for many practical reasons. But there is no reason for the TEI to be strictly based on it, as it is *de facto* now. Technical or pragmatic issues should not determine the choice of a formalization

We believe that our effort can give a contribution to the TEI to envision the shape of its own future.

Thank you for your attention!

fabio.ciotti@uniroma2.it