Journal of the Text Encoding Initiative

Issue 8 (December 2014 - December 2015)
Selected Papers from the 2013 TEI Conference

Fabio Ciotti, Maurizio Lana and Francesca Tomasi

TEI, Ontologies, Linked Open Data: Geolat and Beyond

Electronic reference
Fabio Ciotti, Maurizio Lana and Francesca Tomasi, « TEI, Ontologies, Linked Open Data: Geolat and Beyond », *Journal of the Text Encoding Initiative* [Online], Issue 8 | December 2014 - December 2015, Online since 20 November 2015, connection on 28 November 2015. URL : http://jtei.revues.org/1365 ; DOI : 10.4000/jtei.1365

Publisher: Text Encoding Initiative Consortium
http://jtei.revues.org
http://www.revues.org

Document available online on: http://jtei.revues.org/1365
This PDF document was generated by the journal.

TEI, Ontologies, Linked Open Data: Geolat and Beyond

Fabio Ciotti, Maurizio Lana, and Francesca Tomasi

AUTHOR'S NOTE

This research is supported by Compagnia di San Paolo Foundation. Fabio Ciotti wrote sections 3.1, 3.2, and 4; Maurizio Lana wrote sections 1, 2, and 5; Francesca Tomasi wrote sections 1, 3.3, 3.4, and 3.5; the authors, however, share with each other and with others (Diego Magro, Silvio Pieroni, and Fabio Vitali) the whole scientific responsibility for the project.

1. Introduction

1 The dialogue that TEI started with other semantic models has two aims: data and document interchange as well as the improvement of the editors' ability to formally declare hermeneutical positions. The TEI schemas include most of the elements and attributes (and now classes as well) to provide interpretations, while other non-TEI schemas, such as the EAC-CPF (Encoded Archival Context—Corporate Bodies, Persons and Families) metadata element set, may be employed to

enhance or augment the TEI model. On the one hand, additional schemas could contribute to perfecting the scope of some TEI elements; on the other hand, other semantic models, such as ontologies, could improve the effectiveness of interpretations.

2 Speaking about models means reflecting on the relationships between TEI and ontologies, but also moving towards new emerging semantic paradigms. People working with texts today have some interesting opportunities that the editorial domain is encouraging, such as: describing or enriching texts through TEI encoding; formalizing their knowledge of textual content(s) by using one or more ontologies; offering open access to their work in order to guarantee dissemination; and making their data interoperable by means of linked open data.

3 Moreover, we could say that these options exist because digital libraries are becoming increasingly widespread. The availability of collections has stimulated, as a natural result, the growth of new "movements." Actually, without digital text collections, these opportunities would be mostly theoretical; however, we should also consider the resulting question that arises: What can we do with digital texts that cannot be done with non-digital texts? Each of the opportunities mentioned above entails interesting implications:

1. TEI encoding is sufficiently flexible to allow text to incorporate annotation using many different approaches, including structural, syntactic, and linguistic. Many diverse document and content types may be described using TEI.
2. Ontologies allow the formal description of specific knowledge domains.
3. Open access tries to break existing barriers to access and fosters collaboration and knowledge sharing.
4. The linked open data (LOD) mechanism allows data and projects which were not originally intended to work together to interconnect by implementing a common method for describing and publishing data.

4 All the features described above are fundamental to the aims of *Geolat—Geography for Latin Literature*, a global project for Latin literature annotation which makes use of TEI encoding and other ontologies. The present research therefore intends to address the issues related to TEI collections management from a semantic perspective by presenting a specific case study.

5 The paper is organized as follows: first we introduce the project in order to explain the steps needed to complete it (section 2); then we describe the ontological modeling and the TEI encoding in *Geolat*, by focusing on both places and persons (section 3). We then discuss the annotation data model (section 4). In conclusion, we introduce some future directions (section 5).

2. The Geolat Project

6 Starting from the need to enrich the TEI encoding with an in-depth formalization of the process steps, the *Geolat* project aims to propose an extensible model of a digital library in the literary domain: *Geolat* will semantically annotate texts using many different ontologies, describing places, persons, and textual objects. Additionally, it will publish its content as open-access LOD. The project is led by an interdisciplinary group of scholars, including historians, geographers, classicists, language philosophers, science philosophers, librarians, digital humanities researchers, computer scientists, archeologists, and comparative literature scholars. The main idea is that if places and personal names are formally annotated in a given text collection, those texts can be studied through the knowledge the texts themselves include. It will then be possible, for instance, to answer the question "who was where?"—the places where a given person was at a given time, or the persons who, at a given time, were in a given place—and show the answers not simply with the usual text tools (concordance-like) but also through maps (for places) or graphs (for persons).

7 From a conceptual standpoint, four simple steps are needed to complete the project:

1. the building of a global digital Latin library from the Archaic period up to the fifth century CE, joining different existing available corpora:
    - classical Latin texts from Packard Humanities Institute Classical Latin Texts (PHI) CD-ROM which are now freely accessible, according to Italian and EU laws on intellectual property rights;
    - late Latin texts, from DigilibLT (the Digital Library of Late-Antique Latin Texts);[1]
    - the Latin Grammarians in the collection prepared by Nino Marinone as the basis for *Index Grammaticus: An Index to the Latin Grammar Texts*;[2]
    - possibly, the juridical texts from Bibliotheca Iuris Antiqui(BIA);[3]

- possibly, the Latin poetry from Musisque Deoque: Un archivio digitale di poesia Latina (MQDQ);[4]

2. the building of an ontology. This step will involve a connection between different models starting from some specific issues—describing places and integrating descriptions of people in EAC-CPF (plus the related ontologies illustrating the textual objects)—and from the consideration that roles and functions change in relation to the context. Further, we need to consider that existing geographical ontologies are not (completely) suitable for describing the classical world because, for example, some kinds of ancient places have no immediate equivalent in contemporary conceptualizations (take for instance the notion of *castrum* in Latin culture); a city may be founded by a deity or by a human; or you may be dealing with nomadic populations whose "places" are impossible to describe similarly to those of non-nomadic populations;
3. annotation of every personal and geographical name in the texts using a geographical ontology, which means:

    - assigning IRIs/URIs to textual objects like books, chapters, paragraphs, words;
    - associating Pleiades[5] URIs with geographical names;
    - associating VIAF[6] IDs with personal names for the authority control;
    - integrating a geographical ontology with specific concepts for the classical world and classical texts;
    - integrating the EAC-CPF model and ontology for managing roles and functions of mentioned people.

The documents are marked up with very light TEI/XML encoding that describes document structures and philological phenomena. The textual segments (nouns and nominal phrases) referring to a person or geographical entity are explicitly encoded through specific TEI elements (see section 3). Only instances of individual persons or geographical instances are encoded (while general or abstract geographical concepts are not). The encoding process for places and names is implemented in two phases: first a Named Entity Recognition (NER) system extracts references from primary sources (Isaksen et al. 2012); then these references are edited by human experts.

4. the use of annotated texts in order to facilitate specific research questions related to geography or named persons. This procedure will allow, for example,

    - general users to start reading texts through a geographic interface: for example, a map where they can choose an area, or a person interface through which they can choose from a typology of activities and roles;
    - scholars to discover concepts or information hidden by the standard textual interface usually adopted to access the documents;
    - mobile and augmented reality devices which display passages describing the physical place where you are.

3. Ontological Modeling and TEI Encoding in Geolat

8 The logical architecture of the *Geolat* project is a complex one, requiring many levels of ontological modeling and textual encoding. Some general methodological principles are:

- maintaining the distinction among different levels of abstraction and adopting the most efficient formalism for each level, trying to avoid unneeded complexity;
- minimizing the amount of semantic information directly expressed at the inline markup level in favor of stand-off markup. In this way, it is possible to favor readability, portability, and maintenance of primary resources;
- adopting Semantic Web languages to express semantic information (geographic and prosopographical data) and rigorously defining the annotations' intended semantics with the use of formal ontologies—which allows us to exploit the reasoning capabilities of inference engines and semantic stores;
- allowing the gradual extension and modification of geographical descriptions over time;
- facilitating interoperability with other repositories and sets of geographic data expressed as linked data.

9 In the following sections, the problems related to the description of places (sections 3.1 and 3.2) and persons (sections 3.3, 3.4, and 3.5) will be explained.

3.1 Distinction between Semantic Geographical Data and Geographical Ontology

10 The distinction between semantic geographical data and geographical ontology is analogous to the distinction between assertion box (ABox) and terminological box (TBox) which is made in description logic.

> Description logics and their semantics traditionally split *concepts* and their relationships from the different treatment of *instances* and their attributes and roles, expressed as fact assertions. The concept split is known as the TBox (for *terminological* knowledge, the basis for *T* in *TBox*) and represents the schema or taxonomy of the domain at hand. The TBox is the structural and intensional component of conceptual relationships. It is this construct for which Structure Dynamics generally reserves the term 'ontology.'
>
> The second split of instances is known as the ABox (for *assertions*, the basis for *A* in *ABox*) and describes the attributes of instances (or individuals), the roles between instances, and other assertions about instances regarding their class membership with the TBox concepts. Both the TBox and ABox are consistent with set-theoretic principles.
>
> (Bergman 2009)

11 According to the aforementioned distinction, the geographical semantic data (GSD) in Geolat consists of a set of RDF statements defining the individual geographic features. Each individual entity is:

1. identified by an IRI;
2. associated with one or more membership classes;
3. optionally associated with one or more persons;
4. identified by a set of properties:

    - geographic coordinates in GPS format;
    - placename(s) with chronological information about its/their use over time;
    - itineraries (such as pilgrimage or military expedition) of which the place is a part;
    - historical, geographical, cultural annotations (such as etymology; typology of settlement: city, castrum; reason for being mentioned: battlefield)
    - links to IRI/URI in other data sets like Pleiades, Geonames, and DBPedia etc.

12 The geographic ontology (GO) contains the formal representation of general concepts and relationships. The formalism we have adopted for ontological modeling is OWL 2 RL.

> OWL 2 RL enables the implementation of polynomial time reasoning algorithms using rule-extended database technologies operating directly on RDF triples; it is particularly suitable for applications where relatively lightweight ontologies are used to organize large numbers of individuals and where it is useful or necessary to operate directly on data in the form of RDF triples.
>
> (W3C OWL Working Group 2012, sec. 2.4)

13 Geographic ontology provides concepts such as settlement, city, person, physical location, sea, and river, along with their properties; specifies their taxonomy; and defines relations among the "occurrences of concepts," such as the *located in* relation, which associates the name of a settlement with its physical place, or the *founded by* relation, which relates a city to its founders.

14 The design of this ontology must address various theoretical problems, such as the status of fictional places; the need to draw a distinction between purely fictional places and places whose past existence is acknowledged by the textual tradition but which do not have any certain location nor material vestiges (for instance the city of Alba Longa, according to Titus Livius founded by Ascanio, son of Aeneas); the possibility for certain entities to have distinct ontological properties according to different textual traditions (one city, for instance, can have various foundation stories, with different actors—be they real or fictional).

15 Of course this requires that we state explicitly in the ontology the textual context where each property value is valid.

3.2 Connection between Geographical Data and TEI Markup Vocabulary

16 TEI offers a rich vocabulary for the annotation of textual segments that include references to places and geographical entities (TEI Consortium 2015, ch. 13, Names, Dates, People, and Places). For our annotation goals, the two most important TEI elements are:

1. `<placeName>` identifies a noun or phrase referring to a geopolitical or administrative entity;
2. `<geogName>` identifies a noun or phrase referring to a physical place.

17 Their key attributes are:

1. `@xml:id`: univocally identifies the encoded element; the purpose is to allow inverse references from geographical RDF assertions to the text passages that mention those entities;
2. `@ref`: includes the URI which identifies the geographical entity referred to by the text passage in GSD.

18 See this example from Titus Livius, *Ab urbe condita libri*, 1,3:

```
<milestone unit="section" n="5"/>
pax ita conuenerat ut Etruscis Latinisque fluuius
<geogName ref="http://www.geolat.it/geoDat/aTiber" xml:id="tiber01">
Albula
</geogName>
, quem nunc
<geogName ref="http://www.geolat.it/geoData/Tiber" xml:id="tiber02">
Tiberim
</geogName>
uocant, finis esset.
<milestone unit="section" n="6"/>
```

19 As shown above, the TEI attribute `@ref` allows us to explicitly connect every geographical expression in the text to its RDF description. However, it is also necessary to express the reverse connection from the RDF description to the text. The rationale for this requirement is twofold: from the practical point of view, it is easier to link the query (and inference) results to the relevant text portions; from the methodological point of view, we are sure that the semantic data set is self-consistent and potentially autonomous from the text collection.

20 The simplest method to obtain this result consists in assigning a "textual instance" property to the URI that identifies the entity. The value of this textual instance property is an XPointer expression identifying the corresponding element in the XML/TEI file. This structure is illustrated in the RDF statement below:

1. *subject*: geographical entity URI
2. *predicate*: `textualInstance` assigns a textual location as an instance of the entity
3. *object*: XPointer reference to the element that contains the geographical expression

21 For example:[7]

```
@prefix geolat <http://www.geolat.it/>

geolat:geoData/Tiber geolat:textualInstance
geolat:abUrbeConditam.xml#xpath(//geogName[@xml:id='tiber01']);
geolat:textualInstance
geolat:abUrbeConditam.xml#xpath(//geogName[@xml:id='tiber02']) .
```

3.3 A Model for Managing Persons

22 Among the most significant changes in TEI P5 is the addition of the Biographical and Prosopographical Data features (TEI Consortium 2015, ch. 13.3, Biographical and Prosopographical Data). These features provide elements and attributes for describing individuals and their relationships. In 2006 a "Personography" workgroup (the TEI Personography Task Force)[8] was established to investigate how other existing XML schemas and TEI customizations handle data about people. The result was Wedervang-Jensen and Driscoll's *Report on XML Mark-up of Biographical and Prosopographical Data* (2006).[9]

23 Names of people are identified in TEI with the `<persName>` element. This element, like the `<placeName>` and the `<geogName>`, supports an `@xml:id` attribute for unique identification and a `@ref` attribute to link the name to an external description or URI. At the URI level, many features, as described below, could be used in order to enrich biographical and prosopographical description.

24 The standard bibliographic approach to describing people actually consists in the identification of the unique individuals and in the attribution of an invariant set of features. However, we should never overlook the strong existing connections between people and the textual context. As a result of these connections, roles and functions, intended as features of individuals, may change depending on the context, that is, on the source attesting the individual. It is therefore possible to state that:

1. some features are not only static over time, but also theoretically constant regardless of the context (for example, birth, death, personal name);
2. other features vary depending on date and/or place (for example, age, affiliation, education, event, state);

3. roles (for example, author, actor, editor, speaker) and functions (for example, administrative, organizational, educational) are elements that identify people depending on the context.

25 Thus we can say that a person is a "complex entity," because she/he is connected with different typologies of phenomena: some of these are unchangeable, while others depend on a time period, a place, or a context.

26 In TEI, the <person> element may be associated with different roles or functions. Consider, for example, the digital edition of a literary text. We may use the TEI <person> element to encode information about all the individuals who created and contributed to the digital edition: the author of the analog source, the editor of the printed version, or all the individuals quoted in the text. The concept of person expands the boundaries: although individuals are related to the source, they are also entities with a role enabling a single person to connect either with different resources (that is, documents), or with several other persons (for example, for the sharing of the same role). A three-level relationship therefore arises: among individuals, between a person and a document in which she/he is mentioned, and among a person and other resources. This "three-level relation" model is a concept adopted by a shared standard in the archival domain, used in order to manage authority records: the EAC-CPF Schema (see section 3.4). This approach was chosen in part because it forces a reply to the following questions:

- Why is a person related to another?
- What is written in a document about a person?
- What connection is possible to establish between a person and other resources regarding the same person?

3.4 TEI and EAC-CPF

27 For the annotation of individuals and groups, TEI may be used in conjunction with the EAC-CPF schema, developed in order to formalize the ISAAR (CPF) standard (International Standard Archival Authority Record for Corporate Bodies, Persons and Families),[10] which today is also available as an ontology (Mazzini and Ricci 2011). EAC-CPF contributes to the representation of individuals, emphasizing the importance of both context and relationships. The editorial approach to annotation described here borrows from the domain of archival studies. Archival

science espouses a principle of separation between the description of records (documents) and the description of people (corporate bodies, persons, and families), and emphasizes context (Pitti 2004). The same approach could be implemented in TEI, when the final purpose is to expose data sets of both TEI XML documents and related personographic data to be used by the Web community.

28 The EAC-CPF schema suggests useful ways to extend the `<relation>` concept in TEI. EAC-CPF is based on the concept of an "entity," a corporate body, person, or family that manages relationships among other entities and between one entity and a resource linked at some level. Each relationship could be described, dated, and classified. In addition to elements related to "relation" (`<cpfRelation>` and `<resourceRelation>`), EAC-CPF defines a `<function>` element that "provides information about a function, activity, role, or purpose performed or manifested by the entity being described" on a specific date. The `<functionRelation>` element illustrates a "function related to the described entity.... it includes a `@functionRelationType` attribute that could support a controlled list of type values" (Encoded Archival Context Working Group 2014).

29 A new model of an "authority record"—a complex structure able to document the context in which the identity is attested—could be introduced: the authority not only is generated by the controlled form of the name, and the related parallel forms, but is also the result of relationships generated by the context (that is, the specific document in which the entity is mentioned) to determine a concept.

30 According to the RDF model, it is possible to assert that an identified entity (URI) manages relationships (predicate) with different objects; these objects could be:

1. another entity (URI): another person, place (URI), date (URI), or event (URI);
2. a contextual resource (URI): the document in which the entity is mentioned;
3. an external resource (URI): another object (a document, an image, a video, an audio record, and so on).

31 This procedure could be applied to describe annotations and other contributions to the digital edition, for instance, by identifying a contributor of an annotation who, on a specific date, performed a specific activity (the principle of "provenance" of an assertion, that is, authorship attribution activity). The responsibility (the TEI `<resp>` element) could be intended as a role. Each person is associated with a responsibility statement able to identify the role that the entity

covered in that document, associating persons to documents. The same person may fulfill the same responsibility in other editions. In this way relationships are extended to other documents. Moreover, other individuals who share the same responsibilities may be linked.

32 This process could be declared and exposed as an RDF data set with URIs for the univocal identification and TEI/EAC for classes and predicates in order to build a collection of authorities related to persons that covered either a role or a function in a certain time period and context. By declaring connections as relationships, through the EAC-CPF model, we could develop a knowledge base of people, with a role or function originated by the context.

33 The aforementioned model has already been applied to the description of real individuals who manage activities in relation to documents (Pasin and Bradley 2013). The same approach could be expanded to prosopographical description, with an adequate extension of the function sets to include activities of interest for historiographical research (for instance, military mission, diplomatic activity, arranged marriage, conspiracy). People mentioned, described, or in general cited in sources also assume different roles in different contexts—context being determined by the document, as an historical source, in which a person appears. More specifically, people are related to dates, places, and events, enriching the expressivity of the description.

3.5 People and Places in a "Perspective Function"

34 Place, in particular, could be an interesting key for managing functions. In the *Geolat* project we are able to assert that people have different roles with respect to places. This means that the context (that is, the document) determines the existing relationship between a person and a place. In addition, the role a person plays in a document also changes the kind of the relationship she/he has with the place. A place could be intended not only as a city that a person *was-born-in* or a person *died-in* but also as place that a person *went-to* for a specific event or to *do-something*. A taxonomy of predicates will be developed in order to define the possible relationships between a person and a place.

35 An example of a person description (pseudocode):

```
<persName xml:id="LdM">Lorenzo de‘ Medici</persName>
xml:id=LdM
Access_key: Medici, Lorenzo de’
Dbpedia URI: http://it.dbpedia.org/page/Lorenzo_de%27_Medici
VIAF permalink: http://viaf.org/viaf/54169908
Father-of: Piero de’Medici (URI)
Born-in: Firenze (URI)
Died-in: Firenze (URI)
Born-when: 1449 (xsd:integer)
Died-when: 1492 (xsd:integer)
Visited-what: Roma (URI) –function: diplomatic mission
Visited-what: Venezia (URI) –function: military mission
Visited-when: Roma (URI) 1466
Visited-when: Venezia (URI) 1465
Iconography: http://www.google.com/search?hl=en&q=lorenzo+de
%27+medici&um=1&ie=UTF-8&tbm=isch
&source=og&sa=N&tab=wi&ei=_1vcUOeAJqeI4ASikoCYBQ&biw=1146&bih=709&sei=AVzcUNnjIOXV4gTVloHQCQ
Biography: http://it.wikipedia.org/wiki/Lorenzo_de'_Medici
Author-of: http://it.wikisource.org/wiki/Canti_carnascialeschi_%28Lorenzo_de
%27_Medici%29
Attested-in: http://it.wikiquote.org/wiki/Maria
Provenance: http://www.unibo.it/docenti/francesca.tomasi
```

36 By creating connections between the archival and the philological domains, digital editions open new perspectives on the cultural heritage domain, establishing connections between heterogeneous objects and “creating efficiencies in the re-use of metadata across repositories, and through open linked data resources” (Larson and Janakiraman 2011, 4). Linked data describing persons performing specific roles could be considerably improved by employing analytic description concerning people’s functions, using the context as interpretative key: “the description of personal roles and of the statuses of documents needs to vary in time and according to changing contexts … such roles and statuses need to be handled formally by ontological models” (Peroni, Shotton, and Vitali 2012, 9).

4. A Meta-ontology for the Annotation: Open Annotation Data Model

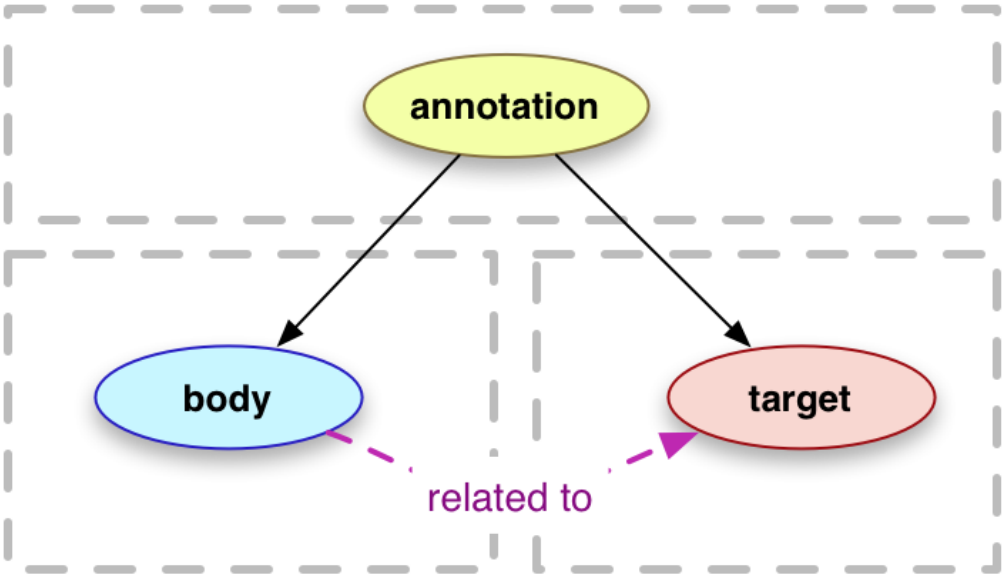
37 The annotation layer that we are connecting to the texts describes semantic features of the textual content. We need, however, to record some meta-properties of the annotations themselves:

- responsibility metadata;
- typological categorization;
- scope or function;
- provenance metadata (responsibilities, certainty level, time).

38 For this reason we have opted to add an intermediate layer to express the relationship between the annotation and the text, based on the Open Annotation data model (OA), which is a plausible model for meeting these requirements:

> An annotation is considered to be a set of connected resources, typically including a body and target, and conveys that the body is related to the target. The exact nature of this relationship changes according to the intention of the annotation, but most frequently conveys that the body is somehow 'about' the target. Other possible relationships include that the body is an identifier for the target, provides a representation of the target, or classifies the target in some way. This perspective results in a basic model with three parts, depicted below....

Figure 1: Annotation, Body and Target



(Sanderson, Ciccarese, and Van de Sompel 2013)

39 In a nutshell, we can define OA as an RDF vocabulary (formally expressed in OWL 2), which allows the expression of the relationship between an annotation and its object and a metadata set for this relationship. The core of OA is the `oa:Annotation` class with the two relationships `oa:hasBody` (that expresses the relationships between an annotation and its content) and `oa:hasTarget` (that expresses the relationships with the object of the annotation).

40 Both the annotation body and target are expressed by URI; other properties of the `oa:Annotation` class are `oa:annotatedBy` (to state responsibility) and `oa:annotatedAt` (to specify the date of the annotation). Using this model, the definition of an annotation in *Geolat* would be as follows (in Turtle syntax):

```
<geolat:titoLivioAnn_1>a oa:Annotation ;
    oa:hasBody geolat:geoData/Tiber ;
    oa:hasTarget  geolat:abUrbeConditam.xml#xpath(//geogName [@xml:id='tiber02'])
    oa:annotatedBy <geolat:agent1> ;
    oa:annotatedAt "2013-09-28T12:00:00Z" ;
```

41 The attribution of a semantic type to this annotation can be achieved with a twofold approach:

- extension of the OA ontology through the introduction of a set of subclasses of the `oa:Annotation` class, in order to formalize this typology;
- use of the `oa:MotivatedBy` property, whose values express the rationale of a given annotation and are members of the `oa:Motivations` class, which is extensible according to the users' needs.

42 The adoption of the OA framework has the advantage of making the annotation metadescriptions homogeneous with the complex semantic *Geolat* architecture, and at the same time using an established standard that assures interoperability with other projects and future software platform change.

5. Closing Remarks and Future Developments

43 In *Geolat* everything is planned and built through free, open-source software. There are several reasons underlying this choice, including the project's need to share its software tools with other researchers. Practical benefits of our adoption of open-source tools include the absence of annual license fees, which can be a real burden for long-term humanities research projects.

44 The entire content is made available under CC licenses. Despite a strong pressure towards CC-BY (which requires only citing the author[s]), *Geolat* chose a CC-BY-NC-SA license which seems more appropriate to stress the value—including from an economic/commercial point of view—of the research outcomes.

45 Semantic annotation of a digital library of textual materials potentially never ends; therefore, a crowdsourcing approach should be adopted. Annotating the entire corpus of Latin literature is undoubtedly a demanding task: as a consequence, a core group of people working over many years is needed. However, being open to crowdsourcing means that the members of the larger research community of Latin scholars can significantly contribute to this effort.

46 Linked open data (LOD) is the most efficient framework to allow semantic interoperability of complex data sets on the Web. In the field of classical studies, several important resources are currently available which publish their data as LOD, including Pleiades,[11] a gazetteer for classical antiquity, and Pelagios,[12] a collaborative initiative to share and link geographic references in ancient cultural artifacts and documents. Notwithstanding its value and usefulness, Pleiades lacks

references to textual passages with place names, which it could gather from *Geolat*. *Geolat* could retrieve from Pleiades selected geographical data (such as GPS coordinates, the ancient names of places, and names in modern languages where available). This mutual sharing of controlled and authoritative data is our primary motivation for using LOD in *Geolat*. This approach contributes to a growing community of scholars sharing data and tools.

47 The *Geolat* model is not language-dependent, and can be adopted for any collection in any language. This is not a purely quantitative approach (the model can be used for *n* different languages): it means that through the model (provided that the same model is adopted for annotating), different texts, from different literatures, can interact and can be read in their intertextual relations: any person or place, mentioned in different texts from different literatures, can be tracked. The aim is to finely analyze and interpret similarities and differences across texts, languages, and literatures.

BIBLIOGRAPHY

Bergman, Mike. 2009. "The Fundamental Importance of Keeping an ABox and TBox Split." Ontology Best Practices for Data-driven Applications, part 2.May 17. http://www.mkbergman.com/489/ontology-best-practices-for-data-driven-applications-part-2/.

Encoded Archival Context Working Group. 2014. *Encoded Archival Context—Corporate Bodies, Persons, and Families (EAC-CPF) Tag Library*. Edition 2014. http://eac.staatsbibliothek-berlin.de/fileadmin/user_upload/schema/cpfTagLibrary.html.

Isaksen, Leif, Elton Barker, Eric C. Kansa, and Kate Byrne. 2012. "GAP: A NeoGeo Approach to Classical Resources." *Leonardo* 45 (1): 82–83. doi:10.1162/LEON_a_00343.

Larson, Ray R., and Krishna Janakiraman. 2011. "Connecting Archival Collections: The Social Networks and Archival Context Project." In *Research and Advanced Technology for Digital Libraries: International Conference on Theory and Practice of Digital Libraries (TPDL 2011)*, edited by Stefan Gradmann, Francesca Borri, Carlo Meghini, and Heiko Schuldt, 3–14. Lecture Notes in Computer Science 6966. Heidelberg, Germany: Springer. doi:10.1007/978-3-642-24469-8_3. http://link.springer.com/chapter/10.1007/978-3-642-24469-8_3.

Mazzini, Silvia, and Francesca Ricci. 2011. "EAC-CPF Ontology and Linked Archival Data." In *Proceedings of the 1st International Workshop on Semantic Digital Archives*, edited by Livia Predoiu, Steffen Hennicke, Andreas Nürnberger, Annett Mitschick, and Seamus Ross, 72–81. N.p.: CEUR Workshop Proceedings.http://ceur-ws.org/Vol-801/.

Pasin, Michele and John Bradley. 2013. "Factoid-based prosopography and computer ontologies: Towards an integrated approach." *Digital Scholarship in the Humanities* 30 (1): 86–97. doi:http://dx.doi.org/10.1093/llc/fqt037.

Peroni, Silvio, David Shotton, and Fabio Vitali. 2012. "Scholarly Publishing and Linked Data: Describing Roles, Statuses, Temporal and Contextual Extents." In *Proceedings of the 8th International Conference on Semantic Systems*, 9–16. New York: ACM. doi:10.1145/2362499.2362502.

Pitti, Daniel V. 2004. "Creator Description: Encoded Archival Context." In *Authority Control in Organizing and Accessing Information: Definition and International Experience*, edited by Arlene G. Taylor and Barbara B. Tillett, with the assistance of Murtha Baca and Mauro Guerrini, 201–26. Binghamton, N.Y.: Haworth Information Press.

Sanderson, Robert, Paolo Ciccarese, and Herbert Van de Sompel, eds. 2013. "Open Annotation Data Model." Community Draft, February 8. World Wide Web Consortium (W3C). http://www.openannotation.org/spec/core/20130208/index.html.

TEI Consortium. 2015. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 2.8.0. Last updated April 6. N.p.: TEI Consortium. http://www.tei-c.org/Vault/P5/2.8.0/doc/tei-p5-doc/en/html/.

W3C OWL Working Group. 2012. *OWL 2 Web Ontology Language Document Overview*, 2nd ed. W3C Recommendation, December 11. http://www.w3.org/tr/owl2-overview/#profiles.

Wedervang-Jensen, Eva, and Matthew Driscoll. 2006. *Report on XML Mark-up of Biographical and Prosopographical data*. February 16. http://www.tei-c.org/activities/workgroups/pers/persw02.xml.

Wittern, Christian, Arianna Ciula, and Conal Tuohy. 2009. "The making of TEI P5." *Literary and Linguistic Computing* 24 (3): 281–296. doi:10.1093/llc/fqp017.

NOTES

1 http://www.digiliblt.unipmn.it/.

2 Edited by Valeria Lomanto and Nino Marinone (Hildesheim: Olms-Weidmann, 1990).

3 http://bia.lex.unict.it/.

4 http://www.mqdq.it/mqdq/.

5 http://pleiades.stoa.org/.

6 Virtual International Authority File (https://viaf.org/).

7 This example, and all the following code snippets of RDF statements, are expressed in Turtle syntax. Cfr. *RDF 1.1 Turtle. Terse RDF Triple Language*. W3C Recommendation 25 February 2014. http://www.w3.org/TR/2014/REC-turtle-20140225/.

8 http://www.tei-c.org/Activities/Workgroups/PERS/index.xml.

9 The actual changes introduced in the TEI Schema to implement the proposals in that document are described in Wittern, Ciula, and Tuhoy (2009).

10 http://www.ica.org/10203/standards/isaar-cpf-international-standard-archival-authority-record-for-corporate-bodies-persons-and-families-2nd-edition.html

11 http://pleiades.stoa.org/

12 http://pelagios.org/

ABSTRACT

This paper presents the rationales and the logical architecture of *Geolat — Geography for Latin Literature*, a project for the enrichment of Latin literature which makes use of a complex mix of TEI markup, Semantic Web technologies and formal ontologies. The purpose of *Geolat* is the annotation of the geographical and personal references in a corpus of Latin TEI encoded texts. These annotations are linked to a set of ontologies that give them formal semantics, and can finally be exposed the as linked open data, in order to improve the documents' interoperability with other existing LOD and to enhance information retrieval possibilities. The paper is organized as follows: first we introduce the project in order to explain the steps needed to complete it (section 2); then we describe the ontological modeling and the TEI encoding in *Geolat*, by focusing on both places and persons (section 3). We then discuss the annotation data model (section 4). In conclusion, we introduce some future directions (section 5).

INDEX

Keywords: geotagging, personography, ontology, Semantic Web, Linked Open Data, digital annotation

AUTHORS

FABIO CIOTTI

Fabio Ciotti is Assistant Professor at the University of Roma Tor Vergata, where he teaches Digital Literary Studies and Theory of Literature. He is President of the Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD, the Italian digital humanities association) and member of the EADH Executive Board. He has been scientific consultant for text encoding and technological infrastructures in several digital libraries and archives projects, and has served in the TEI Technical Council.

MAURIZIO LANA

Maurizio Lana is an assistant professor at Università del Piemonte Orientale where he teaches Library & Information Science; he also coordinates the *geolat-geography for Latin literature* project and, with R. Tabacco, the *digital library of late-Latin texts* digilibLT.

FRANCESCA TOMASI

Francesca Tomasi is assistant professor at the University of Bologna, where she teaches archives and computer science and digital humanities. Her research is mostly focused on ontologies and semantic modeling in the cultural heritage domain. She is in particular editor of the digital collection: Vespasiano da Bisticci, Letters. She is finally President of the School of Humanities library in the University of Bologna.