# Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing

# Schriften des
# Instituts für Dokumentologie und Editorik

herausgegeben von:

Bernhard Assmann
Roman Bleier
Alexander Czmiel
Stefan Dumont
Oliver Duntze
Franz Fischer
Christiane Fritze
Ulrike Henny-Krahmer
Frederike Neuber
Christopher Pollin
Malte Rehbein
Torsten Roeder
Patrick Sahle
Torsten Schaßan
Gerlinde Schneider
Markus Schnöpf
Martina Scholger
Philipp Steinkrüger
Nadine Sutor
Georg Vogeler

Band 15

Schriften des Instituts für Dokumentologie und Editorik — Band 15

# Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing

edited by

Elena Spadini, Francesca Tomasi, Georg Vogeler

2021

BoD, Norderstedt

**Bibliografische Information der Deutschen Nationalbibliothek:**
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über **http://dnb.d-nb.de/** abrufbar.

**Digitale Parallelfassung der gedruckten Publikation zur Archivierung im Kölner Universitäts-Publikations-Server (KUPS). Stand 5. Dezember 2021.**

© 2021

Herstellung und Verlag: Books on Demand GmbH, Norderstedt
ISBN: 978-3-7543-4369-2
Einbandgestaltung: Stefan Dumont nach Vorarbeiten von Johanna Puhl und Katharina Weber
Satz: LuaTEX, Bernhard Assmann

# Contents

## Infrastructures and Technologies

## Formal Models

## Projects and Editions

## Appendices

# Infrastructures and Technologies

# Formal Models

# The Critical Apparatus Ontology (CAO): Modelling the TEI Critical Apparatus as a Knowledge Graph

Francesca Giovannetti

**Abstract**

This paper seeks to explore the use of semantic web technologies to enhance the re-/presentation of the critical apparatus that accompanies a TEI digital scholarly edition. The apparatus is a key instrument for critical editions. Its encoding poses a challenge for researchers, who strive to achieve highly expressive digital representations of their scholarly views. So far, no comprehensive ontology has been developed for the representation of the critical apparatus. This study makes a first step towards filling this gap by proposing the Critical Apparatus Ontology, an OWL ontology for representing the critical apparatus as a knowledge graph.

## 1 Introduction

This study proposes a conceptual model for the representation of the TEI critical apparatus as a knowledge graph. Critical apparatuses are fundamental tools for scholarly editions dealing with works that exist in multiple versions; they provide a window on the editor's workshop, and offer readers the evidence they need in order to evaluate the edition itself; evidence includes, but is not restricted to, variant readings.

When modelling or using a critical apparatus, it is of key importance to understand its purpose: a critical apparatus does not provide a mere list of textual variants, but rather presents textual variants in a way capable of conveying the editor's theory about how the readings and witnesses are related to one another (Damon 2016). In other words, a critical apparatus should put textual variants into context, conceptually sitting at the focal point of the network of texts that participate in the process of reconstruction of a work. Every element of a critical apparatus should be approached as the result of an act of scholarly interpretation (Romanello et al. 2009).

Representing the critical apparatus in TEI can be challenging. It is an operation which involves structuring the critical text and apparatus as hierarchies of ordered, nesting, XML elements. However, as Sperberg-McQueen points out, textual variation represents "a type of textual non-linearity" which hardly fits into a tree data structure (Sperberg-McQueen 1989). This paper argues that the use of a graph model – and of knowledge graphs specifically – instead of a tree model, provides a more intuitive structure for representing a critical apparatus.

*Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing*. Ed. by Elena Spadini, Francesca Tomasi and Georg Vogeler. Schriften des Instituts für Dokumentologie und Editorik 15. Norderstedt: Books on Demand, 2021. 125–139.

The Resource Description Framework (RDF) is a graph data model for capturing and representing knowledge as statements composed of a *subject*, a *predicate*, and an *object* which, when linked to one another and supported by formal semantics (i.e. by one or more ontologies), form a knowledge graph. A number of studies have begun to discuss the use of ontologies to provide TEI-encoded documents with a more formal semantics (e.g., Ciotti & Tomasi 2016; Eide 2014/15; Jordanous et al. 2012; Ciula et al. 2008). Several scholars have recommended that more digital scholarly editions should be published according to the principles of linked open data (LOD), and in collaboration with cultural heritage institutions that are already making their collections available in the LOD cloud (e.g. Daquino & Tomasi 2015; Ore & Eide 2009). Nonetheless, most digital scholarly editions remain document-centric, thus missing out on the opportunities and the greater gains that come with a data-centric approach.

The benefits of using knowledge graphs, in support of TEI XML encoding, to enhance digital scholarly editions outweigh the production costs involved:

1. Entities can be linked to one another via meaningful links to form networks. TEI provides two methods for describing a relationship between elements: one is using specific XML attributes (e.g., `<rdg wit="A">` to express that the reading is witnessed in A); the other is nesting (e.g., `<app><rdg>a</rdg><rdg>b</rdg></app>` to indicate that reading "a" and "b" are alternative readings). However simple, there always remains a margin of ambiguity when interpreting such relationships, as semantics has to be deduced from structure. Knowledge graphs provide a machine-readable method for representing complex connections and their semantics, including textual variation, situations where fragments of the same witnesses are scattered across different libraries, and text transposition (which involves a relationship across distinct locations of the text and, as such, is particularly problematic to represent in a document-centric standard).
2. Entities can be represented according to different, interlinked, perspectives. For example, the concept of reading could be described as the result of a scholarly interpretation, or as a fragment of text.
3. The provenance of an entity or graph can be easily specified by means of dedicated ontological vocabularies. This facilitates the process of assessing information quality.
4. An open-ended number of different, and even conflicting, arguments can be expressed (in such a case, declaring the provenance of information is central).
5. Knowledge graphs offer concrete opportunities for federation. Siloed editions are dismantled as soon as references to external, centralized, repositories are provided. For example, a critical apparatus may relate the edition to an external linked data record of one of the witnesses (e.g., a record held by a cultural institution, such as a library or museum). In this way, interoperability between editions, and the

inclusion of digital scholarly editions in the cultural heritage-linked open data cloud, becomes possible.

6. Ontologies provide the content model with a good degree of flexibility, as users are able to introduce new classes and property without compromising interoperability; the structure of a TEI document can vary on a case-by-case basis, while RDF, with its graph structure, remains consistent across applications.

Some of these objectives are achievable even in an entirely TEI-based environment. However, the use of knowledge graphs allows a more straightforward formalization of complex, interrelated theories and interpretations.

The Critical Apparatus Ontology (CAO) is a conceptual model providing a framework for the representation of the critical apparatus as a knowledge graph. CAO may be used in combination with TEI documents to enhance the edition with formal semantics. An experimental Python script is available for converting specific features of an existing TEI critical apparatus to a CAO knowledge graph.[1]

## 2 Related Works

Previous attempts to represent the phenomenon of textual variation in the form of graphs have led to the development of automatic collation tools such as CollateX (Dekker et al. 2015), whose underlining data structure is the variant graph, i.e. a graph-oriented model for the representation of textual variants, originally developed by Desmond Schmidt (Schmidt & Colomb 2009). Other projects, more concerned with visualization, have moved in a similar direction (e.g. Andrews & Macé 2013; Jänicke et al. 2015).

To date, there has been little work on standard ontologies for the description of the critical apparatus that accompanies an edition of a text. The Linking Ancient World Data (LAWD) ontology was developed for the purpose of enhancing interoperability between data about the ancient world gathered from different projects (Cayless 2015). The LAWD ontology supports the description of single variant readings. As such, expanding the LAWD model with new properties to represent specific relationships between variant readings could have been a viable option. However, LAWD lacks a way of distinctly representing the concept of reading as a scholarly interpretation, and the textual fragment supporting such a reading: a LAWD reading can be both a scholarly interpretation and a text, at the same time. The need for distinguishing the text from its interpretation motivates the choice of creating a new conceptual model for the description of the textual variants and the philological arguments which together form a critical apparatus.

[1] The transformation script can be downloaded from https://github.com/fgiovannetti/cao/tree/master/tei-to-cao.

Consistent with best practice, which encourages reuse of existing ontologies, the Critical Apparatus Ontology (CAO) is based on the TEI abstract model, and incorporates classes and relationships from several conceptual models to solve specific issues: the Web Annotation Data Model (Sanderson et al. 2017) is used for describing scholarly annotations and to link them to the critical text; specific properties of the PROV-O ontology (Lebo et al. 2013) are used to declare the provenance of the interpretations, and to describe specific types of relationship between textual variants; FRBRoo (Bekiari et al. 2015) is used to describe witnesses according to different layers of analysis (i.e., work, expression, manifestation and item).

## 3 An Overview of the Critical Apparatus Ontology

The Critical Apparatus Ontology is an OWL ontology for the representation of the critical apparatus that accompanies a digital scholarly edition. In line with the philosophy of the TEI (TEI Consortium 2019, Ch. 12), CAO does not imply commitment to any particular school of textual criticism.

This section illustrates how to represent the critical apparatus according to the model provided by CAO, from the creation of a new apparatus entry, to the description of the readings and their relationships, and to the formalization of the process of scholarly interpretation, which lays at the foundation of the development of any type of critical edition. The following section does not provide a full description of the model (for which, see the online specification),[2] but, rather, aims to illustrate its expressive potential through a number of examples.

## 4 Creating a New Scholarly Annotation

The creation of a new CAO scholarly annotation (which corresponds to the introduction of a new entry in the critical apparatus) is done using the Web Annotation standard.[3] Each scholarly annotation is composed of a *body* (the content of the apparatus entry) and a *target* (the locus of variation in the text). The Web Annotation standard allows the declaration of the date of creation and of the creator, making it possible to attach provenance information to every single scholarly annotation.

```
@base <http://example.org/> .
@prefix cao: <http://w3id.org/cao/> .
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/> .
@prefix dcterms: <http://w3id.org/dc/terms/> .
@prefix frbroo: <http://iflastandards.info/ns/fr/frbr/frbroo/> .
@prefix oa: <http://www.w3.org/ns/oa#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
```

[2] The full specification of CAO is available at http://w3id.org/cao.

[3] The same approach is adopted by LAWD (see the section “Related works”).

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .[4]

<annotation/anno01> a oa:Annotation ;
  oa:hasBody <app/app01> ;
  oa:hasTarget <varloc/vl01> ;
  dcterms:creator <person/fgiovan> ;
  dcterms:created "2019-03-09"^^xsd:date .
```

# 5 The Body of the Annotation

## 5.1 The Apparatus Entry

Suppose, for example, that the scholarly annotation we just created above aims to present the reader with a set of three possible variant readings for a specific locus of variation within the critical text.

The class `cao:VariationUnit` represents a cluster of variant readings or unit of variation. The property `cao:hasReading` links a unit of variation (`cao:VariationUnit`) to *n* variant readings. Therefore, in this example, the RDF description of the unit of variation would be as follows:

```
<app/app01> a cao:VariationUnit ;
  cao:hasReading <rdg/anno01-rdg01> ,
    <rdg/anno01-rdg02> ,
    <rdg/anno01-rdg03> .
```

## 5.2 The Readings

An apparatus entry may feature two distinct types of variant readings: generic readings, equivalent to the TEI element `<rdg>` (represented by the class `cao:Reading`), and base readings, equivalent to the TEI element `<lem>` (`cao:BaseReading`). The class `cao:BaseReading` is a subclass of `cao:Reading`; it therefore inherits its characteristics. A unit of variation may relate to a maximum of 1 base reading. The text of each reading is introduced by the property `rdf:value`. Continuing our example:

```
<rdg/anno01-rdg01> a cao:BaseReading ;
  rdf:value "diffundi"^^xsd:string.

<rdg/anno01-rdg02> a cao:Reading ;
  rdf:value "diffudit"^^xsd:string .

<rdg/anno01-rdg03> a cao:Reading ;
  rdf:value "diffundit"^^xsd:string .
```

[4] Prefixes are declared only once at the beginning of this section, but they apply to all examples.

### 5.3 Classifying the Readings by Type and Cause

Textual variants can be classified by category. In TEI, the `@type` attribute used on the elements `<rdg>` or `<lem>` serves this precise function. For example, an editor may classify a variant as an omission, a conjecture, an addition, or as a copyist's mistake. CAO represents reading types as individuals of the class `cao:ReadingType` (which is a subclass of `crm:E55_Type`). Possible values include:

- addition (the type for a reading that is an addition, i.e., a syntactic and/or semantic expansion);
- conjecture (the type for a reading that is a conjecture, i.e., an editorial reconstruction or hypothesis of reading);
- correction (the type for a reading that is a correction, i.e., an authorial or scribal correction of a previous textual fragment);
- deletion (the type for a reading that is a deletion, i.e., text marked or somehow indicated by the author or scribe as deleted);
- omission (the type for a reading that is an omission (i.e. the author or scribe did not include the reading in the witness; if intentional, an omission may also classify as a correction);
- transposition (the type for a reading that was transposed from another location in the text, including the flipping of the order of two words or phrases).

Wherever possible, each reading type has been aligned with the SKOS classification scheme for readings provided by LAWD.

For reading causes, encoded in TEI using the `@cause` attribute, CAO provides a new classification scheme which includes:

- dittography (the unintentional repetition of a letter, word or phrase);
- haplography (the accidental omission of a letter, word, or phrase; homeoarchy, homeoteleuton, and saut du même au même are particular types of haplography);
- homeoarchy (the accidental skipping of a word or phrase having the same beginning; a polyptoton is a particular type of homeoarchy);
- polyptoton (the accidental skipping of a word or phrase forming a polyptoton, i.e., presenting the same root word);
- homeoteleuton (the accidental skipping of a word or phrase having the same ending);
- saut du même au même (the accidental skipping of some text in between two similar words or phrases);
- incorporation (the accidental incorporation of materials, such as marginalia).

The properties `cao:hasReadingType` and `cao:hasReadingCause` are used to relate a `cao:Reading` to a type, and a cause, respectively. For example, the RDF description of an omitted reading would be as follows:

```
<rdg/anno01-rdg03> a cao:Reading ;
  cao:hasReadingType cao:omission ;
  cao:hasReadingCause cao:homeoteleuton ;
  rdf:value ""^^xsd:string .
```

## 5.4 Relating the Readings to One Another

The possibility of defining semantic relationships among entities is one of the greatest advantages of graph data modelling. CAO features various types of relationship between readings and, on a case-by-case basis, it is easy to extend the model to accommodate all kinds of unforeseen connections. For example, a chain of authorial revisions may be described as shown below using the property `cao:correctedTo` to identify the relationship between a reading and its modified version, forming a sequence of corrections. So, the following TEI-encoded apparatus entry

```
<app xml:id="anno08">
  <rdg varSeq="1" xml:id="anno08-rdg01">
    <del>this</del>
  </rdg>
  <rdg varSeq="2" xml:id="anno08-rdg02">
    <del><add>such a</add></del>
  </rdg>
  <rdg varSeq="3" xml:id="anno08-rdg03">
    <add>a</add>
  </rdg>
</app>
```

would become:

```
<app/app08> a cao:VariationUnit ;
  cao:hasReading <rdg/anno08-rdg01> ,
    <rdg/anno08-rdg02> ,
    <rdg/anno08-rdg03> .

<rdg/anno08-rdg01> a cao:Reading ;
  cao:correctedTo <rdg/anno08-rdg02> ;
  cao:hasReadingType cao:deletion .

<rdg/anno08-rdg02> a cao:Reading ;
  cao:correctedTo <rdg/anno08-rdg03> ;
  cao:hasReadingType cao:deletion .

<rdg/anno08-rdg03> a cao:Reading ;
  cao:hasReadingType cao:addition .
```

Other CAO properties to describe the relationships among readings are: `cao:follows`, `cao:hasVariant`, `prov:wasDerivedFrom`, and `prov:wasRevisionOf`.

In TEI, the `@varSeq` attribute conveys information about the chronological sequence in which variants have appeared over time. In the same way, the property `cao:follows` is used to chain the readings to link the readings together in chronological sequence.

The property `cao:hasVariant` links a `cao:BaseReading` to its alternative readings (which are members of the superclass `cao:Reading`). The property `prov:wasDerivedFrom` denotes "a transformation of an entity [i.e. a reading] into another, an update of an entity resulting in a new one, or the construction of a new entity based on a pre-existing entity" (Lebo et al. 2013).

The property `prov:wasRevisionOf` is used for specific cases of derivation of a reading from another one: it is a subproperty of `prov:wasDerivedFrom` and defines "a derivation for which the resulting entity is a revised version of some original. The implication here is that the resulting entity contains substantial content from the original" (Lebo et al. 2013). The domain and range are instances of the class `cao:Reading`. Below, are some examples:

```
<rdg/anno01-rdg02> a cao:Reading ;
  cao:follows <rdg/anno01-rdg03> .

<rdg/anno01-rdg02> a cao:Reading ;
  prov:wasDerivedFrom <rdg/anno01-rdg03> .

<rdg/anno01-rdg02> a cao:Reading ;
  prov:wasRevisionOf <rdg/anno01-rdg03> .
```

## 5.5 Citing Other Scholars' Readings

In TEI, the `@source` attribute is used on `<rdg>` to indicate responsibility for the claim that a witness supports a particular reading. The `@source` attribute contains a reference to an external source, normally a published edition. In a fully developed LOD scenario, it would be possible to navigate from one critical edition to another via the links between readings, which would act like bridges across editions. In CAO, a cited reading is related to its original source via the property `prov:hadPrimarySource`. For example:

```
<rdg/anno01-rdg02> a cao:Reading ;
  rdf:value "diffudit"^^xsd:string ;
  prov:hadPrimarySource <http://example.org> .
```

## 5.6 From the Reading to the Witness

Texts reach us in multiple versions. Each textual version represents a different realization of the same work. As anticipated in the introduction, CAO reuses concepts from the FRBRoo ontology to represent a text according to different levels of analysis. An FRBR expression, i.e., a text, is modelled using the class `frbroo:F2_Expression`, which defines "the specific intellectual or artistic form that a work takes each time it is realized" (IFLA 2009). A reading is a scholarly claim: the outcome of an act of interpretation, which is supported by a particular fragment of a witness, a frbroo:F23_Expression_ Fragment.

The cao:witnessedBy property relates a reading to an expression, i.e., to the text of the witness (either the full text, or the specific textual fragment that supports the reading within a witness), as shown in the example below:

```
<rdg/anno01-rdg01> a cao:BaseReading ;
  rdf:value "diffundi"^^xsd:string ;
  cao:isWitnessedBy <wit-fragment/anno01-wfrag01> ,
  <wit-fragment/anno01-wfrag02> .

<rdg/anno01-rdg02> a cao:Reading ;
  rdf:value "diffudit"^^xsd:string ;
  cao:isWitnessedBy <wit-fragment/anno01-wfrag03> .

<rdg/anno01-rdg03> a cao:Reading ;
  rdf:value "diffundit"^^xsd:string ;
  cao:isWitnessedBy <wit-fragment/anno01-wfrag04> .
```

The property frbroo:R15i_is_fragment_of connects an expression fragment to the expression to which it belongs:

```
<wit-fragment/anno01-wfrag01> a frbroo:F23_Expression_Fragment ;
  frbroo:R15i_is_fragment_of <wit-expression/wexp01> .

<wit-fragment/anno01-wfrag02> a frbroo:F23_Expression_Fragment ;
  frbroo:R15i_is_fragment_of <wit-expression/wexp02> .
```

Expressions cannot exist without a carrier. For example, a handwritten note could not exist without the piece of paper on which it is written. A carrier is, in FRBR terms, a manifestation. FRBRoo distinguishes between two classes of manifestation: Manifestation Singleton and Manifestation Product Type. A Manifestation Singleton belongs to the realm of physical things: it is a unique, physical object such as a manuscript, and cannot be replicated. On the other hand, a Manifestation Product Type is an abstract notion: it comprehends publications, which can exist in multiple physical copies.

Among the type of witnesses that a scholar may encounter are manuscripts, either handwritten or digital, which fall into the category of manifestation singletons, and printed editions, which belong to the class of manifestation product types. The properties crm:P128i_is_carried_by and frbroo:R4_carriers_provided_by relates an Expression to a Manifestation Singleton and a Manifestation Product Type, respectively:

```
<wit-expression/wexp01> a frbroo:F2_Expression ;
  crm:P128i_is_carried_by <wit/wit01> .

<wit-expression/wexp01> a frbroo:F2_Expression ;
  frbroo:R4_carriers_provided_by <wit/wit02> .
```

At this stage, possibilities of connection with other datasets open up. There are, indeed, two main ways to describe the witnesses involved in the reconstruction of a critical text: describing each witness within the model in detail, or relying on external descriptions, such as those provided by museums, cultural institutions, and other datasets (e.g., the British Museum, Pleiades, GeoNames, Worldcat, VIAF, DBpedia). CRM-FRBRoo is fully equipped for the description of manuscripts as museum artefacts, as well as for publications.

For example, the URI for the object of the property `crm:P128i_is_carried_by` may link to information belonging to external cultural heritage datasets: the URI itself can be directly reused from already existing records; alternatively, project-specific URIs can be paired with external representations of the same objects by means of the property `owl:sameAs`, as follows:

```
<wit/wit01> a frbroo:F4_Manifestation_Singleton ;
  owl:sameAs <http://collection.britishmuseum.org/id/object/exampleID> .
```

To sum up, a reading (`cao:Reading`) is a scholarly interpretation of a witness fragment (`frbroo:F23_Expression_Fragment`) that belongs to a witness (`frbroo:F2_Expression`), which is carried by a physical document (`frbroo:F4_Manifestation_Singleton`) or an edition (`frbroo:F3_Manifestation_Product_Type`). Other types of witnesses may also be described by using other elements of FRBRoo.

## 6 A Special Case of Variation: Transposition

Transposition occurs when text is transferred from one location to another. As transposition involves a relationship between a transposed text and (at least) two distinct locations, its representation requires that the annotation be linked to multiple targets:

```
<annotation/anno02> a oa:Annotation ;
  oa:hasBody <app/app02> ;
  oa:hasTarget <varloc/vl02>, <varloc/vl03> ;
  dcterms:creator <person/fgiovan> ;
  dcterms:created "2019-03-09"^^xsd:date .
```

The transposed text is then related to the original and the new location via the properties `cao:wasTransposedFrom` and `cao:wasTransposedTo`, respectively:

```
<rdg/anno02-rdg02> a cao: Reading ;
  cao:hasReadingType cao:transposition ;
  cao:wasTransposedFrom <varloc/vl02> ;
  cao:wasTransposedTo <varloc/vl03> .
```

## 7 The Target of the Annotation: Linking the Apparatus to the TEI Document

The example provided below shows a method for linking the apparatus to a specific base text within the TEI document containing the edition. The Web Annotation class oa:RangeSelector allows the identification of the beginning and of the end of the location of the variant using XPath expressions:

```
<l>
... freta rapidisque <anchor xml:id="varloc-s01"/>diffundi<anchor xml:id="varloc-
    e01"/> ...
</l>

<varloc/vl01> a cao:VariationLocation ;
  oa:hasSource <example.xml> ;
  oa:hasSelector [
    a oa:RangeSelector ;
      oa:hasStartSelector [
        a oa:XPathSelector ;
        rdf:value "//l/anchor[@xml:id='varloc-s01']" ] ;
        oa:hasEndSelector [
          a oa:XPathSelector ;
          rdf:value "//l/anchor[@xml:id='varloc-e01']" ] ] ] .
```

XPointer may alternatively be used to specify the location of the variant:

```
<varloc/vl01> a cao:VariationLocation ;
  oa:hasSource <example.xml> ;
   oa:hasSelector [
     a oa:FragmentSelector ;
      dcterms:conformsTo <http://tools.ietf.org/rfc/rfc3023> ;
      rdf:value "xpointer(//anchor[@xml:id='varloc-s01']/range-to(//anchor
[@xml:id='varloc-e01'])" ] .
```

The TEI double-end-point-attached method solves the issue of overlapping lemmata. There are, however, other methods which can be used for linking an RDF apparatus to the text. For example, if the location of the variant is tagged using a generic TEI element such as <seg>, the apparatus entry can directly point to the unique identifier specified for that element:

```
<l>... freta rapidisque <seg xml:id="varloc-01">diffundi</seg> ...</l>
<varloc/vl01> a cao:VariationLocation ;
  oa:hasSource <example.xml> ;
   oa:hasSelector [
     a oa:XPathSelector ;
     rdf:value "//l/seg[@xml:id="varloc-01]" ] .
```

It is also feasible to replace the lemma with an empty element in the TEI document, leaving a gap in the critical text (this would allow the editor not to choose a base or preferred text). Such a method, however, may result in loss of information if the TEI document containing the edition and the RDF critical apparatus are not processed together.

A Web Annotation class oa:TextPositionSelector is also available for describing ranges of characters, making it possible to link an RDF critical apparatus to a plain text document. Using this method, however, would make the connection between the apparatus and the text more fragile as compared to using @xml:ids for building the URIs representing the location of variation.

For a more detailed overview of the Critical Apparatus Ontology, please visit the ontology specification at http://w3id.org/cao.

## 8 Generating, Visualizing, and Querying an RDF Critical Apparatus

This study set out to explore the use of a conceptual model, the Critical Apparatus Ontology (CAO), in combination with TEI text encoding to describe the critical apparatus, and discussed the benefits of this approach. However, there are practical and methodological issues such as how to generate, visualize, and query an RDF critical apparatus, which deserve consideration: the lack of user-friendly tools for working with RDF and ontologies is a barrier for the adoption and diffusion of Semantic Web technologies in the context of digital scholarly editing (Pierazzo 2016).

There are different ways to generate an RDF critical apparatus. A first method is to directly extract the set of RDF statements from an existing TEI critical apparatus by means of an existing Python script which employs lxml, a library for manipulating XML documents (see Behnel 2019), and RDFLib, a package for working with RDF (see RDFLib Team 2013). A limitation of this approach is that the TEI encoding model applied to the critical apparatus must follow a specific structure and set of guidelines; otherwise it will not be possible for the script to convert the critical apparatus in its entirety. For example, in case of author's or copyist's additions, the script requires
`<rdg xml:id="rdg032" type="cao:addition">some added text</rdg>`[5]
rather than

```
<rdg xml:id="rdg032">
  <add>some added text</add>
</rdg>
```

Another possible strategy for generating an RDF critical apparatus is represented by RDFa, which allows the embedding of RDF into TEI documents through attributes, and comes with ready-for-use extraction tools (see, for example, Tittel et al. 2018 and Ruiz et al. 2020).

Nonetheless, both options present the same limitation: there is not a standard way of encoding certain features of a critical apparatus, such as specific types of relationships between variant readings, with TEI.[6] Therefore, it is not possible to

[5] Note that any vocabulary for the classification of reading types is allowed, as long as its prefix is defined in a `<prefixDef>` element.

[6] The only type of relationships that have a standard method of TEI encoding (the `@varSeq` attribute) are generic sequential relationships, translated into CAO as `cao:follows` relationships.

automatically extract such features unless alternative encoding methods, such as the `<graph>` or `<relation>` elements, are used.[7]

That being said, the exercise of human judgement (and manual intervention) on the generated triples remains central to the development of reliable critical apparatuses.

Studies on visualization are, without a doubt, indispensable for enabling regular web users to make sense of the RDF data presented, and for the diffusion of the LOD paradigm. There are several existing tools for the visualization of semantic knowledge bases which could transform RDF critical apparatuses into meaningful graph visualizations (for example, see Bastian et al. 2009). Alternatively, ad hoc visualizations may be created by means of the aforementioned RDFLib (SPARQL queries are used to extract relevant information from the knowledge graph) or the more familiar XSLT. For example, we could imagine a base text or, in its absence, a user-selected text with points of variation highlighted, and made interactive. The user can ask for different kinds of visualizations, such as variant graphs, variants in context, correction processes, outgoing links, raw RDF statements, and so forth.

A web scholarly edition enhanced by means of semantic graphs should also provide a SPARQL endpoint to allow advanced searching and extraction of the data. End users should be able to compose queries in abstract terms, without needing to know any specific formal query language (Ciotti 2018).

## 9 Conclusion and Future Work

Representing the critical apparatus within a document-driven environment such as TEI XML can be difficult, as a critical apparatus sits at the intersection of multiple texts, and establishes a complex network of relationships among these texts. RDF is a data-driven modelling framework. Compared to XML, RDF is closer to the way natural language communication is structured. The adoption of RDF knowledge graphs in combination with TEI may: facilitate the development of software for the visualization of variant graphs; provide scholars with more powerful ways of expressing their analysis and interpretations; allow users to perform semantic queries; increase interoperability between digital scholarly editions and the cultural heritage linked open data cloud.

The Critical Apparatus Ontology (CAO), currently in its $0.9^{th}$ version, aims to contribute to the publication of data-centric digital scholarly editions. Future work will focus on two main areas: carrying out the necessary tests on existing editorial projects before a stable release of the model; and, research on future updates – such

[7] Encoding RDF triples within `<graph>` or `<relation>` represents a case of tag abuse. However, if TEI truly is “whatever you make it” (Cummings 2012), a widespread tag misuse might quickly become a standard.

as the representation of reading types as classes rather than individuals – which will increase the expressiveness of the model.

## Bibliography

Andrews, Tara L., and Caroline Macé, 'Beyond the Tree of Texts: Building an Empirical Model of Scribal Variation through Graph Analysis of Texts and Stemmata', *Literary and Linguistic Computing*, 28.4 (2013), 504–521

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy, 'Gephi: An Open Source Software for Exploring and Manipulating Networks', in *Third International AAAI Conference on Weblogs and Social Media*, 2009

Behnel, Stefan, *Lxml - Processing XML and HTML with Python*, 2019 <https://lxml.de/4.3/>

Bekiari, Chryssoula, Martin Doerr, Patrick Le Bœuf, and Pat Riva, *Definition of FRBRoo: A Conceptual Model for Bibliographic Information in Object-Oriented Formalism* (IFLA, 2015)

Cayless, Hugh, *Linked Ancient World Data (LAWD) Ontology* (Linked Ancient World Data Initiative (LAWDI), 2015) <https://github.com/lawdi/LAWD>

Ciotti, Fabio, 'A Formal Ontology for the Text Encoding Initiative', *Umanistica Digitale*, 2.3 (2018)

Ciotti, Fabio, and Francesca Tomasi, 'Formal Ontologies, Linked Data, and TEI Semantics', *Journal of the Text Encoding Initiative*, 9 (2016) <https://doi.org/10.4000/jtei.1480>

Ciula, Arianna, Paul Spence, and José Miguel Vieira, 'Expressing Complex Associations in Medieval Historical Documents: The Henry III Fine Rolls Project', *LLC*, 23 (2008), 311–25 <https://doi.org/10.1093/llc/fqn018>

Cummings, James, 'TEI, What Else?', 2012 <https://prezi.com/_y7t4nulanba/tei-what-else/>

Dadzie, Aba-Sah, and Matthew Rowe, 'Approaches to Visualising Linked Data: A Survey', *Semantic Web* 2,2 (2011), 89–124 <http://semantic-web-journal.net/content/approaches-visualising-linked-data-survey>

Damon, Cynthia, 'Beyond Variants: Some Digital Desiderata for the Critical Apparatus of Ancient Greek and Latin Texts', in *Digital Scholarly Editing*, ed. by Matthew James Driscoll and Elena Pierazzo, Theories and Practices, 1st edn (Open Book Publishers, 2016), ɪᴠ, 201–18 <https://www.jstor.org/stable/j.ctt1fzhh6v.15>

Daquino, Marilena, and Francesca Tomasi, 'Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects', in *Metadata and Semantics Research*, ed. by Emmanouel Garoufallou, Richard J. Hartley, and Panorea Gaitanou, Communications in Computer and Information Science (Springer International Publishing, 2015), 424–36

Dekker, Ronald Haentjens, Dirk Van Hulle, Gregor Middell, Vincent Neyt, and Joris van Zundert, 'Computer-Supported Collation of Modern Manuscripts: CollateX and the Beckett Digital Manuscript Project', *DSH*, 30 (2015), 452–70 <https://doi.org/10.1093/llc/fqu007>

Eide, Øyvind, 'Ontologies, Data Modeling, and TEI', *Journal of the Text Encoding Initiative*, Issue 8 (2014/15) <https://doi.org/10.4000/jtei.1191>

Giovannetti, Francesca, *The Critical Apparatus Ontology (CAO)*, 2019 <http://w3id.org/cao>

IFLA, *Functional Requirements for Bibliographic Records: Final Report* (IFLA, 2009) <https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf>

Jänicke, Stefan, Annette Geßner, Greta Franzini, Melissa Terras, Simon Mahony, and Gerik Scheuermann, 'TRAViz: A Visualization for Variant Graphs', *DSH*, 30 (2015) <https://doi.org/10.1093/llc/fqv049>

Jordanous, Anna, K. Faith Lawrence, Mark Hedges, and Charlotte Tupman, 'Exploring Manuscripts: Sharing Ancient Wisdoms across the Semantic Web', in *WIMS*, 2012 <https://doi.org/10.1145/2254129.2254184>

Lebo, Timothy, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, and others, *Prov-o: The Prov Ontology*, 2013 <http://www.w3.org/TR/prov-o/>

Light, Michelle, and Tom Hyry, 'Colophons and Annotations: New Directions for the Finding Aid', *The American Archivist*, 65.2 (2002), 216–30 <https://doi.org/10.17723/aarc.65.2.l3h27j5x8716586q>

Ore, Christian-Emil, and Øyvind Eide, 'TEI and Cultural Heritage Ontologies: Exchange of Information?', *LLC*, 24 (2009), 161–72 <https://doi.org/10.1093/llc/fqp010>

Pierazzo, Elena, *Digital Scholarly Editing: Theories, Models and Methods - Elena Pierazzo - Libro in Lingua Inglese - Taylor & Francis Ltd - Digital Research in the Arts and Humanities| IBS* (Routledge, 2016) <https://www.ibs.it/digital-scholarly-editing-theories-models-libro-inglese-elena-pierazzo/e/9781472412119>

RDFLib Team, *RDFLib Documentation*, 2013 <https://rdflib.readthedocs.io/en/stable/>

Romanello, Matteo, Monica Berti, Federico Boschetti, Alison Babeu, and Gregory R. Crane, 'Rethinking Critical Editions of Fragmentary Texts by Ontologies', in *ELPUB*, 2009

Ruiz Fabo, Pablo, Helena Bermúdez Sabel, Clara Isabel Martínez Cantón, Elena González-Blanco García, and Borja Navarro Colorado, 'The Diachronic Spanish Sonnet Corpus (DISCO): TEI and Linked Open Data Encoding, Data Distribution and Metrical Findings', in *DSH*, 2020 <https://doi.org/10.1093/llc/fqaa035>

Sanderson, Robert, Paolo Ciccarese, and Benjamin Young, *Web Annotation Data Model. W3C Recommendation*, 2017

Schmidt, Desmond, and Robert Colomb, 'A Data Structure for Representing Multi-Version Texts Online', *International Journal of Human-Computer Studies*, 67.6 (2009), 497–514 <https://doi.org/10.1016/j.ijhcs.2009.02.001>

Sperberg-McQueen, C. Michael, 'A Directed-Graph Data Structure for Text Manipulation', 1989 <http://cmsmcq.com/1989/rhine-delta-abstract.html>

TEI Consortium, *P5: Guidelines for Electronic Text Encoding and Interchange* release 3.6.0, 2019 <https://tei-c.org/guidelines/>

Tittel, Sabine, Helena Bermúdez-Sabel, and Christian Chiarcos, 'Using RDFa to Link Text and Dictionary Data for Medieval French', in *W23 - 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science*, May 12, 2018 Phoenix Seagaia Conference Center Miyazaki, Japan, ed. by John P. McCrae, Christian Chiarcos, Thierry Declerck, Jorge Gracia, Bettina Klimek (presented at the LDL-2018, Myazaki, Japan, 2018), 7–12 <http://lrec-conf.org/workshops/lrec2018/W23/pdf/10_W23.pdf>

# Projects and Editions

# Appendices