



# DHQ: Digital Humanities Quarterly

2022

Volume 16 Number 2

## Linked data from TEI (LIFT): A Teaching Tool for TEI to Linked Data Transformation

Francesca Giovannetti <francesc\_dot\_giovannett6\_at\_unibo\_dot\_it>, University of Bologna  <https://orcid.org/0000-0001-6007-9118>

Francesca Tomasi <francesca\_dot\_tomasi\_at\_unibo\_dot\_it>, University of Bologna  <https://orcid.org/0000-0002-6631-8607>

### Abstract

The purpose of this paper is to introduce *Linked data from TEI* (LIFT), an open source tool written as a set of Python scripts for generating linked data from TEI-encoded texts. LIFT's goal is to walk users through the transformation process from TEI to linked data step by step, as well as to promote a better understanding of the theoretical and methodological aspects that underpin the transformation. LIFT was created in the context of the University of Bologna's Master Degree in Digital Humanities and Digital Knowledge as a teaching tool for students encountering linked open data for the first time as a method of organizing and publishing cultural knowledge and, specifically, digital scholarly editions on the web in a perspective of data integration.

## 1. Introduction<sup>[1]</sup>

Many and diverse scientific communities, including digital textual scholars, have shown increasing interest in semantic web technologies and linked open data (LOD) as a means of knowledge representation since the early 2000s, when Tim Berners-Lee coined the concept of a web of interconnected data rather than simply documents (see [Spadini et al. 2021]). Nonetheless, there is a lack of user-friendly tools for working with digital scholarly editions and LOD. 1

*Linked data from TEI* (LIFT) is a Digital Humanities tool for generating linked open data (LOD) from TEI-encoded texts. LIFT, which is accessible at <https://projects.dharc.unibo.it/lift>, provides a set of TEI-to-LOD transformation scripts written in Python. The scripts, addressing the transformation of different types of entities to LOD, are thoroughly documented to facilitate their understanding and reuse. The purpose of this paper is to introduce LIFT as a teaching tool for supporting the adoption of linked open data in digital scholarly editing. Our plan is to use LIFT as part of a laboratory activity for the University of Bologna's Master Degree in Digital Humanities and Digital Knowledge, which will combine the knowledge and skills acquired by the students from different modules: text encoding with the TEI, ontology development, Python programming, and linked open data for cultural heritage.<sup>[2]</sup> 2

This paper first provides a brief background on linked open data and their use in digital text representation (see Section 2). It then goes on to consider related work in the field of TEI to LOD transformation (see Section 3) as well as to introduce LIFT from both a technical and methodological perspective (see Section 4). Finally, it discusses the potential application of LIFT in digital humanities learning contexts (see Section 5 and Section 6). 3

## 2. Background

The concept of linked open data refers to a method of publishing structured data – as well as organizing knowledge – on the web in such a way that they can be semantically interlinked to form open clouds of integrated information. As is already well known, LOD are based on the Resource Description Framework (RDF), a graph data model used for making statements about resources. Each RDF statement contains a subject, a predicate, and an object, represented by unique URIs. The subject is the resource being described, the object is either a characteristic of the subject (such as its type) or a resource related to the subject, and the predicate expresses the relationship between the subject and the object (see example 1 below). The meaning of RDF data is defined by machine-readable ontologies, formal descriptions of specific concepts within a domain and the relationships holding between them. 4

```
<http://dbpedia.org/resource/Alexander_the_Great> # Alexander the Great
<http://dbpedia.org/ontology/parent> # has parent
<http://dbpedia.org/resource/Philip_II_of_Macedon> # Philip II of Macedon
```

**Example 1.** An RDF representation of the statement “Alexander the Great has parent Philip II of Macedon” (extracted from DBpedia).

RDF statements can be used to represent information in digital scholarly editions as linked open data.<sup>[3]</sup> To cite just a few examples, the Henry III Fine Rolls edition experiments with linked data to express complex relationships between people in historical documents (see [Ciula et al. 2008]). The Sharing Ancient Wisdoms project, on the other hand, uses linked data to reconstruct text interaction (see [Jordanous et al. 2012]), while the Paolo Bufalini’s Notebook edition uses linked data to represent the interconnections between the various types of fragments – quotations, translations, and annotations – that make up the text (see [Daquino et al. 2019]). Similar information extracted from texts, in addition to administrative, structural, and descriptive metadata, can make a significant contribution to the development of a cultural heritage linked open data cloud: texts do indeed occupy a prominent space in cultural heritage representation. However, in order to reuse and integrate texts into LOD systems, information must first be expressed in RDF according to widely used ontologies in the digital humanities and cultural heritage domains.

5

The Text Encoding Initiative (TEI), an XML vocabulary for encoding texts in the humanities, is the de facto standard for digital scholarly editing.<sup>[4]</sup> The standard defines terms for representing a wide range of textual entities, including personal names, place names, dates, concepts, and events. Unlike RDF, the TEI has traditionally taken a document-centric approach to representing such entities, with texts processed as ordered hierarchies of content objects nesting neatly one inside the other (see [DeRose et al. 1990] and [Renear 1993]). However, transitioning from a document-centric paradigm to a data-centric paradigm, such as LOD, is not an easy task: converting TEI editions into graphs of RDF statements necessitates a conceptual shift in which texts are no longer viewed as documents but as collections of interconnected entities (see [Tomasi 2012]).

6

The movement beyond TEI towards linked open data began in 2004, when a TEI Special Interest Group on Ontologies was formed with the goal of investigating the feasibility of encoding information about persons, dates, events, places and objects outside the text, in parallel with the steps taken by the Museum community, which was working on the development of CIDOC CRM, an ontological vocabulary for the description of museum collections (see [TEI 2019]). The new P5 release of the TEI in 2007 introduced new elements for distinguishing real-world entities from their in-text occurrences [Wittern et al. 2009]. This change, along with the addition of new attributes that could be used to link an entity to external authority records (see example 2 below), made it easier to express relationships between entities in TEI and extract them as RDF statements.

7

```
<person sameAs="http://viaf.org/viaf/101353608">Alexander the Great</person>
```

**Example 2.** The attribute @sameAs encode an equivalence between the person “Alexander the Great” in the TEI document and the personal entity identified by the VIAF URI <http://viaf.org/viaf/101353608>.

Despite these meaningful developments, there are very few user-friendly tools for working with TEI editions and LOD [Pierazzo 2016, 121]. The fact that the TEI guidelines are not intended to be strictly followed complicates matters: annotators can encode the same textual features in different ways, depending on the project. Because no two texts or editions are identical, this flexibility is required; however, the unavoidable drawback is a certain loss of interoperability (see [Andrews 2013, 2]). Enhancing TEI-encoded texts with linked open data represents a possible way of overcoming this limitation.

8

Being willing to introduce DHDK students to the extraction of linked open data from TEI-encoded texts, we ran into a lack of user-friendly tools capable of raising awareness about both the methodology and the technology that underpin such operation. LIFT was created to fill this gap.

9

### 3. Related work

As digital humanities practitioners know, XSLT (Extensible Stylesheet Language Transformations) is a language for transforming XML documents into other XML documents. Because the TEI is an XML vocabulary, digital textual scholars have naturally adopted XSLT as their primary transformation language from TEI to other formats including RDF/XML, which is one of the possible serializations of LOD graphs. The TEI Consortium develops and maintains the TEI Stylesheets, a collection of ready-for-use XSL Transformations scripts that includes a TEI-to-RDF transformation. Despite providing a relatively simple method of extracting RDF statements from TEI documents, the Stylesheets have several shortcomings, particularly when used as a teaching tool. The Stylesheets were not, in fact, designed with a didactic aim in mind and do not offer TEI-to-RDF transformation guidelines: users need to possess good XSLT skills to prepare the TEI document, gain insight into the rationale of the transformation, and adapt it to their specific needs. Also, RDF can be serialized in a variety of ways, with Turtle being by far the most widely used due to its legibility. Since the TEI-to-RDF Stylesheet was created, more powerful methods of linked data production have been released. RDFLib, a Python package for working with RDF, is one of these. The key difference between using XSLT and RDFLib to create RDF triples is that the former requires developers to hardcode RDF triples into the transformation stylesheet (i.e., the developer must choose ahead of time a specific serialization and adhere to its syntax), resulting in a significant loss of flexibility and an increase in human errors. RDFLib, on the other hand, automatically serializes RDF triples, allowing you to add properties to an entity at any point along the way without worrying about syntax and structure (see example 3, which shows a comparison between XSLT and RDFLib as two different methods of generating RDF statements). Finally, while the TEI-to-RDF Stylesheet reuse some concepts from the CIDOC Conceptual Reference Model (see [Crofts et al. 2011]) and the Dublin Core (see [Weibel and Traugott 2000]), it does not make use of the expressive potential of ontologies such as FRBRoo (see [Bekiari et al. 2015]) and PRO (see [Peroni et al. 2012]), which have recently gained attention and application in both text and cultural heritage representation (see [Daquino et al. 2019]).

10

Another tool we considered is XTriples. XTriples is a webservice-based infrastructure for extracting RDF statements from collections of XML documents [Gruntgens and Torsten 2016]. Unlike the Stylesheets, XTriples can produce multiple serializations of the resulting RDF graph and comes with a documentation. The transformation is customizable via an XPath-based configuration file. LOD beginners, however, may find this operation difficult and the tool not immediately usable or replicable.

11

There are also tools, such as Recogito, that enable the extraction of specific features from TEI-encoded texts. Recogito, developed by the Pelagios Network, allows users to add annotations to specific passages of text, which they can then download as RDF statements (see [Simon et al. 2017]). The transformation is not configurable nor explained as this is not the focus of the tool.

12

## TEI to RDF/XML with XSLT

```
<rdf:Description rdf:about="{ $person_uri }">
  <rdf:type rdf:resource="http://schema.org/Person"/>
  <owl:sameAs rdf:resource="{ $same_as_uri }"/>
  <rdfs:label>
    <xsl:attribute name="xml:lang">
      <xsl:value-of select="{ $label_lang }"/>
    </xsl:attribute>
    <xsl:value-of select="{ $label }"/>
  </rdfs:label>
</rdf:Description>
```

## TEI to RDF/XML with RDFLib

```
g.add((person_uri, RDF.type, schema.Person))
g.add((person_uri, OWL.sameAs, same_as_uri))
g.add((person_uri, RDFS.label, Literal(label, lang=label_lang)))
g.serialize(format="xml")
```

**Example 3.** XSLT versus RDFLib TEI-to-LOD transformation; the second method offers a more flexible and less error-prone way of creating RDF statements.

## 4. *Linked data from TEI (LIFT)*<sup>[5]</sup>

Based on the shortcomings discussed above, we decided to develop a TEI-to-LOD transformation tool for our students that would simplify the transformation process while also openly and transparently documenting the workflow, with the goal of making such workflow modifiable and adaptable to new contexts. *Linked data from TEI (LIFT)* is a Python-based, open-source application for creating graphs of RDF statements from TEI documents. LIFT targets digital humanities students as well as textual scholars experimenting for the first times with linked open data in digital scholarly editing. The main component of LIFT is a collection of TEI-to-LOD transformation scripts. The scripts, written in Python, take as input a TEI document and generate an RDF graph. Each script addresses the transformation of specific TEI entities including persons, personal relations, places, and events. A transformation script for the TEI critical apparatus is also available (see Section 4.2). The application is extensively documented in order to facilitate the reuse and adaptation of the transformation workflow. The documentation walks users through the process of preparing their TEI document for transformation, describes the expected RDF graph, and provides step-by-step explanations of the transformation scripts in the form of an interactive Jupyter notebook (see Section 4.3).

Users can access LIFT through a browser or download the collection of scripts for local use. A basic understanding of the TEI and RDF is recommended (the ideal target audience are digital humanities students who are already familiar with digital scholarly editions, the TEI, RDF and ontologies, and basic Python programming).

LIFT's web interface has a straightforward design. When using a browser to access LIFT, users are directed to the homepage, which contains instructions for getting started with the transformation (see Figure 1): by clicking on the 'Quick start' tab, an upload bar for the input TEI document becomes available (see Figure 2). After uploading the TEI document into LIFT, users are directed to another page where they can select a transformation option (see Figure 3). If the transformation is successful, multiple serializations of the resulting RDF graph become available for download (see Figure 4).

### 4.1 Methodology

The creation of linked open data from TEI documents is more than a technical operation. While digital scholarly editions based solely on the TEI standard are document-centric, LOD-based resources are data-centric: the transformation of the information conveyed by the TEI markup into RDF statements forming a graph necessitates careful methodological reflection as the edition is disassembled to form a dataset of interconnected entities, each represented by a URI or a literal.

13

14

15

16

17

In addition to reorganizing tree-structured information into a graph data structure, the transition from TEI to linked data necessitates assigning a formal semantics to the entities by reusing existing ontologies whenever possible. As of today, there is no TEI-specific ontology, so all elements and attributes must be mapped to classes and properties from arbitrary ontologies (see [Eide 2014] and [Ciotti and Tomasi 2016]). There are two major challenges to performing such mapping. First, the chosen ontologies assign a formal semantics to the data, influencing how information is accessed and reused in the future. Users should be aware of the implications of using ready-to-use conversion tools like LIFT: the information conveyed by the original encoding may change when moving from one vocabulary, e.g. the TEI, to another, e.g. FRBRoo, CIDOC CRM, PROV-O, etc. Second, a knowledge graph is composed of entities (i.e. concepts), and links between entities (i.e. relationships between concepts). The distinction between what will become a concept and what will become a relationship is not inconsequential. In TEI, there are at least two ways to represent a relationship. The first approach is to make use of an attribute. The line `<rdg wit="#A">liber</rdg>`, for example, indicates that the reading 'liber' is witnessed by manuscript 'A', where 'is witnessed by' is the relationship between the reading and the manuscript. A second method is to nest elements within each another. The encoding `<app><rdg wit="#A">liber</rdg><rdg wit="#B">libellus</rdg></app>`, for example, indicates that 'libellus' is a variant of 'liber'. Although interpreting the meaning of the relationships in the preceding examples is simple and without ambiguity, this is not always the case.

When developing LIFT, we decided to set some predefined requirements for the encoding of the input TEI document so to ensure a meaningful, error-free restructuring and formalization of the information (see Section 4.2). The encoding prerequisites can be adjusted locally by modifying the scripts. As of semantics, we decided to reuse a set of ontologies that are widely adopted in the fields of cultural heritage and text representation: the CIDOC Conceptual Reference Model (CIDOC CRM) to represent people, places and events; the Agent Relationship Ontology (AgRelOn) (see [Litz et al. 2012]) to describe personal relationships; the Functional Requirements for Bibliographic Records object-oriented (FRBRoo) for texts; the Dublin Core Metadata Terms (DCMI Metadata Terms) to describe bibliographic resources; PROV-O to link the account of an event to its source; the Publishing Roles Ontology (PRO), which enables the reification of roles in such a way that each role is bounded to a specific context (e.g. the role of 'bride' in the context of a particular ceremony); the Time-indexed Value in Context (TVC) (see, again, [Peroni et al. 2012]) to represent time; the property owl:sameAs from the Web Ontology Language (OWL) to interconnect the entities to semantically equivalent resources outside the edition; the property rdfs:label and rdf:value from the Resource Description Framework Schema (RDF Schema), which are used to include human-readable labels and TEI snippets in the graph, respectively. Finally, the Critical Apparatus Ontology (CAO) is leveraged to represent critical apparatus entries (see [Giovannetti 2021]).

A combination of multiple ontologies is required to express the various concepts and relationships underpinning editions, as well as to demonstrate to students how multiple ontologies can be used in the same knowledge graph.

## 4.2 The transformation scripts

LIFT's collection of transformation scripts include:

1. Persons only
2. Persons and events
3. Persons and relations
4. Places only
5. Persons, events, relations, and places
6. Critical apparatus

We decided to write multiple scripts because we wanted the transformation to be modular and the resulting RDF graph's complexity to be incremental. Script 5 combines the previous four into a single transformation option. For the scripts to work, the input TEI document must follow specific guidelines. The TEI, as previously stated, allows for multiple ways to encode the same textual features. For example, in order to markup a personal name, one can use the tag `<persName>`, the tag `<name>`, or even the tag `<rs>` (see <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ND.html>) This is an important feature in terms of flexibility, but it makes creating a universal TEI-to-RDF transformation script difficult. As a result, LIFT's documentation includes a set of encoding guidelines intended to ensure a smooth TEI-to-LOD transformation (example 4 below juxtaposes the input TEI encoding and the resulting RDF statements). In particular, the input TEI document must conform to the following indications:<sup>[6]</sup> 1. Any TEI element being transformed must be assigned a unique identifier within an `@xml:id` attribute; 2. The TEI document must contain a TEI header, even if minimal; the `<person>` element must be used to describe people in the TEI header, and the `<persName>` element must be used to markup in-text occurrences of such people; 3. In a similar way, the `<place>` element must be used to describe places in the TEI header, and the `<placeName>` element must be used to markup

in-text occurrences of such places; 4. People and places described in the TEI header should be assigned a `@sameAs` attribute containing links to external authority records (e.g. VIAF); 5. Any relationship between people should be described in the TEI header within `<listRelation>`; 6. Similarly, events should be described using the `<event>` element within `<person>` or `<place>`. An example input file is also provided when accessing LIFT from the browser. If modifying the TEI encoding to meet the requirements of LIFT is not possible, the documentation in general, and the interactive Jupyter notebook in particular, will help users gain the understanding needed to adapt the transformation to a different input.

#### Input (TEI)

```
<person xml:id="socr" sameAs="http://viaf.org/viaf/88039167">
  <persName xml:lang="en">Socrates</persName>
</person>
[...]
<p xml:id="para01">An example of paragraph mentioning <persName
ref="#socr">Socrates</persName>.</p>
```

#### Output (RDF)

```
@prefix crm:<http://www.cidoc-crm.org/cidoc-crm/> .
@prefix dcterms:<http://purl.org/dc/terms/> .
@prefix frbroo:<http://iflastandards.info/ns/fr/frbr/frbroo/> .
@prefix owl:<http://www.w3.org/2002/07/owl#> .
@prefix rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:<http://www.w3.org/2000/01/rdf-schema#> .

<https://example.org/person/socr> a crm:E21_Person ;
  rdfs:label "Socrates"@en ;
  dcterms:isReferencedBy <https://example.org/text/para01> ;
  owl:sameAs <http://viaf.org/viaf/88039167> .

<https://example.org/text/para01> a frbroo:F23_Expression_Fragment ;
  frbroo:R15i_is_fragment_of <https://example.org/example_v1> ;
  rdf:value ""^rdf:Literal .
```

**Example 4.** The input TEI encoding and the corresponding RDF output.

LIFT, as stated earlier in the introduction, uses Python as its transformation language. Python, which is increasingly being taught in digital humanities courses around the world, offers more efficient and error-free ways of working with RDF and linked open data than XSLT (see Section 3 above). 22

RDFLib is a Python library that specializes in RDF triple creation and management (see the documentation at <https://rdflib.readthedocs.io>). Using RDFLib to create RDF triples is very simple, and the amount of human effort required is minimal. URIs can be saved as variables, and RDF statements can be written on separate lines. RDFLib then generates a single graph with no repetitions. Multiple serializations are possible, including XML, Turtle, N-Triples, and JSON-LD. 23

LIFT uses another Python library, `lxml`, in conjunction with RDFLib to parse TEI documents (see the documentation at <https://lxml.de>). When using XSLT to process XML files, one must be familiar with the structure of the document to be processed. However, using `lxml` and Python, it is possible to work with less well-known inputs. 24

example 4 above shows a portion of a TEI document and the corresponding RDF graph in LIFT. example 5 below provides some components of the transformation script used for generating such output. LIFT searches the TEI document for all occurrences of `<person>`. The value of the `@xml:id` attribute associated with the person is then obtained and used to generate the URI representing that person. LIFT looks also for any in-text reference to the person. To do so, it searches the TEI document for all occurrences of `<persName>` featuring a `@ref` attribute linking to the person. Each person is assigned to the CIDOC CRM class E21 Person. If a `@sameAs` attribute is provided in the TEI document for the `<person>` element, LIFT leverages its value or values to create interconnections to external authority records (using the OWL property `owl:sameAs`). All data used for building the RDF graph are extracted from the TEI document. The content of the `<persName>` element provided within `<person>` is used to generate a human- 25



readable label for the entity. The language is specified by the `@xml:lang` attribute. As the reader can see from this example, LIFT uses a predefined set of ontologies to describe the text (see Section 4.1). In order to modify this behaviour, users must modify the scripts. To facilitate this effort, as already anticipated, LIFT provides a Jupyter notebook that guides users through the scripts step-by-step and interactively (see Section 4.3).

```
for person in root.findall('./tei:person', tei):
    person_id = person.get('{http://www.w3.org/XML/1998/namespace}id')
    person_uri = URIRef(base_uri + '/person/' + person_id)
    person_ref = '#' + person_id
    g.add( (person_uri, RDF.type, crm.E21_Person))

    same_as = person.get('sameAs').split()
    if same_as is not None:
        same_as = same_as.split()
        i = 0
        while i < len(same_as):
            same_as_uri = URIRef(same_as[i])
            g.add( (person_uri, OWL.sameAs, same_as_uri))
            i += 1

    persname = person.find('./tei:persName', tei)
    if persname is not None:
        label = persname.text
        label_lang = persname.get('{http://www.w3.org/XML/1998/namespace}lang')
        if label_lang is not None:
            g.add( (person_uri, RDFS.label, Literal(label, lang=label_lang)))
        else:
            g.add( (person_uri, RDFS.label, Literal(label)))
```

**Example 5.** An excerpt of one of LIFT's transformation script creating RDF triples about persons.

### 4.3 The documentation

In developing LIFT as a tool for supporting the adoption of linked open data in digital scholarly editing, we followed the principle that no tool should be used as a black box. Students/users of LIFT must be provided with the knowledge required to understand the methodology and the technology informing the conversion. LIFT is accompanied by an extensive documentation available at <http://linked-data-from-tei.rtfid.io/> presenting the full Python-based workflow for TEI-to-LOD transformation. LIFT's documentation is divided into four main sections:

26

1. *Prepare your TEI document*, which explains the encoding requirements for the input TEI document;
2. *The RDF graph*, illustrating the structure and semantics of the generated knowledge graph by comparing the TEI input with the corresponding RDF output (this section also discusses why and how users might choose different ontologies for their projects or modify the scripts to work with different input data);
3. *How the scripts work*, where users can access an interactive Jupyter notebook containing LIFT's scripts with instructions and line-by-line explanations that are meant to help users develop the skills required to write their own transformation scripts;<sup>[7]</sup>
4. *Further readings and resources*, which lists relevant publications and provides examples of TEI digital scholarly editions which have been enriched by means of linked open data.

In addition to the documentation, each semantic statement of the resulting RDF graph is accompanied by the original TEI construct that generated it facilitating the comparison between the TEI input and the RDF output: this provides users with indirect support as they can view the input TEI encoding and the output RDF graph side by side and go back to adjust the input TEI document to resolve any transformation issues that may arise (see example 4 above).

31

## 5. LIFT in the classroom

Digital humanities, as a discipline, lie at the crossroads of theory and practice. The learning-by-example paradigm, which entails presenting existing models that can be adapted to address new problems, is particularly effective in

32

similar interdisciplinary contexts (see [Tomasi 2020]). LIFT was created in response to a perceived lack of teaching tools for TEI to LOD transformation based on such paradigm.

LIFT will be used in a text representation laboratory activity for the University of Bologna's Master Degree in Digital Humanities and Digital Knowledge. As anticipated at the beginning of this paper, the activity aims to combine the knowledge and skills acquired by the students from various modules and, particularly, text encoding with the TEI, ontology development, Python programming, and linked open data for cultural heritage. The plan for the activity involves gathering the students in 3-member groups. Each group will choose an existing digital scholarly edition and work on the underlying TEI document to ensure that the encoding requirements of LIFT are met.<sup>[8]</sup> The groups will then use LIFT, either via browser or locally, to produce an RDF graph. Finally, each group will present their findings to the class, highlighting challenges and solutions in transforming each type of entity. Following activities will shift the emphasis away from the input TEI document as students will focus on modifying the transformation scripts to adapt them to the existing digital scholarly editions rather than the other way around. By experimenting with the shift from tree-structured to graph-structured data, and from the TEI vocabulary to RDF using multiple ontologies, students will gain a better understanding of the implications of annotating a text in one way versus another within the context of a real-world task that requires them to combine different skills and competencies that are typically taught separately. Furthermore, because the activity is structured as group work on a specific problem to be solved, students will be able to strengthen their ability to collaborate as a team, which is one of the pillars of digital humanities scholarship and, by extension, should be one of digital humanities pedagogy (see [Licastro et al. 2020]).

33

The laboratory activity will also allow us to evaluate the clarity and effectiveness of LIFT as a teaching tool both directly through an end-of-activity questionnaire and indirectly through the students' presentations.

34

## 6. Conclusion and next steps

This paper introduced LIFT, an open-source tool written as a set of Python scripts for generating linked data from TEI-encoded texts. The use of linked open data for text representation facilitates interoperability with other cultural resources on the web and opens up editions to new areas of cultural heritage research. LIFT's goal is to walk users through the transformation process from TEI to linked data step by step, as well as to promote a better understanding of the theoretical and methodological aspects that underpin the transformation.

35

Some of the key characteristics of the evolving web of data include open-source tools, well-documented applications, and shared ontologies. User-friendly resources that reduce the complexity of transforming digital scholarly editions from a document-centric to a data-centric model are also critical in the development and expansion of a linked open data cloud of cultural heritage. We envisage LIFT both as a teaching tool and a ready-for-use resource for TEI-to-LOD transformation, hoping it will serve as an example of a digital humanities teaching tool that raises awareness about the methodological and theoretical implications of the structural and semantic shift that occurs when transitioning from a TEI-based text representation to LOD.

36

With LIFT, we also aim to encourage further research into the development of open-source, user-friendly tools aiding the mutual integration of digital scholarly editions and the cultural heritage linked open data cloud. Such tools have the potential to make digital humanities, and especially knowledge representation, a more inclusive field of study and research. In order to test this idea, we will use LIFT in the context of the University of Bologna's Master Degree in Digital Humanities and Digital Knowledge, as a teaching tool for students encountering linked open data for the first time as a method of publishing digital scholarly editions on the web in a data integration perspective. We expect that LIFT will help students understand the process not only practically but also, and primarily, in terms of methodology and theory.

37

LIFT is under active development. Features that will be provided in the near future include a transformation script for bibliographic references encoded using specialized elements from the TEI model.bibLike class. Furthermore, the documentation will be expanded with new sections on ontologies for text representation as well as useful resources, particularly examples of digital scholarly editions enhanced through the use of linked open data.

38



# LIFT Linked data from TEI

HOME QUICK START DOCUMENTATION SOURCE CREDITS

LIFT is a Python-based tool for transforming your TEI XML edition into a knowledge graph, ready for publication as linked open data on the web. LIFT comes with a thorough documentation, designed to help you understand and reuse the methodology and technology behind the tool.

- In order to get started, go to [Quick start](#). Make sure you read LIFT's encoding guidelines beforehand (see [Prepare your TEI XML edition for transformation in Documentation](#)).
- Go to [Documentation](#) to access useful information to help you use LIFT, including how to prepare your TEI XML edition for transformation to linked data. A Jupyter notebook illustrating step-by-step how the scripts work is also available through the documentation.
- From [Source](#) you can download the transformation scripts for local use.


 This work is licensed under a [Creative Commons Attribution 4.0 International License](#)

Figure 1. Homepage

The screenshot shows the LIFT website interface. At the top, the title "LIFT Linked data from TEI" is displayed in a large, dark font. Below the title is a dark navigation bar with the following links: HOME, QUICK START, DOCUMENTATION, SOURCE, and CREDITS. The "QUICK START" link is highlighted with a red underline. The main content area features a section titled "Upload the TEI XML source file of your edition:" followed by a file selection button labeled "Scegli file" and the text "Nessun file selezionato", and an "Upload" button. Below this, a note states: "Note: for testing purposes, you can download these test TEI XML files, which you can use as an input for the transformation tool:". A numbered list follows: "1. Persons, places, relationships, and events" and "2. Critical apparatus (using the Critical Apparatus Ontology (CAO))". At the bottom of the page, there is a Creative Commons Attribution 4.0 International License logo and the text "This work is licensed under a Creative Commons Attribution 4.0 International License".


Figure 2. Quick start

# LIFT Linked data from TEI

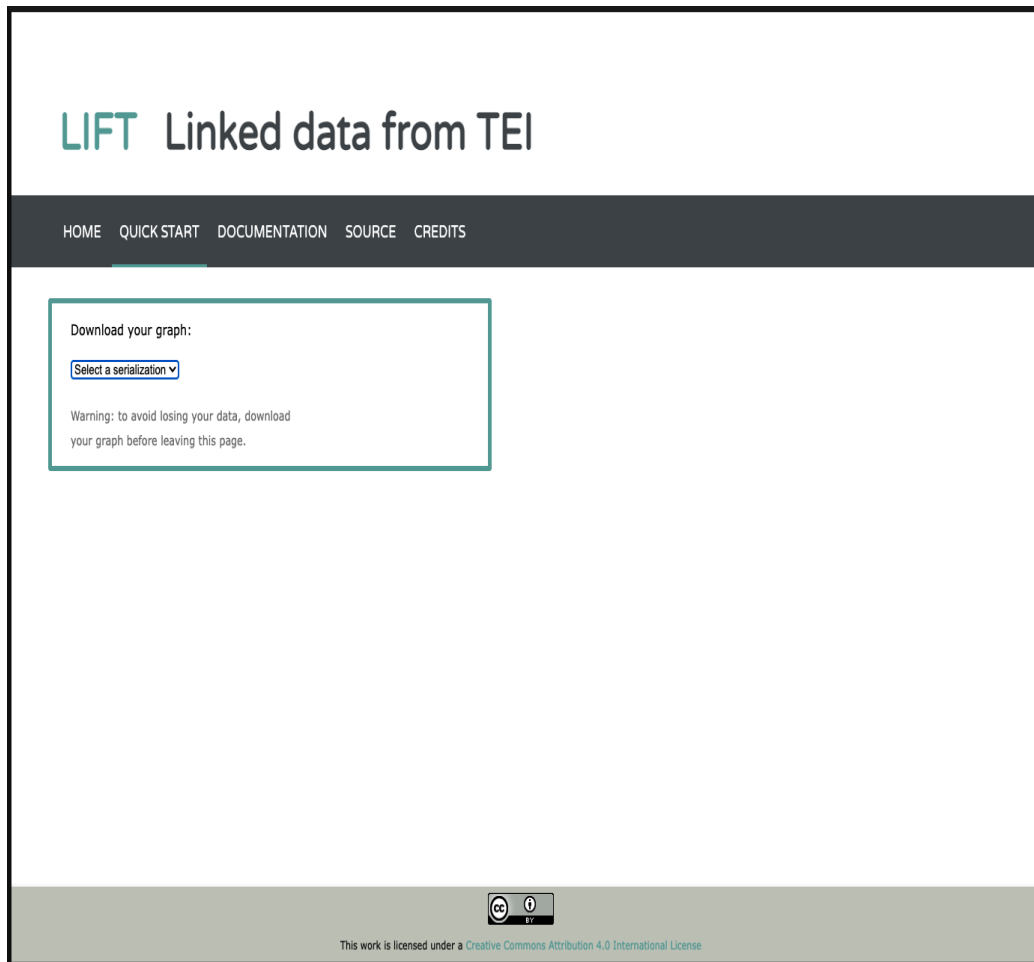
HOME QUICK START DOCUMENTATION SOURCE CREDITS

Select the entities you want to extract as linked open data:

- Persons
- Persons and events
- Persons and relations
- Places
- Persons, events, relations, and places
- Bibliographic references (upcoming)
- Critical apparatus

  
This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

**Figure 3.** Choosing a transformation



**Figure 4.** Downloading the resulting knowledge graph according to different serializations

## Notes

- [1] This paper is the result of a collaborative work. In particular, Francesca Tomasi is responsible for sections 1, 5 and 6, while Francesca Giovannetti is responsible for sections 2, 3 and 4.
- [2] The modules are Digital Text in the Humanities: Theories, Methodologies and Applications; Knowledge representation and extraction; Computational thinking and programming, and Knowledge organization and cultural heritage (see the full descriptions at <https://corsi.unibo.it/2cycle/DigitalHumanitiesKnowledge>).
- [3] For a definition of digital scholarly edition, see [Sahle 2016].
- [4] On the role of the TEI standard in digital scholarly editing, see [MLA 2016], [Price 2008], and [Sahle 2014].
- [5] A Github repository containing all of LIFT's components can be found at <https://github.com/fgiovannetti/lift>. The web application can be found at <https://projects.dharc.unibo.it/lift/>.
- [6] Further requirements apply when transforming a critical apparatus (see, again, [Giovannetti 2021]).
- [7] An online, read-only version of the notebook is available at [https://nbviewer.jupyter.org/github/fgiovannetti/lift/blob/master/jupyter\\_nb/TEItoRDF.ipynb](https://nbviewer.jupyter.org/github/fgiovannetti/lift/blob/master/jupyter_nb/TEItoRDF.ipynb).
- [8] Patrick Sahle's and Greta Franzini's catalogues of digital scholarly editions (<https://v3.digitale-edition.de/>; <https://dig-ed-cat.acdh.oeaw.ac.at/>) will be provided as a starting point, along with RIDE (<https://ride.i-d-e.de/>).

## Works Cited

- Andrews 2013** Andrews, Tara L. 2013. "The Third Way: Philology and Critical Edition in the Digital Age". Application/pdf. <https://doi.org/10.7892/BORIS.43071>.
- Bekiari et al. 2015** Bekiari, Chryssoula, Martin Doerr, Patrick Le Boeuf, and Pat Riva. 2015. "Definition of FRBRoo: A Conceptual Model for Bibliographic Information in Object-Oriented Formalism". 2.4. IFLA.

- Ciotti and Tomasi 2016** Ciotti, Fabio, and Francesca Tomasi. 2016. "Formal Ontologies, Linked Data, and TEI Semantics". *Journal of the Text Encoding Initiative*, no. Issue 9 (September). <https://doi.org/10.4000/jtei.1480>.
- Ciula et al. 2008** Ciula, A., P. Spence, and J. M. Vieira. 2008. "Expressing Complex Associations in Medieval Historical Documents: The Henry III Fine Rolls Project". *Literary and Linguistic Computing* 23 (3): 311–25. <https://doi.org/10.1093/lc/fqn018>.
- Crofts et al. 2011** Crofts, Nick, Martin Doerr, Tony Gill, Stephen Stead, and Matthew Stiff, eds. 2011. "CIDOC CRM: CIDOC Conceptual Reference Model", November. <http://www.cidoc-crm.org/cidoc-crm/>.
- Daquino et al. 2014** Daquino, Marilena, Silvio Peroni, Francesca Tomasi, and Fabio Vitali. 2014. "Political Roles Ontology (PRoles): Enhancing Archival Authority Records through Semantic Web Technologies". *Procedia Computer Science* 38: 60–67. <https://doi.org/10.1016/j.procs.2014.10.012>.
- Daquino et al. 2019** Daquino, Marilena, Francesca Giovannetti, and Francesca Tomasi. 2019. "Linked Data per le edizioni scientifiche digitali. Il workflow di pubblicazione dell'edizione semantica del quaderno di appunti di Paolo Bufalini". *Umanistica Digitale*, December, No 7 (2019). <https://doi.org/10.6092/ISSN.2532-8816/9091>.
- DeRose et al. 1990** DeRose, Steven J., David G. Durand, Elli Mylonas, and Allen H. Renear. 1990. "What Is Text, Really?". *Journal of Computing in Higher Education* 1 (2): 3–26. <https://doi.org/10.1007/BF02941632>.
- Eide 2014** Eide, Øyvind. 2014. "Ontologies, Data Modeling, and TEI". *Journal of the Text Encoding Initiative*, no. Issue 8 (December). <https://doi.org/10.4000/jtei.1191>.
- Giovannetti 2021** Giovannetti, Francesca. 2021. "The Critical Apparatus Ontology (CAO): Modelling the TEI Critical Apparatus as a Knowledge Grap". In *Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing*, edited by Elena Spadini, Francesca Tomasi, and Georg Vogeler, 125–39. Schriften Des Instituts Für Dokumentologie Und Editorik, Band 15. Norderstedt: BoD – Books on Demand.
- Gruntgens and Torsten 2016** Gruntgens, Max, and Torsten Schrade. 2016. "Data Repositories in the Humanities and the Semantic Web: Modelling, Linking, Visualising", 12.
- Jordanous et al. 2012** Jordanous, Anna, K. Faith Lawrence, Mark Hedges, and Charlotte Tupman. 2012. "Exploring Manuscripts: Sharing Ancient Wisdoms across the Semantic Web". In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics - WIMS 12*, 1. Craiova, Romania: ACM Press. <https://doi.org/10.1145/2254129.2254184>.
- Licastro et al. 2020** Licastro, Amanda, Katina Rogers, and Danica Savonick, eds. 2020. "Collaboration". In *Digital Pedagogy in the Humanities: Concepts, Models, and Experiments*. Modern Language Association. <https://digitalpedagogy.hcommons.org/keyword/Collaboration>.
- Litz et al. 2012** Litz, Berenike, Aenne Löhden, Jan Hannemann, and Lars Svensson. 2012. "AgRelOn – An Agent Relationship Ontology". In *Metadata and Semantics Research*, edited by Juan Manuel Doderó, Manuel Palomo-Duarte, and Pythagoras Karampiperis, 343:202–13. Communications in Computer and Information Science. Berlin, Heidelberg: Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-642-35233-1\\_21](https://doi.org/10.1007/978-3-642-35233-1_21).
- MLA 2016** MLA Committee on Scholarly. n.d. "MLA Statement on the Scholarly Edition in the Digital Age". Modern Language Association. Accessed 29 September 2019. <https://www.mla.org/Resources/Research/Surveys-Reports-and-Other-Documents/Publishing-and-Scholarship/Reports-from-the-MLA-Committee-on-Scholarly-Editions/MLA-Statement-on-the-Scholarly-Edition-in-the-Digital-Age>.
- Peroni et al. 2012** 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 - I-SEMANTICS 12*, 9. Graz, Austria: ACM Press. <https://doi.org/10.1145/2362499.2362502>.
- Pierazzo 2016** Pierazzo, Elena. 2016. *Digital Scholarly Editing*. 0 ed. Routledge. <https://doi.org/10.4324/9781315577227>.
- Price 2008** Price, Kenneth M. 2008. "Electronic Scholarly Editions". In *A Companion to Digital Literary Studies*, edited by Susan Schreibman and Ray Siemens.
- Renear 1993** Renear, Allen H., Elli Mylonas, and David G. Durand. 1993. "Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies". *Computer Science*.
- Sahle 2014** Sahle, Patrick. 2014. "Criteria for Reviewing Scholarly Digital Editions". 2014. <https://www.i-d-e.de/publikationen/weitereschriften/criteria-version-1-1/>.
- Sahle 2016** Sahle, Patrick. 2016. "2. What Is a Scholarly Digital Edition?" In *Digital Scholarly Editing: Theories and Practices*, edited by Matthew James Driscoll and Elena Pierazzo, 19–40. Open Book Publishers. <https://doi.org/10.11647/OBP.0095.02>.
- Simon et al. 2017** Simon, Rainer, Elton Barker, Leif Isaksen, and Pau De Soto CaÑamares. 2017. "Linked Data Annotation Without the Pointy Brackets: Introducing Recogito 2". *Journal of Map & Geography Libraries* 13 (1): 111–32. <https://doi.org/10.1080/15420353.2017.1307303>.
- Spadini et al. 2021** Spadini, Elena, Francesca Tomasi, and Georg Vogeler, eds. 2021. *Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing*. Schriften Des Instituts Für Dokumentologie Und Editorik, Band 15. Norderstedt: BoD – Books on Demand.

**TEI 2019** "TEI: Ontologies SIG". 2019. Text Encoding Initiative. 2019. <https://tei-c.org/activities/sig/ontologies>.

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

**Tomasi 2012** Tomasi, Francesca. 2012. "Ledizione Digitale e La Rappresentazione Della Conoscenza. Un Esempio: Vespasiano Da Bisticci e Le Sue Lettere". *Ecdotica* 9 (1): 264–86.

**Tomasi 2020** Tomasi, Francesca. 2020. "Digital Humanities e Organizzazione Della Conoscenza: Una Pratica Di Insegnamento Nel LODLAM". *Aibstudi* 60 (2): 1–15. <https://doi.org/10.2426/aibstudi-12068>.

**Weibel and Traugott 2000** Weibel, Stuart L., and Traugott Koch. 2000. "The Dublin Core Metadata Initiative: Mission, Current Activities, and Future Directions". *D-Lib Magazine* 6 (12). <https://doi.org/10.1045/december2000-weibel>.

**Wittern et al. 2009** Wittern, C., A. Ciula, and C. Tuohy. 2009. "The Making of TEI P5". *Literary and Linguistic Computing* 24 (3): 281–96. <https://doi.org/10.1093/lc/fqp017>.



This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.