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

Schriften des
Instituts für Dokumentologie und Editorik

herausgegeben von:

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.

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

Projects and Editions

Appendices

Infrastructures and Technologies

Towards Resolution Services for Text URIs

Hugh Cayless, Matteo Romanello

Abstract

In this paper we address the lack of fully resolvable URIs for texts and their citable units in the currently emerging Graph of Ancient World Data. We identify three main architectural components that are required to provide resolution services for text URIs: 1) a registry of text services; 2) an identifier resolution service; 3) a document metadata scheme, to represent the relations between texts in the registry, as well as between these texts and related external resources (e.g. library catalogues). After presenting some of the use cases a central registry providing resolvable URIs for texts would enable, we discuss in detail each component. We conclude by considering three examples where the proposed document metadata scheme is used to describe digital texts; this scheme contains a minimum yet extendable set of metadata that can be used to explore and aggregate texts coming from a network of distributed repositories.

1 Introduction

Over the last decade, many projects and institutions in the field of (digital) classics have embraced Linked Open Data as a philosophy of sharing and interconnecting resources about the ancient world (Elliott et al. 2014; Middle 2018), thus leading to the emergence of a Graph of Ancient World Data (Isaksen et al. 2014). The community contributing to the growth of this graph values the use of shared controlled vocabularies based on URIs to refer to *things*. Pleiades has produced an extensive set of URIs to identify ancient geographical locations and, more recently, Pelagios (Isaksen et al. 2014) has aggregated these, along with other gazetteers, and links to related resources. Something similar has been done by another two projects: SNAP (Bodard et al. 2017), and PeriodO (Rabinowitz et al. 2016), for ancient people, and chronological periodizations, respectively. Yet, what is still missing are fully resolvable URIs for texts and their citable units. Such URIs are of the highest importance for this emerging graph of data, considering that cited primary sources are an area where existing datasets often overlap.

Although there exist a number of platforms and identifier schemes for referring to text, there is no central registry for text identifiers and metadata (with a granularity down to the passage level). These kinds of identifiers are crucial in a Linked Data

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. 31–44.

context, because they allow common reference to abstract works (e.g., the *Iliad*) and to specific editions of that work (e.g., the Venetus A manuscript of the *Iliad*, a digital transcription of the Venetus A, etc.). In addition, while there is adequate coverage for canonical Classical texts, the situation is much less well-defined for the many texts outside the canon (e.g., papyri and inscriptions).[1] Further, the existing platforms that deal with identifier resolution do not provide services for minting new identifiers or refactoring existing ones.

Given this gap in the current digital infrastructure, we propose some initial thoughts towards the creation of web services that are able to handle the resolution of URIs pointing to texts and their citable units.[2] These services are aimed at increasing the presence of textual resources in the emerging Graph of Ancient World Data, and at better connecting digital texts with other relevant resources available online.

2 Related Work

The resolution services for text URIs that we propose in this paper would not be even conceivable without the work that has been carried out in recent years in two directly-related research areas. Firstly, the development of Application Programming Interfaces (APIs) to facilitate the exchange of digital texts; and, secondly, the development of standard vocabularies for the semantic description of texts.

Concerning text APIs, some concrete solutions have been developed over the last decade to enable the exchange of structured texts over standard protocols. The Canonical Text Services (CTS) (The Canonical Text Service 2019; Smith 2009) was the first of such protocols to define an API as well as an identifier syntax – the so-called CTS URNs – to retrieve TEI-encoded texts. Unfortunately, some technological factors hindered CTS from becoming a widely-adopted standard for exchanging texts (see section 4.1), and have led to the development of a new API, the Distributed Text Services (DTS) (Distributed Text Services n.d.; Clérice 2017; Clérice et. al. 2017), which aspires at playing for text data, the role that the IIIF API has played for image data. In addition to specifications and working implementations of both CTS and DTS, there exists an ecosystem of tools to work with them, most notably the CapiTainS software suite (Clérice 2017), which provides support for both APIs.

However, the absence of a widely-adopted API to exchange structured texts has also led to the proliferation of ad hoc solutions, whose designs often bear striking similarities with that of CTS and DTS. These include, for example, the API developed

---

1 Trismegistos (Trismegistos : An Interdisciplinary Portal of the Ancient World n.d.) assigns *TM numbers* to ancient documents, and can serve as a kind of gazetteer for some of these types of source text.

2 The work presented in this paper is the result of Linked Texts, a working group recently funded by Mellon Foundation via the *Pelagios Commons* project, which brought together an international group of scholars and research developers with extensive experience in developing and/or working with APIs for text repositories.

for the Scholastic Commentaries and Texts Archive (SCTA) (Witt 2018), the SHINE API developed in the context of the *Research Infrastructure for the Study of Eurasia* project (RISE), or the API that exposes the textual data of the *School of Salamanca* project (Wagner 2019).

Besides initiatives focusing on text APIs, considerable efforts were made to devise vocabularies and ontologies for describing various aspects of texts, with applicability to the ancient world and beyond. The Linked Ancient World Data ontology (LAWD) (Cayless 2016) models some aspects of ancient texts, and, in particular, of text-bearing artefacts such as papyri and inscriptions, while aligning itself with other vocabularies like DC, OAC and CIDOC-CRM. The HuCit Knowledge Base (Romanello & Pasin 2017) builds upon existing ontologies (e.g. CIDOC-CRM and FRBRoo) to link together bibliographic resources about classical canonical texts such as Perseus Catalog and Classical World Knowledge Base (CWKB), and focuses on the citation structures and citable units of these texts. Finally, the SPAR family of ontologies (Peroni & Shotton 2018) provides a set of vocabularies to describe a wide range of aspects around publications, including the characterization of citing behaviors with the Citation Typing Ontology (CiTO) and the representation of bibliographic metadata with the FRBR-aligned Bibliographic Ontology (FaBiO).

Finally, there exist ontologies that could benefit from the availability of resolvable URIs for texts. In addition to the already mentioned CiTO that characterizes citational relationships between texts, ontologies like Sharing Ancient Wisdoms (SAWS) (SAWS The Ontology n.d.) or the Intertextual Relationship Ontology (INTRO) (Oberreither 2019) could be used to describe textual relations between various text passages. These ontologies already provide vocabularies that could be used to create semantic statements about portions of texts, if only there existed resolvable URIs for the text portions one may want to make statements about.

3 Use Cases

A central registry providing resolvable URIs for texts and text passages would enable a wide range of use cases:

1. Services and tools for annotation – be it manual or automatic – could use text URIs to support the enrichment of existing resources. Tools like Recogito (Isaksen et al. 2017) or INCEpTION (Klie et al. 2018) already enable users to annotate texts and images by using existing authority data available via a SPARQL API (e.g. Pleiades, Wikidata, etc.). Recogito users, for example, can annotate geographical places within texts by assigning to each place its corresponding identifier (URI) within a gazetteer (e.g. https://pleiades.stoa.org/places/579885 for Athens). Having resolvable URIs for texts will enable a similar usage scenario for textual materials.

2. Online publishers will use the data to create links between their digital publications and existing digital libraries/text services. Take, for example, the digital version of a book containing citations of primary sources, like the open access publications published by Harvard's Center for Hellenic Studies (Center for Hellenic Studies: Online Publications n.d.). The publisher may want to provide readers with click-through links to the full text of cited sources, while leaving them the freedom to pick a specific edition/translation, should multiple versions of a text or document be available. Also, creating links towards a resolver, rather than to specific digital libraries, can potentially reduce the risk of broken links (as the resolver can be updated to reflect URL changes of digital libraries).
3. Digital Libraries collections could be registered in our text registry, provided that they expose their text data in compliance with at least one of the APIs supported by the registry, with the effect of increasing the accessibility and discoverability of these collections.

4 Text Resolution Services

The main idea of a service able to create resolvable URIs for texts published online is to relieve the publishers of such texts from having to mint URIs for their texts and the citable units in them. If texts are made available via standard APIs and described by sufficient metadata, the creation and resolution of URIs can be delegated to a centralized service, thus making it easier to integrate textual resources into the LOD graph.

In order for such services to work, three main things are required: *identifiers* for texts and their citable units; *machine interfaces* (i.e. APIs) to access and retrieve texts over Web protocols; and *metadata* to describe texts, as well as the relations between these texts and other existing resources (e.g., authority control records).

We believe that three main components are needed to provide resolution services for text URIs:

1. *A registry of text services* to keep track of available text repositories against which text URIs can be resolved.
2. An *identifier resolution service* that can resolve a URI pointing to a section of a text or document (e.g., a line of a papyrus or inscription, or a chapter of a work in prose) to one or more places where a digital version of that text or document can be found.
3. A *document metadata scheme* is also needed in order to represent the relations between texts in the registry, as well as between these texts and related external resources (e.g., library catalogues).

In what follows we discuss the main issues related to each of these three components.

4.1 Text Registry Service

Supported APIs

CTS and DTS are the two main APIs that have emerged in recent years to enable the exchange of structured texts over standard protocols. We identified these two APIs as the most important standards that our text resolution services should support.

Unfortunately, some factors hindered CTS from becoming a widely adopted standard for exchanging texts:

1. its strong commitment to the specificities of canonical texts, which makes it unsuitable for non-canonical texts such as archival documents, papyri or inscriptions;
2. aspects of the design of the API which keep it from scaling to large repositories of texts;
3. the development of the CTS standard, which has been driven by a single project rather than a community.

While the use of the CTS protocol has been limited, the naming system it defines is powerful and has been broadly adopted. The DTS is a new standard whose specifications have recently been published as a first public draft. DTS was designed to overcome the limitations of CTS, discussed above, while remaining retro-compatible with existing CTS URNs. DTS presents a text API modelled as arbitrary collections, with functions for retrieving whole or partial texts, and for discovering the citation schemes for those texts.

Registry Implementation

A suitable model for implementing a text registry that is open to input from the community is w3id.org, where pull requests made to a GitHub repository make it possible to add new entries to the registry. Anyone should be able to add a service to this registry, provided that this service exposes its data via a CTS or DTS API. Moreover, since the DTS API advertises the citation structures and citable units of a text, it can be used to determine, programmatically, which resources URIs will have to be minted for.

4.2 Identifier Resolution Service

Text Identifiers

CTS URNs provide a handy way of creating text identifiers that carry with them information about the text itself and its hierarchy. Yet, there is no place where

these CTS URN identifiers are published, no entities nor community regulating their creation, no central place where they are gathered.

Let us consider an example of how CTS URN work: Proclus' *Elements of Physics* (`urn:cts:greekLit:tlg4036.tlg006`). The Perseus Catalog contains bibliographic information about this text, including a link to an edition available in the Hathi Trust collection (*Perseus Catalog,* Institutio Physica n.d.). The SAWS project (Hedges et al. 2016) has added an edition of this text, and publishes it via its CTS endpoint. Their edition is identified by the URN `urn:cts:greekLit:tlg4036.tlg006.saws01`. Let us imagine now that new editions or translations of the same text are published via a DTS endpoint with compatible CTS URNs. How can all this information be integrated in such a way that it is available to a user looking up the identifier `urn:cts:greekLit:tlg4036.tlg006`?

Resolver or Catalog?

One of the main points of discussion at the Linked Texts working group meeting was how to handle the resolution of text URIs, and whether it was feasible to develop a Handle System, based on the one that is used for resolving Digital Object Identifiers, that could work with DTS and CTS APIs (Almas et al. 2018).

A Handle System works by querying a database for the submitted ID, and then, if a matching URL is found, redirecting the requester to that URL. The situation for DTS is more complex, partially because of the legacy of CTS identifiers, which are URNs. The problems begin with the lack of any central management of CTS IDs. Unlike DOIs, which consist of a namespace identifier and an item identifier, CTS URNs are more hierarchical and more meaningful at each level.

Take `urn:cts:greekLit:tlg0012.tlg001`, for example, which identifies the *Iliad* of Homer. Crucially, it identifies the *work* rather than any specific edition of it. Usefully, a CTS URN allows us to denote a passage either within the abstract work (`urn:cts:greekLit:tlg0012.tlg001.1.1-1.10`, i.e., *Iliad* 1, lines 1–10) or within a specific edition (`urn:cts:greekLit:tlg0012.tlg001.perseus-grc1.1.1-1.10`, i.e. the version of the *Iliad* hosted by the Perseus Project, book 1, lines 1–10).

This means there is a basis for identifying the common abstract notion of the first line of the *Iliad*, and comparing them across all available editions. But what should a resolver do, given a work ID? A possible answer would be for it to list available editions (and translations) of that work, behaving more like a catalog (or a regular DTS Collection endpoint) than a resolver. It gets worse when we consider that edition identifiers are uncontrolled, and that there is no reason that multiple repositories might not contain copies of `urn:cts:greekLit:tlg0012.tlg001.perseus-grc1` (and perhaps even different versions of that).

This complexity means that a *resolver* cannot work like a traditional handle server, because there are cases where there is not a one-to-one relationship between identifier and resource, for example, when a URI corresponds to a non-information resource (e.g., a work in the FRBR hierarchy), or when the URI can resolve to multiple copies of the same resource. In such cases, the resolver will deliver a list of documents (or document fragments), and the decision must be made by the client – be it a human or software agent – about which resource(s) to choose for a given identifier. In other words, despite the attractions of a handle resolution service, the resolver must instead behave in certain respects like a centralized catalog system.

4.3 Document Metadata Scheme

What sorts of information might we want to attach to a digital document? What information/metadata is needed in order to provide useful services on an aggregate of texts coming from a network of distributed repositories?

The kinds of document represented by DTS will usually be editions derived from one or more manuscript or print sources, and so there may be a natural division between data attaching to the source document(s), and to the edition or translation. This division might be handled in DTS by modelling the collections accordingly. Inscriptions, for example, will have information that refers to the primary source document itself: its find-spot, date, provenance, and so on. Then, there may be multiple published editions of that inscription, and these have their own data, including the editor and publication date. Inscriptions may be grouped in various ways, but one very typical way is to organize them geographically: a plausible DTS collection structure for inscriptions might be *location* ⟶ *sources* ⟶ *editions*, with source information attached to the source collection, and edition information to the edition. But this is only one possible organizational scheme. *Canonical* texts will tend to be organized according to author and work, where *author* refers to the creator of the original *work*, which is an abstraction – effectively an agreement that the sources all have the same ancestor. Many source texts may go into the creation of an edition (and not every edition will share the same set of sources). Thus, we may expect both source and edition information to appear with the edition.

Some examples of the information that might be attached to a DTS document follow:

- Source Information
- Author
- Title
- Date / Place of Original Creation
- Repository / Cataloging Info
- Publication History

- Comparanda / Related Documents
- Physical Description
- Language
- Provenance
- Mentioned People / Places / Events
- Surrogates (Editions, Images, Translations)
- Edition Information
- Editor
- Publication Info
- Language
- Source(s)
- Related Editions
- Revision History
- Attestations (may be edition-dependent)

The list is not comprehensive, but represents many of the kinds of *metadata* that can travel along with a TEI document. Much, if not all of it, might be represented in the document itself, but may also usefully be extracted and represented in a more generic form both in order to facilitate discovery in an API such as DTS, and to contextualize the document in a knowledge graph. Moreover, this list of metadata avoids making any strong assumption about the type of textual sources it can be applied to, and ideally aims to be applicable to any text.

If these types of metadata are reasonable candidates for representation in DTS, then we must consider how best to represent them. DTS allows for both basic Dublin Core and for an extended metadata set, with the idea that a client can choose whether to use only the *simple* properties from Dublin Core, or more specialized and sophisticated RDF data. Some possible mappings are listed in Table 1.

Examples

In order to exemplify the usage of the proposed document metadata scheme, let us consider now three examples (the first two speculative, the third one real).

1. A library has decided to publish OCR transcripts in TEI/XML format of its digitized collections as a DTS-compliant repository. These collections contain mostly editions and translations of Greek texts, including the OCR transcript of H. Boese's German translation of Proclus' *Institutio Physica*, entitled *Die mittelalterliche Übersetzung der Stoicheiosis phusike des Proclus*. In this case, source information describes Boese's translation, while edition information relates to the TEI/XML encoding of the OCR transcript.

The `dts:extensions` returned by the DTS document endpoint could provide contextual information about this text by using properties from the document metadata scheme described above:

Metadata	RDF Representation
Source Information	
Author	dc:creator, lawd:responsibleAgent
Title	dc:title
Date of Original Creation / Publication	dc:created, fabio:hasPublicationDate, fabio:has PublicationYear
Place of Original Creation / Publication	lawd:origin, fabio:hasPublicationPlace
Repository / Cataloging Info	dc:identifier, crm:P1_is_identified_by
Publication History (the bibliography for a given primary source, e.g. papyrus or inscription)	dc:bibliographicCitation, list of lawd:Citation
Comparanda / Related Documents	dc:relation
Physical Description	dc:medium
Language	dc:language
Provenance (only applies to text-bearing artefacts)	lawd:foundAt, dc:provenance
Mentioned People / Places / Events	list of lawd:Attestation
Surrogates (Editions, Images, Translations)	dc:hasVersion
FRBR hierarchy of a text (Work, Expression, Manifestation, Item), where applicable	frbr:realization, frbr:embodiment, frbr:exemplar, fabio:isManifestationOf, fabio:isRealizationOf, fabio:isPortrayalOf
Edition Information	
Editor	dc:contributor, lawd:responsibleAgent
Publication Info	dc:bibliographicCitation
Source(s)	dc:source
Related Editions	dc:relation
Revision History	list of prov:Activity

Table 1. Metadata in the Document Metadata Scheme and their RDF representation.

```
"dts:extensions": {
 "@context": {
   "lawd": "http://lawd.info/ontology/",
   "frbr": "http://purl.org/vocab/frbr/core#",
   "fabio": "http://purl.org/spar/fabio",
   "ecrm": "http://erlangen-crm.org/current-version"
 },
 "ecrm:P1_is_identified_by": {
   "rdfs:label": "urn:cts:greekLit:tlg4036.tlg006.boese",
   "ecrm:P2_has_type": "http://purl.org/hucit/kb/types/CTS_URN"
 },
 @id: "https://library.org/texts/urn:cts:greekLit:tlg4036.tlg006.boese",
 @type: "fabio:DigitalItem",
 "fabio:isPortrayalOf": "http://purl.org/hucit/kb/works/3822",
 "frbr:exemplarOf": "https://library.org/editions/
                          urn:cts:greekLit:tlg4036.tlg006.boese",
```

```
  "dc:source": {
    @id: "https://library.org/editions/urn:cts:greekLit:tlg4036.tlg006.boese",
    @type: "lawd:Edition",
    "fabio:hasPublicationYear": "1958",
    "dc:medium": "printed edition",
    "fabio:hasPlaceOfPublication": "http://www.geonames.org/2950159/berlin",
    "dc:language": "de",
    "dc:title": "Die mittelalterliche Übersetzung der Stoicheiosis phusike des
         Proclus",
    "dc:bibliographicCitation": "Boese, H., Die mittelalterliche Übersetzung der
         Stoicheiosis phusike des Proclus (Deutsche Akademie der Wissenschaften
         zu Berlin, Institut für griechisch-römische Altertumskunde,
         Veröffentlichungen 6), Berlin: Akademie Verlag, 1958.",
    "fabio:isManifestationOf": "http://purl.org/hucit/kb/works/3822"
  }
}
```

2. Papyri.info is an aggregator of digital editions of papyri coming from various databases. Any given text may have connected metadata, translations, previous or following editions, and images. Disentangling all of these interrelated physical, print, and digital artefacts can be a real challenge. The Papyri.info system avoids trying to model the precise relationships of the various editions and metadata records that may pertain to a document, simply connecting them using the Dublin Core relation property. But, a more sophisticated representation of these relationships could be developed, perhaps using the CiTO ontology. The example below imagines some of the ways the metadata records for the papyrus P.Col 8 237 could be exposed in JSON-LD as DTS extended metadata:

```
"dts:extensions": {
 "@context": {
   "cito": "http://purl.org/spar/cito",
   "dc": "http://purl.org/dc/terms/",
   "ecrm": "http://erlangen-crm.org/current-version",
   "lawd": "http://lawd.info/ontology/",
 },
 @id: "https://papyri.info/ddbdp/p.col;8;237/source",
 @type: "lawd:Edition",
 "dc:source": {
  @id: "https://papyri.info/ddbdp/p.col;8;237/work", // URI representing the
                                                        print edition
  "cito:isCitedBy": "http://papyri.info/biblio/65200",
  "dc:source" {
   @id: "https://papyri.info/ddbdp/p.col;8;237/original", // URI representing
                                                             the primary source
   @type: "lawd:WrittenWork",
   "rdfs:seeAlso": "http://www.trismegistos.org/text/10553",
   "lawd:origin": "https://pleiades.stoa.org/places/737081",
   "lawd:foundAt": "https://pleiades.stoa.org/places/737081",
   "ecrm:P62i_is_depicted_by": "http://papyri.info/ddbdp/p.col;8;237/images",
   "dc:created": "3 June, 381 or 382"
  }
 }
}
```

3. Beta maṣāḥəft (BM) (Beta Maṣāḥəft: Manuscripts of Ethiopia and Eritrea n.d.) is a research environment for the manuscript tradition of Ethiopia and Eritrea. BM's data model makes a distinction between written artefacts (manuscripts, inscriptions, etc.) on one side, and textual and narrative units on the other.[3] The project exposes all available texts (editions, transcriptions, etc.) by means of a DTS API, and organizes them in a non-hierarchical way. Its DTS API makes use of `dts:extensions` to publish rich metadata about texts in BM in a structured and interoperable format (i.e., JSON-LD).

Inscriptions are also presented in this way, which helps to represent complex situations with minimal effort. For example, the edition of the Greek text of the pseudo-trilingual royal inscriptions (Bausi & Liuzzo 2018) *RIE 270*[4] has, as sources, two inscriptions:

```
"dc:source": [ {
  "fabio:isManifestationOf": "https://betamasaheft.eu/RIE185and270",
  "@type": "lawd:AssembledWork",
  "@id": "urn:dts:betmasMS:RIE185and270"
}, {
  "fabio:isManifestationOf": "https://betamasaheft.eu/RIE185bisand270bis",
  "@type": "lawd:AssembledWork",
  "@id": "urn:dts:betmasMS:RIE185bisand270bis"
} ]
```

Each of these written artefacts has a transcription and the second, RIE185bisand270bis includes a statement that it is a copy of the first.

```
"dts:extensions": {
  "crm:P1_is_identified_by": [ "104961", "RIÉ 270bis", "RIÉ 185bis and 270bis", "
       RIÉ 185bis" ],
  "saws:isDirectCopyOf": "https://betamasaheft.eu/RIE185and270"
}
```

A user could then take the text of the edition that uses only one of the texts from both copies on stone, or part of the transcription of one specific stone.

5 Conclusions and Further Work

In this paper, we addressed the problem of the absence, in the emerging LOD graph, of fully resolvable URIs for texts and their citable units. We sketched out a possible solution consisting of three main components: first, a registry service to keep track of available text repositories against which text URIs can be resolved; second, an identifier resolution service that can resolve a URI pointing to a section of a text or document and to one or more digital versions of that text or document; and third, a

[3] See Orlandi (2013) for Textual and Narrative Units. A further discussion can be found in Liuzzo (2019).

[4] The URI of Beta Masaheft LIT4851greekRoyal is https://betamasaheft.eu/api/dts/collections?id=urn:dts:betmas:LIT4851greekRoyal.

document metadata scheme to represent the relations between texts in the registry, as well as between these texts and related external resources. The next necessary step, before the implementation of these services can start, will be to gather feedback from the wider community. Such a discussion will help us ensure the generalizability of our solution (and especially the document metadata scheme) to potentially any text.

Acknowledgements

The authors wish to thank the *Pelagios Commons* project and the Andrew W. Mellon Foundation for funding the Linked Texts working group, from which this chapter originated. Thanks also to Pietro Maria Liuzzo for his feedback on earlier manuscript drafts, and to the DTS community, with whom we have been discussing many of the issues we address in this paper.

Bibliography

Almas, Bridget, Larry Lannom, and Robert Tupolo-Schneck, *CTS Handles Proposal*, 2018 <https://github.com/rpidproject/cts-handles/blob/master/proposal.md>

Bausi, Alessandro, and Pietro Liuzzo, 'Inscriptions from Ethiopia. Encoding Inscriptions in Beta Maṣāḥəft: From Practice to Discipline', in *Crossing Experiences in Digital Epigraphy* (De Gruyter, 2018), 84–92 <https://doi.org/10.1515/9783110607208-007>

'Beta Maṣāḥəft: Manuscripts of Ethiopia and Eritrea', n.d. <https://betamasaheft.eu/>

'Publications' <https://www.betamasaheft.uni-hamburg.de/publications.html>

Bodard, Gabriel, Hugh Cayless, Mark Depauw, Leif Isaksen, Faith Lawrence, and Sebastian Rahtz, 'Standards for Networking Ancient Person Data: Digital Approaches to Problems in Prosopographical Space', *Digital Classics Online*, 2017, 28–43 <https://doi.org/10.11588/dco.2017.0.37975>

Cayless, Hugh, 'LAWD. An Ontology for Linked Ancient World Data', 2016 <https://github.com/lawdi/LAWD>

'Center for Hellenic Studies: Online Publications', n.d. <https://chs.harvard.edu/CHS/article/display/1166.browse-online-publications>

Clérice, Thibault, 'Les Outils CapiTainS, l'édition Numérique et l'exploitation Des Textes', *Médiévales*, 73.73 (2017), 115–131 <https://doi.org/10.4000/medievales.8211>

Clérice, Thibault, Matthew Munson, and Bridget Almas, *Capitains/Capitains.Github.Io: 2.0.0*, version 2.0.0 (Zenodo, 2017) <https://doi.org/10.5281/zenodo.570516>

De Santis, Annamaria, and Irene Rossi, *Crossing Experiences in Digital Epigraphy, From Practice to Discipline* (Berlin, Boston: De Gruyter, 2019) <https://doi.org/10.1515/9783110607208>

'Distributed Text Services', n.d. <https://w3id.org/dts>

Elliott, Thomas, Sebastian Heath, and John Muccigrosso, 'Prologue and Introduction', *ISAW Papers*, 7.1 (2014) <http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/elliott-heath-muccigrosso/>

Hedges, Mark, Anna Jordanous, K. Lawrence, Charlotte Roueche, and Charlotte Tupman,

'Computer - Assisted Processing of Intertextuality in Ancient Languages', *Journal of Data Mining and Digital Humanities*, 2016

Isaksen, Leif, Rainer Simon, and Elton T.E. Barker, 'Social Semantic Annotation with Recogito 2.', in *DH2017 Conference Abstracts*, 2017 <https://dh2017.adho.org/abstracts/570/570.pdf>

Isaksen, Leif, Rainer Simon, Elton T.E. Barker, and Pau de Soto Cañamares, 'Pelagios and the Emerging Graph of Ancient World Data', in *Proceedings of the 2014 ACM Conference on Web Science - WebSci '14* (New York, New York, USA: ACM Press, 2014), 197–201 <https://doi.org/10.1145/2615569.2615693>

Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych, 'The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation', in *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations* (Association for Computational Linguistics, 2018), 5–9 <http://tubiblio.ulb.tu-darmstadt.de/106270/>

Liuzzo, Pietro Maria, *Digital Approaches to Ethiopian and Erithrean Studies* (Wiesbaden: Harrassowitz Verlag, 2019)

Middle, Sarah, 'Identifying Research Methods for the Use and Production of Linked Ancient World Data', 2018 <https://hcommons.org/deposits/item/hc:22031/>

Oberreither, Bernhard, 'INTRO - an Intertextual Relationships Ontology for Literary Studies', 2019 <https://github.com/BOberreither/INTRO>

Orlandi, Tito, 'A Terminology for the Identification of Coptic Literary Documents', *Journal of Coptic Studies*, 15 (2013), 87–94 <https://doi.org/10.2143/JCS.15.0.3005414rmin>

Peroni, Silvio, and David Shotton, 'The SPAR Ontologies', in *Artificial Intelligence and Soft Computing*, ed. by Leszek Rutkowski, Rafał Scherer, Marcin Korytkowski, Witold Pedrycz, Ryszard Tadeusiewicz, and Jacek M. Zurada (Cham: Springer International Publishing, 2018), 119–36 <https://doi.org/10.1007/978-3-030-00668-6_8>

'Perseus Catalog, Institutio Physica', n.d. <https://catalog.perseus.org/catalog/urn:cts:greekLit:tlg4036.tlg006>

Rabinowitz, Adam, Ryan Shaw, Sarah Buchanan, Patrick Golden, and Eric Kansa, 'Making Sense of the Ways We Make Sense of the Past: The Periodo Project', *Bulletin of the Institute of Classical Studies*, 59.2 (2016), 42–55 <https://doi.org/10.1111/j.2041-5370.2016.12037.x>

Romanello, Matteo, and Michele Pasin, 'Using Linked Open Data to Bootstrap a Knowledge Base of Classical Texts', in *Proceedings of the Second Workshop on Humanities in the Semantic Web (WHiSe II) Co-Located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 22, 2017.*, ed. by Alessandro Adamou, Enrico Daga, and Leif Isaksen, CEUR Workshop Proceedings (CEUR-WS.org, 2017), MMXIV, 3–14 <http://ceur-ws.org/Vol-2014/paper-01.pdf>

'SAWS The Ontology', n.d. <http://www.ancientwisdoms.ac.uk/method/ontology/>

Smith, Neel, 'Citation in Classical Studies', *Digital Humanities Quarterly*, 3.1 (2009) <http://www.digitalhumanities.org/dhq/vol/3/1/000028/000028.html>

'The Canonical Text Service (CTS): The CITE Architecture', 2019 <http://cite-architecture.org/cts/>

'Trismegistos: An Interdisciplinary Portal of the Ancient World', n.d. <https://www.trismegistos.org/>

Wagner, Andreas, 'The School of Salamanca - The Web Application, API Documentation', 2019 <https://github.com/digicademy/svsal/blob/master/docs/API.md>

Witt, Jeffrey C., 'Digital Scholarly Editions and API Consuming Applications', in *Digital Scholarly Editions as Interfaces*, ed. by Roman Bleier, Martina Bürgemeister, Helmut W. Klug, Frederike Neuber, and Gerlinde Schneider (Norderstedt: Books on Demand, 2018), 219–247

Formal Models

Projects and Editions

Appendices

Biographical Notes

Thomas Ahrend (University of Basel, Switzerland – thomas.ahrend@unibas.ch) studied Musicology, Philosophy and Literary Studies in Frankfurt a. M. and Berlin. He received his MA 1996, and his PhD 2005 at Technische Universität Berlin with a dissertation on the instrumental music of Hanns Eisler. 1997–2010 member of the editorial staff of the Hanns Eisler Gesamtausgabe in Berlin. Since September 2010, member of the editorial staff of the Anton Webern Gesamtausgabe at Musikwissenschaftliches Seminar at University of Basel.

Peter Boot (Huygens ING, The Netherlands – peter.boot@huygens.knaw.nl) studied mathematics and Dutch language and literature; he wrote his PhD thesis about annotation in scholarly digital editions and its implications for humanities scholarship. He oversaw the creation of the digital edition of the letters of Vincent van Gogh. He is employed as a senior researcher at the Huygens Institute for the History of the Netherlands where he works, among other things, as a consultant in several edition projects.

Manuel Burghardt (University of Leipzig, Germany – burghardt@informatik.uni-leipzig.de) is head of the Computational Humanities Group at Leipzig University. He is interested in the use of digital tools and computational techniques to explore new modes of doing research in the humanities. His most recent areas of research are Sentiment Analysis in the Humanities, Drametrics, Computational Intertextuality, Computational Analysis of Movies and Series and Music Information Retrieval.

Toby Burrows (University of Oxford, United Kingdom – toby.burrows@oerc.ox.ac.uk) is a Senior Researcher in the Oxford e-Research Centre at the University of Oxford, and a Senior Honorary Research Fellow in the School of Humanities at the University of Western Australia.

Hugh Cayless (Duke University, USA - hugh.cayless@duke.edu) is Senior Digital Humanities Developer at the Duke Collaboratory for Classics Computing. Hugh has over a decade of software engineering expertise in both academic and industrial settings. He also holds a Ph.D. in Classics and a Master's in Information Science. He is one of the founders of the EpiDoc collaborative and currently serves on the Technical Council of the Text Encoding Initiative.

Hans Cools (University of Basel, Switzerland – 1961-2021) had a master degree in medicine and a specialization in orthopaedic surgery and traumatology (Universities of Ghent and Antwerp, Belgium, 1997), a bachelor's degree in physical

therapy, and a standalone degree in informatics (1999). Through various research and project management positions, in both companies and academic institutions, he gained expertise in different aspects of the Semantic Web technologies, focusing particularly on formal data modeling and machine reasoning. Those positions were in internationally collaborative research projects in a biomedical setting, mainly of the 5-7th EU Framework Program. Foremost in these projects were semantic interoperability and reusability of data. Since 2016, he worked in the humanities, as knowledge engineer, ontologist, and Semantic Web technology expert, at the University of Basel, as part of the NIE-INE project, which highlights scholarly editing. He (co-)published several articles, and gave workshops on the implementation of Semantic Web technologies in biomedicine and the humanities. He passed away in April 2021.

Francesca Giovannetti (University of Bologna, Italy – francesc.giovannett6@unibo.it) is a second-year PhD student in Digital Humanities at the Department of Classical Philology and Italian Studies, University of Bologna. She received an MA in Digital Humanities from King's College London and a second cycle degree in Digital Humanities and Digital Knowledge from the University of Bologna. She is interested in combining digital scholarly editing with semantic web technologies and in the use of digital technologies in education.

Matthew Holford (University of Oxford, United Kingdom – matthew.holford@bodleian.ox.ac.uk) is Tolkien Curator of Medieval Manuscripts at the Bodleian Library, University of Oxford.

Marijn Koolen (Royal Netherlands Academy of Arts and Sciences - Humanities Cluster, The Netherlands – marijn.koolen@gmail.coml) studied artificial intelligence and wrote his PhD thesis on using hyperlinks in information retrieval algorithms. He has worked on scholarly annotation for digital humanities research and on annotation-related information behaviour and information systems. He works as a researcher and developer at the Humanities Cluster of the Royal Netherlands Academy of Arts and Sciences, where he leads a project on developing annotation support within the *CLARIAH research infrastructure* project.

David Lewis (University of Oxford, United Kingdom – david.lewis@oerc.ox.ac.uk) is a Research Associate in the Oxford e-Research Centre at the University of Oxford.

Andrew Morrison (University of Oxford, United Kingdom – andrew.morrison@bodleian.ox.ac.uk) is a Software Engineer in the Bodleian Digital Library Systems and Services, Bodleian Library, University of Oxford.

Stefan Münnich (University of Basel, Switzerland – stefan.muennich@unibas.ch) studied musicology and communication science at the Technische Universität Berlin, MA 2011 with a thesis on cantional setting in Heinrich Schütz's Becker-Psalter. 2012 research assistant, 2013–2015 research associate of the Felix Mendelssohn Bartholdy. Sämtliche Briefe edition at University of Leipzig (co-editor of vols. 9 & 12). Since October 2015 research associate of the Anton Webern Gesamtausgabe, Basel; received his Doctorate degree in 2020 at the department of musicology at the University of Basel with a dissertation about music notation and its codes.

Iian Neill (Digital Academy of the Academy of Sciences and Literature, University of Mainz - Iian.Neill@adwmainz.de) is a visiting researcher at the Digital Academy of the Academy of Sciences and Literature Department at the University of Mainz, Germany. He is the creator of Codex, a text annotation environment which uses standoff property annotation to generate entities in a graph meta-model. Codex is currently being used to produce a digital edition of the epistles of Hildegard von Bingen at the Digital Academy in Mainz.

Roberta Padlina (University of Basel, Switzerland – roberta.padlina@unibas.ch) studied medieval philosophy at the University of Fribourg, Switzerland, obtaining a doctoral degree in June 2020. She has twelve years of professional experience in the field of Digital Humanities, thanks to which she has been able to work closely with different actors involved in the online publication of open access research. Roberta has worked for several years for e-codices –Virtual Library of Manuscripts in Switzerland and currently coordinates the National Infrastructure for Editions (NIE-INE) project. Roberta's main focus is on the opportunities and challenges that the digital shift poses for traditional education and research institutions, including developing semantic web strategies for scholarly publications and cultural goods.

Kevin Page (University of Oxford, United Kingdom – kevin.page@oerc.ox.ac.uk) is a Senior Researcher in the Oxford e-Research Centre and Associate Member of Faculty in the Department of Engineering in the University of Oxford.

Miller C. Prosser (University of Chicago, USA – m-prosser@uchicago.edu) earned his Ph.D. in Northwest Semitic Philology from the University of Chicago. His academic interests include the social and economic structure of Late Bronze Age Ras Shamra-Ugarit and the use of computational methods for philological and archaeological research. Miller is the Associate Director of the Digital Studies MA program at the University of Chicago where he teaches courses on Data Management and Data Publication for the Humanities. He also works as a

researcher at the OCHRE Data Service of the Oriental Institute of the University of Chicago where he consults with and supports research projects using the Online Cultural and Historical Research Environment (OCHRE). He has also worked as a tablet photographer for the Mission de Ras Shamra (Ugarit) and the Persepolis Fortification Archive Project, employing advanced digital photographic methods such as reflectance transformation imaging, photogrammetry, and high-resolution digital scanning.

Matteo Romanello (Université de Lausanne, Switzerland - matteo.romanello@unil.ch) is Ambizione SNF Lecturer at the University of Lausanne, where he conducts a project on the commentary tradition of Sophocles' Ajax. Matteo is a Classicist and a Digital Humanities specialist with expertise in various areas of the Humanities, including archaeology and history. After obtaining his PhD from King's College London, he worked as a research scientist at EPFL's DHLAB on the Linked Books and Impresso projects, before moving to his current position. He was also teaching fellow at the University of Rostock, researcher at the German Archaeological Institute, and visiting research scholar at Tufts University.

Sandra Schloen (University of Chicago, USA – sschloen@uchicago.edu) is the Manager of the OCHRE Data Service at the Oriental Institute of the University of Chicago, and is the co-designer and developer of the Online Cultural and Historical Research Environment (OCHRE). Trained in computer science and mathematics (B.Sc. University of Toronto; M.Ed. Harvard University), Sandra has spent over 30 years working with technology as a systems analyst, technical trainer, and software developer. A long association with colleagues in the academic community has enabled her to develop a specialty in solving problems in the Digital Humanities where challenges of data capture, data representation and data management abound. Specifically, she has served extensively as a database manager for several archaeological projects in Israel and Turkey, and supports a wide range of research projects at the Oriental Institute and at other universities.

Desmond Schmidt (University of Bologna - desmond.allan.schmidt@gmail.com) has a background in classical Greek philology, information security and eResearch. He has worked on several scholarly edition projects, including the Vienna Wittgenstein Edition (1990–2001), Digital Variants (2004–2008), the Australian Electronic Scholarly Editions project (2012–2013), the Charles Harpur Critical Archive (2014-) and a pilot edition of Gianfrano Leopardi's Idilli (2018-). He currently works on developing practical web-based tools for making, visualising and publishing digital scholarly editions.

Colin Sippl (University of Regensburg, Germany – colin.sippl@ur.de) is currently a project employee at the University Library of Regensburg. Since 2017, he has been working on extending the open access services of the Electronic Journals Library (EZB). More recently, he has started developing and setting up a digital repository for literature, artefacts and experiments relating to the early life sciences based on the Invenio framework. He specialised in textual data mining and the development of media services in the institutional domain.

Elena Spadini (University of Lausanne - elena.spadini@unil.ch) is a postdoctoral researcher at the University of Lausanne. She holds a Ph.D. in Romance Philology from the University of Rome Sapienza (2016) and a M.A. in Digital Humanities from the École nationale des chartes (2014). She was a Marie Curie fellow in the IT Network DiXiT and co-directed the related volume Advances in Digital Scholarly Editing (Sidestone Press, 2017). She published in international journals and taught specialized courses in various European countries in the field of Digital Philology.

Francesca Tomasi (University of Bologna - francesca.tomasi@unibo.it) is associate professor in Archival Science, Bibliography and Librarianship at the University of Bologna (Italy). Her research is mostly devoted to digital cultural heritage, with a special attention to documentary digital edition, and a focus on knowledge organization methods in archives and libraries. She is member of different scientific committees of both associations and journals. In particular, she is President of the Library of the School of Humanities in the University of Bologna (BDU - Biblioteca di Discipline Umanistiche), Director of the international second cycle degree in Digital Humanities and Digital Knowledge (DHDK), President of the Italian Association of Digital Humanities (AIUCD – Associazione per l'Informatica Umanistica e la Cultura Digitale), and co-head of the Digital Humanities Advanced Research Center (/DH.ARC). She wrote about 100 papers and 4 monographs related to DH topics. She is editor and scientific director of several digital scholarly environments.

Athanasios Velios (University of the Arts London, United Kingdom – a.velios@arts.ac.uk) is Reader in Documentation at the University of the Arts London.

Georg Vogeler (University of Graz - georg.vogeler@uni-graz.at) is professor for Digital Humanities at the University of Graz and scientific director of the Austrian Center for Digital Humanities and Cultural Heritage at the Austrian Academy of Sciences. He is a trained historian (Historical Auxiliary Sciences). He spent several years in Italy (Lecce, Venice). In 2011, he became member of faculty at the Centre for Information Modelling at Graz University, where he was nominated

full professor for Digital Humanities in 2016 and head of department in 2019. His research interests lie in late medieval and early modern administrative records, diplomatics (digital and non digital), digital scholarly editing and the history of Frederic II of Hohenstaufen (1194–1250). He was and is part in several national and international research projects related to his research interests.

Christian Wolff (University of Regensburg, Germany – christian.wolff@ur.de) has been Professor of Media Informatics at the Institute for Information and Media, Language and Culture at the University of Regensburg since 2003. He holds a PhD in information science and is a habilitated computer scientist. His research interests include: human-computer interaction, multimedia and web-based information systems, (multimedia) software engineering and information retrieval (in particular information literacy and social media).