Exploring Manuscripts: Sharing Ancient Wisdoms across the Semantic Web

Anna Jordanous
Centre for e-Research
King's College London
26-29 Drury Lane, London, UK
anna.jordanous@kcl.ac.uk

Mark Hedges
Centre for e-Research
King's College London
26-29 Drury Lane, London, UK
mark.hedges@kcl.ac.uk

K. Faith Lawrence
Department of Digital Humanities
King's College London
26-29 Drury Lane, London, UK
faith.lawrence@kcl.ac.uk

Charlotte Tupman
Department of Digital Humanities
King's College London
26-29 Drury Lane, London, UK
charlotte.tupman@kcl.ac.uk

ABSTRACT

Recent work in digital humanities has seen researchers increasingly producing online editions of texts and manuscripts, particularly in adoption of the TEI XML format for online publishing. The benefits of semantic web techniques are underexplored in such research, however, with a lack of sharing and communication of research information. The Sharing Ancient Wisdoms (SAWS) project applies linked data practices to enhance and expand on what is possible with these digital text editions. Focussing on Greek and Arabic collections of ancient wise sayings, which are often related to each other, we use RDF to annotate and extract semantic information from the TEI documents as RDF triples. This allows researchers to explore the conceptual networks that arise from these interconnected sayings. The SAWS project advocates a semantic-web-based methodology, enhancing rather than replacing current workflow processes, for digital humanities researchers to share their findings and collectively benefit from each other's work.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: On-line Information Services—*Data sharing*; I.7.4 [**Document and Text Processing**]: Electronic Publishing; J.5 [**Arts and Humanities**]: Literature

Keywords

Linked Data, Semantic Web, Digital Humanities, manuscripts, ontology, RDF, TEI XML, gnomologia

1. INTRODUCTION

WIMS'12, June 13-15, 2012 Craiova, Romania

Researchers in digital humanities are increasingly producing online editions of texts and manuscripts, commonly in TEI XML format. Whilst it is beneficial to have these texts more freely available, the benefits of semantic web techniques are currently underexplored in digital humanities research. Hampered by lack of communication and sharing of research information this consolidation rather than expansion of information, the so-called 'digital silo' [33, 21], continues to dominate. The application of RDF and linked data offers the opportunity to address not only this issue in digital humanities research but also to fulfil the promise of this technology. In this paper we describe how RDF has been incorporated into TEI documents for this purpose.

The Sharing Ancient WisdomS (SAWS) project focuses on the tradition of Greek and Arabic wisdom literatures: collections of moral and social advice and/or philosophical ideas. Throughout antiquity and the middle ages, collections of wise sayings were created by scribes. The process of circulating these collections via retranscription opened up opportunities for deliberate editing and alterations by the scribe, as well as occasional transcription errors. These changes, major or minor, often reflected a change of social context, especially when translated between different cultural traditions. A work may have changed as it moved from Greek to Arabic, and then again when translated from Arabic to other languages such as Spanish, or back to Greek.

Many of these collections of sayings have been transcribed and annotated with TEI XML during the first stage of the SAWS project. Our goal is to identify and expose the reuse and evolution patterns of these sayings within their cultural contexts. An extension of the FRBRoo ontology [7] has been developed specifically to describe the transmission of information. The TEI documents are annotated with RDF, using a customised TEI schema incorporating the SAWS ontology. The RDF encodes relationships with the ancient texts on which the collections drew, with later texts which drew on them and also with one another, since such collections were frequently translated or copied by different scribes.

This paper reports how relationships are extracted from the

TEI documents as RDF triples, allowing researchers to explore the conceptual networks that arise from these interconnected sayings, using linked data. We also describe the advantages that this methodology offers to researchers in the digital humanities.

2. BACKGROUND

This section gives an overview of the SAWS project and what it is aiming to achieve. SAWS is discussed in the context of previous work utilising semantic web technologies, especially where related to similar types of data, and on how the SAWS project benefits from the application of linked data.

2.1 Overview of the SAWS Project: Sharing Ancient Wisdoms

The Sharing Ancient Wisdoms (SAWS) project explores and analyses the tradition of wisdom literatures in ancient Greek, Arabic and other languages, by presenting the texts digitally, in TEI XML format, and annotating the TEI documents using RDF to record links and comparisons within and between anthologies, their source texts, and the texts that draw upon them.

Throughout antiquity and the Middle Ages, anthologies of extracts from larger texts, containing wise or useful sayings, were created and circulated widely, as a practical response to the cost and inaccessibility of full texts in an age when these were all in manuscript form. There has long been interest in the study of this literature and the relationships between manuscripts and within collections [11, 26, 25]. The SAWS project focuses on manuscripts that collected moral or social advice, and philosophical ideas, although the methods and tools developed are applicable to other manuscripts of an analogous form (e.g. medieval scientific or medical texts). These sets of wise sayings ('gnomes' or 'gnomic sayings'), are collectively referred to as *gnomologia*.

The key characteristics of these manuscripts are that they are collections of smaller extracts of earlier works, and that, when new collections were created, they were rarely straightforward copies. Rather, sayings were reselected from various other manuscripts, reorganised or reordered, and subtly (or not so subtly) modified or reattributed. The genre also crossed linguistic barriers, in particular being translated into Arabic; again these were rarely a matter of straightforward translations, but tended to be variations. In later centuries, these collections were translated into western European languages, and their significance is underlined by the fact that Caxton's first imprint (the first book ever published in England) was one such collection [3]. Thus the corpus of material can be regarded as a complex directed graph of manuscripts and individual sayings that are interrelated in a great variety of ways. Analysis of these interrelations can reveal a great deal about the dynamics of the cultures that created and used these texts. This scenario lends itself well to being a practically useful application of linked data [12].

One aspect of the project is to produce a digital archive of editions of some of these texts. In itself, this part of the project brings the gnomologia to a wider audience than was previously possible. Traditionally, working with ancient manuscripts requires the researcher to negotiate a number of hurdles during the research process. Access to the physical manuscripts themselves is often limited, to protect the documents; there is also the question of finding time to visit the locations of the manuscripts and plan this time effectively. Researchers often produce their own critical editions of these manuscripts with commentaries and translations, involving a great deal of interpretation and personal choices on the part of the editor. Paper is not the best medium to transmit this information as it presents many challenges in visualising information and presenting it on the page. What we are doing is to create and disseminate critical editions of these manuscripts in digital form, which are heavily enhanced with semantic annotations to make the manuscripts computer-readable whilst maintaining human-readability. Publishing these digital manuscripts online makes them available to a much wider audience than was previously possible.

This publication strategy is one which digital humanities has embraced widely. TEI publications cover a large variety of subject areas and time periods[1]. SAWS is, however, concerned with more than solely making digital editions of the texts accessible. We want to develop current practice in digital humanities, exploiting the potential of linked data to connect together different texts rather than publishing them in isolation. We can also enrich the TEI editions of the texts with more semantic annotations of the relationships and meaningful information contained in the collections.

SAWS aims to exploit digital technologies in order to better understand the gnomologia as well as publish them, tracing cultural dynamics through identifying and marking up relationships and links between and within documents. The research aims of the project, from a humanities perspective, are to record, publish and share knowledge and expertise on the tradition of wisdom literatures in ancient Greek, Arabic and other languages. The project also facilitates exploratory research from manuscript analysis, by presenting them in a manner that enables linking and comparisons within and between anthologies, their source texts, and the texts that draw upon them. The project is thus producing a framework for representing these relationships, using an RDF-based semantic web approach, as well as tools for creating these complex resources, and for visualising, analysing, exploring and publishing them; these comprise the research aims of the project from a Semantic Web perspective.

We also envisage scenarios where other projects will want to link their own materials to these texts. Thus SAWS will provide a hub for future scholarship in this field and in related areas, as a canonical reference point for the digitised gnomologia. The number of manuscripts of this type is large, and we regard the project as creating the kernel of a much larger corpus of interrelated material, being shared and distributed to facilitate and enhance research. Many of the subsequent contributions will be made by others; consequently we will publish a framework of tools and methods

[1]Examples include the Perseus Digital Library: http://www.perseus.tufts.edu/ , Jane Austen's Fiction Manuscripts: http://www.janeausten.ac.uk/ , Inscriptions of Aphrodisias. http://insaph.kcl.ac.uk/iaph2007/ , CervantesVirtual: http://www.cervantesvirtual.com/ , and the Archimedes Palimpsest Project: http://archimedespalimpset.net/

that will enable researchers not only to search or browse this material in a variety of ways, but also to process, analyse and build on the material. Ultimately we are promoting a distributed network of information to the humanist community, comprising a collection of marked-up texts and textual excerpts linked together, to help researchers represent, identify and analyse the flow of knowledge and transmission of ideas through time and across cultures.

The SAWS project involves collaboration between philological scholars from Sweden, Austria and the UK, a group developing the technical (WWW and SW) aspects of the project and digital humanists who are experienced in applying technical solutions to humanities research questions. It is to be stressed here that several members of the SAWS team fall into more than one of the above categories.

2.2 Related Work and Standards

2.2.1 TEI and RDF

In 1990, DeRose and co-authors reflected on how electronic text documents could best be structured for flexibility in use and reuse, concluding that 'text is best represented as an ordered hierarchy of content object ... the hierarchical model can allow future use and reuse of the document as a database, hypertext, or network.'[5, p. 3]. More than twenty years later, Portier and co-authors stressed that '[w]ith no doubt, the possibility to make complex queries is one of the most useful features of electronic documents.' [24, p. 7]

The Text Encoding Initiative (TEI)[2] is an international standard for the exchange of data, particularly for encoding information about texts. It has been widely adopted as the standard encoding for projects marking up textual data with semantic content [29, 19, 23] and has inspired similar XML encoding standards such as MEI, the Music Encoding Initiative [27]. The popularity of TEI within digital humanities research is due to various factors, such as how it allows the researcher to embed structure and metadata within the transcribed text and produce a variety of useful outputs and indices. That the TEI has been adopted as standard by this community means that interoperability with other projects is enhanced if TEI is used for a particular project.

As described above, a key aspect of SAWS is to represent relationships between and within the different collections of gnomic sayings. RDF[3] is appropriate for this purpose, particularly when supported by an ontology of relevant information and knowledge about this kind of data. RDFa allows RDF to be expressed as attributes within markup language documents, primarily in XHTML documents to date. It is desirable, however, to extend the scope of RDF to be applicable to XML documents on a wider scale, including in TEI XML documents [9], so that RDF semantic information can be used for documents such as the TEI XML documents used in SAWS. Previous attempts have been made to accomodate this [13, 31, 4, 22, 14, 16] but none have been adopted as standard by the TEI community.

RDFTEF [31] offers a way of adding RDF markup to TEI files by converting the TEI representation to RDF/XML which replaces the original TEI version. It converts basic structural markup to RDF and then allows for specific ontologies to be adopted if needed for more intricate markup. Written in Java and based around Jena, RDFTEF can export output in either RDF/XML format or (a form of) TEI/XML format. RDFTEF implements a basic markup ontology for structure, syntactically important elements and sequences of entities (varying in granularity) within the text. Additional ontologies can be added as required. Standard XML tools cannot be deployed within the tool, although (relatively complex) SPARQL queries can be used to query the resulting RDF [24]. The XML limitations, alongside that fact that only a prototype implementation is available that seems to have been left unmaintained (last source code update 2007) have led to RDFTEF being dismissed as '[o]nly a "toy" experiment' [24]. For our purposes, there is an added disadvantage: rather than being part of the TEI editing process, RDFTEF introduces a new, separate stage to the editing workflow, with extra software to be implemented and learnt. Given the non-technical nature of our target audience, this is a significant concern to the SAWS project and adds potential barriers to the adoption of our approach. Additionally, RDFTEF does not allow the RDF and TEI structure, data and semantic markup to co-exist within the same document, which causes potential problems with update consistency and quantities of files.

Other tools are available for representing document structure(s) with RDF: EARMARK [22], GODDAG [4] and MCT [13]. EARMARK [22] is essentially an OWL ontology for document structures, hence it enables RDF to be used to express document structures such as TEI markup but does not solve our problem of how to add non-structural RDF semantic information to a TEI file. GODDAG [4] allows graph models to be constructed as representations of an XML document at a single point in time. Hence GODDAG has similar benefits and limitations as EARMARK and also has the issue that a GODDAG cannot easily be updated once created, restricting the information sharing and updating that we wish to promote. MCT [13], which represents document structures using trees, follows a similar pattern, although it is slightly easier to update an MCT model than a GODDAG representation [24]. Essentially these three options use RDF to model structural information, but not to model the text and additional semantic information, so again structure, data and markup become separated.

Previous work by Jewell [14] and Lawrence [16] has explored the enrichment of TEI encoded texts with RDFa to support the automatic extraction of RDF. Focusing initially on performance texts, this work primarily used the OntoMedia (OM) ontology [15] to describe elements including characters, character location, interaction and travel events within the textual narrative and annotate the existing TEI with explicit reference to the ontological class that the event or entity had been typed as. This typing was done on an automatic basis, processing information extracted directly from the TEI via a conceptual mapping between the TEI and OM. A second script was then used to generate RDF linked data from the extended TEI. By drawing on the ontological data held in the RDF as well as the information encapsulated in both the structure and elements of the TEI, sets

[2]TEI: http://www.tei-c.org/

[3]RDF: http://www.w3.org/TR/REC-rdf-syntax/ , RDFa: http://www.w3.org/TR/rdfa-syntax/

of triples were created that could be cross-referenced with each other and to external data resources while retaining a link back to their textual context. Whilst this approach was successful for this project, the scripts were hard-coded with specific OM information and hence were not appropriate to apply more generally to TEI documents.

A recent development within the TEI community[4] sees the <relation> element used to encode RDF triple information to a TEI document. It encodes the Subject-Predicate-Object triple format through the attributes @active (Subject), @ref (Predicate) and @passive (Object). The major advantage of this development is that it allows RDF to be encoded directly within the TEI environment, fitting into the workflow that researchers are already using, rather than requiring the implementation (and learning of) additional software. Additionally, the @resp (responsibility) attribute can be used to indicate provenance information for a triple by associating the triple with the URI of the person asserting that triple. SAWS has adopted this use of the <relation> element; an example is given below in Section 4.4.

2.2.2 Relevant Ontologies

As mentioned previously, the links and relationships within SAWS documents are encapsulated within an ontology common to all SAWS documents. The ontology formally defines the vocabulary being used to express RDF information (for consistency and accuracy) whilst still being versatile, inclusive and extensible. The SAWS ontology mainly reuses the FRBRoo ontology [7] with some extensions. FRBRoo is a combination of the CIDOC-CRM and FRBR ontologies:

- The CIDOC Conceptual Reference Model (CRM) is an ontology of the information and relationships relevant for cultural heritage documentation [6]. As an official ISO standard (ISO 21127) CIDOC-CRM has been recognised as a common vocabulary for discussing information published on cultural heritage, such as by archives, museums or libraries, and mapping them to a digital equivalent representation [6, 28, 10, 1, 9, 32].
- The Functional Requirements for Bibliographic Records model (FRBR) was devised as an entity-relationship model of bibliographic data and publications [17, 30]. It provides a vocabulary to document and distinguish between: the ideas/concepts forming the basis of a *Work*; the *Expression* of such *Works* in a fixed but abstract form; the *Manifestation* of such *Expressions* in physical form; and single *Items* that are exemplars of such *Manifestations*.

The CIDOC and FRBR ontologies were originally developed independently of each other. On recognising the usefulness of combining the two ontologies, the two communities collaboratively produced the FRBRoo ontology [7]. FRBRoo is essentially the FRBR ontology expressed in an object-oriented form more compatible with that of the CIDOC-CRM, implemented such that it extends the CIDOC-CRM with the FRBR vocabulary. Given the relevance of the CIDOC-CRM and FRBR vocabularies overall, particularly in terms of transmission of cultural information through the repeated transmission of ideas expressed in written works, FRBRoo was the most appropriate ontology to use to represent the SAWS vocabulary, with extensions as necessary after consulting experts in manuscript study. In particular, for SAWS purposes, FRBRoo clarifies how the CIDOC *LinguisticObject* class and the FRBR *Expression/Manifestation* classes relate to each other, allowing greater expressive clarity in representing relationships in and between manuscripts.

Other ontologies relevant to SAWS also exist:

- An extension of the CIDOC-CRM, CRM_{dig}, documents digital objects [8]. Whilst it significantly enhances the CIDOC-CRM for current digital documents, it does not add significantly to the documentation of editions of ancient manuscripts such as those being studied in SAWS. Hence the standard form of CIDOC-CRM is more relevant for our purposes in SAWS.
- In addition to the FRBR ontology, we considered other ontologies documenting bibliographic resources[5] as well as an ontology for documenting scholarly works[6]. Though rich in ways to represent manuscript information, unfortunately they generally lacked sufficient depth to map to the content of the SAWS manuscripts in the detail required.
- Conversely, the OntoMedia [15] and Stories ontologies[7] look more at the content of a text document through events and timings in the stories, but at the expense of less focus on information about the document.
- SKOS could be used to represent the hierarchical structure and organisation of information content within the manuscripts whilst Dublin Core metadata allows us to describe information about the manuscript.[8]

Each of these ontologies are relevant in part for the information being modelled in the SAWS project. Rather than combining several ontologies to represent individual aspects of the SAWS manuscripts, though, the combination of the CIDOC-CRM and FRBR ontologies collectively represent most aspects of the SAWS manuscript data, hence FRBRoo was adopted as the base ontology for SAWS.

3. SHARING ANCIENT WISDOMS: DATA

3.1 The Manuscripts

The term 'manuscript' refers to a document which is being transcribed by SAWS editors. Often manuscripts are large

[4]Sourceforge.net discussion: Encoding RDF relationships in TEI - ID: 3309894, at *http* : *//sourceforge.net/tracker/?func* = *detail&aid* = 3309894&$group_id$ = 106328&*atid* = 644065

[5]The Bibliographic Ontology: http://bibliontology.com/specification and the Simplified Ontology for Bibliographic Resources: https://gist.github.com/1331983

[6]Scholarly Works Application Profile: http://www.ukoln.ac.uk/repositories/digirep/index/ Scholarly_Works_Application_Profile

[7]OntoMedia: http://www.contextus.net/ontomedia , Stories: http://contextus.net/stories/

[8]SKOS: http://www.w3.org/TR/skos-reference , DCMI: http://www.dublincore.org/documents/dcmi-terms/

in size and contain more content than the wisdom sayings in which we are interested. Also, a collection of sayings can span several (parts of) different manuscripts. Accordingly, we refer to a collection of sayings (gnomologium) as a 'CompilationInstance', an extension to the FRBRoo ontology (see Section 4.3) that can exist in one or more manuscripts.

Our primary focus is on Greek gnomologia, from the ninth to twelfth centuries AD, and on Arabic collections of sayings from the same period. Examples of (firstly) a simple section of interest inside these gnomologia, and (secondly) an anecdotal section, are:

> One cannot cover a fire with a cloak nor a shameful sin with time.
>
> Diogenes was asked by someone why people give to beggars but not at all to philosophers, and he said, 'Because, perhaps, they expect to become lame or blind but not to become philosophers.'

Within this second section there are two parts of interest to the manuscript scholar: the statement itself, by Diogenes ('Because, perhaps, they expect ...'), and the narrative text surrounding the statement ('Diogenes was asked...'). The TEI schema for SAWS allows scholars to distinguish between these two types in markup where appropriate.

Over the centuries, manuscripts were often transcribed by various scribes. Different compilers organised the collections in different ways; perhaps according to author, or alternatively according to themes within the sayings, and then according to author within each theme. During the transcription process, there were also many discrepancies, misattributions, mistakes in transmission, or sections missed out.

3.2 The Relationships

We wish to explore relationships within a particular collection (between manuscripts and within manuscripts), between collections, between languages, between collections and source texts (e.g. the original transcriptions of the sayings) and between collections and edited literary texts which made use of them. Example relationships include:

> Manuscript *isWrittenAt* Scriptorium
>
> Manuscript *isInLanguage* Language
>
> CompilationInstance *isWrittenBy* Scribe
>
> CompilationInstance *isTranslationOf* CompilationInstance
>
> Section *isSequentiallySimilarTo* Section [i.e. one Section of a CompilationInstance has a slightly different sequence to another Section but is related, for example through editorial decisions made whilst copying]
>
> ContentItem *isShorterVersionOf* ContentItem
>
> ContentItem *isVerbatimOf* ContentItem

Clearly, a text may have several relationship statements which can be made about it. The definition of relationships has been a key activity for the SAWS project; it is however not simply a mechanical process, but one from which the researchers will learn more about their own texts. The overarching aim of such a model is to allow researchers to represent, identify and analyse the flow of knowledge across texts and cultures. This not only enriches the texts themselves but also lays the basis for a study of the cultural dynamics across the centuries of Greek and Arabic thought, and cultural exchange across civilisations. In developing our model, we have built a vocabulary to express not only the relationships among the texts and textual excerpts in our set of texts, but also those that may occur in analogous bodies of material. This vocabulary has been developed through collaboration between the scholarly community in digital humanities and manuscript studies together with information scientists, to ensure the relevance and completeness of the vocabulary whilst observing linked data standards. The vocabulary has been defined formally and shared as an ontology (see Section 4.3) to maintain consistency across annotation whilst keeping the vocabulary extensible and refinable.

The examples below (all translated into English) show how sayings evolve during the transcription process. The saying in Item 1 is attributed to Alexander the Great in an Ancient Greek gnomological manuscript, 'Gnomomologium Vaticanum'. This is followed by Item 2, an extract from Plutarch's 'Life of Alexander' (8.4.1), which may be a potential source of the saying, although of course this may have been mediated by other anthologies. The text is not a direct quotation, but has become somewhat paraphrased.

1. Alexander, asked whom he loved more, Philip or Aristotle, said: 'Both equally, for one gave me the gift of life, the other taught me to live the virtuous life.
2. Alexander admired Aristotle at the start and loved him no less, as he himself said, than his own father, since he had life through his father but the virtuous life through Aristotle.

Another example shows a 'Chinese-Whispers' effect developing over time. Item 3 is a extract from an Arabic anthology of the sayings of Greek philosophers, attributing a saying to Pythagoras (from 'Selections from the Sayings of the Four Philosophers: (B) Pythagoras' saying 18 (ed. Gutas)). In this case the source text seems to be Diogenes Laertius's 'Life of Aristotle' (5.19), although here (Item 4) not only has the saying become more pithy in translation, the attribution of the saying has changed from Aristotle to Pythagoras. Scholars are interested in when this happened, and why.

3 He said: Fathers are the cause of life, but philosophers are the cause of the good life.

4 Aristotle said that educators are more to be honored than mere begetters, for the latter offer life but the former offer the good life.

4. EXPLORING AND EXTRACTING INFORMATION FROM THE MANUSCRIPTS

The SAWS project has three main aspects:

- The encoding and publication of a digital archive of editions of a selected number of these texts;
- The identification and publication of the links between the anthologies and source texts/recipient texts;
- The building of tools to allow scholars outside the SAWS projects to link their texts to ours.

4.1 TEI XML Encoding of Digital Editions

Each of the texts is being marked up in TEI-conformant XML and validated to a customised schema designed at King's College London for the encoding of gnomologia. Our structural markup reflects as closely as possible the way in which the scribe laid out the manuscript. The TEI schema uses the <seg> element to mark up base units of intellectual interest (not necessarily identified as single units by the scribe), such as a saying (statement) together with its surrounding story (narrative). For example:

Alexander, asked whom he loved more, Philip or Aristotle, said: 'Both equally, for one gave me the gift of life, the other taught me to live the virtuous life'.

This contains both a statement and a narrative:

```
<seg type="ContentItem">
    <seg type="narrative">
        Alexander, asked whom he loved more,
        Philip or Aristotle, said:
    </seg>
    <seg type="statement">
        Both equally, for one gave me the gift of
        life, the other taught me to live the
        virtuous life.
   </seg>
</seg>
```

Each of these <seg> elements is allocated an @xml:id to provide a unique identifier (which is automatically generated - see Section 4.2) that differentiates them from all other examples of <seg>, for instance:

```
<seg type="statement" xml:id="AppGnomVat001s2">
```

In other words, it allows each intellectually interesting unit (as identified by the SAWS team's scholars) to be distinguished from each other unit, thus providing the means of referring to a specific, often very brief, section of the text.

4.2 Auto-generation of xml:ids

In TEI it is good practice to assign significant (or even all) elements an @xml:id, to uniquely identify them within the TEI document. These xml:ids can then be used to form URIs for each part of the document, at the stage of constructing RDF triples from the TEI markup (see Section 4.6). As it can be tedious and error-prone to allocate all xml:ids manually, an XSL transform has been written to automatically assign ids to significant elements - structural section divisions and <seg> elements. Allocated ids are generated based on the document id and the structural location and type of the element. Examples can be seen in the xml:ids for the <seg>s given in Section 4.4 and Figure 1.

4.3 Annotation Ontology

Several types of relationships have been identified within and between the manuscripts. These manuscript relationships exist at many different levels of granularity, from links between individual sayings to inter-connections in families of manuscripts. Using this underlying ontology as a basis, links between (or within) manuscripts can be added to the TEI documents using RDF markup.

From initial discussions between domain experts and technical staff, key resources and relations within the manuscripts were identified. Following good practice in ontology design, and to make use of existing ontology knowledge, several ontologies were reviewed as potential matches or base ontologies for the SAWS ontology (as reported in Section 2.2.2).

Having identified FRBRoo as a good choice of base ontology, the initial set of resources and relations were mapped to the FRBRoo model, to construct an OWL ontology for SAWS. Where the domain experts wished a certain vocabulary to be used which conflicted with the vocabulary provided by FRBRoo, this mapping was expressed using owl:equivalentClass or owl:equivalentProperty, as appropriate. In some cases, the domain experts were happy to use the FRBRoo vocabulary in place of their own (for example, using the property FRBRoo:P130.shows_features_of to express some information being common to two textual materials, rather than the more vague and semantically loose saws:isRelatedTo).

In most cases, existing FRBRoo classes were adequate for the SAWS ontology, either as a direct import from FRBRoo (e.g. Actor, Person, Place) or as a mapping from the desired terminology to a corresponding FRBRoo class (e.g. saws:Family owl:equivalentClass FRBRoo:F1.Complex_Work, or saws:PhysicalManuscript owl:equivalentClass FRBRoo:F4.Manifestation_Singleton). For ease of expressing property domains and ranges, it was necessary to create a new class saws:Material which represents any textual material in SAWS. The saws:Material class corresponds to the union of the classes FRBRoo:F2.Expression and FRBRoo:E33.Linguistic_Object. For further expressive power, several subclasses of saws:Material were created:

- saws:Edition [edited materials]
- saws:HypothesisedInstance [a (text of a) manuscript which scholars hypothesise may have existed but which has now been lost]
- saws:ManuscriptText [the text on a manuscript]
- saws:CompilationInstance [a particular collection of sayings being worked on - manuscripts contained several such collections, as well as other material[9]]

[9]SAWS disregards non-gnomologic material, although it can be represented if desired by FRBRoo:F20.Self-Contained_Expression or F23.Expression_Fragment, as appropriate.

- saws:Section [a division of a CompilationInstance into a self-contained expression]
- saws:Segment [a division of a Section equivalent to the <seg> described in Section 4.1]

Corresponding to the TEI schema markup for expressing types of <seg>s, subclasses of saws:Segment were declared:

- saws:DescriptiveItem [decorative element within the Material, either meaning-bearing or not]
- saws:ContentItem [logical unit within the Material as identified by the Scribe]
  - saws:Narrative [text surrounding or immediately preceding or following the Statement (saying), e.g. 'Aristotle says...', 'The frogs asked for a king.']
  - saws:Statement [the actual saying, e.g. 'All men are mortal']
  - saws:Definition [defining a concept or term used in the Material]
  - saws:Comment [comment on a part of the Material, usually from the modern Editor]
  - saws:Other [unit of the Material within an ContentItem, as identified by the Editor, which isn't a Narrative, Statement, Definition or Comment]
- saws:Marginalia [remark made in the margin of the Material by a Scribe, not necessarily the original Scribe]

Subclasses for the FRBRoo class for Person were also declared, to allow the representation of People as AttributedAuthor [original author of a saying], Scribe[copier of manuscripts] or Editor[a scholar studying the texts in modern times].

The properties in the SAWS ontology are too numerous to list here; the interested reader is referred to the published ontology file, available at `http://purl.org/saws/ontology`. Here we highlight the key properties added to the SAWS ontology to link between and within the manuscripts through observations which can be triplified. Each property's domain and range is the union of saws:Section and saws:Segment.

- saws:isInLanguage
- saws:isVerbatimOf [word-for-word]
- saws:isVariantOf [specialised by other relationships - default option]
  - saws:isShorterVersionOf
  - saws:isLongerVersionOf
  - saws:isCloseTranslationOf [one language to another]
  - saws:isLooseTranslationOf[one language to another]
  - saws:isCloseRenderingOf [e.g. poetry into prose, or dialects of the same language]
  - saws:isLooseRenderingOf [e.g. poetry into prose, or dialects of the same language]

Once the mapping had been completed, the resulting ontology was reported back to the domain experts as a vocabulary to use for annotation of the manuscript. Further discussions and formative feedback from the domain experts at this stage were used to evaluate and revise the ontology. The revised ontological relationships were then incorporated into the SAWS TEI schema as potential values for the @ref attribute of <relation> (see Section 4.4), so that TEI manuscripts could be annotated according to the ontology. Relations were also added to the ontology to represent RDF triples derived from the TEI markup (see Section 4.6).

The SAWS ontology can be accessed through the permanent URL `http://purl.org/saws/ontology`. The ontology is currently being evaluated through extensive application for annotation by domain experts (see Section 6).

4.4 Annotating TEI Documents with RDF

As well as representing the texts in a standard digital humanities way, i.e. using TEI to describe the text in detail and to ensure interoperability with other text encoding projects, we want to be able to represent the relationships highlighted in the SAWS ontology. Whilst several relationships are implicitly and explicitly encoded in the TEI, which we can then extract into linked data, adding RDF allows us to encode ontological information not included in the TEI markup, to enhance the semantic value of our texts further.

Having incorporated the ontological relations into the SAWS TEI schema, they are available for annotation use in TEI. 17 manuscripts have been digitised and marked up in TEI, and more to be marked up during the life of the project (and hopefully after the project completion date, as an ongoing process). Most have now had at least partial semantic annotation through the RDF triples, as the scholars transfer their tacet and editorial knowledge into the TEI files.

To annotate the SAWS TEI documents with RDF triples, we use the xml:id given to the TEI section of interest as a URI. The TEI element <relation> is used in the SAWS documents, using the @resp attribute to acknowledge that many of the links being highlighted are subjectively identified and (a matter of expert) opinion. The @ref attribute states the relationship type, taken from the list of relationships included in the ontology described above. In keeping with existing TEI attributes, the @active attribute points to the URI of the object entity that is being linked from, and the @passive attribute points to the URI of the subject entity that is being linked to. An example from our texts is:

```
[File 1]
<seg type="statement"'
xml:id="K_al-Haraka_tr_c1_s1">
Every body is moved by something else. Therefore it
is only moved via the soul in it and therefore the
soul only becomes intelligent by the intellect in
it. If motion were a characteristic of the body,
every body would have to be moved.
</seg>

[File 2 - see Figure 1]
```

<seg type="statement" xml:id="K_al-Haraka_c1_s1">
كل جسم يتحرك من غيره وذلك أنه إنما يتحرك من قبل النفس التي هي فيه
وذلك أن النفس إنما هي عاقلة بالعقل الذي فيها
ولو كانت الحركة من خاصّة الجسم لوجب أن يكون كل جسم متحركاً
</seg>

Figure 1: Arabic text corresponding to the saying referred to in the example RDF triple (Kitab al-Haraka) in Section 4.4

```
[File 1*]
<relation active="K_al-Haraka_tr_c1_s1"
ref="saws:isCloseTranslationOf'"
passive="K_al-Haraka_ci_s1"
resp= "http://www.scm.uni-halle.de/gsscm/personen/
alumni/dr._elvira_wakelnig/"/>
```

*For convenience, editors have generally added their <relation> annotations directly after or close to the active seg (i.e. the seg that is the subject of the triple); however the <relation> element can be added to any file and do not have to be attached to the original file.

This is equivalent to saying: "The narrative segment identified as $K\_al - Haraka\_tr\_c1\_s1$ is a close translation of the segment identified as $K\_al - Haraka\_c1\_s1$ (Figure 1). Relationship identified by Elvira Wakelnig".

The editors annotating the TEI documents are domain experts, recording and publishing their knowledge. This is an ongoing process, as more manuscripts become digitised for the project. (Currently 17 manuscripts are in TEI/XML format and are being annotated with RDF).

Converting these TEI relations to RDF triples, the xml:ids for <segs> are converted to hash URIs during the transformation of the TEI/XML document, by attaching them to an appropriate namespace declaration in the format `http://www.ancientwisdoms.ac.uk/mss/msName#segId` e.g. `http://www.ancientwisdoms.ac.uk/mss/K_al-Haraka_tr#c1_s1`

As all annotations (i.e. the ontological relationships) include the @resp attribute of the <relation> element, it is easy to trace provenance or responsibility (@resp) of these interpretations of the text. Use of the @resp attribute also allows alternative opinions to be expressed by others, marking their opinions accordingly with the @resp attribute, or for previous assertions by past scholars to be added as annotations. The @resp value is a URI for the person making the assertion behind the annotation.

4.5 Linking our data into the Semantic Web

A key part of the SAWS project is to identify and publish data and inter-relations between data in the manuscripts as Linked Data on the Semantic Web. To this end, we both provide URIs to link into the SAWS data and link out from the SAWS data to external data sources.

Using the auto-generated xml:ids for the document parts, and unique names (namespace appended with filename) for the manuscripts, we can generate URIs as described above. These URIs provide access to the SAWS data at a fine-grained level, for other linked data collections to link into.

Externally, SAWS triples link out to URIs from several data sources on the ancient world, as well as more modern data collections. For URIs for ancient places and people, we use the Pleiades historical gazetteer of ancient places, and an online collection of people mentioned in the Prosopography of the Byzantine World database. If people or places are not included in Pleiades/PBW then we compensate by linking to Geonames/DBpedia respectively. To refer to languages used in the documents, we use the ISO-639-2 standard. To refer to manuscripts and other texts outside the SAWS collections, or parts of these documents, there is no single canonical point of reference; currently we are evaluating different options during annotation, from the Perseus Digital Library, the Thesaurus Linguae Graecae (TLG) and the Canonical Text Service (CTS) URNs.[10]

By linking to other sources in this way we encourage more sharing of our scholars' data, provide access points to the data such as for people interested in linked entities e.g. Aristotle, and make our Linked Data part of the Semantic Web. Additionally, the current version of the ontology underlying the semantic annotation of the manuscripts can be obtained from `http://purl.org/saws/ontology` . We publish our Linked Data under the Open Data Commons license [18].

4.6 Publishing TEI+RDF online

We use *kiln* (`https://github.com/kcl-ddh/kiln`) to publish our dynamic pages, supplemented with a Django CMS for static pages. Kiln provides a framework to publish TEI (and other XML), through stylesheets. It incorporates Solr (`http://lucene.apache.org/solr/`) for search and indexing of the manuscripts, and includes a Sesame plugin for ease of access to a RDF triple-store within the site.

To publish the TEI files, they are transformed from XML to XHTML+RDFa using an adaptation of the standard TEI stylesheet (`http://www.tei-c.org/Tools/Stylesheets`). The adapted stylesheet transfers RDF triples from the TEI file to the displayed XHTML by converting the information in the <relation> elements to RDFa within the XHTML.

We use an XSLT to extract ontology-specified relations. The transform extracts triples that have been directly encoded by the domain experts as a TEI <relation>. It also retrieves the semantic information already represented in the TEI encoding, which can be automatically generated from the TEI markup. The transform builds up a collection of triples in RDF/XML format, to be stored in the Sesame triple-store.

5. ONLINE PUBLICATIONS AND OUTPUT

The three aspects of the SAWS project (see Section 4) will reflect in three types of online publications:[11]

[10]Pleiades:http://pleiades.stoa.org/ , PBW: http://www.pbw.kcl.ac.uk/ , Geonames: http://www.geonames.org/ , DBpedia: http://dbpedia.org/ , ISO-639 standards: http://id.loc.gov/vocabulary/iso639-2/ , Perseus:http://www.perseus.tufts.edu/ , TLG: http://www.tlg.uci.edu, CTS: http://cts3.sourceforge.net/

[11]At http://www.ancientwisdoms.ac.uk

- Digital editions: publication of semantically-enhanced digital editions, through TEI+RDF files published as XHTML+RDFa, to be read, browsed and searched
  - Selected manuscripts (approx 4 - 6) are to be presented in a demonstration digital edition to gather evaluative feedback from manuscript scholars (not necessarily digital humanists) in June 2013. These are linked together via the RDF and TEI markup and are searchable using Solr. Following this gathering of feedback, the demonstration will be refined in line with received comments, then expanded to a full implementation with several more manuscripts. Currently, 17 collections of gnomologia have been digitised and annotated in preparation for this stage, with more being prepared on an ongoing basis during the life of the project. It is hoped that after June 2013, the completion date for the SAWS project, the process of adding annotated TEI files to the SAWS repository will be ongoing, not only by SAWS scholars, but by external researchers.
- Semantic Web: Publication of facts and information from manuscripts as Linked Data, through a Sesame triple-store and SPARQL endpoint, and of the underlying ontology
  - The SAWS extension of the FRBRoo ontology is available as an OWL file at `http://purl.org/saws/ontology`
  - RDF triples from the demonstration gnomologia will also be presented at the June 2013 demo to manuscript scholars. As our SAWS scholars complete annotating a TEI manuscript, the RDF triples are extracted from the TEI via a XSLT transform and added to a Sesame triple-store. For the June 2013 demo the triple-store will be queryable via a SPARQL endpoint. Mindful of the usability requirements of our primary target audience, however, we are researching more user-friendly alternatives for a non-technical audience.
- Online tools: tools for 'doing SAWS'
  - This will allow researchers outside the immediate SAWS team to have access to tools to create, mark-up, edit and semantically annotate digital versions of manuscripts and add them to the SAWS repository and triple-store. Additionally, XSLT stylesheets, the TEI schema RNG file and the ontology OWL file will be available from this part of the site. Development of this part of the site is a future stage of the SAWS project (although certain parts of its content, such as the OWL ontology file, are made available already). As such, development of this content for the SAWS website will be discussed in Section 8.

6. EVALUATION

Overall, the project will be deemed to be successful if upon completion of the project (June 2013), we have achieved:

- Digital edition of manuscripts published using TEI and RDF annotations.
- Manuscripts to be navigatable through structural and semantic links.
- Semantic content in manuscripts to be searchable and queryable through extraction of RDF information.
- Positive impact to the philological community, particularly those researching medieval manuscripts (this is further discussed at the end of this section)

Currently, we are preparing a demonstration of a digital edition of selected manuscripts and its corresponding RDF store. This demo will be presented to a workshop of domain experts, for feedback and to engage more interest in adoption of SAWS approach. The feedback will be two-fold: for those scholars who have been involved in the SAWS process, marking up texts in TEI and adding RDF, this will be a chance for them to see how Semantic Web technology allows them a new form of access to the information in the manuscripts. Accordingly, their comments will be likely to centre around the usability, functionality and presentation of the demo site. For the scholars not connected to the SAWS project (the majority of the audience), we expect that feedback will be more wide-ranging, feeding back comments on the whole approach as well as specific comments on the demo. It is hoped that for both audiences, the demonstration of Linked Data shows how the Semantic Web can be used to publish expert knowledge and to present information in new ways. A particular note of success will sound if the demo highlights connections between the manuscripts that are novel discoveries for (some of) the audience, especially if these come from reasoning inferences rather than directly-encoded relations.[12]

In evaluating the SAWS ontology, the logical consistency of the resulting OWL ontology was checked using the Protege tool. Another stage of evaluation of the ontology occurred when the original ontological requirements from the collaboration between domain experts and technical observers was mapped to the existing ontology FRBRoo, which has undergone extensive review from both the CIDOC and FRBR community as well as users of FRBRoo. After this mapping process, the resulting ontology was presented to the domain experts in the form of a vocabulary they could use to express relationships between and within the manuscripts they studied. Formative feedback solicited from the domain experts at this stage was used to refine the ontology further. Currently, the ontological relations have been made available for annotation use in TEI through the TEI schema, as described above in Section 4.4. The application of the ontology for practical annotation is ongoing, resulting in further feedback on discrepancies between what the experts want to express and what vocabulary the ontology provides, at a much finer-grained level of detail. Critically, the ontology is successful to the degree that it allows domain experts to record their tacet knowledge and expertise in digital form; this is largely being demonstrated and the ontology is in its latter stages of refinement (notwithstanding future extension and reuse of the SAWS ontology by others).

[12]It is hoped that presenting the SAWS demo to a more technical audience such as at WIMS'12 will provide feedback on more technical aspects of the project, as has already been seen from the helpful feedback from the WIMS'12 reviewers.

Further evaluation of the relationships has been undertaken by presenting the ontology resources and relations to domain experts outside of the SAWS team, and soliciting feedback. Doing this has both improved the ontology and, where the SAWS domain experts are presenting the relationships, helped them to understand the ontology better. For example, on receiving feedback that the ontology was too focussed towards medieval anthologies of wise sayings, one of the SAWS domain experts reported this feedback to the technical team. As the underlying FRBRoo ontology provides ways to deal with other types of texts, this issue could then be treated as a way of improving the presentation of the ontology (by including more of the underlying FRBRoo ontology) rather than improving the ontology.

The SAWS project has received many indications of interest from the philological community (those who research historical texts). On a longer term basis, success of the SAWS approach will be demonstrated in the future and ongoing adoption of a SAWS-style approach by others across this community, for editing, annotation and publishing of digital manuscript editions. A particular marker for success will be the linking to and from SAWS manuscripts by scholars outside of the SAWS research team, particularly if the SAWS digital editions become the canonical reference point for the manuscripts digitised during SAWS. Another indicator will be if the SAWS approach is adopted by researchers outside the immediate target audience of manuscript scholars, for example those studying modern texts or other objects representable by TEI, for example the MEI (Music Encoding Initiative) community [27].

7. ADVANTAGES OF TAKING THE SAWS APPROACH

Through marking up the manuscripts in TEI XML we have made these collections of wise sayings available in digital form with structured content, removing the accessibility problems to the original physical manuscripts. The text of the manuscripts has been supplemented with expert knowledge, much as would happen when producing a critical edition.

The mark-up process has been undertaken both by experts in this area and non-experts supervised by experts and given brief training. Especially for larger-scale mark-up projects, the mark-up process can be time-consuming and it is useful to be able to share this workload without needing to recruit several people with detailed expert knowledge. The process of tagging the electronic versions of the collections is modular and can be performed in a distributed way, across a number of people, with the experts being able to add more detail from their specialist knowledge whilst sharing the more repetitive mark-up with others.

The markup provided through the SAWS TEI schema caters for different stages of annotation, from quick annotations 'out in the field',[13] when researchers are actually at the physical location where the manuscript is kept, to initial editing of structure and brief observations, through detailed analysis to the publication of a critical edition of the manuscript. This models the analytical processes and stages that such researchers are already familiar with in their work.

[13]As described by Charlotte Roueché, P.I. for SAWS.

The inclusion of RDF in TEI documents is a current area of interest in the TEI community [10;, 9, SIG: http://www.tei-c.org/Activities/SIG/Ontologies] as there is a growing desire to make more of the XML documents by including relationship information within the TEI markup itself. Information on how documents are related and how links exist within documents is extracted from analysis to be included in the TEI editions of the manuscripts. This enhances the semantic content of these electronic versions. This process is supported by the extension and reuse of a well-designed, flexible ontology on cultural heritage bibliographical object documentation, the FRBRoo ontology [7].

8. FUTURE WORK

Once documents have been fully tagged up with RDF annotations, this will facilitate more automated knowledge extraction from the documents such as sequential orderings of sections of the text within collections.[14] Such orderings are significant in manuscript analysis as scribes would often take some editorial liberties with the texts they were transcribing, re-ordering them as they saw best. Identifying such relationships automatically will be of great help to the digital humanities researcher, particularly where manual identification is possible but time-consuming, or where such relationships may be overlooked.

The SAWS approach allows us to extract triples from the marked up TEI documents, to be stored in a triple store and queried with SPARQL. With the data in a queryable form, this opens up a whole host of exciting possibilities. The primary aim is to enable the creation of digital analysis and information extraction tools for the immediate target audience (digital humanities researchers), to collect information. Outside the immediate audience, data on wise sayings and how they have evolved over transcription and transmission would also be of interest to linguists, social scientists and historians. The collections of sayings could also be exploited for potential 'pop'-applications outside the academic sphere of interest, such as online or mobile apps to generate wise sayings in appropriate contexts.

In SAWS we are creating a framework for others to use and extend; a growing network of interconnected information. As the body of material of interest in this field is potentially very large, we do not view the project as creating *just* a digital, online edition, although this will be *one* result of the project, but rather as creating the kernel of a much larger corpus of interrelated digital editions. We envisage this as a SAWS 'hub' for enabling related projects to annotate and link their own texts. The research value of such a corpus would be much greater than the sum of its parts, and would increase dramatically once a 'critical mass' was reached.

Many of the subsequent contributions to this corpus will, of course, be carried out by other researchers. If these projects are to be able to interoperate and contribute to the wider corpus, rather than existing as a collection of separate editions (which would be of much less value to researchers), it is important that everyone 'speaks the same language' regarding how this material is represented in digital form, both

[14]To this end, the SAWS project is a use case for a current project at FH Worms, Germany, on similarity detection in texts across single and/or multiple languages.

semantically and technically; herein lies a significant part of our long-term provision for future scholarship. Moreover, these contributions are likely to be made over a long duration, certainly long in relation to the speed of technical developments, so our approach must be such as to allow migration, without loss of information, as the technological environment changes. We hope our adoption of current standards and ontology reuse assists us here to some extent.

The development of user-facing tools is another area in which it will be essential to work closely with the scholars who will (or who may) use the tools. We cannot assume that the users will be au fait with the technology, neither can we assume that all scholars will have access to specialists in this area, so the tools must be usable with the help only of standard on-line help and documentation. This offers us an excellent opportunity to identify, test and respond to the needs of this scholarly community. As we are taking a long-term view, and to ensure that the tools and interface are usable and support the requirements, we will take an evolutionary approach to development, involving incremental cycles of prototype implementation and evaluation in collaboration with potential users and other developers.

As part of this we will develop lightweight tools for this broader community, or as far as possible reuse existing tools, that are simple to use, maintain and enhance. As one example, we are collaborating with a team at FH Worms, Germany, on an editor which allows links to be added between two text documents. This editor, the Text-Text-Link-Editor, will supplement and enhance the Text-Image-Link-Editor that was developed as part of the TextGrid project [20]. We also intend to make use of tools developed during the TEXTvre project [2] to enable easier collaborative sharing and editing of texts through the TextGridLab. One last example deals with the scenario where new relationships will be identified after editing is complete; indeed these may be identified by specialists in other areas, such as philosophy. Therefore we will require tools to add new RDF triples and also to maintain and revise the ontology version.

9. CONCLUSIONS

The Sharing Ancient Wisdoms project allows us to exploit semantic web technologies for a better understanding of the cultural dynamics of gnomologia (collections of wise sayings). This is achieved by means of:

- New editions being published online, with open access
- Through the identification of points of interest in texts and relationships between texts
- A methodology to be used by others analysing and publishing similar material

In SAWS, we are producing digital editions of some (neglected) texts, identifying a network of relationships between texts and providing a framework for other to continue to build upon this network. Essentially the aim of SAWS is to produce critical electronic editions of these manuscripts which are heavily enhanced with semantic annotations to make the manuscripts computer-readable whilst remaining human-readable.

We advocate a methodology for bringing these manuscripts to life online, making them accessible in a way not previously possible and offering tools to support the researcher in studying the collections. We hope that by using these texts and the relationships between them to analyse the flow of knowledge between texts and cultures, SAWS will give us a better understanding of the processes of cultural exchange between civilisations, and in particular of the cultural dynamics across the centuries of Greek and Arabic thought.

10. ACKNOWLEDGMENTS

The SAWS project is funded by HERA (Humanities in the European Research Area) as part of a programme to investigate cultural dynamics in Europe. It is composed of teams at the Department of Digital Humanities and the Centre for e-Research at King's College London, The Newman Institute Uppsala in Sweden and the University of Vienna.

11. REFERENCES

[1] C. Binding, K. May, and D. Tudhope. Semantic interoperability in archaeological datasets: Data mapping and extraction via the CIDOC CRM. In *Research and Advanced Technology for Digital Libraries*, volume 5173 of *Lecture Notes in Computer Science*, pages 280–290. Springer, 2008.

[2] T. Blanke and M. Hedges. Humanities e-Science: From systematic investigations to institutional infrastructures. In *Proceedings of the 6th IEEE e-Science conference*, Brisbane, Australia, 2010.

[3] W. Caxton. *The Dictes and Wise Sayings of the Philosophers*. (originally published London, 1477), reprinted 1877 (Elliot Stock, London), 1877.

[4] A. Dekhtyar and I. E. Iacob. A framework for management of concurrent XML markup q. *Text*, 52(2):185–208, 2005.

[5] S. J. DeRose, D. G. Durand, E. Mylonas, and A. H. Renear. What is text, really? *Journal of Computing in Higher Education*, I(2):3–26, 1990.

[6] M. Doerr. The CIDOC CRM - an ontological approach to semantic interoperability of metadata. *AI Magazine*, 24(3):75–92, 2003.

[7] M. Doerr and P. LeBoeuf. Modelling intellectual processes: The frbr - crm harmonization. In C. Thanos, F. Borri, and L. Candela, editors, *Digital Libraries: Research and Development*, volume 4877 of *Lecture Notes in Computer Science*, pages 114–123. Springer, Berlin / Heidelberg, 2007.

[8] M. Doerr and M. Theodoridou. crm_{dig}: A generic digital provenance model for scientific observation. In *Proceedings of TaPP'11: 3rd USENIX Workshop on the Theory and Practice of Provenance*, Heraklion, Greece, 2011.

[9] O. Eide, A. Felicetti, C. Ore, A. D'Andrea, and J. Holmen. Encoding Cultural Heritage Information for the Semantic Web. In *EPOCH Conference on Open Digital Cultural Heritage Systems*, Rome, Italy, 2008.

[10] O. Eide and C.-E. S. Ore. From TEI to a CIDOC-CRM conforming model: Towards a better integration between text collections and other sources

of cultural historical documentation. In *Digital Humanities (Poster contribution)*, Urbana-Champaign, Illinois, 2007.

[11] D. Gutas. Classical Arabic wisdom literature: Nature and scope. *Journal of the American Oriental Society*, 101(1):49–86, 1981.

[12] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool., 2011.

[13] H. V. Jagadish, L. V. S. Lakshmanan, M. Scannapieco, D. Srivastava, and N. Wiwatwattana. Colorful XML: One Hierarchy Isn't Enough. In *Proc ACM SIGMOD Int Conf on Management of Data*, volume 1, pages 251–262. ACM Press, 2004.

[14] M. O. Jewell. Semantic Screenplays: Preparing TEI for Linked Data. In *Proceedings of Digital Humanities*, London, UK, 2010.

[15] K. F. Lawrence. *The Web of Community Trust - Amateur Fiction Online: A Case Study in Community Focused Design for the Semantic Web. Part IV.* PhD thesis, University of Southampton, UK, 2007. Available at *http : //eprints.ecs.soton.ac.uk/*14704/.

[16] K. F. Lawrence. Wherefore Art Thou? - Crowdsourcing Linked Data from Shakespeare to Dr Who. In *Proceedings of Web Science*, Koblenz, Germany, 2011.

[17] O. M. A. Madison. The IFLA functional requirements for bibliographic records: International standards for universal bibliographic control. *Library Resources & Technical Services*, 44(3):153–159, 2000.

[18] P. Miller and R. Styles. Open data commons, a license for open data. In *Workshop about Linked Data*, 2008.

[19] E. Mylonas and A. Renear. The text encoding initiative at 10: Not just an interchange format anymore - but a new research community. *Computers and the Humanities*, 33(1):1–9, 1999.

[20] H. Neuroth, F. Lohmeier, and K. M. Smith. Textgrid - virtual research environment for the humanities. *International Journal of Digital Curation*, 6(2):222–231, 2011.

[21] S. Nichols. Time to change our thinking: Dismantling the silo model of digital scholarship. *Ariadne*, 58, 2009.

[22] S. Peroni and F. Vitali. Annotations with EARMARK for arbitrary, overlapping and out-of order markup. In *Proceedings of the 9th ACM symposium on Document engineering*, pages 171–180, Munich, Germany, 2009.

[23] E. Pierazzo. A rationale of digital documentary editions. *Literary and Linguistic Computing*, 26(4):463–477, 2011.

[24] P. Portier, N. Chatti, S. Calabretto, E. Egyed-Zsigmond, and J. Pinon. Modeling, encoding and querying multi-structured documents. *Information Processing & Management*.

[25] M. Richard. Florilèges grecs. In *Dictionnaire de Spiritualité V*. 1962. cols. 475-512.

[26] F. Rodríguez Adrados. *Greek wisdom literature and the Middle Ages: the lost Greek models and Their Arabic and Castilian Translations*, pages 91–97. English translation by Joyce Greer (2009), 2001.

[27] P. Roland. The Music Encoding Initiative (MEI). In *Proceedings of the First International Conference on Musical Applications Using XML*, pages 55–59, 2002.

[28] P. Sinclair, M. Addis, F. Choi, M. Doerr, P. Lewis, and K. Martinez. The use of CRM core in multimedia annotation. In *Proceedings of First International Workshop on Semantic Web Annotations for Multimedia (SWAMM 2006), part of the 15th World Wide Web Conference*, Edinburgh, Scotland, 2006.

[29] C. M. Sperberg-McQueen. Text in the electronic age: Texual study and textual study and text encoding, with examples from medieval texts. *Literary and Linguistic Computing*, 6(1):34–46, 1991.

[30] B. Tillett. What is FRBR? a conceptual model for the bibliographic universe. *Library of Congress Cataloging Distribution Service*, 25(5):1–8, 2004.

[31] G. Tummarello, C. Morbidoni, and E. Pierazzo. Toward textual encoding based on RDF. In *Proceeding of the 9 th International Conference on Electronic Publishing (ELPUB 2005), Kath. Univ. Leuven*, number June, pages 57–63. Citeseer, 2005.

[32] R. Varnienè-Janssen and J. Juskys. Strategic, methodological and technical solutions for the creation of seamless cultural heritage content: Lithuanian approach. In *Proceedings of Summer School in the Study of Historical Manuscripts*, Zadar, Croatia, 2011.

[33] D. M. Zorich. A survey of digital humanities centers in the United States. Technical Report 143, Council on Library and Information Resources, Washington, DC, November 2008.