# Affordances and limitations of algorithmic criticism

Verhaar, P.A.F.

Chapter 7

# Data Representation

## 7.1. Introduction

In *The Art of Literary Research*, Richard Altick advises students of literature to make systematic notes of all the primary and secondary sources which are examined. Such a collection of annotations can form "a repository of factual data and ideas drawn from your sources".[446] Using the captured notes as a basis, scholars can reflect on the various linkages and on the potential contradictions within the various facts and ideas, and such processes can eventually culminate in the formation of scholarly claims. In a study into the scholarly information practices of researchers in the sciences, the social sciences and the humanities, Palmer et al. concluded that "notetaking" can be viewed as one of the core scholarly primitives. The authors indicate that making notes can form an important part of searching and reading, and note additionally that it is often viewed as a preliminary stage in generating new original texts. [447] These findings were corroborated by Chu, who has proposed a schematic description of the scholarly process followed by literary scholars, on the basis of data acquired from structured interviews and surveys held among more than 150 researchers from different schools of literary criticism. Chu observes that literary scholars generally begin their research by producing notes about the primary and secondary materials they have selected. Such annotations point to the "quotations, answers, images, themes"[448] that are relevant in the light of a specific research question, and mostly consist of brief additional texts which clarify the relevance or the meaning of that text fragment. After the first reading, individual annotations can be related to each other, and through such associations, certain patterns may emerge. The overall argumentation generally begins to take shape when scholars become aware of certain forms of repetitions, or when they observe specific anomalies with respect to the occurrence of particular phenomena.

In a conventional setting, annotations are habitually intended for personal use only.[449] Such analogue research annotations are typically unstructured and the

[446] Richard Altick, *The Art of Literary Research* (New York: Norton 1963), p. 177.

[447] Palmer Carole L. Palmer, Lauren C. Teffeau & Carrie M. Pirmann, *Scholarly Information Practices in the Online Environment: Themes from the Literature and Implications for Library Service Development.*, p. 30.

[448] Clara M. Chu, "Literary Critics at Work and Their Information Needs: A Research-Phases Model", in: *Library & Information Science Research*, 21:2 (January 1999), p. 260.

[449] About his bibliographic slips, Altick explains that "nobody sees those slips but me". See Richard Altick, *The Art of Literary Research*, p. 173.

terminology is rarely fully standardised. When annotations are created digitally, however, this can often yield a number of benefits. One of the advantages is that the various facts and ideas can be searched. This advantage applies even more strongly when the annotations are captured in a structured format. When research is carried out with the aid of digital tools, this generally leads to a greater standardisation of research results. Such standardisation is typically a function of the fact that digital tools demand consistent and predictable data as input. When structured annotations can be processed algorithmically, connections, correlations and differences can often be identified more methodically.

An additional advantage of structured digital data is that these can be shared beyond the research project in which they were originally created. It is often assumed that, when researchers share some of their resources, peers can combine the various data sets that are available in order to carry out broader and more resourceful studies. In particular areas of research, such extensive data sets are indispensable. To make valid claims about developments in global literary history, for instance, it is necessary to collect details about different regions and different eras. As the vast data sets which are needed for such expansive studies can impossibly be produced by individual scholars, Franco Moretti has claimed that "quantitative work is truly cooperation".[450] When data are consolidated using a fixed format, this also enables peers to replicate the analyses that have been performed and to verify the main claims that were made on the basis of these analyses.

Open access to research data, combined with transparency about the way in which data were produced, can help to terminate the isolationist nature of humanities research and may genuinely transform it into a collaborative endeavour. Academic work in the humanities traditionally concentrates on the formation of ideas produced by individual scholars, and it is highly exceptional for a single theory or idea to be spawned by a team of researchers. Borgman notes that the humanities largely have "the lowest rate of co-authorship and collaboration".[451] Conventional scholarship frequently confirms the stereotype of the "solitary humanist; the ideal, derived from the Romantic Era, of the great mind communing with itself".[452] Computer-based scholarship, conversely, is often collaborative. [453] Tools and data sets are often constructed by interdisciplinary groups of scholars, and conferences on the use of technologies within the humanities attract increasingly large audiences. Scheinfeldt observes that scholars in the digital hu-

[450] Franco Moretti, *Graphs, Maps, Trees: Abstract Models for a Literary History* (Verso 2005), p. 5.
[451] Christine Borgman, *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*, pp. 219–220.
[452] Cathy N. Davidson, "What If Scholars in the Humanities Worked Together, in a Lab?", *The Chronicle of Higher Education*, 28 May 1999, n.pag.
[453] Cathy N. Davidson, "Humanities 2.0: Promise, Perils, Predictions", in: Matthew Gold (ed.), *Debates in the Digital Humanities*, Minneapolis: University of Minnesota Press 2012.

manities are often viewed as "nice", and that the field can be characterised using bywords such as "'collegiality', 'openness', and 'collaboration'".[454] The numerous online platforms on which scholars exchange ideas clearly illustrate the notion that there is a strong sense of a community.

Sharing research data also poses a number of obstacles, however. Moretti assumes that data constitute a transferrable commodity, and that, once produced, they can have an existence independently from their creator. A prerequisite for a shared use of electronic resources, however, is that these are consistently available in a standardised format, and that there can be a degree of technical and semantic interoperability. Standardisation does not always seem possible in the case of humanities research. Studies usually focus on the myriad of ways in which human beings have expressed themselves, and the interpretation of such highly diverse artistic utterances cannot always be formalised unequivocally. In addition, standardisation inevitably imposes certain limitations. When standards are being developed, the aim is usually to describe a domain and to allow room for as many aspects as possible. Once a standard is adopted, it loses some of it flexibility and users of the standard may need to adjust their own descriptive practices to the standard. A standard also restricts what can be said and how things can be said, and full expressiveness may need to be sacrificed for the higher goal of being able to collaborate.

Despite these potential difficulties, structured annotations can be valuable for a variety of reasons, and, within the case study that was conducted for this thesis, a decision was also taken to capture all the generated data using a standardised data format. As part of the case study, a software application was developed which can generate annotations about literary phenomena such as rhyme, metre, assonance and alliteration. This application has been used to produce secondary data about the poetry of Louis MacNeice. In total, 246.660 observations have been generated about the 311 poems that were analysed. As it seemed crucial to ensure that these observations could all be captured accurately, the selection of the data format was based on three requirements. Firstly, the format needed to provide support for a sufficiently rich ontology which can accommodate the various phenomena this study has focused on.[455] A second requirement was that it needed to be possible to connect the terms from the ontology to specific text fragments. The representation language in itself could not impose any practical limitations in connecting terms from the ontology to specific text fragments. Thirdly, when a term was connected to a text fragment, it also had to be possible for a scholar to claim responsibility for this act. The act of describing a literary text is, almost inevitably, interpretative. Systems for the representation of data ought to be based on the assumption that

[454] Tom Scheinfeldt, "Why Digital Humanities is "Nice"", in: Matthew K. Gold (ed.), *Debates in the Digital Humanities*, Minneapolis: University of Minnesota Press 2012, p. 59.

[455] For a discussion of the term ontology, see section 4.2 of this thesis.

there are no objective facts about human artefacts in themselves. The only class of facts that may be said to exist is that a certain scholar, working at a particular moment in time, made a statement about a particular artefact. Data formats must be capable of recording such controversy.

As was also discussed in Chapter 6, the data that were produced within the case study have partly been captured using The Text Encoding Initiative (TEI). The standard provides a valuable method for describing the logical structure of the text, amongst other aspects. Because it was found, however, that the standard did not fully meet the three requirements which were formulated above, the majority of the observations about MacNeice's poems have been recorded using the Open Annotation Collaboration (OAC). This chapter offers a motivation for the two data formats which were chosen and additionally discusses the strengths and the shortcomings of these two formats. Importantly, the description of the weaknesses of TEI in section 7.2 differs in a qualitative sense from the discussion of the affordances of OAC in section 7.3. TEI has already been in use for more than 25 years,[456] and the practical details of its use have already been discussed extensively elsewhere. The criticism of the standard is mostly of a fundamental and theoretic nature. OAC, on the contrary, is currently an emerging technology, and there are still very few texts on the ways in which the protocol can be used, especially in the domain of literary studies. The discussion of the strengths and weaknesses of OAC is, for this reason, much more practical and concentrates more specifically on the techniques that have been applied in the case study of this thesis.

## 7.2. The Text Encoding Initiative

Unsworth has explained that digital models are based on ontologies.[457] The suitability of digital surrogates depends for a large part on the nature of the underlying ontology. To gauge the utility of the TEI for the case study in this thesis, it is necessary to consider the standard's facilities for the description of textual aspects that are typically associated with poetic texts.[458] The <l> and <lg> elements, both from the TEI core set, may be used to mark up the structural division of a poem into stanzas and into verse lines. Within <lg>, the @type attribute can be used, so that a distinction can be made between stanzas of different lengths. Both <lg> and <l> may be combined with a number of attributes that can offer information on rhyme schemes and metrical patterns. More specifically, the @met attribute can be

[456] James Cummings explains that "The TEI grew out of a recognized need for the creation of international standards for textual markup that resulted in a conference at Vassar College, Poughkeepsie, in November 1987." See James Cummings, "The Text Encoding Initiative and the Study of Literature", p. 451.

[457] John Unsworth, "What Is Humanities Computing and What Is Not?".

[458] Most of these are described in section 6 of the P5 TEI guidelines, <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/VE.html> (16 March 2013)

used to record either a term descriptive of the metrical pattern (e.g. "iambic pentameter") or a schematic representation of the metre (e.g. "-X-X-X-X-X"). The @rhyme attribute is generally used at the level of the <lg> element to indicate the rhyme scheme of the lines contained within the line group. On the individual lines, the words that rhyme may potentially be encoding using the <rhyme> element. On the basis of this mechanism, both final rhyme and internal rhyme may be recorded.[459] Importantly, since the @rhyme attribute cannot be used more than once within an <lg> element, encoders can only provide information about a single form of rhyme. Many of MacNeice's poems contain adroit combinations of perfect rhyme, slant rhyme or semi-rhyme, however. Examples include "The Habits", "Homage to Clichés" and "Solstices". By default, the TEI standard offers insufficient support for the description of such concurrent instances of different forms of rhyme.

While some aspects of metre and of rhyme may be captured within attributes of the <l> and the <lg> elements, the TEI crucially lacks specialised vocabulary for the description of figures of speech or of figures of thought. About the annotation of figurative language, the guidelines propose that, given "the great richness of modern metaphor theory", any single proposal "would have seemed objectionable to some and excessively restrictive to many". The guidelines suggest that, when particular descriptive terms are not supplied by the standard, encoders ought to add these terms themselves. During the development of the Myopia application, which is a visualisation environment aimed to support the close reading of poetry, Chaturvedi et al. have addressed these limitations of the TEI via the definition of bespoke elements and attributes for particular literary phenomena.[460] The elements <ambiguity> and <connotation> were created, firstly, to enable scholars to record particular words associated with the tokens in the poetry. Figurative language could also be described via a <metaphor> element with associated @tenor and @vehicle attributes. To identify verse feet, a <foot> element was added, and the syllables contained in these could be categorised as iambic, spondaic and pyrrhic patterns via the @type attribute. <subject>, <verb> and <object> were added to record the syntactic structure of sentences. Figures of speech based on acoustic patterns could be characterised using <consonance>, <assonance>, and <alliteration>.

An alternative approach, next to the creation of additional terms in customised TEI schemas, is to make use of the existing <interp> element. This element may be used to associate a user-defined descriptive term with a unique identifier. This identifier can subsequently be used as the value of the @ana attribute. Amongst many other elements, this attribute may be used within a <seg> element, which

[459] The TEI P5 guidelines coincidentally use Louis MacNeice's poem "The Sunlight on the Garden" to illustrate the facilities for encoding occurrences of internal rhyme.

[460] Manish Chaturvedi et al., "Myopia: A Visualization Tool in Support of Close Reading", (2012), n.pag.

can be used to delineate a specific sequence of words, without immediately specifying the nature of the selected text fragment. Kate Singer has argued that such expansiveness is very beneficent from the viewpoint of literary criticism. Within the TEI standard, there is clearly a dearth of terms for the description of literary devices. Such paucity, however, stimulates encoders to "interrogate poetics terms and interpret texts in descriptive rather than prescriptive ways precisely because they do not automatically resort to classical literary terminology".[461] Descriptive encoding also implies a form of classification. Singer speculates that, if the TEI consortium had supplied a highly developed system for the classification of literary devices, encoders would have been prompted to apply these terms mechanically, and to overlook phenomena which lie outside predefined categories. She surmises additionally that the ability to supply new terms via the @ana attribute encourages scholars to reflect critically on the unique properties of specific fragments, and to devise singular terms to describe their qualities. Such new terms may describe the connotations of specific words or they may characterise the nature of recurrent tropes. The terms that are used traditionally with literary criticism are, to a large degree, historically contingent, but the TEI also enables scholars to supply local or idiosyncratic variants for the received terminology and to effectuate a degree of acclimatisation.

Whereas TEI encoding can in theory be added both manually and via text mining algorithms, the mark up is incorporated most frequently via the former method. The encoding is commonly the reflection of a close reading of the text. Close reading usually consists of deeply attentive forms of engaging with texts, and readers may respond to the words they encounter in widely diverse ways. It may be argued that the technical structure of the TEI mirrors such variability. Via its lenient ontology and via its extensibility, the TEI offers support for idiosyncratic and irregular ways of interacting with texts. In quantitative analyses of large collections of texts, however, it is generally important to explore the broader patterns and to establish the interconnections between distinct texts. For such applications, it is essential to ensure that related phenomena have been identified identically throughout the entire corpus. In many cases, such consistency can only be achieved if terms have been described via an ontology which has been applied strictly.

The TEI Schema, which is maintained by the Text Encoding Initiative Consortium, may be considered a "loose" ontology, in two important senses. Firstly, the standard clearly does provide a vocabulary which can be used to describe and to encode specific characteristics of texts. In the introductory chapters of the TEI guidelines, it is stated that the standard is based on an "abstract model" and that each element has a specific "semantic function" which encoders ought to

[461] Kate Singer, "Digital Close Reading: TEI for Teaching Poetic Vocabularies", in: *Journal of Interactive Technology and Pedagogy*, 3, n.pag.

respect.[462] In a formal and strict ontology, relations between entities have been defined explicitly.[463] A limitation of the TEI standard, and of XML-based languages for the addition of in-line mark up in general, is that the semantic relations that can exist between elements cannot be stated explicitly. DTDs and XML Schemas can specify that one element ought to be contained or nested within another element, but no information can be provided on the precise meaning of this nesting. Renear et al. observe that all XML-based markup languages share a degree of ambiguity. There are no facilities for the description of "the fundamental semantic relationships amongst document components and features in a systematic machine-processable way".[464] The intended meaning of a specific element cannot be inferred from the way it is used in the document, and knowledge about the intention of these elements must be built into the software applications that query these texts. Willet notes that "SGML/XML markup itself is not a "data model", as it exclusively "serializes a data structure".[465] The technique may be viewed as controlled vocabulary, rather than as a formal ontology.

Secondly, the TEI ontology may also be viewed flexible or loose because of its extensibility and its leniency. Encoders are empowered to modify existing elements and attributes and to add new terms according to their own needs, thereby undermining the concept of a central ontology. In an important sense, the TEI enables scholars to give expression to a phenomenology, as elements frequently reflect both a scholar's idiosyncratic understanding of a descriptive category, and the subjective views on the textual aspect to which the category is applied. Research projects which have identified lacunae in the guidelines or in the technical possibilities of the TEI standard are often forced to develop project-specific elements and attributes. If the needs for extensions to the standard are sufficiently widespread, SIGs can be established to institutionalise ontological views of particular sets of phenomena. In spite of the fact that the TEI endorses specific ontological commitments, there is generally no widespread consensus on how and under which circumstances specific elements should be applied. Since the aim from the onset had been to accommodate hugely diverse research areas, the standard provides "a maximum of comprehensibility, flexibility, and extensibility".[466]

---

[462] "TEI Guidelines for Electronic Text Encoding and Interchange", <http://www.tei-c.org/Guidelines/P5/> (12 March 2013)

[463] Lee Lacy, *OWL: Representing Information Using the Web Ontology Language*, p. 36.

[464] Allen Renear, David Dubin & C. M. Sperberg-McQueen, "Towards a semantics for XML markup", in: *Proceedings of the 2002 ACM symposium on Document engineering - DocEng '02*, (New York, New York, USA: ACM Press, 2002), p. 119.

[465] Allen H. Renear, "Text Encoding", in: *A Blackwell Companion to Digital Humanities*, Oxford: Blackwell 2002, p. 237.

[466] TEI Guidelines for Electronic Text Encoding and Interchange, <http://www.tei-c.org/Guidelines/P5/> (12 March 2013)

The idiosyncrasy of encoding practices, and the extensibility of the vocabulary can obviously jeopardise the reuse of encoded texts beyond the project in which they were originally created. This aspect is not necessarily problematic when cross-project interoperability is not an objective, and when scholars principally aim to capture the annotations about texts in a systematic manner for the purpose of their own research. In some cases, however, the data format itself may also hinder an effective representation of research annotations. One particular problem inherent in XML-based encoding systems is that the format may lead to conflicting hierarchies. The W3C recommendations stipulate that XML documents must follow a strict hierarchical structure. Documents, more specifically, ought to have exactly one root element, and all other elements need to be contained hierarchically within this single root.[467] Various theorists of textuality correspondingly surmised that texts are typically composed of distinct constituent parts, which are related to each other in a hierarchical manner. This theory views texts as "Ordered Hierarchies of Content Objects" (OHCO).[468] A poem, for instance, may be composed of stanzas, and these stanzas, in turn, may consist of lines. Various authors have also argued that the OHCO theory is flawed, since there are also numerous examples of textual phenomena which do not nest neatly. Collisions of hierarchies occur, for instance, in verse texts which contain enjambments. The verse lines may be encoded using the <l> element, but, when scholars also want to encode grammatical sentences using <seg>, this evidently results in invalid XML. Phenomena that appear in poetry, such as the tropology, the syntax, the metre and the literary devices based on sound frequently overlap, too. A metaphor, for instance, may start at one line, and end on the line that follows.[469]

Renear et al. have argued that the OHCO theory may be redeemed if it is accepted that a text may be seen as "system of concurrent perspectives which decompose into concurrent sub-perspectives which in turn can be decomposed".[470] Physical, prosodic or syntactic analyses of a text all produce distinct hierarchies, but, since content objects within a single hierarchy always seem to nest properly, an overlap of content objects should be taken as an indication of the fact that these two objects belong to different analytic perspectives. For the purpose of the Myopia viewer, for instance, the scholars produced four separate files, and each of these focused on a separate hierarchy. It can easily be envisaged that duplicating a source text multiple times introduces difficulties when this source text, for whichever

[467] W3C, "Extensible Markup Language (XML) 1.0 (Fifth Edition)", 2008, <http://www.w3.org/TR/2008/REC-xml-20081126/> (16 July 2014)

[468] Allen Renear, David Durand & Elli Mylonas, "Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies", in: *Research in Humanities Computing*, Oxford: Oxford University Press 1995.

[469] Manish Chaturvedi et al., "Myopia: A Visualization Tool in Support of Close Reading".

[470] Allen Renear, David Durand & Elli Mylonas, "Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies", n.pag.

reason, needs to be updated. An alternative approach is to select a single hierarchy as the primary hierarchy within a file, and to flatten other hierarchies, by using milestone tags, or elements without a body. This approach, however, can only be effective when the data to be recorded consists solely of the identification of a particular location within the text. This may be the case for a page break or a line break. When specific words actually need to be delineated and categorised, as in the case of assonance or of metaphors, the problem of conflicting hierarchies persists.

In his article "Digital Representation and the Text Model", Dino Buzzetti draws attention to an additional obstacle inherent in descriptive mark up. Buzzetti distinguishes between the expression and the context of the text. The former term is "the linear order of the succession of codified characters" and the latter term is used to refer to "that which the various strings of characters signify".[471] Cesar Segre clarifies that literary theorists have variously used the term 'content' to refer to "themes, composition, and genres" or to "ideas, feelings and inspirations".[472] These concepts or sentiments collectively form a "semiotic product"[473] which critics may extract from the concrete expression by evaluating, for instance, the connotative effects of words. The structure of the content, however, does not necessarily coincide with the structure of the expression. In many cases, the central ideas of a text cannot be connected unequivocally to particular sequences of characters. Data on the meaning of an extended metaphor, the general setting of a text, the central thematic concerns, or the intertextual connections to other texts cannot always be captured using a standard which focuses exclusively on the linear string of characters. MacNeice's poem "Charon", for instance, is basically an extended metaphor in which a bus trip across London is portrayed as a voyage across the river Styx. The hands of the bus driver are "black with money", his "eyes are dead" and the passengers also note his "varicose veins". Various words in the poem collaborate to produce a nightmarish effect. Brown argues that the repetition of the phrase "we just jogged on" conveys "the irreversible process of life towards death".[474] This central subtext of a death in life, however, cannot be associated exclusively to a single sequence of words within the text. Since the extended metaphor cannot be connected to a single vehicle, it cannot be encoded effectively using TEI-based mark up.

Next to the difficulties that result from multiple hierarchies, and from the fact that occurrences of phenomena cannot always be situated in a single location within the text, a third difficulty that inheres in TEI-based encoding is that descrip-

---

[471] Dino Buzzetti, "Digital Representation and the Text Model", in: *New Literary History*, 33:1 (2002), p. 68.

[472] Cesare Segre, *Introduction to the Analysis of the Literary Text* (Bloomington: Indiana University Press 1988), p. 41.

[473] Dino Buzzetti, "Digital Representation and the Text Model", p. 88.

[474] Terence Brown, *Louis MacNeice: Sceptical Vision*, p. 68.

tive terms can generally be applied only once. As was discussed, the @ana attribute may be used to record, for instance, a term indicating the connotation of a word. Once a word or a sequence of words has been marked up, it cannot be encoded at the same hierarchical level with another element. Within a single file, any decision to apply a particular element effectively suppresses all opposing views. Many of the statements that are made via the TEI are of a subjective nature, however. Huitfeldt argues that "[t]here are no facts about a text which are objective in the sense of not being interpretational". A scholar encodes the text "in accordance with his or her interpretation", and "the transcriber's interpretation is not theory-independent".[475] Eggert writes likewise that "every electronic representation of text is an interpretation". Encoding is "doomed to remain problematic, incomplete and perspectival".[476] Aspects pertaining to conflicting hierarchies cannot be recorded within a single TEI file, and the same is true for the registration of disagreement. In XML documents, it is generally arduous to express multiple equivalent opinions about a single fragment. A particular encoding monopolises a single opinion, and scholars who wish to record alternative perspectives need to create a new manifestation of the text. In Chapter 3, plain texts and annotations about texts were described as different types of data, but, in a TEI file the two typed of data are intermingled in a single file. Within this format, a new set of annotations can only be created by duplicating the full primary data, implying an inefficient use of primary data.

In Chapter 3, a distinction was also discussed between annotations which describe aspects which are explicit and observable, and annotations which describe implicit textual aspects. Data in the latter category are often speculative and interpretative. Meister stresses that "interpretation is an interpretation if and only if at least one alternative to it exists".[477] When tags are inserted that characterise the use of metaphors, or that locate instances of other literary devices such as onomatopoeia or synaesthesia, the files generally reflect the interests and the interpretations of one particular scholar. Describing the semantic contents of the text is crucially an open-ended process, as scholars may interpret and re-interpret their sources virtually limitlessly. For this reason, it seems inadequate to incorporate such observations directly within the primary text. Annotations about the logical structure or the typography, conversely, possess a degree of objectivity, as they can be verified via a consultation of the original sources.[478] As observations

[475] Claus Huitfeldt, "Multi-Dimensional Texts in a One-Dimensional Medium", in: *Computers and the Humanities*, 28 (1994), pp. 237–239.

[476] Paul Eggert, "The Book, the E-Text and the "Work-Site"", in: Marilyn Deegan & Kathryn Sutherland (eds.), *Text Editing, Print and the Digital World*, Farnham: Ashgate 2009, pp. 67–73.

[477] Jan Christoph Meister et al., "Crowdsourcing meaning: a hands-on introduction to CLÉA, the Collaborative Literature Éxploration and Annotation Environment", in: *Digital Humanities 2012*, (Hamburg: 2012), n.pag.

[478] The observable typographical features of a text may be described though encoding, but their meaning may still be open to interpretation.

about the logical structure of the text may be assumed to be relatively factual and stable, proximate inclusion into the plain text seems less problematic.

The TEI can reasonably be used for descriptive annotations about the logical components of the expression, but the standard seems less suitable for the consolidation of conjectures about the text's content, as these are often contentious and idiosyncratic. Subjective and interpretative observations can arguably be captured more appropriately in a separate document and independently from the document that contains the full text. To some extent, such a division was implemented in the Just in Time Markup (JITM) project, which was developed at the Australian Scholarly Editions Centre at the Australian Defence Force Academy. In this project, the type of separation that was effectuated however, was more rigorous. Berrie explains that, in the JITM system, the bare transcriptions are stored separately from all scholarly annotations of these texts. At the request of a user, a specific set of mark up tags can be added to a transcription, thus allowing the creation of "on demand user-customised versions of electronic editions".[479] Since the interpretative mark up is not inserted into the actual transcription file, this latter resource remains authentic. It also ensures that transcriptions can be reused. Marked-up files generally privilege a particular perspective on the text, but if the mark-up is recorded separately from the text, it also becomes possible to record opposing views on the text. The transcription file, once completed, is assumed to remain static, and a reference scheme is used to insert mark up into the text. As a result, "the electronic edition can become an evolving work of scholarship based on the work of many hands".[480] In JITM, transcription files are tokenised by making use of the spaces that appear in the document. The spaces are assumed to delineate words, and these words in turn consist of characters. The file which records the mark up works with position codes such as "17.001" or "28.006". In this particular case, these codes refer to the first character of the 17th word or the sixth character of the 28th word. Mark up codes can be inserted dynamically at the positions which are recorded in this manner. This system of capturing positions appears to be fragile, as positions may evidently change when the transcription file changes at some point.

A related form of stand-off mark up was implemented in the CATMA application, which was developed at the University of Hamburg. CATMA is described as "a practical and intuitive tool for literary scholars, students and other parties with an interest in text analysis and literary research".[481] Scholars can initially supply a collection of TEI-encoded texts. Within the tool's interface, scholars can select specific fragments and describe these using their own tags. These tags are stored in a separate *User Mark up* document. This document contains references to the

[479] Phillip William Berrie, *Just In Time Markup for Electronic Editions*, n.pag.
[480] Ibid.
[481] <http://www.catma.de/> (12 April 2014)

fragments these tags are applied to. As the descriptive tags and the tags that explain the logical structure of the text are separated, there is no risk of overlapping or contradictory mark up. Several User Mark up files can be created for a single source document. Tags can also be structured hierarchically, as users can create tags and "subtags". When a user searches for a "supertag", the subtags are retrieved as well. One difficulty is that the tags are tied to the CATMA application, and that these cannot be used outside of the environment. It is also difficult to ensure cross-project interoperability. The functionalities of the CATMA tool are further expanded within *Collaborative Literature Éxploration and Annotation*,[482] which aims to "supplement Google Books with a web based collaborative text exploration, markup and analysis environment".[483]

## 7.3. Semantic web

The Worldwide web has been described as "the most powerful communication medium the world has ever known".[484] It is used by millions of people on a daily basis, it determines much of our social and our cultural lives, and it is also one of the largest drivers of our global economy. Tim Berners-Lee, who is commonly viewed as the progenitor of the Web, has explained that the current ubiquity of the Web was largely the result of its flexibility and of its decentralised nature. Berners-Lee recognised that the web would never have developed into the popular global phenomenon that it is today if it had imposed strict and rigid rules on how to structure web pages and on how to create hyperlinks. To ensure a widespread adoption, "the Web had to throw away the ideal of total consistency", and, on the Web, "anything can link to anything".[485] The Web generally lacks a formal and consistent structure, and this means that its contents cannot be searched in the same way a well-designed database can. The effectiveness of searches is often hampered by the fact that the content consists of texts which were designed for human readers. Web pages are typically created using HTML, which is a standard mainly for the specification of the typographic appearance of web pages. This focus on presentational aspects jeopardises the accessibility of the data that is contained in these documents, as search engines can usually carry out full text searches only.

[482] <http://www.catma.de/webfm_send/22> (12 April 2014)

[483] Jan Christoph Meister et al., "Crowdsourcing Meaning: A Hands-on Introduction to CLÉA, the Collaborative Literature Éxploration and Annotation Environment".

[484] World Wide Web Foundation, "History of the Web", <http://webfoundation.org/about/vision/history-of-the-web/> (14 May 2013)

[485] Tim Berners-Lee, James Hendler & Ora Lasilla, "The Semantic Web: A New Form of Web Content That Is Meaningful to Computers Will Unleash a Revolution of New Possibilities", in: *Scientific American*, (2001), pp. 73–74.

Since the current web predominantly consists of "multimedia human-readable material", it basically functions as a "glorified television channel".[486]

The Semantic Web consists of a collection of techniques which aim to ensure that the content on the Web can be processed more systematically by machines. Berners-Lee (2002) explains that the Semantic Web was already part of the original vision for the web as it was conceived in the late 1980s. The techniques that were envisioned aimed to "bring structure to the meaningful content of Web pages".[487] The objective of the semantic web was not to replace the current web, but to extend it with an additional layer through which information can be given a "well-defined meaning, better enabling computers and people to work in co-operation". The aim of the Semantic Web is to publish those structured data collections directly on the Web, thus precluding the need to translate these data into human-readable HTML files first. Using semantic web techniques, machines can exploit data which is currently "in relational databases, XML documents, spreadsheets, and proprietary format data files".[488] The Semantic Web makes use of a number of central components. Concepts, importantly, need to be referred to using identifiers, rather than via words. The Semantic Web makes use of the general web architecture, which, according to W3C, is "an information space in which the items of interest, referred to as resources, are identified by global identifiers called Uniform Resource Identifiers (URI)". A second central component of the Semantic Web is the *Resource Description Framework* (RDF).[489] URIs and RDF can be used in combination to describe concrete and abstract concepts, together with the manifold relationships that can exist between these concepts.

RDF offers a framework which can be used to make claims about a particular domain, and, as such, it can serve as an alternative to the TEI. To create structured data about text fragments, it is necessary, as a first step, to unambiguously delineate and identify the fragments which need to be annotated. To be able to include these fragments into an RDF assertion, they additionally need to be defined as URIs. Joel Kalvesmaki explains that, throughout the history of textual scholarship, systems of canonical references have been used to label specific sections of works. These references enable scholars to cite particular parts of a text, without having to indicate a particular edition.[490] Large works such as *The Bible* or Homer's *Illiad*, for instance, have been segmented into units which can be referred to separately using codes such as *1 John 4:19* or *Homer, Iliad 1.1*. This system of

[486] Tim Berners-Lee, "Foreword", in: Dieter Fensel (ed.), *Spinning the Semantic Web*, Cambridge: MIT Press 2005, pp. xi–xiv.

[487] Tim Berners-Lee, James Hendler & Ora Lasilla, "The Semantic Web: A New Form of Web Content that is Meaningful to Computers will Unleash a Revolution of New Possibilities", n.pag.

[488] Ibid.

[489] RDF has been discussed earlier in Chapter 3.

[490] Joel Kalvesmaki, "Canonical References in Electronic Texts: Rationale and Best Practices", in: *Digital Humanities Quarterly*, 008:2 (2014), 1.

canonical references can usefully be extended into the digital realm. Kalvesmaki stresses, however, that there are currently no widely accepted interoperable protocols for defining and resolving canonical references. In texts that are encoded using the TEI, individual units can be associated, for instance, with a canonical reference using the @cRef attribute. These references can subsequently be incorporated in a URI. To ensure that references can function correctly, the server which hosts the TEI file that contains these fragments needs to be able to return the associated fragment. An alternative is to make use of the *Canonical Text Services* (CTS) protocol. The protocol defines a URN scheme in which the work and the fragment can both be identified. The identifier for the passage is generally derived from an identifier that is specified in a TEI document. To ensure the uniqueness of the URNS, a CTS URN also demands a declaration of a CTS URN namespace.

Since, within the current study, no authoritative CTS URN namespace had been registered yet, a decision was taken to base the referencing system on the URLs of the online files. In the case study, all poems have been encoded in TEI. The TEI encoding focuses mostly on the logical structure of the texts, and the different stanzas, lines and words have all been given unique identifiers. A URI may be created by combining the URL of the poem with a unique fragment identifier, derived from the values of the @n attributes within the TEI files. The second line of the poem "Belfast", for instance, may be referenced using the URI `<http://www.bookandbyte.org/macneice/belfast.xml#l2>`. The same principle can be used to refer to individual words, as these have been numbered similarly.

Ranges of words may be referenced by making use of the XPointer standard. XPointer offers a provision for the identification of longer fragments of texts, in which the identifiers of the opening point and the closing point need to be supplied.[491] This technique can be applied usefully for the description of instances of enjambment, for instance. One example of this device can be found in the second canto of *Autumn Journal*: "But tonight is quintessential dark forbidding / Anyone beside or below me". To refer to the full sentence, the following URI may be constructed: `<http://www.bookandbyte.org/macneice/AutumnJournalII.xml#xpointer(id("w123")/range-to(id("w145")))>`. In this example, "w123" is the value of the @n attribute which was assigned to the first word in the grammatical sentence. In this study, it was assumed that the individual word constitutes the smallest unit that needs to be addressed. If, in other studies, it is necessary to comment on individual characters, an alternative solution may be chosen, in which the TEI encoding also assigns identifiers to characters.

The descriptive terms that are used in a Semantic Web application are generally derived from a series of ontologies. A broad variety of ontologies have

[491] <http://docstore.mik.ua/orelly/xml/xmlnut/ch11_07.htm> (21 June 2014)

been made available already, but new terms can be created using the Web Ontology Language (OWL). OWL, more specifically, provides a mechanism for describing particular classes of objects. The technique can be used to mint identifiers for such classes, and to record some of their formal characteristics. The declaration of a class can also include a textual definition of the term, which may enable human readers to evaluate its suitability.[492] When the OWL-based ontology is made publicly accessible, the classes that have been declared can be used within RDF assertions.

Within the case study that was conducted in this thesis, data have been produced about a wide range of literary devices such as alliteration, assonance, onomatopoeia, rhyme and metre. Since it was found that the majority of the phenomena investigated have not been described yet in any existing ontology,[493] a new ontology of literary terms was developed, using OWL. The classes "Poem", "Stanza", "Line" were created to characterise the structural components of poetic texts. The classes "Couplet", "Tercet", "Quatrain", "Quintain" and "Sestet" were established to enable scholars to distinguish between stanzas of different lengths. Importantly, classes were also defined for the description of the literary devices that were investigated in the case study.[494] Terms were defined more specifically, for "perfectRhyme, "AssonanceRhyme", "consonanceRhyme", "SemiRhyme", "Alliteration", "Assonance", "Consonance", "InternalRhyme", "DeibhideRhyme", "AicillRhyme" and "InternalConsonanceRhyme". It was also specified that assonance rhyme and consonance rhyme are both specific cases of slant rhyme. In the human-readable descriptions of these terms, definitions were cited from the *Princeton Encyclopedia of Poetry and Poetics*, and from Abram's *Glossary of Literary Terms*. The generic class "Rhyme" was declared, additionally, to be able to cluster the various more specific forms of rhyme. The number of terms and the types of relations between these terms may be expanded at a later stage. It would be possible, for instance, to define a hierarchical structure in which figures of speech are distinguished from figures of thought. For the purpose of the current case study, however, a declaration of the basic terms was considered sufficient.[495]

Next to having a suitable ontology, a method needs to be chosen to connect terms from this ontology to selected text fragments. Such structured annotations can be created using a number of technologies. This section will focus on the

[492] Lee Lacy, *OWL: Representing Information Using the Web Ontology Language*, p. 173.

[493] Searches for these terms in the linked open data search engine *Falcons*, available at <http://ws.nju.edu.cn/falcons/objectsearch/index.jsp >, did not produce any results.

[494] The ontology that was created followed the classification that was discussed in Chapter 2 of this thesis.

[495] Information about the full ontology that was implemented can be found in Appendix B.

nanopublications protocol and on the *Open Annotation Collaboration* (OAC).[496] Nanopublications, firstly, were originally conceived of by members of the Concept Web Alliance.[497] The technique entails a mechanism for making individual academic findings available separately, in a machine-readable format. A nanopublication is viewed as the "smallest unit of publishable information".[498] The technique may be used to publish "quantitative and qualitative data, as well as hypotheses, claims, and negative results that usually go unpublished".[499] Since each nanopublication is made available under a single URI, they can also be cited, for instance, in textual publications. The concept of nanopublications was largely developed in response to a perceived information overload. Textual scholarly publications typically contain a multitude of statements expressed in natural language, and the sheer quantities in which such textual publications are made available at present clearly complicate staying abreast. Findings which are made available as nanopublications may be processed efficiently by machines, and Mons and Velterop expect that computer applications can produce "real time alerts"[500] for the benefit of researchers interested in specific concepts.

A nanopublication, more concretely, can consist of three components. The central component is the assertion. It is an RDF triple which represents the main finding that is made available. Since it was recognised that findings frequently need to be understood within a specific context, a second component was defined in which researchers can describe the conditions under which the assertion is considered to be true or relevant. Thirdly, and importantly, the provenance of the assertion can also be captured. In this third section, data can be supplied about the creator of the assertion, i.e. the person who is responsible for its intellectual contents, and about the time and date on which the assertion was made. Data about the provenance should enable peers to evaluate the trustworthiness of the statement. One notable characteristic of the provenance section is, additionally, that researchers may record the type of evidence. Velterop and Mons discuss a number of evidence types. The assertion that is captured in a nanopublication may be "derived from observation or measurement". Alternatively, it may be "derived as

[496] Sanderson et al. note that there have been a number of other technologies for the registration of structured annotations, including Annotea, Hypothesis and iAnnotate. Many of these technologies have shortcoming or are no longer continued. The two technologies which are discussed in this section are considered to be the most appriate solutions in the academic domain. See Robert Sanderson, Robert Sanderson, Bernhard Haslhofer, Rainer Simon, et al., "The Open Annotation Collaboration (OAC) Model", <https://arxiv.org/pdf/1106.5178.pdf>

[497] The CWA is "an open collaborative community that is actively addressing the challenges associated with the production, management, interoperability and analysis of unprecedented volumes of data"

[498] Erik Schultes & Mark Thompson, "Using Nanopublications to Incentivize the Semantic Exposure of Life Science Information", p. 1.

[499] <http://nanopub.org/guidelines/working_draft/> (2 November 2014)

[500] Barend Mons & Jan Velterop, "Nano-Publication in the e-science era", (2009), n.pag.

a prediction based on a model or theory".[501] In this current case study, recording the evidence type is highly relevant. Using this provision, a distinction can be made between data which were generated algorithmically on the one hand, and data which have been created or edited manually on the other.

The code below gives an impression of how the nanopublication framework can be applied to express the observation that the second line of Louis MacNeice's poem "Belfast" contains alliteration.[502]

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix nanopub: <http://www.nanopub.org/nschema#>.
@prefix nphbvar: <http://www.nanopub.org/nanopubs/hbvar#>.
@prefix dcterms: <http://purl.org/dc/terms/>.

{
   nphbvar:n0 nanopub:hasAssertion nphbvar:n0assertion.
   nphbvar:n0 nanopub:hasProvenance nphbvar:n0provenance.
   nphbvar:n0 rdf:type nanopub:Nanopublication.
}

nphbvar:n0assertion {
       http://www.bookandbyte.org/macneice/#l2 lit:contains
       lit:Alliteration.
}

nphbvar:n0provenance {
       nphbvar:n0assertion
       dcterms:created "2014-12-12"^^xsd:date.
       nphbvar:n0assertion nanopub:authorID info:eu-
       repo/dai/nl/32959737X.
}
```

As can be seen, this example does not make use of the context component. In general, this component can be used to qualify a statement, and to explain that the claim applies exclusively under specific conditions. This provision, however, does not seem applicable to largely subjective statements about literary texts. The provenance section, however, is highly germane, as it offers a solution to one of the crucial limitations of the TEI format. If there is disagreement about the observation

[501] Barend Mons & Jan Velterop, "Nano-Publication in the e-science era", n.pag.
[502] The RDF graph is represented via the Turtle notation.

that is made, contrastive claims can be published as separate nanopublications. Each alternative claim can be timestamped and attributed individually.

The nanopublication framework also presents a number of difficulties. As was discussed, the method that was developed within the case study for the recognition of alliteration does not only signal the occurrence of this device. It also produces a pattern representing the sounds that are repeated. Crucially, since a nanopublication can only be associated with one assertion,[503] it does not seem possible to record this pattern, in addition to the fact that the passage contains alliteration in itself. In some cases, it also seems difficult to describe literary phenomena exclusively via a tripartite structure. Lines 47 and 48 of the eleventh canto of "Autumn Journal" contain an instance of aicill rhyme: "Who know that truth is nothing in abstraction / That action makes both wish and principle come true". Since the individual words were marked up using the <w> element, and since each <w> element has been assigned an @n attribute, words can be identified separately. The subject of this assertion, nevertheless, consists of two separate parts. The words "abstraction" and "action" collectively produce the internal rhyme, but the nanopublications framework does not allow for assertions which contain multiple subjects.

A number of these difficulties can be addressed by making use of the OAC. It is a protocol that can be used to capture structured annotations.[504] The framework enables annotators to express a "relationship between two or more resources, and their metadata". The OAC proposes a framework in which system-independent annotations can be published in accordance with linked open data standards. The OAC data model propose three central classes. The "Annotation" itself consists of a "Target", which represents the object that is annotated. The "Body" is a comment or any other resource which offers information about this target. OAC is primarily a method for the management of annotations, and its data model does not propose any terms beyond those terms that are needed to construct the annotation itself, and to describe its provenance.

The descriptive terms to be used in the annotation are not stipulated in the model itself, however.[505] In the OAC guidelines, it is explained that scholars can classify particular text fragments using terms derived from an external ontology. This is viewed, more specifically, as an example of semantic tagging. In OAC, the ontology terms can be supplied directly within the body of the annotation, and this tag must be associated with the class oa:SemanticTag. The OAC guidelines also encourage annotators to explain the reasons for creating an annotation, using the property a oa:motivatedBy. Within the context of this current study, the term

[503] <http://nanopub.org/guidelines/working_draft/> (18 September 2013)
[504] <http://www.openannotation.org/spec/core/> (18 September 2013)
[505] Dirk Roorda & Charles van den Heuvel, "Annotation as a New Paradigm in Research Archiving", pp. 4–5.

oa:classifying seem most appropriate. It entails "the assignment of a classification type, typically from a controlled vocabulary, to the target resource(s)". As is the case for the nanopublications data model, the OAC provides a mechanism for the description of the provenance of annotations. Such statements can be "useful for determining the trustworthiness of the Annotation, potentially based on reputation models". Both the person and the time at which an annotation was created can be recorded. Provenance information can be attached to the Annotation, to the Body, and to the Target.

There are a number of important differences between the nanopublications framework and the OAC. One important difference is that, with the latter framework, it is possible to further modify the body of the annotation. The example below illustrates the manner in which the rhyme scheme of a stanza may be captured via OAC.[506] As can be seen, the Body of the annotation consists of the semantic tag lit:PerfectRhyme. This tag is classified as an oa:SemanticTag, and, additionally, it is associated with a pattern which represents the rhyme scheme. Next to the classes from the OAC and the ontology of literary terms, the code below also makes use of terms from FOAF, which is an ontology which can be used to describe properties of persons and of the relations between persons.[507] The person who takes responsibility for the annotation is identified using the oa:annotatedBy property.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix lit: <http://www.bookandbyte.org/olt/0.1/#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix oa: <http://www.w3.org/ns/oa#> .

<http://www.bookandbyte.org/macneice/annotations/anno1>
   a oa:Annotation ;
   oa:annotatedBy <info:eu-repo/dai/nl/32959737X> ;
   oa:hasTarget <AContact.xml#s2> ;
   oa:hasBody lit:PerfectRhyme ;
   oa:motivatedBy oa:classifying .

<lit:PerfectRhyme>
   a oa:SemanticTag, skos:Concept .
   lit:hasPattern "1 - 1" .
```

[506] Advice on how to express the data that were produced in this research project via OAC has kindly been given by Rob Sanderson and by Tim Cole.

[507] <http://www.foaf-project.org/> (26 April 2014)

```
oa:classifying
   a oa:Motivation .

<info:eu-repo/dai/nl/32959737X>
   foaf:name "Peter Verhaar" .
```

A second important difference between OAC and nanopublications is that, within OAC, it is also possible to create annotations which consist of multiple Targets. An occurrence of internal rhyme may be described using the following structure.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix lit: <http://www.bookandbyte.org/olt/0.1/#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix oa: <http://www.w3.org/ns/oa#> .

<http://www.bookandbyte.org/macneice/annotations/anno1>
   a oa:Annotation ;
   oa:annotatedBy <info:eu-repo/dai/nl/32959737X> ;
   oa:hasTarget <AutumnJournal-II.xml#w47-4> ;
   oa:hasTarget <AutumnJournal-II.xml#w48-2> ;
   oa:hasBody lit:AichillRhyme ;
   oa:motivatedBy oa:classifying .
```

One important consequence of the fact that OAC closely follows the central principles of the semantic web is that it can ultimately become possible to share annotations across applications and across technical platforms. Van den Heuvel and Roorda also identify a number of important challenges. At present, there is no widespread consensus concerning the manner in which text fragments may be targeted. Projects have often implemented idiosyncratic solutions, and this potentially complicates an exchange and a reuse of annotations. Secondly, since the OAC model does not prescribe a terminology that can be used within the body of an annotation, assertions created within different projects may also contain distinct vocabularies, undermining their interoperability.[508] The use of semantic web technologies in general equally implies a range of difficulties. Veltman notes that the semantic web "deals with meaning in a very restricted sense and offers static solutions".[509] The technologies may be relevant for relatively straightforward fact-

[508] Dirk Roorda & Charles van den Heuvel, "Annotation as a New Paradigm in Research Archiving", p. 4.
[509] Kim H. Veltman, "Towards a Semantic Web for Culture", in: *Journal of digital information*, 4:4 (2006), p. 2.

finding purposes in scientific applications, but the humanities generally deal with less formalised information. Within humanities research, it is essential to take historical and geographical dimensions into account, as the meaning of terms may change both synchronically and diachronically. Meroño-Peñuela explains similarly that the sources which are maintained by cultural heritage institutions are often "messy and heterogeneous", and that such complexity can strongly complicate longitudinal queries. Humanistic ontologies, for this reason, demand "dynamic concept formalizations instead of static ones, especially for contested, open-textured or ambiguous concepts".[510] In many cases, it is necessary to capture data about the historical and the social contexts in which terms are used, and to allow for multiple, potentially contradictory conceptualisations of terms.

Since annotations based on semantic web technology often make use of pre-defined ontologies, scholars who use such languages need to describe the literary phenomena that occur in a text via fixed categories. Semantic enhancement generally implies a form of simplification, as the limited list of terms which were anticipated in an ontology does not necessarily match the diversity of the actual phenomena which are encountered in the literary text. A central difficulty that inheres almost inevitably in all classification systems is that there is a conflict between the demands that are posed by the need to process texts systematically on the one hand and the need to describe texts flexibly and responsively on the other. When a particular collection of texts illuminates the inadequacy of an existing vocabulary, scholars may respond to this by creating a new ontology, but the use of idiosyncratic terms may jeopardise the ability to query text collections consistently.

Andy Clark's thesis of the extended mind states that the technologies that we use to capture and to process information can take over crucial functions of the human brain. The semantic web may be viewed as one of the technologies which scholars can use to extend their minds and to stimulate and to enhance their thinking. If it is accepted that particular technologies also result in particular types of thinking, it may be expected that the processing based on semantic web technologies inhibits the development of radically new ideas. It places phenomena in categories which have been envisaged beforehand, and plaintively disregards the phenomena for which such categories do not apply. These difficulties, which can arise from the rigidity of ontologies, have partly been addressed within the Pliny tool, which was developed by Bradley and Prusin. Pliny can be used "in the pre-ontological context".[511] The tool provides support for the creation of annotations at the moment when they are still largely unstructured. Pliny firstly lets its users record any ideas that occur during the reading of the text that is being studied, in a

[510] Albert Meroño-Peñuela, "Semantic web for the humanities", in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, (Springer, 2013), p. 646.

[511] J. Bradley, "Thinking about Interpretation: Pliny and Scholarship in the Humanities", in: *Literary and Linguistic Computing*, 23:3 (5 September 2008), p. 271.

format that is not governed by an existing ontology. Once captured, users can rearrange and reshuffle the notes in order to "discover previously unrecognized patterns and relationships and to stimulate new ideas",[512] thus inspiring new interpretations. Pliny aims to provide a solution to the difficulty that generic preconceived ontologies rarely do justice to the particularities of an individual work of literature. Personalised annotations can be valuable for the interpretation of separate texts, but quantitative research performed on large corpora of texts typically demands that all the phenomena which are similar or comparable can also be referred to consistently using a stable set of descriptive terms. The conflict between generic descriptive terms and terms which are geared towards the particularities of a work of literature continues to pose an important conundrum.

Despite the challenges that can potentially complicate an exchange and a reuse of the data, The OAC data model appears to meet most of the requirements that were discussed in the introduction of this chapter. It can be used in combination with an ontology which supplies descriptive terms for the phenomena which have been discussed in Chapter 6 of this thesis. The data format in itself does not pose any technical difficulties. The targets of the annotations were formed by fragment URIs, and these are based on identifiers which are declared in the TEI files. The locations that are delineated by these identifiers can overlap, and the problem of conflicting hierarchies does not pose itself. Importantly, it is possible to express disagreement, as statements can be attributed exclusively to a single scholar. In cases of controversy, scholars who disagree may create separate annotations. A decision has been taken, for these reason, to record all the secondary data of the case study via the OAC.

## 7.4. Conclusion

This chapter has compared two types of data formats on the basis of three criteria. Formats were compared by considering the expressiveness of the ontology they are based on. Secondly, the analysis focused on the technical constraints that were connected to the syntax of the format. Thirdly, the possibilities for expressing disagreement were also taken into account.

With respect to the TEI, it was found that the default ontology it is based on[513] offers limited possibilities for the characterisation of aspects of poetic language. The standard, by default, does offer extensive support for the description of the logical structure of texts, for the identification of place names, geographic terms, bibliographic titles and for the description of the various witnesses of historical

[512] J. Bradley, "Thinking about Interpretation: Pliny and Scholarship in the Humanities", p. 266.

[513] In this context, the phrase "basic ontology" essentially refers to the collection of elements and the relationships between these elements that have been defined in the canonical tei_all schema. This core set of elements has been extended within specific research projects, but it is understood that such customisations do not belong to the standard's basic ontology.

texts. The standard is consequently very suitable in the context of textual criticism. It is less effective in studies which aim to explore literary phenomena across different texts on the basis of a quantitative method. The canonical TEI schema can obviously be modified, and those terms which were found missing can be added in a customised schema. Although such modifications may solve the problem of limited expressivity, the technical format of inline XML also poses a number of complications. Annotations generally target different textual fragments, and some of these targets clash. The strict hierarchical structure of XML prohibits a flexible delineation of such overlapping text fragments. Additionally, it is mostly impossible to record contrastive descriptions of text fragments within a single TEI document.

It was also shown that, when structural annotations are recorded using Semantic Web technologies, some of these complications can be avoided. Statements in RDF can derive their descriptive terms from an ontology, and when the terms which are necessary are not supplied by an existing ontology, a new conceptualisation of a domain may be supplied by making use of OWL. In this study, a new ontology was proposed for the description of literary phenomena. In a sense, an OWL-based ontology does not differ in a qualitative sense from a customised TEI schema. Both types of ontologies can, after they have been developed and tested within individual research projects, be made publicly available, and, after their publication, they may or may not attract a broader community of users. In the case of annotations stored via the OAC, however, most of the practical challenges associated with the TEI can be avoided. The targets of annotations are essentially references to specific locations within texts, and, if necessary, these locations may also overlap. The problem of conflicting hierarchies, consequently, does not present itself. Secondly, a single fragment can be described in many different annotations, and these annotations can all be associated with different scholars. The OAC can consequently be used to express polyvocality in literary criticism.

It may be observed that, within digital humanities studies, there is often a bifurcated position towards ontologies. In computer-based research, scholars typically use the descriptive terms supplied by a standard to describe their individual response to a literary text. The descriptions that are created are often time and location specific. They are often coloured heavily by the tenets of specific schools of criticism. At the same time, the ontologies which supply descriptive terms are typically perspectival as well. They are invariably developed for a specific purpose and within a critical tradition. As the form of usage which the standard accommodates does not necessarily match the needs of individual scholars, there is frequently an urge to manipulate the central ontology. The TEI standard can clearly be made responsive to idiosyncratic scholarly propensities. In large scale quantitative analyses of texts, however, it must be ensured that all instances of the phenomena which are studied are consistently identified as such. Quantitative research crucially requires the consistent application of a fixed ontology. In digital

humanities research, it is often pivotal to strike a delicate balance between rigidity and extensibility.

An important characteristic of OWL is that it enables scholars to define relations between different ontologies. By making use of the property “sameAs”, it can be recorded, for instance that a term such as “lineGroup” from one conceptualisation is semantically equivalent to a concept which is referred to elsewhere as “stanza”. Similarly, the property “differentFrom” can be used to explain the disparities between ontologies. The aim to develop a single encompassing ontology seems a utopian quest. By making use of formal semantics, however, materials that have been formatted according to dissimilar methodological practices can still be integrated. When interconnections among distributed data collections have been made explicit via semantic web technologies, descriptions of distinct domains, or distinct descriptions of a single domain may still be reconciled.

In this study, a decision was taken to combine the TEI and the OAC framework to use the strengths of each. The interpretative observations have largely been captured via OAC, as it was estimated that the protocol allows both for strictness and for a degree of flexibility. This chapter has focused mostly on the representation of structured annotations. The data that are captured, however, obviously serve a specific purpose. They are created in order to perform specific analyses. The more precise ways in which the data can be analysed will be discussed in the next chapter.