Transcending Textual Borders? Digitising a Middle-English Lunary from British Library Egerton MS 827 and a brief introduction to XML markup in the Humanities.[*]

Carrie Griffin and Julianne Nyhan, University College Cork

1 Introduction

It is not until recently that students of Middle English utilitarian writings and scientific material have had access to, and information on, this large corpus of work. As a result of recent advances, attempts have also been made to classify and examine different aspects of this material.[1] The wide range of texts that comprise this corpus undoubtedly offer illuminating insights into daily life and work, the nature and variety of beliefs and practices, and the growing popularity of such material in the late medieval period. However, in order to facilitate further research in this area, it is necessary that unedited texts are made more widely available and accessible, and an effective way of doing this is to produce electronic editions of such works. This paper will examine, firstly, the nature of the manuscript, British Library Egerton 827, and will describe

[*] This paper was delivered by the authors at *Borderlines VII Postgraduate Conference*, University College Cork, April 2003, and at *The Book as Artefact* Conference, Marsh's Library, Dublin, February 2004.

[1] For an overview of the advances in scholarship to date see Taavitsainen, I. and P. Pahta (eds.) *Medical and Scientific Writing in Late Medieval English*. (Cambridge 2004) 1-22.

and classify the previously unedited text, *XXX Dies Lune* transcribed from it. A brief introduction to XML markup for humanities scholars will then be presented, and the process of creating the digital edition of *XXX Dies Lune* discussed.

2 London, British Library, MS Egerton 827

London, British Library, MS Egerton 827 dates from the early 1400s.[2] It has been described as 'an astrological and medical book'[3] of 50 folios, containing nine prose treatises, including a tract on the nature of women; nativities; the *Wise book of Philosophy and Astronomy*, which is a tract on the influence of the planets on the behaviour and characteristics of men; tables of the moon and astrological diagrams. It also features a lapidary, some charms and recipes, and concludes with a short astrological tract. The manuscript is a small, vellum, quarto volume, measuring 185 x 122 mm, and is written by one scribe, in a textura book hand, except for the recipes found on ff 45-50^{v}, which are copied in various later hands.[4] The manuscript has, apart from some lunar charts and astrological calendars, no decoration, giving it a distinctly workmanlike appearance. Indeed, its relatively small dimensions (presumably for portability), the plain, functional appearance of the codex, and the signature, 'Welles leche', which occurs at f 50^{v}, strongly indicate that it was originally intended for practical use and regular consultation.

3 Description and Contextualisation

The text that has been digitised from this manuscript is a previously unedited lunary entitled *XXX Dies Lune*. This is the only extant version of this particular text, and it occurs between ff 10^{r} -13^{v} of the codex.[5]

Lunaries offer perpetual prognostications for the lunar month, dividing it into thirty, twenty-eight or twelve sections according to the days of the

[2] Keiser, George R., *The Manual of Writings in Middle English: Volume 10: Works of Science and Information*, (New Haven 1998) 3774.

[3] Taavitsainen, I. *Middle English Lunaries: A Study of the Genre*. (Helsinki: 1988) 71.

[4] ibid.

[5] ibid. 71-2.

moon, the mansions of the moon, or the moon's passage throughout the signs of the zodiac respectively.[6]

Texts of the first classification, that is, those that predict from one new moon to the next, over thirty days, or a synodic month, are known as 'lunaries proper', and the text we are concerned with falls into this category.[7] The predictions of lunaries are, at their most basic level, based on a system of astrology: they contain a large number of prognostications on more than one subject, advising people how to act, or what to expect, given certain cosmic circumstances. Writings such as these can also include elements that are foreign, such as biblical motifs, which 'enforce the fictional aspect' of such texts, but these, too, are symptomatic of a 'tendency towards encyclopaedic knowledge...essential to this mode of thought.'[8] These elements of *XXX Dies Lune* will be discussed more fully below.

The content of texts of this genre have been described by Max Förster as '[the] answers to seven questions', and have not been seen to stray too far from this formula except to omit certain types of predictions. Lunary texts generally, according to Förster, advise how to act on a given day and continue with all or some of the following: predictions on fugitives and lost or stolen property, predictions for children born on said day, and for those who fall ill; what a dream on the day signifies, and whether or not the day is favourable for bloodletting. A statement indicating the character of the time, whether good or bad, often serves as an introduction to the predictions for each day.[9]

XXX Dies Lune, is a collective lunary and therefore generally agrees with this formula. The prediction for each day opens with a brief statement on the nature of the day and, for twenty-five of the thirty days, mentions a biblical person, or refers to a biblical event. So, for example:

> *on þe fyrste day of þe mone was Adam made/ þat day is good and profytabele all werkes to werke [f 11^r].*

[6] For treatment of the genre of ME lunary texts, see Taavitsainen Taavitsainen, I. *Middle English Lunaries* and Means, Laurel. *Medieval Lunar Astrology: A Collection of Representative Texts.* (Lewiston 1993).

[7] Taavitsainen, *Middle English* Lunaries, 23, 47.

[8] ibid. 97.

[9] Föster, 'Vom Fortleben', 32-33, in Taavitsainen, *Middle English Lunaries*, 98.

Advice is also concerned with sources of livelihood; predictions may tell whether the time is good for buying and selling, for travelling or for agricultural pursuits; apparently, day 7 of the moon is 'good to gelde borys and daunt bestes' [f 11^r]. Nativities in this text are brief, and consist of the character of a child, often commenting on his/her professional career and length of his/her life; the prognostication for day fourteen, for example, tells us that:

> *A chyld þat is born þat day schal be a marchaunt proud and hardy, but he schal noght leue longe. [f 12^v].*

Prognoses for illness are important components of lunaries and the scope of such predications is generally broad, not least in this text. Information on how long the illness will last, if an illness will result in death, how to prevent death and whether or not medical care will be effective. One or two predictions even advocate moving house as a prerequisite for healing; for example, the prediction for day four of the moon advises that:

> *he þat falleth seek þat day schal sone dye but he chaunge hys place of hys dwellyng*

Dream prognostics also feature prominently in this lunary, and are in the main concise and specific, as is this one for day eight:

> *A drem schal be soth but look þu telle it no man but comaunde þi selfe to God [f 11^v].* [f 12^v]

Advice regarding phlebotomy is given at the end of each prediction, offering guidance as to the right timing of treatment; however, these statements do not volunteer more than whether it is 'it is good/not good to bleed that day', or they tell us to bleed early or late on a given day. As mentioned above, additional materials are incorporated into the text, as is the case with most lunaries. Biblical motifs are present in most of the predictions found in *XXX Dies Lune*, and 'the implication is that the events mentioned...had happened under similar positions of the skies'.[10] Parallel consequences, therefore, could be expected to

[10] Taavitsainen, *Middle English Lunaries*, 101.

occur under similar conditions. The events and persons referred to are borrowed from the Old Testament, and exclusively from the books of Genesis and Exodus. Further, line fragments, in Latin, from the first thirty psalms, to correspond with the thirty days, are interpolated with each prognostication; these are identified in the footnotes to our digital edition. It is most likely that such features were included to remind the reader of the benefits of prayer or, perhaps, to raise the status of the literature, imbuing it with a sense of authority. Taavitsainen suggests that such material would have 'replaced references to Greek gods and philosophers'.[11] Some scholars have also suggested that it was a common superstition that sentences or phrases from the Bible could act as amulets or charms of a sort, protecting the reader from evil.[12] As it happens, the psalm references throughout directly follow predictions on dreams, which frequently advocate prayer:

> *A drem þat day schal be chewyd with inne ij dayes or ijjj. Tel it no man but pray to god*

This practice of recommendation of a psalm is also frequently observed in other Middle English lunaries. It is possible therefore that, whilst the predictions seem to be finite, regular devotion and prayer could help to prevent, or lessen, the consequences of a bad dream.

The growth in popularity of the lunary text in England during the first half of the fifteenth century is most certainly connected with the growth of literacy, and to the emergence of a new class of readers from the middle layers of society. These pragmatic times demanded information, instruction and entertainment, and in many ways, lunaries were able to provide for all of these needs. The types of codices in which they are found reveal their position within medieval English society; they occur alongside astrological literature and didactic writings, in medical manuscripts and almanacs, as well as in household miscellanies and commonplace books. There is, of course, considerable overlap between these categories, since medical manuscripts, for instance, often contain computational material, just as astrological miscellanies have advice on matters of health, like phlebotomy. Because boundaries are blurred in the general field of scientific and medical writing in Middle English, a strict

[11] ibid. 101.

[12] ibid. 104.

classification of the manuscript contexts of lunaries is impossible, but it seems likely that such literature, as mentioned above, would have been primarily utilitarian in nature. Egerton 827 appears to contain additional aids, in the form of diagrams and tables, to assist in determining the state of the skies at any given moment; these would appear to provide a backup of sorts to the lunary text, which offers no technical support regarding astronomy. The *Wise Book of Philosophy and Astronomy*, immediately preceding the text of *XXX Dies Lune* in the manuscript, gives simple computational advice on how to measure a planet's reign, by 'reciting twice the psalms with the litany from a planet's last position'; this would, according to Means, take about an hour, and is a simple methodology that would have required very little knowledge of astronomy.[13] Further on in the codex, at f 30^r, we hear that it is profitable:

> *...meche to wete in what sygne þe mone is for he is ner þe ground þan ony oþer planete and þerfore alle worldly þingys it lendeth more of hys effect.*

Therefore, the works of this manuscript would appear to complement one another, thereby enhancing its utilitarian function. Bearing this in mind, it is perhaps surprising that a 'closely-related' verse lunary, also entitled *The Thrytty Daies of the Mone*, and also dating from the early fifteenth-century, purports to have been:

> *wryttyn ... for our profit For our solas and oure delijt.*[14]

This verse text, dating from 1417 is, arguably, the closest Middle English verse counterpart to the prose lunary discussed here; the predictions and biblical motifs agree, and both manuscripts boast similar contents, most notably computational tables and diagrams and copies of the *Wise Book of Philosophy and Astronomy*. The Digby text does, however, address its audience, lending some clues as to the possible usage/readership of texts like these. It's three-fold claim—to provide profit, *solas*, and delight—is, arguably, achieved in the text. The information provided is certainly

[13] Means, *Medieval Lunar Astrology*, 56.

[14] Taavitsainen, *Middle English Lunaries*, 122; 'The Thrytty Daies of the Mone' is found in 9 mss, including Oxford, Bodleian Library, MS Digby 88, ff64-75, Ashmole 189, ff212-215^v, and Huntington Libr. HM 64, ff84-95 (each of which also contain a copy of *The Wise Book of Philosophy and Astronomy*).

profitable, in that the text is a manual, of sorts, for timing one's actions for the best possible outcome. Comfort is provided by references to psalms and the reassurance that prayer will help the onset of a fated outcome; delight may come with the recognition of well-known biblical events and names. Middle English lunary prognostic texts were also disseminated in printed books, and were copied and printed well into the seventeenth century; their popularity in the late-fourteenth through to the fifteenth-century, however, suggests that they may have transcended their primary utilitarian role and may have been read too for interest and pleasure.

4 Editorial Policy

The text of *XXX Dies Lune* is transcribed diplomatically from London, British Library, MS Egerton 827, ff 10^r-13^v. The Middle English spelling is retained, as are 'thorn' and 'yogh' (where they occur in the ms). All abbreviations are expanded. In order ensure that the electronic edition of *XXX Dies Lune* is a searchable, stable document that can be used in scholarly research it has been 'marked-up' or 'encoded' in XML. Before moving on to discuss the process of encoding a document in XML, some observations will be made on the concept of markup.

5 What is markup?

Markup has been defined in a pithy statement by the Text Encoding Initiative 'as any means of making explicit an interpretation of a text'.[15] Allen Renear has described the process of marking up a digital text as 'storing, in the computer's memory, codes that represent the linguistic content (typically alphabetic characters and punctuation) and additional information related to this content, such as intended or observed formatting or layout effects and explicit identification of sections of text as being footnotes, titles ... . A text-encoding system is the system of codes

[15] C M Sperberg-McQueen and Lou Burnard (eds) *Guidelines for electronic text encoding and interchange* 4th ed. (Oxford 2002) 13.

which effect such a representation'.[16]

The concept of markup is not a new or modern one, and it has been employed in one form or another since texts were first committed to hard-copy representations. Indeed, McGann has argued that 'there is no such thing as an unmarked text, and the markup system laid upon documents to facilitate computerised analyses are marking orders laid upon already marked up material. (Thus all texts implicitly record a cultural history of the artifactuality)'. [17] Examples of forms of markup used in the pre-digital age include the presentational features that were frequently employed in manuscripts. In the earliest Western manuscripts, for example, presentational markup was applied to texts in the form of different coloured inks: brick-red ink was frequently used to distinguish or emphasise notable portions of text, and in the early middle ages, the main text of a manuscript was sometimes written in red ink with the accompanying commentary clearly delineated in black.[18] Forms of markup can also be observed in British Library Egerton MS 827, for example, the use of decorated initials to herald new sections of text. It can also be argued that textual features of this lunary, such as pericopes, function in a similar manner to that of a hyperlink in a modern day electronic document. Clearly, one function of the pericopes in this lunary is to refer the reader to a related piece of information that is contained in another document (though in this case, the information is more likely to be in the memory of the listener/reader rather than in another hard-copy or electronic collection).

6 Markup applied to electronic texts

With regard to electronic texts, three main divisions of markup have been identified and expounded upon by theorists: presentational, procedural

[16] Allen Renear, 'Out of praxis: three (meta)theories of textuality', Katherine Sutherland (ed), *Electronic text: investigations in method and theory*, (Oxford 1997) 107-126: 108-9.

[17] Jerome McGann, *Radiant textuality: literature after the world wide web*, (New York 2004) 138.

[18] Bernhard Bischoff, *Latin palaeography in antiquity and the middle ages*, tr. Dibh Crinn and David Ganz (Cambridge 1990) 17.

and descriptive.[19]

7 Presentational and Procedural markup

Presentational markup is applied to a text to describe its appearance[20] and can be used to indicate, for example, white space and font changes.[21] In some text processing and preparation systems, for example, LaTeX and OpenOffice, presentational markup is frequently indicated through the use of procedural markup.[22]

Procedural markup usually comprises a set of instructions given to a piece of software to specify how a portion of a document should be processed. From an academic encoding perspective, presentational and procedural markup is frequently unsatisfactory: it does not describe the content of a document, and consequently, such a document cannot easily be searched for descriptive information, for example, all personal names.[23] Users who encode documents with procedural markup in one editor can encounter considerable difficulties if they try to edit or open their document in another editor. Furthermore, because procedural and presentational instructions are usually contained in the same document as the information that is being marked-up, changes made to the formatting of the document, for example, will result in an extra layer of editing to ensure that no errors are introduced to the base text every time an aspect of markup is changed.

[19] cf. Allen Renear, 'The descriptive/procedural distinction is flawed' *Markup Languages: theory and practice* 2 no. 4 (Cambridge, MA, USA 2000); Wendell Piez, 'Beyond "the descriptive vs. procedural" distinction' *Markup Languages: theory and practice* 3 no. 2 (2001) 141-172.

[20] James. H. Coombs, Allen H. Renear and Steven J. DeRose, 'Markup systems and the future of scholarly text processing', *Communications of the Association of Computer Manufacturers* 30 no. 11 November (1987) 4. http://www.oasis-open.org/cover/coombs.html.

[21] Renear, *Three (Meta)Theories*, 113.

[22] Cournane has noted that LaTeX uses both procedural markup and declarative markup. Declarative markup specifies structural aspects of a text without giving any details about how that markup should be processed. Mavis Cournane, *The application of SGML/TEI to the processing of complex multilingual historical texts*, unpublished PhD diss. University College Cork (1997) 26.

[23] Coombs et al. *Markup systems*, 4.

8 Descriptive markup

Descriptive markup is applied to a document to indicate what each part of a document is, for example, a lemma, a quotation or a sobriquet. Charles Goldfarb has stated in regard to descriptive markup that 'the markup process stops at the first step: the user locates each significant element of the document and marks it with the mnemonic name ... that he feels best characterised it'.'[24] The process can be described as stopping at the first step because all further processing that is applied to a document, such as formatting, is specified in another document.

Descriptive markup has many advantages over presentational and procedural markup: if applied correctly, in most cases it ensures that documents are searchable, portable, usually vendor independent, easily transformed into other formats, and support multiple views of data without ever changing the master file.[25]

9 What is XML and why use it?

The acronym XML stands for *eXtensible Markup Language*. XML is a set of guidelines, as outlined in the *XML 1.0 Specification*,[26] that can be used to identify information both within and about texts. It is essentially a subset of SGML. In terms of markup languages, HTML on one hand, and SGML and XML on the other, occupy different conceptual layers. Thus, HTML is described as a markup language, while XML and SGML are described as meta-markup languages.[27]

The markup language HTML consists of a fixed corpus of tags, that can be applied to a document in order to reflect three different types of

[24] Charles G. Goldfarb, *Annex A of ISO 8879, the SGML international standard Introduction to generalised markup*, (Geneva 1986) 2.

[25] See especially: Renear *Three (Meta)Theories*.

[26] Tim Bray, Jean Paoli, C.M. Sperberg-McQueen, Eve Mailer (second edition) and Francois Yergeau (third edition), *Extensible Markup Language (XML) 1.0*. http://www.w3.org/TR/2004/REC-xml-20040204.

[27] The history of the meta-markup language called the Generalized Markup Language (GML) whence SGML was derived has been well documented, see especially: Cournane, op. cit.; Charles F. Goldfarb, *The roots of SGML* (1996); idem. *Design considerations for integrated text processing systems*, IBM Cambridge Scientific Centre (1973). http://www.sgmlsource.com/history/G320-2094.htm.

information, that have been categorised by Flynn as structural, content descriptive and visual.[28] Thus, an encoder can use elements to describe the structure of a document, for example, its headings, paragraphs and lists. Content descriptive tags are also available to the encoder, who can, for example, use the element `<strong>` to emphasise information. Formatting tags can also be used to indicate that a portion of text should be rendered in bold type or Times New Roman font, or whatever.

Languages such as SGML and XML are more properly described as meta-markup languages.[29] Thus, XML is not a language in itself, but rather a set of instructions that can be combined in order to create specialised markup languages. A project using XML rather than HTML as a format to encode its master documents is not limited to the number of tags that are set out in the HTML specification, but rather can create a specialised markup language to reflect the unique requirements of that project. Further, all the advantages of descriptive markup outlined above are associated with XML. It is a stable standard that is unlikely to become obsolete in the foreseeable future. It is both vendor and platform independent, thus ensuring that XML documents are not owned by a particular group or company and that creators of XML documents are not dependent upon a particular piece of software in order to view, edit, analyse or transform them. Indeed the father of XML has stated:

> XML derives from a philosophy that data belongs to its creators and that content providers are best served by a data format that does not bind them to particular script languages, authoring tools, and delivery engines but provides a standardised, vendor-independent, level playing field upon which different authoring and delivery tools may freely compete.[30]

[28] Peter Flynn, *Making more use of markup*, SGML '95 Boston, MA. http://imbolc.ucc.ie/ pflynn/articles/moreuse.html.

[29] Bosak notes that this is a generalisation and that the SGML layer 'is not as abstract as a true metalanguage like Bachus/Naur Form (BNF) which is used to define programming languages'. Jon Bosak, Media-independent publishing: four myths about XML', *IEEE Computer* 31 no. 10 (1998). http://www.xml.coverpages.org.

[30] An infamous example of the potential consequences of depending on proprietary software or hardware is the electronic version of the *Domesday Book*. Created as part of a multi-million pound BBC Domesday project, it is now unreadable because the medium it was published in (12 inch video discs) is no longer supported. See, for example, Duckworth, J., 'Future proof', *Laboratory News*, (Aug. 2002) 13-14.

XML documents can be easily transformed and output in a range of formats including pdf, HTML, XHTML, another xml vocabulary, and plain text. The instructions that transform a document's format, structure, and appearance are contained in a separate file from the infomation that is being transformed, thus ensuring the data in a master file is not compromised. From an academic encoding perspective, XML markup can be designed to encode extremely complex information and thus documents can support superior search and interrogation facilities that can be harnessed to support a wide range of applications from research to e-learning.

Humanities computing projects who choose to encode their data in XML have two primary choices: they can either encode their data in TEI conformant XML or devise their own XML encoding scheme.

10 The Text Encoding Initiative

One of the most successful scholarly applications of SGML/XML has been developed by the Text Encoding Initiative (hereafter TEI), established in 1987. The TEI has been funded by the Association for Computers in the Humanities (ACH); the Association for Computational Linguistics (ACL); the Association for Literary and Linguistic Computing (ALLC); the U.S. National Endowment for the Humanities (NEH); the European Community; the Mellon Foundation; and the Social Science and Humanities Research Council of Canada.[31]

Up to 1987 texts created by museums, libraries, publishing houses as well as Universities and individual scholars were being marked up in a variety of computer languages. The goal of the TEI was the creation of an encoding scheme that was stable, international, interdisciplinary and capable of supporting data exchange.[32]

Members of the TEI had clear ideas and theories about the criteria that an electronic scholarly text should meet,[33] and guidelines were written

[31] TEI web-site, accessed on 1/10/05. http://www.tei-c.org/.

[32] Cournane, *The application of SGML*, 44.

[33] See especially Sperberg-McQueen, 'Text in the electronic age: textual study and text encoding, with examples from medieval texts', *Literary and Lingusitic Computing* 6, (1991) 34-46.; idem. *Textual criticism and the Text Encoding Initiative*. http://www.tei-c.org/Vault/XX/mla94.html.

to define the markup of, *inter alia*, a text's physical appearance, content, structure, and bibliographical information. As the name suggests, the guidelines are not a series of prescriptive statements about how markup can be applied to a text. While the process of TEI conformance has been clearly defined, scholars are able to select the markup they require from a number of sets of encoding schemes and it is possible for projects to modify existing TEI markup if necessary.[34]

The TEI has been criticised by some scholars, Olsen argues that it simply allows too much variation and flexibility: 'The editors of the TEI are writing a data interchange format while at the same time working out a mechanism to support theoretically informed encoding specifications for just about any textual object that scholars in a wide variety of disciplines might encounter. Unfortunately, the resulting drafts of the TEI specification(s) reflect this underlying confusion of the task at hand'.[35] McGann, on the other hand has argued that the TEI's focus and interpretation of text is too narrow.[36] While many of these criticisms are valid and important, the TEI is a widely accepted international guideline that many academic projects[37] adhere to, because in many cases the significant advantages associated with using it outweigh the disadvantages. The TEI guidelines support the exchange of information and tools between all kinds of projects and the encoding of many forms and categories of data, including poetry, drama, lexicography and manuscript description. Furthermore, the TEI boasts a large support community and members are able to draw on existing tools and documentation to assist in the implementation of the TEI guidelines, as well as subscribe to mailing lists and attend conferences and specialist seminars.

[34] See for example Susan Rennie, 'The Electronic Scottish National Dictionary (eSND): Work in Progress', *Literary and Linguistic Computing* 16 no. 2 (2001) 153-160; Gregory Toner and Maxim Fomin 'Digitizing a Dictionary of medieval Irish: the eDIL Project', *Literary and Linguistic Computing*, 21 no. 1 (2006) 83-90; Stephen R. Parkinson and Antnio H. A. Emiliano 'Encoding medieval abbreviations for computer analysis', *Literary and Linguistic Computing* 17 no. 3 (2002) 345-360.

[35] Olsen, Mark, *Text theory and coding practice: assessing the TEI*, 1 (1996). http://barkov.uchicago.edu/talks/ACH96.TEI.talk.html.

[36] McGann, *Radiant textuality*, 187-191; 193-207.

[37] The TEI projects page. (accessed on 1/12/05). http://www.tei-c.org/Applications/.

11 Encoding *XXX Dies*

Another option available to projects is to design their own XML encoding scheme to describe their specific data. The markup that was designed to encode *XXX Dies Lune*, will now be discussed in order to present a basic introduction to some aspects of XML for humanities scholars.

The encoding scheme we devised for *XXX Dies Lune* is a simple yet functional one that consists of elements and attributes that describe the structure and content of the document. The elements and attributes created to encode *XXX Dies Lune* are set out in Figure 1. In terms of XML, an element is a semantic label for a unit of information, and it is delimited by start and end tags. For example the element `<ps>Adam</ps>` describes Adam as a personal name. An attribute can be used to specify further information about an element. In this example, the attribute specifies additional information about the category a personal name: `<ps type="biblical">Adam</ps>`, The question of when to encode information as an attribute rather than an element is one that is frequently discussed by encoders. While answers to this question will vary based on the context, a good rule of thumb is to encode essential information in elements and non-essential or additional information in attributes.[38]

Figure 1: elements and attributes used in *XXX Dies Lune*

ELEMENT `<folio>` indicates an MS folio
ATTRIBUTE `n` specifies a folio number
ELEMENT `<lb>` indicates a line break
ATTRIBUTE `n` specifies the number of a line break
ELEMENT `<ps>` indicates a personal name
ATTRIBUTE type can be used to categorise a personal name e.g. biblical or comical
ELEMENT `<abbrev> indicates an abbreviation\\`ELEMENT `<sup>` indicates additional information supplied by an editor of a text
ATTRIBUTE `resp` allows the initials of that editor to be specified
ELEMENT `<frn>` indicates a text string in a language other than the

[38] See especially: Robin Cover, *SGML/XML: Using elements and attributes*. http://xml.coverpages.org/elementsAndAttrs.html and Uche Ogbuji *Principles of XML design: When to use elements versus attributes*. http://www-128.ibm.com/developerworks/xml/library/x-eleatt.html.

language used in the main body of a text
ATTRIBUTE `lang` allows a foreign language to be specified
ATTRIBUTE `type` allows a language register to be specified e.g. scientific
ELEMENT `<pn>` indicates a place name
ELEMENT `<cert>` indicates information that is uncertain
ATTRIBUTE `level` indicates the level of uncertainty e.g. 50%
ELEMENT `<sic>` indicates that information is ’thus’
ATTRIBUTE `resp` allows the initials of an editor who encoded the information as ’thus’ to be specified

12 The Document Type Definition

You may be wondering at this point how a computer with no innate intelligence can recognise the myriad of languages that can be created with XML? The definition of XML markup is recorded in a DTD. The acronym DTD stands for *Document Type Definition*, and is described in the *XML 1.0 Specification* as a ‘grammar’.[39] Thus, as the term grammar suggests, the DTD lists the elements, attributes, entities and notations that form an XML document, and specifies the various types of relationships that exist between them.

13 The document type declaration

The most simple form of an XML file must comprise a processing instruction, and a root element, as illustrated in Figure 2.

Figure 2

```
<?xml version="1.0" standalone="yes"?>
<greeting>
Hello World!
<greeting>
```

While the processing instruction tells the parser which version of XML is being used and specifies the character encoding (numeric values that make up the character set of one language or another) of a particular

[39] Tim Bray et al., *XML Spec.* §2.8.

document, the root element completely contains all the other elements in a file.

A project may create a number of DTDs in order to encode and describe different document instances. Depending on the requirements of a particular project, an encoder can associate a DTD with a document instance in one of three ways: as an internal subset (where the DTD is contained in the same file as the markup it is describing), an external subset (where the DTD is contained in a separate file from the markup it is describing) or as a combination of both. In the case of *XXX Dies Lune*, the document is stored as an external subset. One advantage of storing a DTD as an external subset is that it does not have to be used exclusively with one document, but can be used to define a number of documents in a corpus. In order to associate a specific DTD with a specific XML document, the *Document type declaration* is invoked, 'the XML document type declaration contains or points to markup declarations that provide a grammar [DTD] for a class of documents'. In Figure 3 below, an excerpt from *XXX Dies lunes*, the document type declaration points to the document's DTD:

Figure 3

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE document SYSTEM "dies.dtd">
```

14 Declaring an Element

A DTD can contain a number of declarations and each declaration consists of two parts: a generic identifier and a content model.

Figure 4

```
<!ELEMENT sup   (#PCDATA)>
↑                  ↑
```

generic identifier   content model

The generic identifier[40] gives the name of the element being declared:

[40] ibid. §3.

in this case <sup>, as illustrated in Figure 4. The content model[41] defines what is permitted to be contained in that element. In this case, the element <sup> may contain PCDATA or parsed character data (text only).

The next example illustrates a slightly more complex content model, and is an example of a model group.

Figure 5

```
<!ELEMENT lb (#PCDATA|ps|abbrev|pn|frn|sup|sic|cert)*>
↑                            ↑
```

generic identifier   content model

The declaration in Figure 5 describes the permitted content of an <lb> element, and furthermore describes the relationship between the elements <lb> is permitted to contain, by harnessing connectors such as the union operator | (indicating 'or') and the occurrence indicator * (indicating zero or more occurrences). In this example, the model group permits <lb> to contain text or one of seven other elements including <ps> (personal name) and <pn> place name. The occurrence indicator further refines the relationship between the members of this model group, and the * then indicates that this model group can occur zero more times in a text.

15 Declaring an attribute

The syntax for declaring an attribute is very similar to that of declaring an element, one of the obvious differences between the two types of declarations is the use of the term !ATTLIST for attributes.

After specifying the name of the attribute the next necessary piece of information pertains to the type of attribute.

Figure 6

```
<!ELEMENT frn  (#PCDATA|abbrev)*>
<!ATTLIST frn  lang (la|en|fr|gk) "la"
type (pericope) "pericope">
```

[41] ibid. §3.2.1.

The attribute declaration in Figure 6 permits two attributes to be used with `<frn>`: ‘lang’ to encode foreign language information and ‘type’ to encode a type of language use. In the case of the ‘lang’ attribute four options are specified: the language codes for Latin, English, French and Greek and the language code for Latin is set as the default if no code is specified.

16 Character references

The final aspect of XML to be covered here is that of character references.

Figure 7

```
<!ENTITY thorn "&#x00fe;">
<!ENTITY THORN "&#xde;">
<!ENTITY yogh &#x21d;>
<!ENTITY YOGH &#x21c;>
```

The example above illustrates some of the character references that are used in the *XXX Dies* DTD. By default, all XML documents are encoded in Unicode, which currently defines more than 40,000 different characters in a range of languages. [42] However, when an XML document is used in a variety of ways in different organisations and across a number of platforms, one of the issues that can arise is the possibility that not all applications support Unicode.[43] The TEI consortium give the example of the character , that can be represented in an XML document with default encoding as the Unicode character with value `OOE9`. Representing character references with numeric values ensures that documents can be passed between different applications, that may or may not support Unicode, without resulting in a loss of data. For example, the character may nor be available in a non-Unicode character set, but loss of data can be prevented by representing the character with its hexadecimal value or character entity reference: `&#x00E9;` or value OOE9.[44]

[42] The Unicode Consortium, *The Unicode Standard, Version 4.1.0, defined by: The Unicode Standard, Version 4.0* (Boston 2003).

[43] See especially, Markus Kuhn, *UTF-8 and Unicode FAQ for Unix/Linux*, (2005). http://www.cl.cam.ac.uk/ mgk25/unicode.html.

[44] C. M. Sperberg-McQueen et al. *TEI Guidelines*, (2002) 17.

A document with a number of hexadecimal references embedded in it is not very legible,[45] and an encoder may find themselves having to refer continually to the Unicode book in order remind themselves which values represent which character. XML, therefore, defines a set of entities, that can be used to represent characters from the Unicode set. In Figure 7 above, for example, the mnemonic Auml is mapped to its hexadecimal reference, and the entity can then be used in the document in the form of `&Auml;` to represent, Ä, a capital A with an umlaut, hexadecimal value `&#x00C4;`

17 Conclusion

Many practical advantages of encoding a text in XML have been discussed in this paper. On a more theoretical level, the process of transcribing and digitising *XXX Dies Lune* raised many questions for us about the nature of text and artifact. For example, does the medium that a text is represented in (such as vellum, paper, facsimile image or e-text) have a significant impact on the ways that a text may be interpreted? If a manuscript version of a text is considered to be an artifact, does it become an electronic artifact when digitised? Indeed, can an electronic text become an artefact? It is hoped that questions such as these will be explored in a further paper.

[45] Legibility is a stated goal of XML, 'XML documents should be human-legible and reasonably clear', Tim Bray et al. *XML 1.0.* §1.1.