**STOKES — Problem**

<5 figures>

**The Problem of Digital Dating: A Model for Uncertainty in Medieval Documents**

Stokes, P. A.

The dates of medieval documents are often not precisely known and so are catalogued with labels such as ‘early 12th century’. While these labels are useful and meaningful for a medievalist, they present difficulties in a digital context in terms of searching, sorting, and aggregating existing descriptions. This paper will examine these challenges, propose an alternative way of modelling dates, and then make some suggestions for representing these in user interfaces. It draws on work for Models of Authority, a new project on Scottish charters of the 12th century that is funded by the UK Arts and Humanities Research Council and that uses and extends the DigiPal framework (DigiPal, 2010–14).

# The Problem

# Scholars have developed conventions for recording the various degrees of uncertainty in dates of manuscripts and documents.1 To take an example from Models of Authority, National Library of Scotland GD55/32 was written between 1189 and 1196; by convention this is indicated ‘1189×1196’ (POMS document no. 3/14/3). However, one or both dates in this range may be approximate or uncertain (‘circa 1192 × 24 March 1201’: POMS 3/486/8), or different date ranges may be possible (‘A.D. 670 × 671 [? A.D. 681]’: eSawyer S.1168). Alternatively, only a general date might be known (‘late 12th / early 13th century’: POMS no. 3/590/9). Other possibilities include ‘early’, ‘mid’, or ‘late’ in the 12th century; the first or second half of the century; and sometimes the second quarter, the first third, and so on (see further examples in DigiPal, 2010–14; eSawyer, 2010; as well as DigiPal, ‘Glossary’).

This system of dating has served medievalists well. However, there is no obvious way of searching for or ordering material labelled in this way. When exactly does ‘late 12th century’ begin and end? Does it include the last quarter of the century? Does the first quarter of the 12th century come before or after the ‘early’ 12th century? Where does ‘A.D. 670 × 671 [? A.D. 681]’ fit on a timeline? We can decide answers to all of these questions in any given application, as indeed has been done (Stokes, 2012; see also the ‘search by date’ function in Manuscripts Online, and DigiPal’s faceted search, among many others). However, the result will necessarily be arbitrary and therefore difficult for others to understand. It will also inevitably be inconsistent with practice in other projects, and this in turn will lead to problems with linked data or aggregating sites (examples of which in this context include Biblissima, MESA, and Manuscripts Online). One could try to recover the intentions of the original cataloguer(s), but this is often lost and, even if recoverable, is unlikely to be consistent from one source to the next. Models of uncertainty do exist that go some way toward capturing these formulae, such as that in the TEI Guidlines (§§13.1.2 and 21.2), and those described by Grossner (2013), as well as projects such as TopoTime (Meeks and Grossner, 2013–) and PeriodO (Rabinowitz et al., n.d.). However, they still leave open these problems of searching and presenting the material, are very early in development, and/or do still not seem to adequately capture the specifics of manuscript studies. Instead, an alternative is required.

# The Model

It is argued here that asking if ‘early 12th century’ includes the year 1100 is legitimate in a digital context, but to medievalists the question is in many ways meaningless, since if cataloguers knew that the manuscript was written after 1100, then they would have specified this. ‘Early 12th century’ does not mean ‘no earlier than midnight 1 January 1100 and no later than midnight 31 December 1115’, but rather something closer to ‘probably some time in the first 15 years or so of the century, but perhaps a little later or earlier’. To indicate this difference more concretely, it is useful to use probability density functions. Summarising crudely, a probability density function (pdf) represents the likelihood that a given variable has a given value. For instance, if we know when a document was written, then it has 100% probability of being written then and 0% probability of being written any other point of time; the resulting pdf is the Dirac delta function and is conventionally represented as shown in Figure 1. The assumption implicit in many search interfaces that ‘early 12th century’ has a fixed and firm beginning and end—say 1100×1115—and can be represented using the rectangular distribution shown in Figure 2. In contrast, judging intuitively from my own experience with manuscripts, ‘early 12th century’ is probably captured more accurately by something like a normal distribution (Figure 3). We can also combine these for more complex cases like ‘A.D. 670 × 671 [? A.D. 681]’ (Figure 4). Indeed, the range of possible curves is essentially limitless, and the model is sufficiently general to allow for any curve that best represents the particular case at hand.

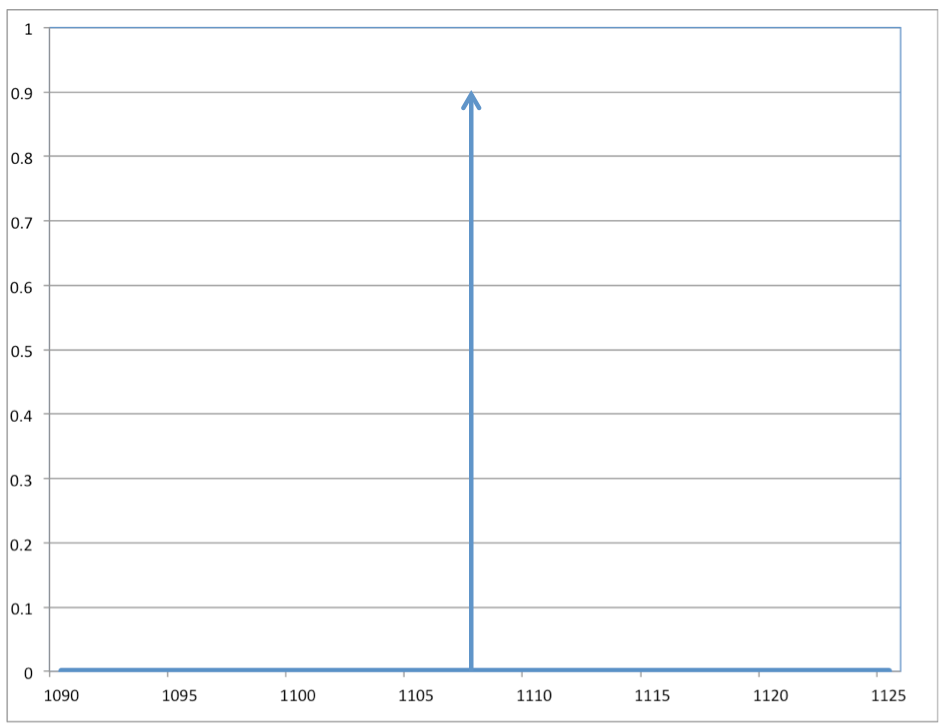


Figure 1. Dirac delta function representing ‘1107’.

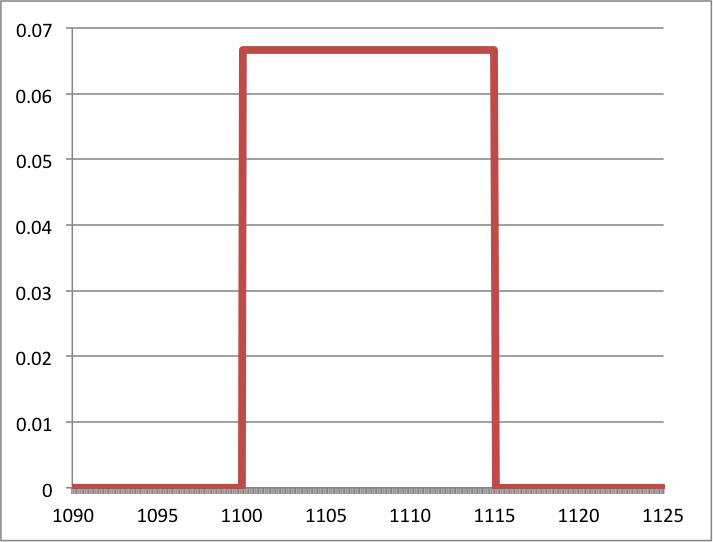


Figure 2. Rectangular distribution representing ‘1100×1115’.

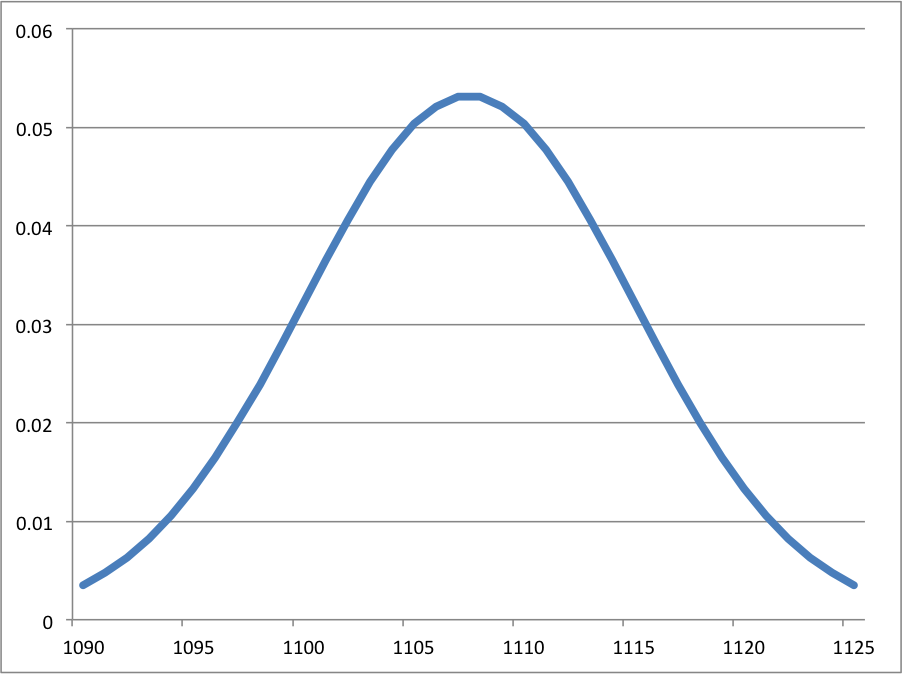


Figure 3. Normal distribution representing ‘early 12th century’.

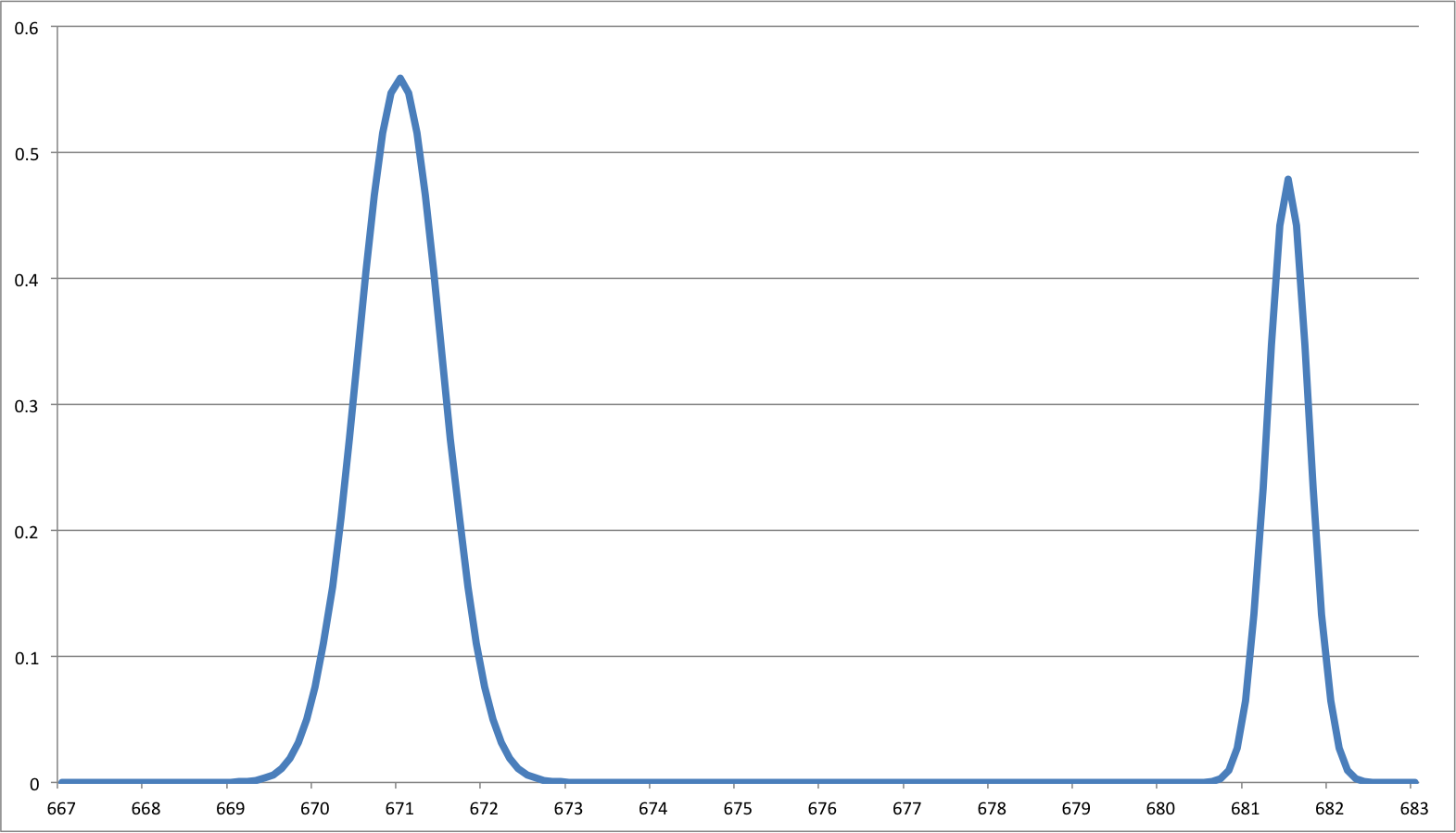


Figure 4. Hybrid distribution representing ‘A.D. 670 × 671 [? A.D. 681]’.

It is important to emphasise that these curves do not necessarily represent the mathematical likelihood of the date of writing, or even of the cataloguer’s judgment, since accurately quantifying human impressions is difficult and problematic. Indeed, the resulting curves need not even necessarily be true pdfs. They may be mathematically valid and numerically accurate in a given implementation, but this assumes a reliable statistical model and so seems only appropriate for those projects that are sufficiently quantitative that such a model is available (examples may include DEEDS and work described by Smit [2011] or Wolf [2015]). However, it would be incumbent on people using this approach to demonstrate the validity of their statistical representation (for cautions against which, see Sculley and Pasanek [2008]; Stokes [2009]; Hassner [2013]). The point is in fact the opposite: rather than providing exact figures, they are intended instead to represent more meaningfully in digital form when the scholar in question considered the document to have been written. For instance, if I want to communicate the approximate frequency of a given letter-form in time, then instead of using a simple timeline like that shown above, I can instead calculate the sum of the distribution functions of all the scribal hands that show this form. Figure 5 shows the resulting curve for all occurrences in the DigiPal database of the tall-**e** form of the letter **æ**, and this seems to effectively capture the received view that the letter-form was common early in the 11th century but went out of use soon after (Ker, 1957; Stokes, 2014). Alternative representations could include lines of varying colour, adjusting the value or saturation according to the value and thereby allowing easier comparison of different categories, or perhaps even transparency, where users adjust the date and, according to their distribution, images of the relevant letters fade in and out of view. The curves could also be used to provide a significance value for search results by taking the integral of the curve across the time interval that the user has specified, or again adjusting the transparency of images representing these forms.

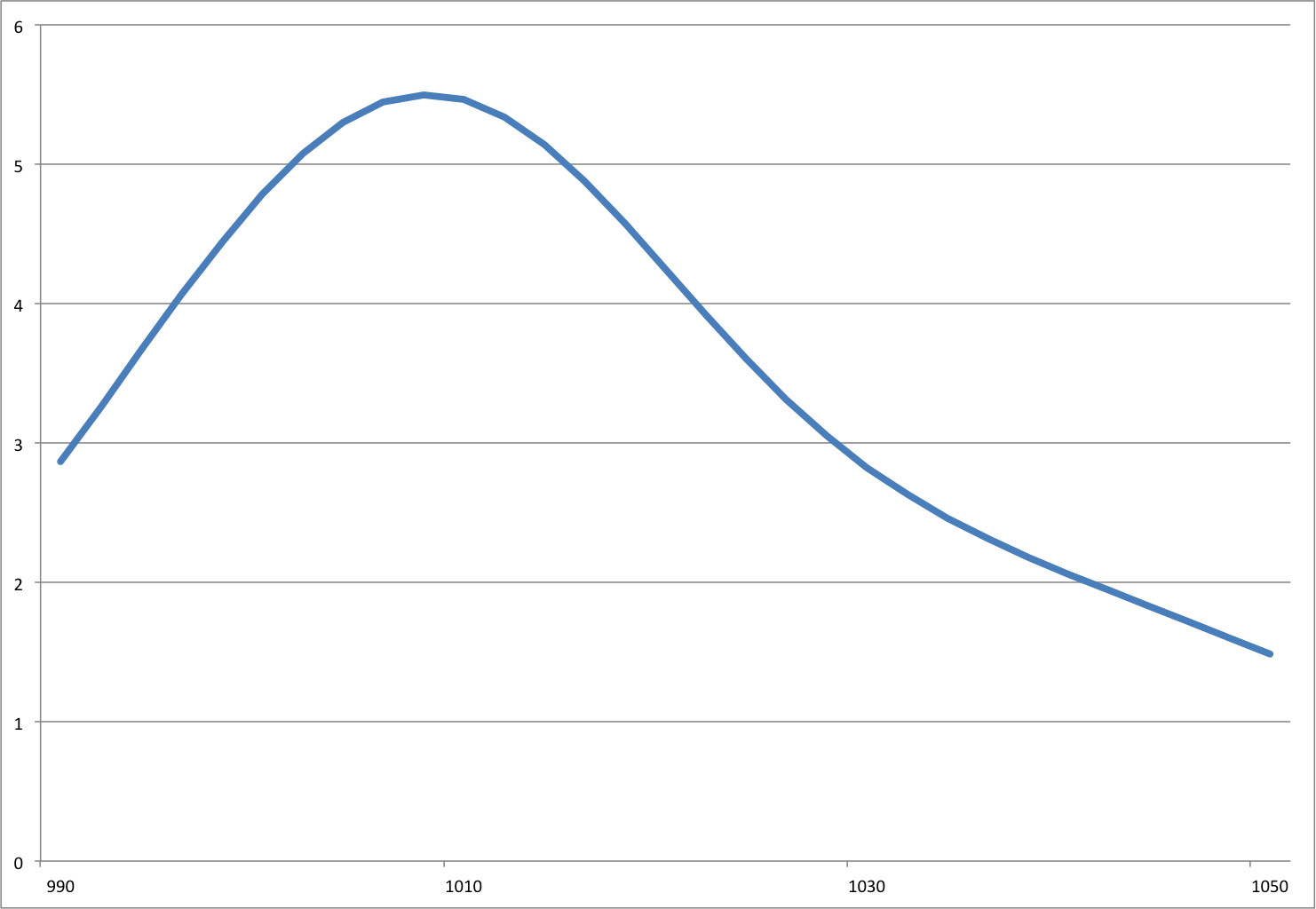


Figure 5. Distribution curve for occurrences of the tall-e form of æ.

# 

# Conclusion

As noted above, the highly quantitative nature of this model does not bring any more certainty, nor does it address all of the concerns raised at the start of this paper. However, it does seem to promise a more intuitive and meaningful way of interacting with the content. It has been observed that the computer should not be used to provide firm answers or proving the truth of hypotheses regarding historical fact, but should instead provide means for interaction, visualisation, and knowledge creation (Clement et al., 2009; Jessop, 2008; Sculley and Pasanek, 2008). Paradoxically, then, the benefit of the model proposed here could lie not in its mathematical accuracy but the opposite: by highlighting the ‘fuzziness’ of the content it may help to break down the illusion of certainty that the computer typically brings, and may instead present a more useful and meaningful interface. Initial experiments and informal surveys with medievalists to date have suggested very strong support indeed for this approach. No doubt that other, better models will be developed by others in due course, but the point remains for now that the representation of dates in databases of manuscripts and documents is too narrow, and more imagination is required if we are to make the best use of what we have.

# Funding

This work was supported by the Arts and Humanities Research Council [AH/L008041/1].

# 

**Note**

1. This section of the proposal draws extensively from Stokes (2012).

# References

**Biblissima.** (n.d.). Patrimoine écrit du Moyen Age et de la Renaissance, http://www.biblissima-condorcet.fr.

**Clement, T., et al.** (2009). How Not to Read a Million Books. Rutgers University, New Brunswick, NJ, http://www3.isrl.uiuc.edu/~unsworth/hownot2read.rutgers.html.

**DEEDS.** (n.d.). The Documents of Early English Data Set. University of Toronto, http://www.utoronto.ca/deeds.

**DigiPal.** (2010–14). Digital Resource and Database of Palaeography, Manuscript Studies and Diplomatic. London: King’s College, http://www.digipal.eu.

**eSawyer.** (2010). The Electronic Sawyer: Online Catalogue of Anglo-Saxon Charters. London: King’s College, http://www.esawyer.org.uk/.

**Grossner, K.** (2013). A Space-Time Datatype for Historical Place? https://cga-download.hmdc.harvard.edu/publish\_web/SpaceTime/DH2013\_Grossner\_SpaceTimeDatatype.pdf.

**Hassner, T., et al.** (2013). Computation and Palaeography: Potentials and Limits. *Dagstuhl Manifestos,* **2:** 14–35, doi:10.4230/DagMan.2.1.14.

**Jessop, M.** (2008). Digital Visualisation. *Literary and Linguistic Computing,* **23**(3): 281–93, doi:10.1093/llc/fqn016.

**Ker, N. R.** (1957). *Catalogue of Manuscripts Containing Anglo-Saxon.* Clarendon Press, Oxford.

**Levy, N., et al.** (forthcoming 2015). Consolidating the Results of Automatic Search in Large-Scale Digital Collections. In Brookes, S., Rehbein, M. and Stokes, P. A. (eds), *Digital Palaeography.* Aldershot: Ashgate.

**Manuscripts Online.** (n.d.). Written Culture, 1000 to 1500. Version 1.0, http://www.manuscriptsonline.org.

**Meeks, E. and Grossner, K.** (2013–). TopoTime Gallery and Sandbox. Stanford University Libraries, Stanford, CA, http://dh.stanford.edu/topotime/.

**MESA**. (n.d.). Medieval Electronic Scholarly Alliance. http://www.mesa-medieval.org.

**POMS**. (n.d.). People of Medieval Scotland, 1093–1314. King’s College, London, http://www.poms.ac.uk.

**Rabinowitz, A., et al.** (n.d.). PeriodO: A Gazetteer of Period Assertions for Linking and Visualising Data. http://perio.do/.

**Sculley, D. and Pasanek, B. M.** (2008). Meaning and Mining: The Impact of Implicit Assumptions in Data Mining for the Humanities. *Literary and Linguistic Computing,* **23**(4): 409–24, doi:10.1093/llc/fqn019.

**Smit, J.** (2011). The Death of the Palaeographer? Experiences with the Groningen Intelligent Writer Identification System (GWIS). *Archiv für Diplomatik, Schriftgeschichte, Siegel- und Wappenkunde,* **57**: 413–25.

**Stokes, P. A.** (2009). Computer-Aided Palaeography: Present and Future. In Rehbein, M., et al. (eds), *Kodikologie und Paläographie im Digitalen Zeitalter—Codicology and Palaeography in the Digital Age.* Norderstedt: Books on Demand, pp. 309–38, http://kups.ub.uni-koeln.de/volltexte/2009/2978/pdf/KPDZ\_I\_Stokes.pdf.

**Stokes, P. A.** (2012). The Problem of Digital Dating, Part I. In DigiPal (q.v.), http://www.digipal.eu/blog/the-problem-of-digital-dating-part-i/.

**Stokes, P. A.** (2014). *English Vernacular Minuscule from Æthelred to Cnut, circa 990–circa 1035.* D. S. Brewer, Cambridge.

**TEI: The Text Encoding Initiative.** (2014). P5: Guidelines for Electronic Text Encoding and Interchange, Version 2.7.0. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ (last updated 16 September 2014).