ECHOE Corpus Documentation

Paul Langeslag

revision of August 4, 2024

Contents

License

Corpus Definition and Taxonomy

ECHOE: Electronic Corpus of Anonymous Homilies in Old English is a text corpus, native to TEI P5-conformant XML but here additionally supplied in a flattened plaintext transformation, containing manual transcriptions, newly produced for ECHOE, of all extant witnesses of anonymous and Wulfstanian prose homilies predating c. 1200 (excluding the late collections known as the Lambeth and Trinity homilies) and whose dominant language is Old English, as well as long-format anonymous Old English prose saints' lives from the same period. Defined another way, it transcribes all manuscript copies of all non-Ælfrician Old English prose homilies and saints' lives. Following the initial release of its XML source code in October 2024, the corpus is subject to revision on a rolling release basis. Corrections and technical queries may be directed to Paul Langeslag, while Winfried Rudolf as the principal investigator is the project's contact person for strategic matters and

press requests.

The definition of homiletic manuscript articles (hereafter "versions") follows N. R. Ker's 1957 *Catalogue of Manuscripts Containing Anglo-Saxon* for materials known to Ker.[1] Each version is contained in its own file (`018.40.xml`, `018.42.xml`, etc.), which is numbered following Ker's manuscript and article system, i.e. the first element represents a manuscript or fragment sorted alphabetically by location and library, while the element after the period is normally Ker's sequential article number (for longer articles) or letter (for shorter articles) within the manuscript. Materials that have come to light since the publication of Ker's *Catalogue* are numbered in accordance with the same principles but with `x` appended to Ker's manuscript number to allow its inclusion in the alphabetically appropriate position within the system (e.g. 284x for Westminster Muniment 67209, alphabetically between London, Lambeth Palace on the one hand and Oxford, Bodleian Library on the other; the other new entry is 386x for the Taunton fragments). Articles within such newly discovered manuscripts and fragments have been defined in accordance with Ker's principles, with the one difference that articles that have no peers within their fragment are given an article identifier `a` where Ker would have omitted such an identifier: thus 284x.a for the sole article contained in Westminster Muniment 67209, the superfluous article identifier facilitating uniform processing. In Ker's system, a single article may contain the same or similar material twice, first in Latin and then in Old English. Since ECHOE is a language-specific corpus, however, it excludes complete Latin compositions of this kind. In such cases, we have added to the Ker numbering system by (implicitly) affixing `a` to the article number for the Latin and (explicitly) `b` for the Old English. Thus Ker's 49B, art. 29 is the same Wulfstan homily contained twice, once in Latin (Bethurum Xb, *De christianitate*) and once in English (Bethurum Xc, *Be cristendome*); we have included only the latter in our main corpus, under the identifier 49B.29b. The Hatton 113 witnesses to the same text follow a similar pattern: here 331.10a is the Latin (not in our main corpus), 331.10b is an Old English marginal translation of part of the Latin, and `331.10c` is the full Old English text that follows the Latin in the manuscript. Where Ker subdivided his articles, this subdivision too is appended to the article number, as in 186.19a–186.19l for the range of brief compositions in Cotton Tiberius A.iii, or 338.01xix for the last of the sections into which Ker subdivides article 1 of Junius 121.

NOTE

ECHOE **file names** use leading zeroes for both manuscript and article levels in the format `001.01.xml`, but no such leading zeroes are used in any other context, whether it be scholarly reference or `@xml:id`. Thus the ECHOE version contained in `049B.01.xml` is referenced as ECHOE 49B.1, while its sentences are encoded `s49B.1.1`. etc. The other discrepancy is in `336.02-04.xml`, which for ease of processing is referenced as `336.2` in the file's metadata but is best referenced as ECHOE 336.02–04. When processing the files' content, we recommend ignoring the file names while keeping the 336 discrepancy in mind.

At the time of initial release, acceptable images of London, British Library Cotton Otho A. viii, B. x, and Vitellius D. xvii were not available to us, so that the twelve Old English homiletic and hagiographical items therein contained are not at present included in ECHOE. Service disruptions at the British Library in 2020–2024 additionally prevented some final proofing of further London transcriptions.

[1] N. R. Ker, *A Catalogue of Manuscripts Containing Anglo-Saxon* (Oxford: Clarendon, 1957).

Citing ECHOE

Although any representation of an ECHOE version is derived from the underlying XML transcription, most users will want to cite the online edition. Please keep in mind that (1) the data are subject to rolling updates, lending relevance to the date of access; and (2) any one transcription may be rendered with different display settings, so that one and the same version or extract may be quoted in a range of realizations. We recommend users feed their citation managers the following values:

```
@online{echoe,
    title = {ECHOE},
    subtitle = {Electronic Corpus of Anonymous Homilies in Old English},
    editor = {Winfried Rudolf and Thomas N. Hall and Paul S. Langeslag and
    Grant L. Simpson and Charles D. Wright and Susan Irvine and Julia Josfeld
    and Ruby Wai Ying Ku and Hasan M. M. Aldhahi and Sergio Gonzalez Orjuela and
    Julie Kraft and Esther M. Lemmerz and Bente Offereins-Gummelt and Sabine
    Ines Rauch and Carolina Ruthenbürger and Melanie Vollbrecht and others},
    location = {Göttingen},
    publisher = {ECHOE Project},
    date = {2024/},
    url = {https://echoe.uni-goettingen.de},
    urldate = {(date of access here)}
}
```

Depending on your stylesheet, this may yield a citation like the footnote citation here embedded.[2] See the Works Cited at the end of this document for a bibliographical reference following the same Chicago Manual of Style conventions.

Once the web edition as a whole has been cited, specific content may be referenced e.g. as follows:

- (ECHOE 35.4) [no leading zeroes; see note above]
- "Cyrice is þære sawle scip and sceld on domesdæg" (ECHOE 68.11.8).

Even when citing a specific version or other particular content only, we recommend providing only the root URL of the web edition upon first citation rather than deeplinking individual entries. References to particular content may then be incorporated as per the above examples, the pattern "ECHOE 35.4" identifying a version and "ECHOE 68.11.8" referencing a single sentence-like segment.

We recommend that the distributed corpus (i.e. the XML source code repository accommodating the present document) be cited in addition to the web edition only where some feature not available in the web edition is relevant to the discussion, e.g. in natural language processing research or any work evaluating ECHOE metadata directly. You may cite the full repository as follows:

```
@online{echoe-repo,
    title = {ECHOE Repository},
    editor = {Winfried Rudolf and Thomas N. Hall and Paul S. Langeslag and
    Grant L. Simpson and Charles D. Wright and Susan Irvine and Julia Josfeld
```

[2] Winfried Rudolf et al., eds., "ECHOE: Electronic Corpus of Anonymous Homilies in Old English" (Göttingen: ECHOE Project, 2024–), https://echoe.uni-goettingen.de.

```
    and Ruby Wai Ying Ku and Hasan M. M. Aldhahi and Sergio Gonzalez Orjuela and
    Julie Kraft and Esther M. Lemmerz and Bente Offereins-Gummelt and Sabine
    Ines Rauch and Carolina Ruthenbürger and Melanie Vollbrecht and others},
    location = {Göttingen},
    publisher = {ECHOE Project},
    date = {2024/},
    url = {https://github.com/ECHOEProject/echoe},
    urldate = {(date of access here)}
}
```

These data fields may yield a citation like the footnote citation here embedded.[3] See the Works Cited at the end of this document for a bibliographical reference following the same Chicago Manual of Style conventions.

Repository Contents

The ECHOE repository at https://github.com/ECHOEProject/echoe hosts the following components:

- The XML corpus
- A flattened, plaintext transformation
- An ebook of the corpus based on the same transformation settings
- This user manual
- A handful of metadata files extracted from the corpus and used for ECHOE Online
- A `misc/` folder with (1) the ODD containing ECHOE's TEI schema and documentation; (2) several sample XSLT transformation stylesheets to assist users with their own transformation efforts; and (3) a CSS stylesheet allowing users to use the XML corpus as a reading corpus locally in their web browsers, with a fair selection of features.

Please observe that neither the plaintext corpus nor the ebook retains the richness of the XML corpus, as these attempt to render only the readings intended by the most recent scribal revisor prior to c. 1200, but also silently emend readings following the editors' understanding of the text. Punctuation, capitalization, and virtually all metadata are absent from these output formats. The rendering produced by our CSS stylesheet similarly is just one possible way of representing a somewhat wider selection of the data, and limited somewhat by the capabilities of CSS; users may find that spacing in particular is occasionally imperfectly rendered as CSS has no direct way of disregarding internode spacing.

Character Encoding

For its entity declarations, ECHOE relies on MUFI recommendations. Insofar as these extend beyond the standard unicode character set into the Private Use Area, particularly in the realm of punctuation, nonnormalized transformations of the corpus should only be undertaken with MUFI-compliant typefaces such as Junicode. In addition, we encode one letter-shape not currently in MUFI: *ƒ*-shaped *y* <ꝛ> as it occurs in the Vercelli Book, and more rarely in CCCC 41, Hatton 115, and Blickling. This letter-form has been available in

[3] Winfried Rudolf et al., eds., "ECHOE Repository" (Göttingen: ECHOE Project, 2024–), https://github.com/ECHOEProject/echoe.

Junicode 2 since June 2023, but only as an OpenType character variant (`cv50:3` if counting from 1). Accordingly, we have encoded it inline using <g> from TEI's `gaiji` module and defined it in the TEI header using <charDecl>. For it to be rendered mimicking its manuscript form, a stylesheet will have to refer it back to the appropriate stylistic variant (`font-feature-settings: 'cv50' 3`). For most purposes, it may be rendered simply <y> (i.e. <g> may be rendered or transformed without further instructions).

The TEI Header

☞ ECHOE's TEI header follows TEI P5 recommendations and makes substantial use of their provisions for manuscript description. In the present documentation, we will only draw attention to specific resources and conventions referenced.

Conventional Titles

Multiple instances of `fileDesc/titleStmt/title` encode a range of received titles for the composition. `@type="main"` is reserved for what is considered the leading convention, while a series of `@type="alt"` entries may record additional titles, including titles found in manuscripts.

Scribal Hands

`fileDesc/sourceDesc/msDesc/physDesc/handDesc` identifies the scribal hands distinguished in the version using instances of <handNote>, each of which provides (as a value of `@xml:id`) an identifier to which the transcription refers as different hands intervene, as well as a prose description of the hand in question. Where available, hand identifiers follow DigiPal (recognizable from identifiers beginning with `DP`), or else Scragg's *Conspectus of Scribal Hands Writing English, 960–1100* (using the prefix `SC`); new identifiers may be recognized by the prefix `EC` (followed by a number refencing an adjacent DigiPal hand as well as an alphabetical suffix). Where DigiPal assigns a hand to a scribe, we have additionally supplied the attribute `@scribeRef` and used not DigiPal's scribal identifier intended for human processing but instead their database identifier: thus rather than give a value `G.108`, we supply `DP0078`, which may be transformed into https://digital.eu/digipal/scribes/78/ to redirect to DigiPal's scribal entry for scribe G.108.[4] If a scribe's name is known, we encode it as a value of `@scribe`. In the running text, the attribute `@hand` is most notably found on the elements <add> and <del>. Uncertain hands are there marked `uncertain`, and lack header entries.

Any notes not on a specific hand are recorded as a <p> node directly within <physDesc>. As TEI does not permit such a paragraph to follow the <handDesc> node, it precedes the descriptions of specific hands in the document tree.

Dating, Origin, and Provenance

Manuscript dating as recorded under `fileDesc/sourceDesc/msDesc/history/origin/origDate` follows Ker, unless a more accurate estimation has come to the editors' attention. To allow sorting by date of manuscript composition, attributes `@notBefore` and `@notAfter`

[4] This practice departs slightly from TEI recommendations, which have `@scribeRef` point "typically" to a document-internal description.

have been populated on the basis of Francis Leneghan's analysis of Ker's dating system.[5] We recommend using the mean between the two values as the basis for chronological comparison between manuscripts.

Fields `origPlace` and `provenance` have likewise been provided as descendants of `fileDesc/sourceDesc/msDesc/history`.

Sigla

Where available, ECHOE records Scragg's sigla as used in his *Vercelli Homilies*.[6] The siglum of each version's containing manuscript is encoded at `fileDesc/sourceDesc/msDesc/msIdentifier/idno[@type='siglum']`. For manuscripts not in Scragg we have opted to leave this node empty.

Bibliographical References

We use `fileDesc/sourceDesc/bibl` to encode four kinds of reference:

- The Ker identifier (which may in most cases also be reconstructed from the ECHOE identifier for versions known to Ker);
- The identifier from Gneuss and Lapidge's *Anglo-Saxon Manuscripts*[7] where applicable;
- The *Dictionary of Old English*[8] "long" short title;
- The Cameron number, as likewise documented in *DOE*.

As the Cameron/*DOE* taxonomy indexes texts rather than versions, these last two identifiers should be understood as "nearest text" approximations rather than exact matches to the ECHOE version. In addition to the choice of base text, textual boundaries may also differ. For texts edited by Napier and Bethurum in particular, we generally refer to Napier over Bethurum in cases where Bethurum departs from the textual boundaries established by Napier and followed by Ker.

Motifs and Liturgical Occasion

Each version's participation in a set of common motifs is recorded in a sequence of `<term>` nodes under `profileDesc/textClass/keywords`. A master list of motifs may be obtained from the project repository, while ECHOE Online provides the same as a discovery filter.

Insofar as may confidently be inferred from each version and/or its sources and peers, a text's intended (preaching) occasion is recorded under `profileDesc/settingDesc/setting`, and wherever such an occasion is noted, a broad categorization is supplied in `@ana` into one of the values `temporale`, `sanctorale`, and `other`, that last value covering such more broadly understood contexts as baptism, confession, the teaching of Creed and Paternoster, and church dedication sermons as well less clearly understood settings. In hagiographical prose not clearly centred on a specific feast day (e.g. a text on the Virgin

[5] Francis Leneghan, "Making Sense of Ker's Dates," *Proceedings of the Manchester Centre for Anglo-Saxon Studies Postgraduate Conference* 1 (2005): 2–13.

[6] Donald Scragg, ed., *The Vercelli Homilies*, EETS 300 (Oxford: Oxford University Press, 1992).

[7] Helmut Gneuss and Michael Lapidge, *Anglo-Saxon Manuscripts: A Bibliographical Handlist of Manuscripts and Manuscript Fragments Written or Owned in England up to 1100*, Toronto Anglo-Saxon Series 15 (Toronto: University of Toronto Press, 2014).

[8] Angus Cameron et al., eds., "Dictionary of Old English: A to I" (Toronto: Dictionary of Old English Project, 2007), https://doe.utoronto.ca/.

Mary that is not primarily associated with one of Nativity, Purification, Annunciation, or Assumption), this field records just the name of the saint, and no header identifies Michaelmas or Martinmas specifically even though Blickling rubrics do. Of the Marian feast days, the Annunciation is grouped with the temporale, the remainder with the sanctorale.

Methods

☞ ECHOE users may benefit from understanding how the corpus was created. This section offers a brief account of the approach used.

The text itself was transcribed directly into XML by human transcribers, along with much of the basic markup: script color and size, abbreviations, scribal interventions, and some emendations and biblical source references. A template of the TEI header was also manually populated at this time. Transcribers typically worked in teams of two proofreading each other's transcriptions. Transcriptions were subsequently proofread by a revision team consisting of a junior revising editior and a senior revising editor. In most cases, both revisers proofread the whole of the raw XML directly against the manuscript images again, so that the vast majority of transcriptions were subjected to a minimum of three proofreading stages. The final form of the transcriptions reflects the single-pair-of-eyes principle, with the senior revising editor, who has been involved in every decision on transcription and encoding alike since the start of the project, attempting to subject all transcriptions to the same formatting standard at the end of the revision process. This also involved adding further emendations and rejecting some emendations previously encoded, as well as adding and rejecting perceived scribal interventions. Illness prevented the senior revising editor from subjecting a modest proportion of the corpus to a final check ahead of the initial release, but this is expected to be redressed in the near future.

Sentence segmentation and the cross-referencing of cognate clusters of sentences was likewise carried out manually by the senior revising editor in collaboration with the principal investigator. This work predates the development of our "echo tool" for laying bare similarities between sentence-like segments across the corpus, so further connections may yet be discovered. However, segmentation is envisioned to remain unchanged after the initial release, as any subsequent changes to segmentation would render the scholarly citation of specific segments or their deeplinking by other resources unreliable. Accordingly, any cognate material within the corpus discovered in future will have to be cross-referenced with segment boundaries remaining as they are, even if parallel witnesses differ in their segmentation strategies and the similarity of the material may as a consequence remain somewhat obscured.

Explicit word segmentation (tokenization) and the markup of names and numerals were carried out in an automated fashion and then manually corrected. TEI headers were proofread along with the text (i.e. several times), but global changes were repeatedly made to their structure both during and after revision, so not every part of the header has seen equal amounts of proofreading. Here, however, faulty metadata was typically quickly spotted in the proofing environment and corrected as needed. Rhetorical markup was conducted by hand in a separate stage. A range of cross-corpus structural changes to such matters as the encoding of transpositions, glosses, and Romanized Greek were made at a late stage using a combination of mechanical and manual intervention.

The sequence of stages here laid out lacks a dedicated text-critical examination. Generally

speaking, the transcription and first proofreading stages focused primarily on ensuring an accurate letter-for-letter representation of the manuscript text, whereas subsequent revision was carried out with more of an eye to ensuring readings are plausible and grammatical; but at every stage of proofreading the quality of the text was only one among many points of attention. Whereas some of our transcribers and early-stage proofreaders worked with leading critical editions at hand, others did not, and final revision was carried out with reference to textual scholarship for suspect readings only. Time constraints in the face of the amount of text to be processed precluded a more thorough textual analysis, so that the text of ECHOE, though generally well considered, is not consistently as authoritative as that of leading editions. Having said that, we are confident that ECHOE improves on received readings in many places, particularly where text has been erased or the manuscript damaged. For more on our text-critical approach, see Emendation and Conjecture below.

Selected Principles

☞ This section describes key transcription and encoding strategies.

Abbreviation

Abbreviations have been encoded using two pairs of parallel environments as child elements of **<choice>**: **<abbr>** and **<expan>** where the entire word is encoded as an abbreviation, or **<am>** and **<ex>** where a single contraction or suspension is so encoded. An **<abbr>** enviroment will often contain **<am>** as well, and **<expan>** always contains **<ex>**. An effort has been made to restrict the use of **<abbr>**/**<expan>** to the encoding of full words as per TEI recommendations, though there may be some inconsistency in this regard. None of the elements involved carry any attributes.

Scribal Intervention

Simple deletions and additions have been encoded as isolated **<del>** and **<add>** elements respectively, but with considerable metadata. Both identify the intervening hand as a value of `@hand`; **<del>** furthermore has `@rend` to encode the manner of deletion with a value `erasure`, `expunction`, `lettermod`, `overwriting`, `strikethrough`, `underlining`, `unfinished`, or `unmarked`, whereas **<add>** has `@type` with a value `augm`, `subst`, `altern`, or `gloss`, as well as `@place`, encoded as `above`, `below`, `inline`, `onerasure`, `lettermod`, `overwriting`, `tmargin`, `lmargin`, `rmargin`, `bmargin`, or `unfinished`. The element **<subst>** is used without attributes to encode parallel pairs of **<add>** and **<del>** with attribute values to match. Where one or more letters have been deleted alongside a word, we have judged from context whether the deletion is to be considered part of the word or a separate element, as reflected in its **<w>** assignment; thus even an illegible erasure at the end of a now uninflected adjective may be understood as a one-time inflectional ending if the syntax encourages such a reading.

Where a scribe has added a word or phrase as a gloss or alternative to the antecedent reading, this has been encoded following TEI recommendations by giving each reading an identifier and populating a **<standOff>** environment at the end of the file with an **<altGrp>** list of **<alt/>** mappings identifying which readings belong together and by what weighting one is to be preferred over the other. We have defaulted to equal weighting for same-language alternatives but assigned the full weight to the original reading in case of (Latin) glosses. Where the original reading is a single word, the `@xml:id` attribute has been attached to **<w>**,

or else to an enclosing <seg>; the new reading attaches the attribute to <add>. In rare cases (331.10c), more than one alternative reading has been supplied. To facilitate processing these variants without necessarily accessing the <alt> mappings registered in the <standOff> environment, we have used identifiers containing a numeral for the series, followed by an alphabetical for each member of the series, so alt2a alt2b alt2c offer variant readings for the same word, while alt3a alt3b alt3c compete for the reading of the next glossed word. We have used identifiers like gloss1a gloss1b for Latin glosses that could not be substituted in for the glossed word. Since the original reading is always encoded as part of the main text and additionally given an identifier alt (or gloss) + [numeral] + a, while later readings are identified as <add type="altern"> (or <add type="gloss">) and given successive identifiers that start with the same sequence, both CSS and XSLT stylesheets may flexibly be employed to style or filter out these glosses without referring to the <standOff> documentation (see Notes for Further Processing below).

TEI P5 accounts for the documentation and display of interventions aimed at the transposition of words, but not for reordering the affected elements at the transformation stage following the annotations of the intervening scribe. Hence while we have followed the TEI practice of documenting the revised order in a <standOff> environment populated with references back to the affected elements by way of their identifiers, we have additionally facilitated the reordering of elements by (1) containing the group of affected elements in <seg type="transpose"> (with a @hand value attributing agency) and (2) giving each element an attribute @ana with a value like pos1. This facilitates reordering in XSLT transformation (see Notes for Further Processing below).

Emendation and Conjecture

In our text-critical approach, we have tended to favor the manuscript reading insofar as we have attempted to let broadly plausible readings stand even where variants offer superior readings. If, for instance, a single later witness in a larger tradition omits several of the pains of hell, it may seem natural to declare the copy in question faulty due to eyeskip; but if the resulting text is not ungrammatical or nonsensical, we have often retained it, with no more than an XML comment to draw attention to the discrepancy. The extent to which we have adhered to this approach may be illustrated with the following sequence from the Vercelli *Saint Martin*:[9]

> Þā hē ðā sanctus Martīnus þæt geseah, þā arn hē sōna up on þæt hūs and gestōd hē ongēn þām winde. (394.20.67)

In the source as in the variants, Martin confronts the fire, not the wind; this is a clear-cut case of eyeskip from "ongēan þām līge" to "ongēan þām winde" in the following sentence (cf. 382.17.92–93). Since the resulting text remains broadly plausible, however, we have opted not to emend. With the help of ECHOE Online's multiversion view and source analysis pane, users can study the differences for themselves, so we have no need to encode variation into our individual transcriptions where the text is not indefensibly corrupt.

Conversely, following the principle expressed by Michael Lapidge that (to paraphrase rather freely) an uncertain conjecture is better than a corrupt manuscript reading inasmuch as it stimulates thought,[10] we have not been shy to use <supplied> to offer conjectural reconstructions of lost or damaged text, the more so since users are free to display or suppress

[9] For further examples see 310.29.67–68; 310.64.113–114; 394.4.51.

[10] Michael Lapidge, "Textual Criticism and the Literature of Anglo-Saxon England," *Bulletin of the John Rylands University Library of Manchester* 73, no. 1 (1991): 17–45.

such nodes at will. Thus in the case of the badly cut up Westminster Muniment fragment (ECHOE 284x) we have seen fit to follow Ray Page[11] in supplying more than half of the text with the help of other witnesses, as without this material the Westminster text would have lacked any sentence structure. In other cases, we have supplied missing words based on badly faded text (186.19f), a Latin source (182.4), or context (ibid.). However, as one of the largest projects revisiting Old English manuscript readings since the advent of high-resolution imaging, we also offer many damaged readings with confidence and with no such markup where we believed we could make out enough of the letters in question to identify them, while we have enclosed in <**unclear**> any readings where strokes and traces were less firmly attributable. Both <**unclear**> and <**supplied**> encode the cause of uncertainty as a value of `@reason` (`erasure`, `trimming`, `damage`, `reagent`, `fading`, `initial`, `binding`, `omitted`, `loss`), while for <**supplied**> we have furthermore added a value for `@evidence` that is some combination of the values `internal`, `external`, and `conjecture`; these values may be found proposed in the TEI P5 guidelines but have not proved of great value in our project. The value `external` here suggests other witnesses and/or a Latin source contributes to the reading, whereas `internal` suggests internal logic speaks for it; `conjecture` expresses a greater degree of speculation. We have provided up to two space-separated values where appropriate. For examples of recovered and/or conjectural readings, see 130.b, 130.d, 182.4, and 186.19f, as well as isolated passages in e.g. 38.4, 38.37, 182.2a, 196.19l, and 331.43.

Named Entities

Proper nouns and demonyms have been encoded in three categories:

1. **Personal names** have been tagged in the format <**persName** key="Paul">. A list of keys may be found in the project repository, while ECHOE Online provides the same as a discovery filter. Please note that the titles *crist*, *antecrist*, and *farao* have been included among personal names (the first of these identified as `@ref="Jesus"`), but such lexemes as *god*, and *hælend* have not. Names of individuals recorded in PASE[12] carry a `@ref` attribute pointing to what was once the URL of the relevant entry, but PASE has since disabled deeplinking, so a seamless integration of the two resources is not at present possible.
2. **Place names** have been tagged in the format <**placeName** key="#Jerusalem">, with references to fuller descriptions (where necessary) and an approximate geolocation (where available) situated in the <**standOff**> environment at the end of the document. The place name category has been broadly interpreted, including rivers, cities, countries, and continents. ECHOE Online provides a discovery filter allowing the identification of versions mentioning any one particular place name.
3. **Demonyms** have been tagged in the format <**name** type="demonym" key="Egyptian">. The category has been broadly interpreted, including associations with cities, countries, and sometimes families such as Pontius. A list of keys may be found in the project repository, while ECHOE Online provides the same as a discovery filter.

Please note the following complications:

- In view of the fact that Old English often formulates country-scale place names with reference to peoples, we have striven to encode such phrases as *egypta land* as including both a place name and a demonym:

[11] R. I. Page, "An Old English Fragment from Westminster Abbey," *Anglo-Saxon England* 25 (1996): 201–7.
[12] "Prosopography of Anglo-Saxon England," accessed July 1, 2024, https://pase.ac.uk/.

```
<placeName key="#Egypt">
    <name type="demonym" key="Egyptian"><w>egypta</w></name>
 <w>land</w></placeName>
```

- `isra(h)el` has been taken as a demonym where possible, even in the phrase *filii israhel* where we could alternatively have applied the same nesting logic as above, or even interpreted it as a reference to Jacob. In rare cases it is undeniably a place name ("on israele lande") and has been marked up accordingly. We have followed the terminology of our sources in the distinction between Hebrew, Israelite, and Jewish.

Numerals

Numerals have been marked up in the format `<num n="1">`. The number *one* has been excluded where it has been deemed to serve as an adverb (most frequently in the form *āna* "only, alone"), but we have treated it as a numeral in the sequence *nā þæt ān* "not only," where it is properly adjectival. The form *ōþer* (potentially an ordinal "second; other"; sometimes "first" and occasionally "third" in a sequence) has been excluded from numeral markup, and thus treated as a regular adjective "other," where it is not found alongside an indication of sequence, such as *ān* "one" or *forma/ōþer* "first," or in a phrase like *on þone ōþerne dæg*; but where it is part of such a sequence it has generally been marked up even if its natural translation is "other" or "next" rather than a numeral. Numerals include fractions (*tēoþa*: `0.1`; *se þridda dǣl*: `0.33`), cardinals, ordinals (we have not distinguished between these two types of numeral in the markup), and a range of words rooted in numerals, such as *þrynnes* "Trinity" and Septuagesima, but excluded from markup are zero (*nān* "none" and the like), *ānræd* and derivatives, dual pronouns, *betwēonum* "between" whether the object follows or splits the phrase (*be ūs twēonum*), or such month names as September.

Some occurrences of numerals may have been overlooked, as this type of markup was initiated during an advanced stage of the proofreading process.

Script

In the text body only (but not in rubrics), we distinguish between the scripts vernacular minuscule (`vernacMinusc`), Carolingian minuscule (`carolMinusc`), and hybrid (`hybrid`). The manuscripts also regularly contain Romanized Greek for (parts of) *āmen* and of *nomina sacra* that transliterate into "cr" (OE) or "chr" (Latin).[13] In the *nomina sacra* these are always abbreviations and could thus be accommodated in parallel abbreviation environments by transcribing the abbreviation `xpc` and the expanded form `christus` (while additionally encoding abbreviation markers and expanded content). Only *āmen* is not an abbreviation, and so we have dealt with this exceptional case by using a parallel `<choice>` environment with Romanized Greek encoded as `<orig>` and Latin transliteration in `<reg>`. In addition, in both abbreviated content and for *āmen* we have encoded the specific letter sequences that may be called Romanized Greek using `<seg type="romanizedGreek">`:

```
<w><choice>
    <orig><seg type="romanizedGreek">AMHN</seg></orig>
    <reg>AMEN</reg>
  </choice></w>
```

[13] We are aware of Pierre Chaplais, "The Spelling of Christ's Name in Medieval Anglo-Latin: 'Christus' or 'Cristus'?" *Journal of the Society of Archivists* 8, no. 4 (1987): 261–80 but have, for the moment, decided that a corpus like ours is not the place to canonize bold inferences.

In this way, it is possible to configure whether the output should be transliterated or not, irrespective of whether abbreviations are resolved. To transliterate the abbreviated *nomina sacra*, one would only have to replace `x`, `p`, `c`, and `h` (and their uppercase counterparts) in Romanized text nodes with `ch` (or `c`, depending on language and informed opinion),[14] as well as determining how to handle the combination of letter and abbreviation marker `ħ` in `iħs`.

Tokenization

Determinations of what constitutes a word have been made on the basis of the (draft, non-public) *DOE* headword list in conjunction with reasoning inferred from published *DOE* entries. The most straightforward consequences of this policy include our printing e.g. *oþþæt* not *oþ þæt*, *forþām þe* not *for þām þe*, and our treating *helle*-compounds as one word except where *DOE* favors their separation. It also means that where *DOE* recognizes both phrasal and integrated forms of a collocation or phrasal verb (the integrated variant typically associated more with glosses than with continuous prose), we have attempted to apply the same logic, so that we have read *intō* with a place but *in tō* with a person, and *in gān* not *ingān* where possible but not in the instance in 344.5 cited under *in-gān* in *DOE*. We have attempted to force doxologies into the tokenization *on worulda woruld ā būtan ende* where possible, including where scribes have used word spacing that might suggest a different reading, but a few occurrences have necessitated that we follow *DOE* in adopting readings with *āworuld* or *abūtan*.

Accents

Accents appearing over letters or words have been encoded by marking the containing `<w>` nodes with `@notation="accentuated"`. Although this means we have neither been able to specify the nearest letter over which an accent occurs, nor to identify the hand, it is precisely because of the difficulty of ascertaining these particulars that we have opted to encode them in this way. We have attempted to hold to a cutoff date of c. 1200, but in practice these metadata are more likely to reflect the whole period of the texts' active use.

Hyphenation

Hyphenation has been excluded from transcription.

Rhetorical Markup

Within the ECHOE subset of Blickling, Vercelli, and their variant witnesses, we have striven to record the rhetorical structures of binomial pair and enumeration, while instances of audience address have been marked up throughout the corpus. These features have been recorded as `<seg>` nodes with the attribute `@ana` carrying a value `binomial`, `enumeration`, or `audAddress`. (Contrast uses of the same element with `@xml:id`, used to mark up sequences for which scribes have offered alternative readings or glosses, and `@type`, used to mark transposition sequences and Romanized Greek script. Please note that where transposition elements extend beyond the word level, this too involves `<seg>` elements with `@ana` attributes; these can be distinguished from rhetorical markup by their value `pos1/pos2`.)

[14] See again Chaplais.

Audience Address

Various forms of generic audience address, such as *men þā lēofestan*, *brōþor mīne*, and *man*, have been marked up throughout the corpus. Address has also been encoded where it appears within a narrative related by the main voice, because addresses by cited authorities and some biblical characters may functionally blend with the preacher's address of his audience:

- "Ac uton wē, *men ðā lēofestan*, cwæð se hālga Ysodorus, ēaðmōdlīce biddan. . ." (394.27.139).

Binomials

Binomial markup follows formal as well as semantic criteria and may therefore not be consistent by any one criterion. Formally, markup has been provided for any sequence comprising two words of the same part of speech and fulfilling the same syntactic function, especially where they rhyme or alliterate (e.g. *tō āsecgenne and tō āreccenne*, *strang and staþolfæst*). One member may occur in postposition (*ge gōde dæde ge yfele*; *nihtes fyrst oþþe dæges*). Phrases with modifiers have largely been excluded, with the following exceptions:

- Demonstrative pronouns and possessive adjectives if they have the same referent, e.g.:
  - *se ārlēasa and se synfulla*; *þā ēadignesse and þā fægernesse*
  - *þēos mengu and þis folc*
  - *ēowre tēaras and ēowre hrēowsunga*; "Uncra synna hē sworrete and uncra scylda" (32.9.84)
- Adverbs modifying verbs or adjectives if they contribute to the shared meaning, e.g.:
  - "Hē bið *ā wesende and ǣghwǣr andweard*" (382.2.61)
- Adverbs modifying adjectives or adverbs if they are identical:
  - *swīþe ārfæstlīce and swīþe wundorlīce*
- Attributive adjectives if they are identical and the two elements of the binomial pair are particularly closely related (formally or semantically), e.g.:
  - *eallum mōde and eallum mægne*
  - *ēce līf and ēce reste*

Prepositional phrases have only been included when using the same preposition and governed by the same verb, e.g.:

- *ge mid wordum ge mid dǣdum*
- "Hīo ālȳseð þone mannan *fram dēaþe and fram wītum*" (394.3.109).

Semantically, binomials usually feature a close semantic connection, i.e. they may be tautologies (*gescēop and geworhte*), antonyms (*ge cȳles ge hǣto*) or more generally part of the same semantic field (*widewan and stēopcildru*; *lilian and rōsan*). Semantics cannot be judged entirely objectively, however, and sometimes the elements' syntactic closeness has been considered enough to tag them as a binomial:

- "Se ēadiga wer swā gesigefæstod wearð þæt hē þā bysmornysse forhogode *heora lāra and heora costunga*" (344.5.139).
- "And ðās twā ceaftra monna cynne from fruman middangeardes oð worulde ende *timbreð and fylleð*" (32.12.4).

Enumerations

Enumeration markup has been provided for sequences consisting of at least three elements of the same word class which are connected by at least one conjunction and fulfill the same syntactic function. Most enumerations so marked up consist of noun phrases, which may include modifiers. Simple sequences of three or more verbs, adjectives, or adverbs have also been included. Examples:

- "Hie nǣfre *hungor and þurst and cȳle* ne ālǣtað" (32.9.26).
- "þæt wē *magon and mōton and cunnen*" (394.19.49)
- "se ēadiga Sīmeon wæs *sōðfæst and clǣne and godfyrht* on his life" (394.19.42)
- "*ge gōde dǣde ge yfele ge worda gesprecenra ge worca gedōnra ge geþōhtra geþanca and ēarena gehȳrnesse*" (394.4.124)
- "*on godfæder ælmihtigne and on his sunu and on þone hālgan gāst and on þā untōdǣledan þrynnysse and on ðā þurhwuniendan ānnysse*" (394.3.3)

Numbered sequences of any length and sequences of verb phrases with a form of *bēon/wesan* as head have likewise been counted as enumerations. In these cases, the segment tag may have been placed around (a) whole sentence(s):

- "*Þonne syndon þrȳ dēaðas liornode on bōcum. Þæt is þonne se ǣresta dēað hēr on worulde þæt se man mid mānegum synnum oferhealden bið; þonne is se æftera dēaþ þǣre sāwle gescēad and līchoman; þonne is se þridda dēað þæt þā sāwla sculon eardigan on helle.*" (394.11.20)
- "*Þǣr is ēce mēd, and þǣr is līf būtan dēaðe and gēoguð būtan ylden and lēoht būtan þȳstrum and gefēa būton unrōtnyssen and sibb būton unþwærnysse and orsorhnyss būton dēaðes ēge tō libbanne, and þǣr is ēce gesǣlignes mid fæder and mid þām suna and mid þām hāligan gāste ā būton ende, ĀMEN*" (38.35.95).

Binomials within enumerations have only been tagged if they rhyme, alliterate or feature a particularly close semantic connection (e.g. *līchama and sāwle*):

- "*on mǣgenþrymme and on mihte* and on godcundnysse" (394.24.3)
- "*ǣgðer ge on golde ge on seolfre* ge on fela ōðra dēorwurðra þinga" (38.36.105)

Consistency

☞ Several transcription policies have necessarily been aspirational rather than objectively consistent. The most important such aspects are here discussed as a courtesy to the user.

Scribal Intervention

A particularly challenging part of revision has been to find a consistent line on perceived scribal interventions where it is uncertain whether the visible text hides a prior erasure or whether the (main) scribe has altered a letter after first writing (a portion of) it, including the reconstruction of what the original reading may have been or when to count a false onset as a letter. Our transcribers have typically encoded a great deal of detail in these regards, so that the senior reviser's task has often been to find a more conservative line by cutting back on perceived interventions. The below recommendation to ignore minor revisions in the main hand when transforming reflects an awareness that some surplus and/or inconsistent detail may remain.

Large Initials and Majuscule Letters

While we have attempted to mark capitals as uppercase letters based on letter size and/or majuscule form, the former criterion in particular lacks a clear distinction. In rubrics, running titles, and *āmen*, majuscule and minuscule forms often appear side by side; in these cases we have transcribed uppercase or lowercase by majority principle. We have transcribed all Roman numerals in lowercase where they appear in the text body, while in rubrics we have determined their case from context or, where no preceding context is available, by their form. Due to repeated course changes, the treatment of slightly large or majuscule letters in places other than sentence- or name-initially is less than consistent. There are thus several distinct reasons for the below recommendation to case fold the corpus in processing.

Punctuation

Punctuation has been reproduced to the point the MUFI 4.0 character recommendation allows; in some cases one character or combination of characters stands in for one that was otherwise unavailable. We have attempted to discern what hand is responsible for what punctuation, but such attempts are necessarily tentative. That caveat is all the more emphatic for the specific business of describing when and how a main hand's punctuation was altered in a second hand, e.g. turning *punctus* into *punctus elevati*.

Manuscript punctuation may serve a number of purposes and follow systems that cannot always be defined or predicted exactly. For the purposes of ECHOE, we may distinguish between the sort of punctuation that is typically found at the ends of clauses or phrases (generally thought of as serving a function on a spectrum between rhetorical and syntactic) and the sort of punctuation that serves to set off individual (part-)numerals and words, especially one-letter words and abbreviations such as *ā* "always," *ǣ* "law," and *ē* for *est*, but occasionally employed to indicate word boundaries where words have been written together. In the transcription and encoding of punctuation, ECHOE adheres to the following policy:

1. All punctuation characters except intraword instances and abbreviation markers are encoded as <pc>.
2. All *punctus* have been encoded as middle dots regardless of their distance from the baseline, as it would have been impracticable to distinguish between baseline dots and middle dots, let alone intermediate levels.
3. All rhetorical/syntactic punctuation has been transcribed and encoded with a space on either side of the <pc> node (but not at the end of a line of XML).
4. Punctuation intended to set off numerals or one-letter words or abbreviations has been transcribed and encoded immediately adjacent to the element it was intended to isolate, without intervening space. (For most purposes, spacing between XML elements is irrelevant, but intraword punctuation cannot be encoded as its own punctuation element.)
5. It is not ECHOE policy to record punctuation intended to indicate word separation between words written together, but a degree of inconsistency should be expected in this regard as this function cannot always easily be distinguished from rhetorical/syntactic markup.

Sentences

Though the sentence may seem like a natural syntactic unit to readers who have grown up with predictable punctuation, borderline cases may be produced in any language. The

paratactic structure of much of the surviving corpus of Old English makes this problem substantially more complex. Moreover, if sentences could be objectively defined but turned out to be unwieldy for our purposes, then the resulting unit would be of limited value to ECHOE, which seeks to make shorter shorter segments available for comparison. Accordingly, the <s>-unit employed by ECHOE is best understood as occupying a place between the levels of the clause and the sentence, and informed by the following classes of information:

- Syntax
- Content
- Length
- The extent of a corresponding Bible verse or other Latin source
- Direct speech
- Manuscript punctuation
- Manuscript majusculization

A typical unit seeks to cover one syntactic clause or a short sentence. *Inquit*-clauses are set off from direct speech unless this results in very short units or syntactically incomplete sequences: thus "hē cwæð" is not left stranded, but "þā spræc hē and cwæð" has typically been assigned its own segment. The thinking here is that there are use cases for which it may be desirable to isolate direct speech. However, Bible verses as defined today typically include inquit-clauses and direct speech in the same verse, and such Bible formatting is typically mimicked in ECHOE so as to make it easier to compare multiple renderings and echoes of a given Bible verse. Conversely, the division of biblical books into verses sometimes violates conventional content constraints either by including multiple speakers in a single verse or by separating subject and verb in the Latin in a way that would be undesirable in Old English, e.g. by starting a verse with "dicens" with the subject left stranded in the preceding verse. Wherever the biblical division is undesirable for these sorts of reasons, ECHOE reverts to its syntax-first approach and cites the appropriate verses in multiple units, or multiple verses in the appropriate units, where necessary, even if that means multiple verses have to be cross-referenced in the biblical source reference.

A degree of inconsistency should be expected on a number of counts, including shifting insights over the seven years it took to define <s>-units across the corpus and the difficulty of modifying segmentation after the fact once the source reports had begun referring explicitly to sentence identifiers already assigned. Hence the below recommendation to rely on rolling context windows for analyses in which consistency of context window matters. In one case (331.53), it has been necessary to define "Leofan men" as its own unit because manuscript markup marks the transition between 331.52 and 331.53 as optional, and setting it off as such in the XML required that the <s>-node should end at that point.

Tokenization

Where a scribe has substituted one reading for another, we have tried to give each reading its own <w>-tag where possible. Generally this means that if a full word was replaced with another full word, we will have tagged it twice so it could be lemmatized twice if we spotted it during manual review of the semiautomated tokenization process. However, where a modification of a substring of a word resulted in a new reading, it has been technically impractical to mark it up twice without encoding the entire word as having been deleted. The latter approach was nevertheless eventually adopted in the interest of future lemmatization in such cases where material is actually replaced, but not until a late stage in the project.

Accordingly, such cases may be found to have been encoded in three distinct ways: (1) the whole word is marked as deleted even if only a substring was in fact altered, so that both words could be tokenized in full; (2) the deletion and substitution represent parallel word parts (using `<w part="I">`, so they could be separately lemmatized while still accurately representing what letters were modified; (3) the original and new reading are enclosed in a single `<w>` element, on the assumption that a lemmatizer will process either the original or the revised form, not both. Words that have material added to turn them into different words, e.g. the addition of the prefix *un-*, are most frequently contained in a single `<w>` element only and can accordingly not be twice lemmatized or POS-tagged on the basis of the XML corpus without relying on concatenated lemmatization values such as the space-delimited string `snotorlic unsnotorlic`.

Named Entities and Numerals

Please take note of the issues with category boundaries outlined above. Although these types of markup were conducted systematically, they were carried out comparatively late and have accordingly gone through less proofreading.

Incomplete Features

☞ A modest number of features could not be completed by the time of this release. These are described in this section.

British Library Manuscripts

Digitizations of the damaged manuscripts London, British Library, Cotton Otho A. viii, B. x, and Vitellius D. xvii (Ker/ECHOE MSS 168, 177A, and 222) were not available to us during the initial drafting of ECHOE, while the Library's restricted access policies during the succession of pandemic and ransomware attack (2020–2024) further hindered our access to these materials. Given the state of these manuscripts, we determined that it was not worth transcribing them on the basis of microfilm copies; they will be added once high-resolution images become available.

When the ransomware attack brought down BL services, we had recently completed final general proofreading for nearly all London manuscripts. However, some specific work remained outstanding, notably hand attribution. Consequently, Cleopatra B. xiii, Tiberius A. iii, Vespasian D. xiv, and Vespasian D. xxi (Ker/ECHOE MSS 144, 186, 209, and 344) are the main offenders still attributing scribal interventions to `DP0000`, our code for hands yet to be identified. Conjectural readings in the particularly damaged portions of Otho C. i (notably *Evil Tongues*, ECHOE 182.4, but also *Malchus*, 182.2c) as well as the fragment BL Additional 38651 (a Wulfstan autograph) were left for reassessment too long and have been left with editorial XML comments until such time as access is restored.

Biblical Source References

In ECHOE's main drafting stage, the sources of biblical quotations and paraphrases were encoded on a running basis during drafting and/or proofreading. When the drafting of separate source reports (not included in this release) was begun, this earlier practice was put on hold in order to avoid the duplication of work, and it was decided that biblical sources could later be copied in from the source reports. Since source reports were completed

for fewer than half of ECHOE versions by the time of this release, a small proportion of biblically sourced corpus segments continue to go unreferenced.

Rhetorical Markup

The markup of binomial pairs and enumerations was completed only for Blickling, Vercelli, and their manuscript variants.

Paragraphing

An unsystematic start has been made dividing the XML corpus into paragraph-like <ab> ("anonymous block")-segments. These are intended to facilitate reading e.g. in ebook format or in some other user-friendly transformation. Since we have hitherto been unable to allocate time for this task, however, the work remains far from complete and witnesses to the same text should not at this stage be expected to exhibit the same division.

Lemmatization

While lemmatization was never part of ECHOE's initial objective, we have long intended to undertake a follow-up project of this nature. In its anticipation, a very few <w> nodes containing unusual forms have been manually marked up with lemmatization data merely to prevent their being incorrectly assigned in future. This documentation accordingly does not document lemmatization markup.

Notes for Further Processing

☞ When transforming (parts of) the corpus or otherwise using it as a basis for further processing, you may want to keep the following in mind.

General Principles

In our encoding strategy, we have anticipated a range of desired outputs not by encoding normalized, semidiplomatic, and diplomatic transcriptions in parallel, but by enabling a range of renderings through a finer grain of processing instructions, some of which are demonstrated by the display sliders in ECHOE Online. The most notable of the underlying variables are set out in Table 2.

Table 2: Notable variables available for output formatting

Variable	Effect
<sic>	When enabled, this element prints scribal errors as encountered. Exclusively used in a <choice> environment with a sibling <corr>.
<corr>	When enabled, this element prints editorial emendations. Exclusively used in a <choice> environment with a sibling <sic>.
<surplus>	When enabled, this element prints scribal errors whose simple omission improves the reading.
<supplied>	When enabled, this element prints editorial emendations whose simple insertion improves the reading.

Variable	Effect
`<abbr>`	When enabled, this element prints abbreviated forms as encountered. Exclusively used in a `<choice>` environment with a sibling `<expan>`. Toggle its inclusion in tandem with `<am>`, and to the exclusion of `<expan>` and `<ex>`.
`<expan>`	When enabled, this element prints the editorially reconstructed expansion of a scribal abbreviation. Exclusively used in a `<choice>` environment with a sibling `<abbr>`, and a child `<ex>` that may be styled to display which part of the word has been supplied by the editor.
`<am>`	This element occurs not only as a child of `<abbr>`, where it requires no special treatment, but also as a direct child of `<choice>` with a sibling `<ex>`, normally where the editor saw fit to single out only part of a word as abbreviated. Enabling it thus prints abbreviated forms as encountered.
`<ex>`	This element occurs not only as a child of `<expan>`, as described above, but also as a direct child of `<choice>` with a sibling `<am>`, normally where the editor saw fit to single out only part of a word as abbreviated. Enabling it thus prints the editorially reconstructed expansion of a scribal abbreviation.
`<pc>`	When enabled, this element prints medieval punctuation. (Note that in some 250 cases, a punctus <·> has been interpreted as an abbreviation marker instead and thus encoded within `<am>`, which prohibits a child `<pc>`.)
`<seg type="transpose">`	When left as is, transposed material is printed in its original order. To adopt the reviser's intended order, the elements contained must be sorted by value of `@ana` as described under Scribal Interventions below.
Special characters	Special characters may be normalized using substitution functions (e.g. `functx:replace-multi` in XSLT). This concerns glyphs, such as <p>, <ẏ>, <ſ>, and <·>, as well as such entities as `&slongdes;` and `&punctelev;`. To filter out ȝ (*ƒ*-shaped <y>), simply rendering `<g>` elements suffices.
Case	Inconsistencies in the distribution of uppercase characters are best avoided by substituting lowercase counterparts for all capitals.

As should become clear from Table 2, a normalized rendering of an ECHOE version is simply one produced using transformation or styling instructions that render elements like `<corr>`, `<expan>`, `<ex>`, and `<supplied>` while bypassing elements like `<sic>`, `<abbr>`, and `<surplus>`, as well as carrying out a substring substitution function to filter out special characters. ECHOE contains no information on suggested modern punctuation or capitalization, so users looking to normalize face two choices here: fold to lowercase or keep the slightly inconsistent capitalization of the transcriptions, and strip away all punctuation or substitute modern or simplified marks in all those places where the transcriptions encode more pluriform marks. A diplomatic rendering is produced with the opposite settings. A stylesheet should normally render only one of `<abbr>` and `<expan>`, and only one of `<am>` and `<ex>`, but it is reasonable for (semi)diplomatic renderings to offer information

about both original and emended readings using some scheme to visualize which is which (e.g. ~~deleted~~ \added/). It is possible to provide for a range of outputs in a single XSLT stylesheet using parameters and **<choose>** environments with **<when>**-tests for different parameter values.

Scribal Interventions

Individual letter modifications, including instant modifications in the main hand, have been recorded where we felt we could do so with some confidence.

Elements marked for transposition may be rendered in whichever way is desired: they are encoded in their original order and thus will by default be so rendered, but the following XSLT template will reorder the elements as intended by the intervening scribe:

```
<xsl:template match="tei:seg">
    <xsl:choose>
        <xsl:when test="@type = 'transpose'">
            <xsl:apply-templates>
                <xsl:sort select="@ana"/>
            </xsl:apply-templates>
        </xsl:when>
        <xsl:otherwise>
            <xsl:apply-templates/>
        </xsl:otherwise>
    </xsl:choose>
</xsl:template>
```

Punctuation

Punctuation has been encoded with hand attribution, but this is an uncertain business. We caution our users against statistically associating punctuation tendencies with specific scribes on the basis of our data; indeed, for many purposes punctuation (**<pc>**, but also look out for **<am>**·**</am>**) may as well be suppressed in transformation.

Numerals

Where possible, the markup of numerals in **<num>** excludes such adverbial uses as *āna* "only." Because this practice involves a grey zone, and many instances of *ān* may functionally be understood as indefinite articles, you may choose to exclude `@n="1"` from analysis depending on your aims.

Named Entities

In view of issues with category boundaries outlined above, for some purposes it may be worth defining selection criteria that cover a concept across multiple categories, e.g. `match="tei:placeName[@key='#Egypt'] or tei:name[@key='Egyptian']"`. Users interested in New Testament characters named Mary or James will know to study our character conflation policy closely by inspecting the results associated with each key.

Capitalization

Since our scribes may use majuscules or large minuscules in unexpected places, and we have attempted, insofar this may be consistently done, to mimic this practice using capitalization, many applications will benefit from lowercasing all letters as part of the normalization routine.

Style

We have elected to refer our XML documents to a sample CSS stylesheet as a way of accommodating users desiring to use ECHOE as a reading corpus locally in a web browser. This stylesheet declaration should not be seen as a canonical part of the corpus, and may be disregarded in processing.

Transcription and Metadata

ECHOE adheres to a comparatively diplomatic standard of transcription. Included are the following palaeographical and interpretive data and metadata:

Table 3: ECHOE in-text metadata

Aspect	Specification
Letter-forms	æ, ꝺ, ð, ę, œ, ꝛ, ſ, ʃ, þ, ƥ, ꝑ, ẏ, ꝕ, &; other characters normalized (but not ȝ or punctuation, which see below).
Manuscript leaves	<pb/> (page beginning), including manifest reference and canvas index.
Manuscript columns	<cb/> (column beginning).
Manuscript lines	<lb/> (line beginning).
Punctuation	Emulated to the extent MUFI allows.
Scribal interventions	Included up to c. 1200.
Additional scribal performances	Included up to c. 1200.
Scribal hands	Identified using <handNote> and <handShift/>.
Scribal identity	Identified within <handNote> where available.
Script	Identified using <handShift/>.
Language	Identified using @xml:lang.
Biblical echoes	Identified using <quote> with @source.
Textual units	<ab> (anonymous block) as suggested by content and/or scribal annotation; <s> (sentence-like segment) elements informed by factors of syntax, length, content, direct speech, and Bible verse; see further above.
Linguistic data	<w> (word), as yet without further specification but primed for lemmatization (and potentially part of speech) in future.
Named entities	Identified using <name>, <persName>, <placeName>, with since-broken @ref linking to PASE where applicable.
Numerals	Identified using <num>.
Rhetorics	Encoded as <seg> with a categorization as a value of @ana (audAddress, binomial, or enumeration).
Rubrics	Encoded using <head> and, more rarely, <trailer>.

Aspect	Specification
Running titles	Encoded using <fw> (forme work).
Marginal content	Encoded using <note> or <add> as appropriate.

Notably absent from ECHOE's transcription standard are the following palaeographical and codicological data:

- Manuscript damage not interfering with readings (e.g. holes predating the text)
- Scribal word spacing (e.g. the separation of prefixes; the affixation of <ȝ>)
- Letter-forms beyond the ones specified above
- Scribal performances postdating c. 1200 (notably absent is the Tremulous hand of Worcester)

Annotated Indices of Body Elements and Universal Attributes

The alphabetical list of TEI elements in Table 4 offers notes on transcription and encoding policy within the <body> environment on a per-element basis. Excluded are the TEI header (for which see above) and the <standOff> environment (which requires no explanation beyond the TEI Guidelines).

Table 4: TEI body elements used in ECHOE

Element	Usage in ECHOE
<ab>	Paragraph-like chunk determined by content and/or manuscript highlighting. Not at present consistently implemented.
<abbr>	Abbreviated words and phrases.
<add>	Scribal addition; see Scribal Intervention above.
<am>	Abbreviations marker (typically a macron in Old English contexts, even where the manuscript form is more diverse; in Latin contexts we have retained a wider spectrum of markers).
<cb/>	Marks the start of a column (limited to 310.82).
<choice>	Accommodates parallel environments; see under Abbreviation above, and <corr> in this index.
<corr>	Where we have seen fit to emend, we have used parallel <sic> and <corr> elements within <choice>, except where it has been more economical to use <surplus> or <supplied>, which see.
<del>	Scribal cancellation; see Scribal Intervention above.
	Used only in versions with interim and/or closing rubrics, dividing the document into sections to permit multiple instances of TEI's <head> and <trailer> elements.
<ex>	Material supplied by the editor to resolve an abbreviation.
<expan>	The resolved abbreviation in full, comprising both visible and supplied material.

Element	Usage in ECHOE
`<foreign>`	Foreign text (Latin, Greek, or Hebrew), encoded as a value `la`, `grc`, or `hbo` of `@xml:lang`. Where possible, the `@xml:lang` attribute has instead been attached directly to another element, such as `<quote>` or `<s>`.
`<fw>`	Running manuscript header ("forme work").
`<g>`	Encodes a glyph for which there is no MUFI unicode point. The only character so encoded in ECHOE is *f*-shaped *y* (see Character Encoding above.
`<gap>`	Lacuna due to damage, erasure, or modification. We have confined ourselves to the `@unit` values `chars`, `lines`, and `leaves`, while `@reason` may read `damage`, `erasure`, `fading`, `reagent`, `trimming`, `binding`, `lettermod`, `loss`, `unfinished`, `overwriting`, `omitted`.
`<handShift/>`	Registers either the start of a new scribal hand (as a value of `@new`) or a (temporary) shift to a different style of script (`@script`). All transcriptions are encoded after the rubric as starting in `vernacMinusc`, but most Latin quotations are in either `carolMinusc` or `hybrid`. Please note that we have excluded rubrics from our analysis of hand or script. For Romanized Greek see under Script above.
`<head>`	Rubric, with optional attributes `@rend` and `@xml:lang`.
`<hi>`	Encodes manuscript highlighting in terms of script size and color. Uses the TEI-recommended syntax `@rend="size(2.5) color(red)"`, the size being measured in line height counting the space between two rulings. Script size has only been recorded for letters two full lines tall and up. Color identifiers used are limited to `black`, `blue`, `gold`, `green`, `purple`, and `red`, as well as the combinations `black-red`, `black-gold`, `black-green`, `red-black`, `red-blue`, `red-green`, and `green-red` to indicate letters with ornamental coloring. In the compound coloring scheme, the first color is the letter's main color, while the second is ornamental. It should be noted that the colors `blue` and `purple` may sometimes be the result of the degrading of another ink, though we have attempted to record the intended color.
`<lb/>`	Manuscript line beginnings, counting written lines only; line breaks in marginal content are not counted except in such items in which (as in CCCC 41) the complete version is recorded in the margin. Where a line break falls within a parallel environment, we have proceeded as follows: with mutually exclusive content (notably `<abbr>` and `<expan>`), `<lb>` is recorded twice with the same value of `@n`; within `<subst>`, whose parallel contained elements may both be printed, it is recorded only once (in `<add>`, not `<del>`). Runover content (whether of rubrics or folio- or text-finally) has not been marked as occupying an additional line.

Element	Usage in ECHOE
`<name>`	Used only to encode demonyms (`<name type="demonym">`), broadly defined to include nouns (*israhela* in *israhela bearn*) as well as adjectives (*israhelisc*), but also groups like *Jewish* (but not *Christian* or *Pharisee*), family names and geographic cognomens (*Pilate*, *Magdalene*, *Iscariot*, *Nazarene*). Because Old English will sooner refer to *the land of the English* than to a strictly geographical *England*, phrases like *engla land* are doubly tagged with the first word a demonym and the full phrase a place name (cf. `<placeName>` below).
`<note>`	Marginal content with a `@type` value of `marker`, `commentary`, or `unrelated`. Most are cross-shaped markers. Contrast marginal additions, for which see `<add>`.
`<num>`	Numerals, identified as a value of `@n`. Excluded are forms of *ān* with adverbial force (“only, alone”) as well as anything amounting to zero (such as *nān*). Included are fractions (a tithe being `0.1`, a third share `0.33`) and integers, with complex numerals such as *þūsend þūsend* resolved to `1000000` and the like. Where complex numerals are interrupted by non-numerals (*nigon hund wintra and þrītig wintra*), the span goes up to and including the last numeral but cuts off before the remainder of the phrase, so that the cited example would include “wintra and” in the numeral span but exclude the second instance of “wintra.” Also included are words transparently based on and semantically referring to numbers, such as *twifeald* “twofold” and *þrynnes* “Trinity,” but a case like *ānfeald* has been excluded where it means “simple” without relevance to the number one, while words like *betwēonum* “between” have been excluded even in constructions like *be ūs twēonum* (“between us”), and the same is true for month names like *September*.
`<orig>`	Encodes the Romanized Greek manuscript form “AMHN”, in parallel with `<reg>` in a `<choice>` environment.
`<pb/>`	Marks the start of a folio side, and gives the relevant IIIF manifest and canvas index as a value of `@ref`.
`<pc>`	Marks punctuation except where *punctus* have been encoded as abbreviation markers, as TEI does not permit `<pc>` within `<am>`.
`<persName>`	Personal names, as well as a select few titles used in much the same way: *crist*/*christus* (keyed to `Jesus`), *antecrist*, and *farao* (but not e.g. *hǣlend*).

Element	Usage in ECHOE
<placeName>	Geographical names broadly conceived, including the names of settlements, countries, rivers, seas, continents, and Paradise (the earthly and the heavenly grouped under a single key). Because Old English will sooner refer to *the land of the English* than to a strictly geographical *England*, phrases like *engla land* are doubly tagged with the first word a demonym and the full phrase a place name (cf. <name> above). The @key value points to further detail in a <standOff> environment, with approximate geographical coordinates where applicable.
<quote>	Used to set off a Latin or Old English quotation, paraphrase, or echo of a biblical passage (registered under @source). Nonbiblical sources, as well as a discussion of biblical sources, may be found in the separate source reports available at ECHOE Online. By encoding biblical renderings directly in the XML, the corpus allows for guided searches and comparisons of such renderings for specific verses. Please note that TEI describes the element as setting off material "attributed by the narrator or author to some agency external to the text," but we have opted to mark up reused material irrespective of authorial attribution. Where multiple sources are recorded, they are space-delimited, and this may mean either that all were used or that it is difficult to determine which (e.g. which gospel) was used.
<redo>	Expresses that the same intervention was carried out twice, e.g. an interlinear correction followed by an identical inline correction or vice versa. Since functionally this verges on retracing, which we have chosen not to encode, not all scribal reaffirmations are so represented in the transcriptions, and the element only occurs just over a dozen times.
<reg>	Encodes the normalized form of Romanized Greek "AMHN", in parallel with <orig> in a <choice> environment.
<restore>	Encodes deletions that have been undone. Since such cases are in our corpus typically caught up in a substitution operation, we have more commonly had to resort to <undo/> instead. Like all interventions, it carries @hand and @rend, the latter with a value erasure (of the cancellation) or addition (of the textual restoration).

Element	Usage in ECHOE
`<s>`	Sentence-like segmentss, as defined using a weighted consideration of the factors syntax, length, content, direct speech, and Bible verse. Ideally, each unit comprises a longer clause or a short sentence to facilitate the manual comparison of multiple witnesses, biblical renderings, or similar passages. Each unit receives a unique identifier (`@xml:id`) with a value consisting of `s`, the ECHOE version identifier, and the sentence number (e.g. `s394.11.1`), and known connections with other witnesses have been manually encoded by referencing all known witnesses but one to the remaining witness as a value of `@n` with a value starting `x`; thus e.g. `<s xml:id="s190A.b.1" n="x49B.1.1">`.
`<seg>`	Used for three purposes: (1) to add identifiers to sequences marked up for scribal revision (see `<alt>`, and the discussion of transpositions under Scribal Intervention above); (2) to set off Romanized Greek script (see Script above); (3) for thematic markup, using the attribute `@ana` to identify the motif (see Rhetorical Markup above).
`<sic>`	Records an erroneous scribal reading; see `<corr>`.
`<subst>`	Records a substitution of scribal readings; see Scribal Intervention above.
`<supplied>`	Editorial addition.
`<surplus>`	Dittography and other material deemed to harm the reading.
`<text>`	Each file contains a single `<text>` element accommodating the transcription of a single version.
`<trailer>`	Concluding rubric (rare), a counterpart to `<head>`.
`<unclear>`	Letters and passages of which we thought we could see traces, which we have accordingly transcribed, but which we have been unable to identify with confidence. Expect to see it alongside `<gap>` and/or `<supplied>` with the semantic differentiation that the three element types imply. Where we had greater confidence in our readings, we have simply transcribed them as text nodes without marking them `<unclear>`.
`<undo/>`	Registers that a scribe has canceled an earlier intervention, typically a substitution or an addition. For straightforward undeletions see `<restore>`. The element carries `@hand` and `@rend` in addition to a `@target` pointer to the affected element.

Element	Usage in ECHOE
`<w>`	Marks a word token following *DOE*'s headword list. We have attempted to enclose parallel content (MS readings and emendations; original readings and scribal revisions) in separate `<w>` nodes to facilitate their separate processing for lemmatization or POS-tagging, but we have not so enclosed any lacunae unless we were able at least conjecturally to make out a minimum of one letter (in which case we have sometimes had to guess at the length of the word). Partial word substitutions may have been encoded in one of three ways, as described under Tokenization above.

A selection of universal or widely used attributes are noted in Table 5.

Table 5: Selected TEI attributes used in ECHOE

Attribute	Application in ECHOE
`@ana`	Thematic markup (e.g. rhetorical categories); transposition order.
`@notation`	Associated with `<w>` where a scribe before c. 1200 has placed an accent over part of the word (given the value `accentuated`).
`@rend`	Manuscript highlighting; cf. `<hi>` in Table 4.
`@xml:id`	Used for the following purposes: (1) sentence identification, on `<s>`; (2) hand identification, on `<handNote>` in the TEI header; (3) place name identification in the `<standOff>` environment; (4) identification of nodes marked for transposition or in connection with `<undo>` or `<redo>`, or to identify the members of a group of alternative readings or lemma/gloss pairs.
`@xml:lang`	Language; cf. `<foreign>` in Table 4.

Bibliography

Cameron, Angus, Ashley Crandell Amos, Antonette diPaolo Healey, et al., eds. "Dictionary of Old English: A to I." Toronto: Dictionary of Old English Project, 2007. https://doe.utoronto.ca/.

Chaplais, Pierre. "The Spelling of Christ's Name in Medieval Anglo-Latin: 'Christus' or 'Cristus'?" *Journal of the Society of Archivists* 8, no. 4 (1987): 261–80.

Gneuss, Helmut, and Michael Lapidge. *Anglo-Saxon Manuscripts: A Bibliographical Handlist of Manuscripts and Manuscript Fragments Written or Owned in England up to 1100*. Toronto Anglo-Saxon Series 15. Toronto: University of Toronto Press, 2014.

Ker, N. R. *A Catalogue of Manuscripts Containing Anglo-Saxon*. Oxford: Clarendon, 1957.

Lapidge, Michael. "Textual Criticism and the Literature of Anglo-Saxon England." *Bulletin of the John Rylands University Library of Manchester* 73, no. 1 (1991): 17–45.

Leneghan, Francis. "Making Sense of Ker's Dates." *Proceedings of the Manchester Centre for Anglo-Saxon Studies Postgraduate Conference* 1 (2005): 2–13.

Page, R. I. "An Old English Fragment from Westminster Abbey." *Anglo-Saxon England* 25 (1996): 201–7.
"Prosopography of Anglo-Saxon England." Accessed July 1, 2024. https://pase.ac.uk/.
Rudolf, Winfried, Thomas N. Hall, Paul S. Langeslag, Grant L. Simpson, Lena DeYoung, Susan Irvine, Julia Josfeld, et al., eds. "ECHOE: Electronic Corpus of Anonymous Homilies in Old English." Göttingen: ECHOE Project, 2024–. https://echoe.uni-goettingen.de.
———, et al., eds. "ECHOE Repository." Göttingen: ECHOE Project, 2024–. https://github.com/ECHOEProject/echoe.
Scragg, Donald, ed. *The Vercelli Homilies*. EETS 300. Oxford: Oxford University Press, 1992.