



Teaching the Text Encoding Initiative: Context, Community and Collaboration

YASMIN FAGHIHI 

MATTHEW HOLFORD

HUW JONES 

*Author affiliations can be found in the back matter of this article

RESEARCH PAPER

]u[ubiquity press

ABSTRACT

In common with many technical aspects of digital humanities, the TEI has a reputation for being difficult to teach and difficult to learn, with potential practitioners put off by the large and (at first sight) intimidating set of guidelines, the seemingly complex hierarchical structure and the profusion of angle brackets. One-to-one or small group teaching in the context of a specific project is often the preferred method, where the short but steep learning curve required to engage with the TEI can be addressed in a way which is relevant to the aims and experience of the learner. This, however, is not a particularly efficient way of teaching. In this article, the authors discuss their experience of teaching (and learning) the TEI, and how lessons learned in contexts relating to specific projects might feed into the teaching of TEI in a more general setting – the Digital Humanities at Oxford Summer School being the prime example.

CORRESPONDING AUTHOR:

Huw Jones

University Library, Cambridge,
UK

hej23@cam.ac.uk

KEYWORDS:

text encoding; tei; pedagogy;
xml; manuscripts

TO CITE THIS ARTICLE:

Faghihi, Y., Holford, M., &
Jones, H. (2022). Teaching
the Text Encoding Initiative:
Context, Community and
Collaboration. *Journal of
Open Humanities Data*, 8: 15,
pp. 1–14. DOI: [https://doi.
org/10.5334/johd.72](https://doi.org/10.5334/johd.72)

The Text Encoding Initiative (TEI) is, according to its own homepage, ‘a consortium which collectively develops and maintains a standard for the representation of texts in digital form’. It has been a major driver for and influence on text-based digital humanities for over thirty years, and is ‘widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation’ (ibid.). In common with many technical aspects of digital humanities, the TEI has a reputation for being difficult to teach and difficult to learn, with potential practitioners put off by the large and (at first sight) intimidating set of guidelines, the seemingly complex hierarchical structure and the profusion of angle brackets. One-to-one or small group teaching in the context of a specific project is often the preferred method, where the short but steep learning curve required to engage with the TEI can be addressed in a way which is relevant to the aims and experience of the learner. This, however, is not a particularly efficient way of teaching.

The most visible part of the TEI are the guidelines but, as James Cummings points out, the TEI is at heart ‘a community of users and volunteers that produces a freely available manual of regularly maintained and updated recommendations for encoding digital text’ (Cummings, 2019a). Even at a comparatively early stage in its history, Elli Mylonas and Allen Renear argued that the TEI’s dual functions as an interchange format and data description language ‘pale before’ its role as a research community – whose subject matter is no less than what is stated in the following sentence:

‘textual communication, with the principal goal of improving our general theoretical understanding of textual representation, and the auxiliary practical goal of using that improved understanding to develop methods, tools, and techniques that will be valuable to other fields and will support practical applications in publishing, archives, and libraries’ (Mylonas & Renear, 1999).

To this, we might add that TEI is ‘an analytical framework for deep reading’. Certainly, our experience of teaching the TEI is that it is something to engage with rather than ‘learn’ – and that the ‘something’ is not just the practice of encoding with TEI, or even the current community of users of and contributors to the TEI, but also those that have gone before them. Cummings again: ‘it formalizes a history of the community’s concerns for textual distinctions and exemplifies understandings of how to encode them and how these have developed over its existence; it acts as a slowly developing consensus-based method of structuring those distinctions’ (Cummings, 2019a). If you were to write a history of text-based digital humanities in the late 20th and early 21st centuries, the TEI guidelines would be a good place to start.

Twelve years ago, an article was published which highlighted some of the problems encountered in teaching the TEI (Terras *et al.*, 2009) – the extensive and general nature of the guidelines, the broad range of uses and users, and the sometimes difficult connection between the theory of TEI (and text encoding) and how it operates in a project or institution. The authors proposed a series of online learning resources that would teach the TEI by example – in the context of specific fields or activities.¹ Susanna Allés-Torrent and Gimena del Rio Riande have gone on to point out that in addition to the specific requirements of projects and materials, TEI teaching and learning is also affected by different cultural and linguistic contexts (Allés-Torrent & Riande, 2019). It seems that when teaching the TEI, context is everything, which poses particular challenges for the ‘general’ TEI course, online or in person.

The writers of this present article have been teaching the TEI in a variety of contexts for the past 12 years – both in relation to specific projects or sets of material (the *FIHRIST Catalogue of Islamicate Manuscripts*, *Medieval Manuscripts in Oxford Libraries*, *Cambridge Digital Library*) and in a more general setting – the Digital Humanities at Oxford Summer School being the prime example. In this article, we will use four case studies to talk about our experience of teaching the TEI. We will then outline some general conclusions on what it means to teach the TEI and discuss how some aspects of teaching in a specific learning environment might feed into the teaching of the TEI in a more general context.

1 <https://teibyexample.org/>.

CASE STUDIES

1. MANUSCRIPT DESCRIPTION AND FIHRIST

FIHRIST is a union catalogue incorporating descriptive manuscript data from nineteen institutions in the UK and Ireland.² It uses TEI as its data format, with descriptions created by contributing institutions either as part of projects or on an ongoing basis. Having begun in 2009 as the *Oxford and Cambridge Online Islamic Manuscript Catalogue*, an online catalogue for Arabic and Persian manuscripts held at the Bodleian and Cambridge University Library, it secured further funding in 2011 to broaden its scope and become a union catalogue for manuscripts from the Islamic world. The name ‘Fihrist’ simply means ‘catalogue’ in Arabic, Persian and Turkish and was inspired by the famous 10th-century work by *Ibn al-Nadim*, the *Kitab al-Fihrist*. Countless specialised studies have used the FIHRIST as a source of data. Because it includes the titles of a large number of works that are no longer extant, as well as biographical information on little-known early authors, it throws light on otherwise obscure facets of mediaeval Islamic intellectual history in many fields (Stewart, 2007). Moreover, *al-Nadim’s Fihrist* presents a complex structure of organising information, which requires a non-linear, (i.e. not exclusively chronological) reading of the work. By embedding gaps, he introduces the idea of an eternal work in progress. Our aspiration when creating FIHRIST was to include descriptions of all manuscripts held in the British Isles belonging to the widely defined Near and Middle East, Africa and the various parts of Asia which relate to Islamic and Eastern Christian cultures by language and script. By implementing a detailed data-model and encoding standard, we envisaged the potential of working with the data on various levels beyond discovery and access. The technical infrastructure is maintained by the Bodleian Libraries in Oxford, and the Editor Yasmin Faghihi, at Cambridge University Library, directs work on the content. Oversight is provided by a board whose remit includes advice and training on the creation of TEI descriptions. A fundamental aspect of teaching the TEI for FIHRIST is the fact that it is a work in progress, both in terms of the scope of the content and the nature of the descriptions themselves, which evolve in response to new research approaches and new materials.

The decision to use TEI as a data format was based on a number of factors. Initially, our prime goal was to work towards a sustainable solution for creating descriptive data for manuscripts that not only facilitated access to collections, but also provided the groundwork for a digital infrastructure which would generate further research. Manuscript descriptions from this (loosely defined) corpus, had until then been locked in library card catalogues and/or out-of-print monographs, which had been surpassed by evolving metadata standards for cataloguing other materials such as printed books and journals. The complexity of manuscripts as research objects demanded a more flexible and extensive metadata format. The focus of most online public access catalogues has been on creating machine-readable cataloguing records for standardised bibliographic descriptions including basic physical features in accordance with established cataloguing rules. Traditional manuscript catalogues on the other hand often consist of a series of paragraphs of prose, which include historical information (dates, relationships), analytical information (contents, identities of people, script styles) and structural information (such as size and material), often in no prescriptive order. While largely consistent within single catalogues, the organisation and rendering of this information varies significantly across catalogues.

The inclusion of manuscript collections into digital library catalogues had fallen into the gap that divided the text-based format of humanities research from the highly structural approach of library and information science. The evolution of Digital Humanities and the TEI, especially the introduction of the Manuscript Description module,³ presented a new opportunity. As a community-based and text-focused standard, TEI was the most cutting-edge metadata format for transforming primary research on manuscripts and manuscript collections into computable data without losing the integrity of its textual features. The manuscript description module (called msDesc) offers a wide range of relevant descriptor codes, which facilitate a structural approach to describing the text-bearing object in detail. These include top-level elements containing content related information, physical description, history, additional information and more, with a range of sub-elements available within each. Our decision to embed the

2 <https://www.fihrist.org.uk/>.

3 <https://tei-c.org/release/doc/tei-p5-doc/en/html/MS.html>.

msDesc element into the header rather than the body of the TEI document also opened up the inclusion of transcriptions at a later stage. The starting point for the FIHRIST data model was a schema for the description of Islamic manuscripts which had been customised from the ENRICH⁴ project for describing Western Medieval manuscripts (Pierazzo, 2010).

Amongst the challenges of teaching text encoding for FIHRIST is a constant engagement with the TEI on three conceptual levels: data modelling to maximise the encoding of research data, the integration of existing library standards for data compatibility and export, and developing best practice for training, reuse and documentation. To meet these challenges, all manuscript descriptions in FIHRIST are created in the raw TEI/XML using the Oxygen XML Editor, which allows for direct engagement with the process of creating and modelling data. Data modelling and the development of best practice were extremely important in the initial experimentation phase, and these processes feed directly into the teaching and learning of TEI for contributors to FIHRIST.

The teaching of the TEI for FIHRIST fell into four broad stages. In the initial phase, core project members learned the TEI principles and practices at the same time as developing the data model, an iterative process which is fundamental to many TEI-using projects. At the start of the project, no examples of descriptions for Islamic manuscripts in TEI existed, so the process of learning went hand-in-hand with the process of modelling. Following a two-day general introduction to the TEI, learning was mainly autodidactic with project members exploring the guidelines alongside the manuscript data. The initial learning stage of working with XML, the Oxygen Editor and the TEI guidelines, without concrete examples for best practice, was a steep learning curve. A year later more expertise was gained when Faghihi attended the TEI strand of the Digital Humanities at Oxford Summer School, which we now convene.

As FIHRIST expanded to become a union catalogue, the demand for TEI training for contributing institutions grew. Teaching mainly occurred in one-to-one or small group settings, dealing directly with the materials at hand, and learning through project work rather than through exercises. Teaching and learning continued to go hand-in-hand with data modelling, with new materials and approaches requiring adjustments to the schema and to documentation. While time-consuming, this approach proved very effective as a learning method, allowing us to teach in a real-world context, and cementing the communal and collaborative network on which FIHRIST relies.

A third phase followed a major data-consolidation project at Oxford, when the wider infrastructure for hosting various other manuscript catalogues (including the Medieval Manuscript Catalogue) was established and improved by a generic schema for all TEI catalogues, along with a more complex and consistent approach to authority files. Data and code were now hosted on GitHub,⁵ and the editing supported by use of the GitHub client. This new phase of development was preceded by a major data clean up, correcting the many idiosyncrasies and inconsistencies generated in the initial phase and setting new parameters for improved practices. As there were now a significant number of contributors to FIHRIST, a new approach to teaching and learning was required. Training was delivered in a series of structured workshops where the creation of TEI descriptions, with a particular focus on use of the authority files (lists of standard forms for certain entities in the data such as names and works), was embedded in a complete workflow involving collaborative working with GitHub. The use of the GitHub client along with the Oxygen XML Editor for managing and committing files, and the use of GitHub functionality for raising issues and resolving conflicts were included, ensuring better transparency and improved communication. Efforts were made to consolidate training documentation and to encourage the sharing of knowledge across the network.

With most contributors now either competent TEI users or with access to support in their own institutions, current TEI support for FIHRIST mainly occurs in a more ad hoc and collaborative form – through GitHub itself and through an email list. Issues arising from new materials or new approaches to description are raised and discussed, sometimes leading to alterations to the data model and documentation and/or further enquiry on the main TEI discussion forums.

⁴ <https://digital.humanities.ox.ac.uk/project/enrich>; <http://www.manuscriptorium.com/en/tei-p5-enrich-schema-en>.

⁵ <https://github.com/fihristorg/fihrist-mss>.

In this sense, the constant interplay between the process of learning and data modelling in response to real-world challenges arising from both materials and research continues to enrich our approach to manuscript description. The results of these reiterative processes generate further content for training new contributors, nonetheless by creating explicit examples. Interpersonal teaching, however, remains key to ensuring new users acquire conceptual skills of converting traditional descriptions (or approaches to descriptions) into computer readable (TEI-) data, in accordance with an approved best practice. Moreover, the growing community of experienced users generates increased capacity for training, reflecting the collaborative and reiterative potential of TEI.

Our experience in teaching the TEI for FIHRIST has produced a number of general conclusions about learning TEI for manuscript description. While the TEI has a reputation for being difficult to learn, most elements and attributes (computer tokens) include natural language components, for example `decoDesc` for Decoration Description or `history` for history etc. and are intuitively understood or easily memorable. Certainly, in an English-speaking context, this proved to be a major advantage, especially when teaching students unfamiliar with coding or other mark-up languages. It was nevertheless essential to begin with an introduction to the rules governing XML, such as well formedness, nesting, the role of elements, attributes and values, and the functionality of the XML editor.

However, one of the complications of teaching TEI for manuscript description is that technical skills and a knowledge of XML are not all that is required. A combination of technical, language and subject skills is required to create descriptions, especially if creating descriptions from the object in hand, when expertise in manuscript studies including codicology together with knowledge of the language is essential. A background knowledge of skills and standards more commonly associated with libraries and archives is also important. One of the challenges of creating descriptive data for items held in research library catalogues, regardless of the prime objective to build an independent union catalogue, is to ensure data compatibility. In other words, we must bridge the gap between applied library standards for key components such as names, titles, transliteration, subject keywords and the idiosyncrasies of extant legacy data. One of the great benefits of the TEI is its inclusiveness, not only in scope through the extension of existing modules and adoption of new ones, but also in the capacity to incorporate existing standards. Thus, we were able to adhere to and include library standards for names, subjects, languages, and the other key components, and include identifiers and links to external authority files. As a result, an introduction to relevant standards is vital before proceeding to show how some of these are encoded within each manuscript record. An example familiar to most users was the application of the Library of Congress Subject Headings commonly used in UK and US libraries and beyond.

Adding provenance information to each record is a vital component of data credibility. As contributors are instructed to encode institutional information in the description of the electronic file nested under `fileDesc`, a mandatory component of `msHeader`, each individual is encouraged to encode their name and details of their contribution in a separate element in the revision description `revisionDesc`. In cases where the manuscript description is born digital or substantially enriched from legacy data, the contributors are taught to include their names in the source element of the manuscript description `msDesc`, where the intellectual content is directly credited to the person conducting the research. In our teaching, we point out the significance of data provenance in relation to data reuse and that by observing this practice, the name of the record creator will become an integral part of the provenance of both research and data.

An integral part of the new set up was the visibility and accessibility of all relevant data in GitHub. Cataloguers were now able to edit their data and contribute directly to the repository. With the transfer of data and the new practice of communication, a new workflow was introduced which required additional training. To compensate for the lack of a data-inputting tool, an XSLT conversion template was produced, which generated a HTML preview of the TEI document and proved very helpful for instant proofreading. As a result, teaching included not only the theory and practice of TEI and how to encode new aspects of manuscript descriptions, but also required training in using the various tools (GitHub client, XSLT conversion) in order to ensure that a consistent workflow was observed. As communication was also transferred to

GitHub, reducing the traffic on the mailing list, engagement with the tools became a core part of contributing to FIHRIST.

In conclusion, the TEI training for FIHRIST was and remains a hybrid approach. As with many Digital Humanities methods, a multitude of skills and experience is required to arrive at the successful creation of a manuscript description, and these must be taught in tandem with the technique of encoding itself. Teaching manuscript description in TEI is most effective if set in the context of a use case and adjusted to the skill set of the user. When working from legacy data with some prior experience in encoding, cataloguers often manage to become self-sufficient learners after an introductory session, relying on documentation and working by example with occasional support. However, when creating TEI records as the direct result of research conducted on the manuscript, training is an ongoing and iterative exercise. This approach of examining the manuscript in parallel with testing the boundaries of the TEI to encode the results has been a learning experience for both the teacher and the student, with benefits for the wider enterprise and for the TEI community as a whole.

2. MEDIEVAL MANUSCRIPTS

An important arena for teaching TEI in the Bodleian Libraries is as part of ongoing work on the manuscript catalogue, *Medieval Manuscripts in Oxford Libraries*.⁶ This is a union catalogue of Western medieval manuscripts in the Bodleian and a number of Oxford college libraries. It consists of manuscript descriptions and authority files encoded in TEI according to a project-specific schema. As with FIHRIST, the catalogue data is stored in GitHub,⁷ which also provides a relatively simple process for data submission and review. TEI editing is carried out using the Oxygen editor,⁸ for which the University of Oxford has an institutional licence.

Training typically takes place as part of specific projects for catalogue enhancement hosted both by the Bodleian and by individual Oxford colleges. These have focussed on retroconversion of existing printed catalogues and on the enhancement of catalogue records for manuscripts about to be digitised. Retroconversion projects have been led by Merton, Exeter and St John's colleges; the main recent digitization project has been Manuscripts from German-Speaking Lands.⁹ The project officers are typically archivists or librarians at an early stage in their careers who have domain expertise in medieval manuscripts and cataloguing but do not have extensive pre-existing experience with TEI or markup. Training is provided by staff at the Bodleian and aims principally to enable the project officers to fulfil the requirements of the particular project. It is, nevertheless, the aim of the Bodleian Libraries and colleges that project officers should receive a good grounding in XML and TEI as a transferable skill and part of their professional development.

Most project officers are not familiar with Oxygen or GitHub and require initial training in the use of these interfaces, which is provided via a combination of written documentation and one-to-one training. Initial training in TEI encoding itself covers the essentials of XML (elements, attributes, well-formedness, the schema) and the broad structure of the TEI msDesc module. This is again provided on a one-to-one basis with intentionally high-level documentation.¹⁰ The full set of project encoding guidelines is made available for reference but no attempt is made in training to go through every eventuality it covers. Officers are usually then ready to begin independent encoding, usually working from a template pre-populated with basic metadata such as institution and collection details, bibliographical information on the printed catalogue in the case of retroconversion, and with key elements of the msDesc module in place (two similar templates are available in the msDesc repository, one for manuscripts of a single codicological unit and one for multi-unit manuscripts). They are encouraged to raise questions via email and their work is reviewed particularly thoroughly for the first few weeks. After review (usually via GitHub) the contributions are merged into the master branch and indexed for display on

6 <https://medieval.bodleian.ox.ac.uk/>.

7 <https://github.com/bodleian/medieval-mss>.

8 <https://www.oxygenxml.com/>.

9 <https://hab.bodleian.ox.ac.uk/>.

10 <https://github.com/msDesc/medieval/blob/master/quick-start.md>.

the project website. The web interface is updated weekly and officers are able to explore their records in the context of the whole catalogue.

It is difficult to generalise about the aspects of encoding that new officers find challenging. One area that often requires particular attention, however, is the transcription of text from manuscripts ('representation of primary sources' in the TEI guidelines). This is an area where the original TEI guidelines can seem particularly complex, for example in the variety of potential solutions for encoding abbreviations, and where care needs to be taken in the use of distinct but related elements such as *damage*, *gap* and *supplied* to correctly encode missing or damaged text. The somewhat simplified project guidelines can still be confusing at first.

It must be emphasised, however, despite the existence of detailed project guidelines, that project officers are not simply passive encoders. Due to the variations in scholarly practice in the description of manuscripts, the encoding of a manuscript description or the retroconversion of a printed manuscript catalogue will often raise issues that have not yet been addressed in the project guidelines, or to which better solutions may emerge. The project guidelines are regularly updated, and officers thereby gain a sense of the TEI as an evolving community of practice rather than a fossilised collection of rules.

Teaching TEI in the context of a single active project has several advantages. It undoubtedly simplifies the training process. Half a day's one-on-one training is typically sufficient for a project officer to begin independent encoding. Contributing to an active project is, in addition, motivating. Seeing their work publicly available, and seeing how their encoding enables search and browse functionality, is a powerful incentive for the effort spent in encoding. To a significant extent officers learn by doing, but they do so in the context of a detailed project schema which provides immediate feedback for most encoding errors. They are also shown how to use an XSLT transformation within Oxygen to generate an HTML preview of a catalogue entry. Experience shows that this is essential for proofreading and for understanding the relationship between markup and the final display of the record.

Teaching TEI in this way does mean that project officers become familiar in detail with only a part of the TEI; indeed, they are working with reference to project guidelines rather than the TEI P5 guidelines. Are they indeed 'learning TEI' in the fullest sense? We would argue that they are. 'Learning TEI' rarely, if ever, means learning *all* the TEI; more usually it means learning what is relevant in a particular context, while, ideally, gaining a sense of other contexts in which TEI is used, ideally, learning that the TEI can (and should) be customised, and learning that the guidelines change and evolve. In addition to the manuscript description module, officers do also become familiar with several other aspects of TEI: the header; the modules for names, dates, people and places; certainty, precision and responsibility; and with the representation of primary sources (since almost all cataloguing involves transcriptions of text from manuscripts). Through a close focus on one particular aspect of TEI officers leave with a solid grounding in 'the TEI' more generally which can be taken on to other projects.

3. DAVID JONES

In June and July 2021, Jones and Faghihi co-convened two week-long workshops which taught text encoding and the TEI through the work of the poet and artist David Jones (1895–1974). Both were hybrid workshops with the instructors physically present (with access to the material archive), and participants joining online. Both were co-taught by DH practitioners and literary scholars – Faghihi and Jones from a DH background and Laura McKormick Kilbride, Tom Berenato and Anna Svensden on the literary studies side. Each was successful in ways that were interesting and new to all involved – participants and teachers, literary scholars and DH practitioners. In this section, we will describe the organisation of the workshops, and what they taught us about learning and applying text encoding and the TEI in a tightly-focussed research context based on a limited corpus of material.

The first workshop, sponsored by the Cambridge Humanities Research Grants scheme and organised through Cambridge Digital Humanities, was held in Cambridge and concentrated on Jones's correspondence with Jim Ede, the owner and curator of the Cambridge art gallery called Kettle's Yard. There were 20 participants from a variety of backgrounds (students, researchers, librarians, archivists), and while there was no requirement for any knowledge or experience

of text encoding or the TEI, there was an approximate 50-50 split between those from a DH background and those who approached the workshop through an interest in David Jones. In preparation for the workshop, Jones's correspondence with Ede had been digitised, high resolution images made available on Cambridge Digital Library, and course materials uploaded to a CDH Moodle site. The week-long workshop began on Monday morning with a virtual tour of Kettle's Yard, followed by a session on Jones and his work. In the afternoon, we embarked on an intense three-hour introduction to text encoding and the TEI which left participants and teachers alike a little exhausted. Then each participant was assigned two letters, and given outline TEI records to work with, and unleashed on the business of text encoding.

For the following three days, participants were expected to work semi-independently. There were two hour-long drop-in sessions each day (one technical, one on Jones and his work – though in practice these boundaries blurred), and participants had access to an online forum to discuss issues and to access course materials for reference and support. Completed work was submitted on Thursday afternoon and uploaded to the Digital Library test site for a public event on the Friday afternoon. This broader event showcased Jones and his work, the physical archive, the workshop itself, and the published outputs.

The second workshop followed a very similar format, though we adjusted the schedule to spread the intense TEI learning over the first two days. This was followed by independent working and drop-in sessions, and finished with a public event and the publication of the workshop's outputs. For this workshop, the focus was on a single draft of an unpublished poem *The Book of Balaam's Ass* from the collections of the National Library of Wales – an abandoned manuscript, fragments of which appear in Jones' final book, *The Sleeping Lord* (1974) (Berenato, 2021). The draft was divided up into manageable chunks, and each participant was allocated five pages to encode. The participants for this workshop were heavily skewed towards literary scholarship, being mainly researchers with existing expertise in Jones and his work. Again, the physical archive was an important part of the introductory sessions and the public event, with Jones's *Cara Wallia Derelicta* serving as the backdrop on both days.

The workshops were very successful, not just in terms of the learning and material outputs but also in the creation of a sense of community and a strong group dynamic – something which we did not necessarily expect in the hybrid format. We learned some significant things about text encoding, the TEI, and what it means to teach them.

The first point that emerged was the role of text encoding not just as a route to publication or analysis, but as a form of close reading. As Kate Singer points out '... encoding—and teaching encoding—might be a valuable pedagogical tool, to enhance 'close reading' and, additionally, to refocus reading as method of evaluative labeling, categorization, and selection of discrete bits of text' (Singer, 2013). The focus on a discrete set of materials, the co-teaching of the workshops by experts in the material alongside DH practitioners, and the sense that the workshop would produce real published outputs (rather than just exercises) meant that we were teaching text encoding more as a framework for the exploration of the text than as a technical standard. The application of the TEI brought about new insights into Jones's work and his working process – for instance in the classification of his editorial symbols -- in the attempt to identify and order different acts of writing, and in the movement from prose to poetry in the text. Participants and teachers alike were surprised at the extent to which text encoding generated new perspectives on material which some of them had been studying for many years.

The second point was the success of the hybrid format in generating the collaborative community approach that is fundamental to the TEI. This was facilitated by the mixture of literary scholars and DH practitioners in the teaching team – providing an immediate example of interdisciplinary collaboration. Also significant was the idea that the workshop was not a series of exercises, but a process which would produce useful and published outputs which would be of concrete benefit to Jones studies. Working collaboratively on something real fostered a strong communal and collaborative spirit. The emphasis on drop-in sessions rather than formal taught sessions also helped to generate a sense of everyone working together – in particular, the very lively sessions where the group attempted to decipher Jones's handwriting. The comparatively low contact time did not seem to hinder the participants, who made effective use of the Moodle forum to work collaboratively in a hybrid space. We learned from

the intense and rather exhausting TEI introductory session on the first day of the first workshop, and spread these sessions out over two days for the second.

The third point was the framing of text encoding as an approach or methodology independent of its particular manifestation in the TEI. Participants and teachers quickly moved to the position of text encoding as a way of thinking about texts – and an approach which could be adopted usefully as a thought experiment as well as in its practical application. As one of the workshop participants said, for gaining a deeper understanding it would be worth encoding a text even if you immediately threw the results of your efforts away.

Finally, the most surprising element of both workshops was how participants with little or no knowledge of text encoding or the TEI on Monday were able to produce TEI documents fit for publication (with some editing) by the end of Thursday. When planning the workshops we felt it was important to put forward the idea of publishing the outputs as a challenge, but were concerned that we were being too ambitious. As Fukushima and Bourrier point out ‘... it is much easier to give a student a B+ on a paper that shows effort and promise than it is to send back multiple TEI letters that show effort and promise but must be perfect for publication’ (Fukushima and Bourrier, 2019). We are not entirely sure why the participants were able to pick the TEI up so quickly, but we think it has something to do with directly engaging with the texts and materials at a very early stage in the learning process, with the emphasis on independent working in a collaborative framework, and with the fact that real work was being produced rather than merely practice exercises. Concentrating on specific texts and materials allowed us to focus on a relevant subset of the TEI in a way which effectively reproduced use of the TEI in real project workflows. Equally encouraging was the number of participants who expressed an interest in continuing to work on the materials after the workshops had finished.

Apostolo, Börner and Hechtel (2019) write persuasively about the challenges of teaching using real materials rather than practice examples and the need for specialist knowledge and the sometimes daunting and complex nature of archival materials. However, in our experience, the use of real material in the context of a live project, working in a collaborative environment and with access to subject specialists, is the fastest route to arriving at their concept of the aim of a genetic edition, which is ‘... understand[ing] what might have happened in the author’s mind while he or she was writing’ (Apostolo, Börner, & Hechtel, 2019). It was this feeling of close engagement with the author and the writing process which provided the main inspiration for fast and effective learning of the TEI.

4. DIGITAL HUMANITIES AT OXFORD SUMMER SCHOOL TEI STRAND

All the authors of this article have a long and varied history of engagement with the TEI strand of the Digital Humanities at Oxford Summer School (for a comparative case study of the development of the Summer School see Cummings, 2019b). Jones attended as a learner in 2009, when it was still the TEI Summer School, and in the same year, Faghihi presented FIHRIST (then in its infancy) as a test case. Faghihi herself attended as a learner in 2011, and Jones and Holford have each at different times run the manuscript description session of the strand. In 2019 the three of us took over as co-convenors of the strand, running one in-person Summer School in 2019 over five days before COVID brought two cut-down online versions in 2020 (one hour) and 2021 (three hours).

The main challenge of the Summer School lies in the cohort, which consists of around thirty participants from a variety of backgrounds (research, libraries and archives being the most common), a variety of disciplines, and with varying levels of technical experience. Some come with a general interest in the TEI, and others have very specific projects or activities in mind. The course is currently filled on a first-come-first-served basis with no prerequisites or selection criteria, meaning that we have to assume that everything needs to be taught from scratch. Furthermore, to provide a useful foundation the course needs to cover not only encoding but also analysis, publication and schema consolidation. In this context, even the full week of the in-person event does not seem like a very long time to get to grips with the TEI.

Our first consideration when we took over the running of the strand was to establish what we could realistically hope to achieve given the nature of the cohort and the timeframe. We wanted to make a distinction between text encoding as an activity and the TEI as a way of doing text

encoding. We wanted to concentrate on outcomes that made sense to learners: description, transcription, publication and analysis. We wanted to give them enough technical background to make them self-sufficient learners of the TEI without getting too bogged down in the code. We wanted to introduce the TEI as a living community of practice and practitioners of text encoding. In addition, more than anything else, we wanted to show how text encoding and the TEI could be relevant (and transformational) to the ideas and projects of the participants.

Starting with the in-person event in 2019, first we asked group members about their interests and ambitions for the course in order to illustrate the diversity of the field, and to help us to make the content more relevant. Our first exercise was to encode physically a passage of text using marker pens. Our aim here was to engage with text encoding as an activity outside of any particular technology or standard – a process Victor Del Hierro takes to its logical conclusion in writing out TEI tags in biro onto a physical copy of a poem (Ives *et al.*, 2013). We also wanted to reassure the less technically experienced members of the group that the main principles of the course – the *point* of text encoding – would be still relevant to them even if they struggled to get up to speed with some of the encoding itself.

We then embarked on a crash course in XML and an introduction to the TEI guidelines, emphasising how the modular nature of the TEI reflects the many different activities for which it has been used. For the practical elements of the sessions, we recommended the Oxygen XML Editor, which has specific support for TEI and is the most commonly used editor for text encoding. Oxygen is licensed software, and while many participants took advantage of the 30-day free trial for the duration of the School, those who wished to pursue text encoding would have to pay or take advantage of institutional licences. Participants were free to use open-source alternatives, such as the one explored by Mike Hawkins in his *Text and Pointy Brackets* blog (Hawkins, 2020), but with the proviso that we would not be able to provide technical support.

The middle three days of the course concentrated on core parts of the TEI: description, transcription and the use of the `correspDesc` module. As well as a workshop covering practical aspects, on each day we invited a guest speaker who was actively involved in a relevant project to give a concrete idea of what it is like to work with the TEI in a real-life environment and to convey the sense of the TEI as a community of practitioners. Given our own research backgrounds, the course was inevitably slightly skewed towards manuscript studies. On the final day, we covered schemas and the customisation of the TEI for the needs of specific projects, and touched on XSLT and on routes to publication.

The two online workshops in 2020 and 2021 were more limited in their ambitions. Here we could only really hope to give a sense of what the TEI is and why people might want to use it. We altered our approach from 2020 where we gave three one-hour workshops to 2021 where we gave a larger, single, three-hour workshop which seemed to work much more effectively and allowed participants to complete exercises and gave time for discussion of their experience. The major piece of feedback from these sessions was that even online it was much easier to engage with concrete examples than with theory.

What can we learn from these experiences? First, that teaching the TEI in a general context with a mixed cohort is much more challenging than teaching in a specific research context. People learn best when presented with concrete examples which are relevant to them, and it is difficult to do this with a large group with varied interests. Concentrating on three core areas seemed an effective approach, but it might be that a wider variety of examples would be helpful for future sessions. Varying levels of technical expertise were somewhat challenging, but the group quickly became quite self-supporting, with more experienced members helping those who were struggling.

Perhaps the most successful aspect of the in-person course was the presentation of the TEI as a community, and the development of a community in the cohort itself. This was greatly helped by the presence of the invited speakers, many of whom stayed beyond their sessions to join in and help out. There was a real sense of communal learning, with participants working together and discussing approaches and ideas. We even had people dropping in from other strands of the Summer School, leading to interesting interdisciplinary discussions and cross-connections with other areas of DH. Our efforts to relate the TEI to the ideas and ambitions of

the group seemed successful, and a number of participants contacted us after the course for help with setting up projects (The David Jones Digital Archive, the Mary Hamilton Papers and the Correspondence of Giacomo Leopardi are three examples of this).

The most challenging areas were the sessions on customising the TEI through the schema and routes to publication. Each of these areas would probably have benefited from being introduced earlier in the week, and particularly the session on schema customisation, which is essential to the use of the TEI in specific contexts. Publication and analysis might have been better presented as a theme throughout the week rather than a separate session, covering some commonly used tools and scripts in the context of specific use cases.

CONCLUSION

What can we conclude about good ways of teaching the TEI in a general context? This is a question that will come into focus very soon in our own institutions with both Cambridge and Oxford launching digital humanities Masters courses in 2022/23, in which text encoding and the TEI will be key components. The most common profile of an attendee at a TEI course or workshop is someone who has heard of the TEI in some kind of general sense, and who thinks it may be useful to them either in their specific institutional or research context or as a valuable skill. When they actually encounter TEI, they often have that feeling peculiar to technical training that everything should be possible on a conceptual level, but is completely impossible on a practical basis. What can we do to allow them to understand what the TEI is, what it does, and also give them a route to using it?

An important starting point is to distinguish between text encoding as an activity, XML as a data standard and the TEI as a standard for text encoding; these are three things that are often merged in an unhelpful way in the minds of learners. By looking at the practice of text encoding as distinct from its expression in XML and/or TEI, we can talk more clearly about how and why it is used – as a route to publication, as the basis for analysis, and as a framework for engaging with the text – before embarking on the more technical, and possibly more challenging, parts of the training.

A basic understanding of XML is certainly fundamental to getting going with the TEI – and it is helpful if it is introduced generally as a data format in which text encoding can be done (rather than only in the context of its use by the TEI). Going through the building blocks of XML – elements, attributes, comments, namespaces, hierarchy and schemas – grounds learners in the actual work of text encoding, and reduces the uncertainty which some experience when confronted with a forest of angle brackets. Just as important is the concept of XML files as text files like any other, readable both by machines and by humans, which reside in directories (also known as folders) on your computer and do not rely on any particular software to edit them or maintain them. This addresses an unhelpful merging between the concepts of file, editor and interface. We often come across the misapprehension that the XML file is somehow ‘in’ the editing software or ‘in’ the platform. In this context, a general introduction to the basic concepts of directories and files, and how they interact with software would be a useful addition to the preliminary materials.

In addition to emphasising the sustainability and portability of XML files, one of the most important points for learners is that the process of text encoding happens immediately in front of them and in a space they can understand and control rather than in some other space such as the web, the cloud, a piece of software, or a database. It is very helpful for learners to conceive of their outputs as a distinct dataset with multiple uses, separate from the interfaces which present it, the tools which analyse it, or the software which edits it. The privileging of display over other uses is a particular problem, as Turska, Cumming and Rahtz point out – ‘in digital editions the encoded texts themselves are the most important long-term outcome of the project, while their initial presentation within a particular application should be considered only a single perspective on the data’ (Turska, Cummings, & Rahtz, 2017). Focussing on outputs as a dataset rather than simply a source of presentation addresses one of the major general obstacles in teaching digital technologies, which is that learners can see the point and grasp the principles, but find it difficult to grasp ‘where’ it is all happening.

In this context, the TEI can be introduced as a framework or standard for text encoding, which (currently) is expressed in XML, and which has outputs which might be used in multiple ways. On a practical level, the large scope of TEI is best engaged with through the specific activities it has been used for – firstly through the modules, which give a good general overview of the current coverage of the TEI, and secondly through existing projects. Sessions led by TEI practitioners who are currently engaged in work covering core areas of TEI (such as digital editions, correspondence, and manuscript description) are a very good way for learners to get a feel for what is possible, and also to engage with the TEI as both an implementable standard and as a community of practice. The guest speaker sessions at the 2019 Summer School were particularly well received, with students able to relate their own ideas and ambitions to concrete examples. Here, issues such as schemas and the customisation of the TEI for the specific needs of projects, and the benefits of working directly with the XML against the efficiencies of editing tools arise naturally in a real-world context which makes sense to the learner.

Our experience with the Summer School tells us that a focus on the concrete outputs of text encoding – sustainable and interoperable datasets for analysis and publication – needs to be present throughout the learning process in order to give context to the sometimes detailed and repetitive activities needed to get to grips with the TEI. Many of the responses to Stella Dee’s survey of teachers and learners of the TEI can be summed up as ‘Now I have my TEI document, what can I do with it?’ (Dee 2014), with the primary assumed purpose being publication on the web. This naturally leads to the question ‘Why go through all these pains, when you can save your text as HTML or PDF and put them on the Internet and Google will find them?’. In order to make sense of the TEI we need to be constantly referring back to the *why* as well as the *how*, with an emphasis on TEI files as a dataset as well as a source for publication.

An emphasis on outputs also puts some key features and challenges of the TEI in context. One of the most common complaints about the TEI is that there are multiple ways of doing the same thing. This can be particularly confusing for learners from a library or archives background who are used to highly prescriptive standards such as the AACR2 cataloguing rules and Encoded Archival Description (EAD) – though it should be pointed out that that even here there are considerable variations in practice. Learners should be aware that the multiple approaches enabled by TEI reflect multiple needs, and that the subjective nature of text encoding makes sense in the context of a strong (and iterative) tie between encoding methods and research questions and their outputs. This helps learners to make sense of encoding with the TEI as a developing framework for exploration rather than the implementation of a set standard.

On a more conceptual level, a fundamental point to communicate is the TEI as a community of practice and practitioners. Our experience is that this can happen at a micro level within the group, where participants start to communally explore the possibilities offered by the TEI in response to research questions, and at a macro level as participants realise that the TEI is something they themselves can engage with and contribute to. One example is the long history of the TEI trying to reflect their community’s views on sex, gender, and gender identity.¹¹ The concept of the TEI as a repository and record of practice in text-based Digital Humanities work is something which makes particular sense to researchers, who see an opportunity to make a methodological as well as a scholarly contribution to the field.

One of the most important conclusions to emerge from our experience in teaching the TEI is that the actual process of text encoding is in itself probably the most important output. The realisation that the act of encoding text has value in itself – as a method of deep reading, as a framework for engaging with the text, as a methodology for experimentation and interrogation – is the biggest ‘eureka moment’ that we see in learners. The TEI invites you to explore texts in the context of the approaches, methods, questions and idiosyncrasies of other scholars, and to contribute your own.

Finally, a large factor in the success of workshops and training that take place in a specific research context is being able to do real work in the learning environment, rather than asking participants to complete practice exercises. The idea that the outputs of the learning process will be real and useful contributions to the field or project seems to be the major factor in people engaging quickly and effectively with the TEI in the learning environment, and persevering with

11 see issues 367, 426, 2189 and 2190 on the TEI GitHub site <https://github.com/TEIC/TEI/issues>.

the TEI after the end of the training. As Julia Flanders and colleagues noted, ‘the ability to see their work realized as a readable edition was a crucial motivator for students in pushing through the process of learning and debugging their TEI/XML encoding’ while at the same time being careful to ‘to avoid creating orthodoxies in TEI encoding arising from display outcomes that appear authoritative’ (Flanders et al., 2020). This approach generates a high level of motivation and cohesion within the group, with everyone pulling together to try to complete a real-world task within a time limit, and it provides a good example of what it’s really like to work on a project. Tackling the kinds of problems thrown up by real-life materials demonstrates how text encoding can act as both a framework for discussion and exploration as well as a method for generating outputs.

This fundamental sense of working on something real is very difficult to replicate in a general context like the Summer School. It would be possible to ask people to come with examples from their own work, but then you would miss the communal and collaborative aspects of the work upon which all can engage at once. One solution would be to reframe the Summer School as a series of workshops which gave a general introduction to the TEI through a particular set of sample materials, which if carefully selected would allow us to cover the major themes and concepts while replicating some of the sense of deep engagement and excitement which comes from tackling real world problems. We look forward to discussing how this might work in our planning for future training in the TEI.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR INFORMATIONS

Yasmin Faghihi: Conceptualization; Investigation; Methodology; Writing – original draft; Writing – review & editing

Matthew Holford: Conceptualization; Investigation; Methodology; Writing – original draft; Writing – review & editing

Huw Jones: Conceptualization; Investigation; Methodology; Writing – original draft; Writing – review & editing

AUTHOR AFFILIATIONS

Yasmin Faghihi  orcid.org/0000-0001-5556-168X
University Library, Cambridge, UK

Matthew Holford
Bodleian Libraries, Oxford, UK

Huw Jones  orcid.org/0000-0002-8533-9083
University Library, Cambridge, UK

REFERENCES

- Allés-Torrent, S., & Riande, G. D.** (2019). The Switchover: Teaching and Learning the Text Encoding Initiative in Spanish. *Journal of the Text Encoding Initiative*, 12. DOI: <https://doi.org/10.4000/jtei.2994>
- Apostolo, S., Börner, I., & Hecht, A.** (2019). Collaborative Encoding of Text Genesis: A Pedagogical Approach for Teaching Genetic Encoding with the TEI. *Journal of the Text Encoding Initiative*, 12. DOI: <https://doi.org/10.4000/jtei.2926>
- Berenato, T.** (2021). David Jones’s ‘Balaam business’: The Poetics of Forgiveness after Passchendaele. In Feldman, M., Svendsen, A., & Tønning, E. (Eds.), *Historicizing Modernists: Approaches to ‘Archivalism’* (pp. 153–172). London: Bloomsbury Academic. DOI: <https://doi.org/10.5040/9781350215078.ch-8>
- Cummings, J.** (2019a). A world of difference: Myths and misconceptions about the TEI. *Digital Scholarship in the Humanities*, 34(Supplement 1), 58–79. DOI: <https://doi.org/10.1093/lc/fqy071>
- Cummings, J.** (2019b). Building DH training events. In Crompton et al. (Eds.) *Doing more Digital Humanities*. London: Routledge. DOI: <https://doi.org/10.4324/9780429353048-18>
- Dee, S.** (2014). Learning the TEI in a Digital Environment. *Journal of the Text Encoding Initiative*, 7. DOI: <https://doi.org/10.4000/jtei.968>

- FIHRIST Union Catalogue of Manuscripts from the Islamicate World.** Retrieved from <https://www.fihrist.org.uk/> (last accessed: 24th December 2021)
- Flanders, J., Bauman, S., Clark, A., Doyle, B., Hamlin, S., & Quinn, W.** (2019). TEI Pedagogy and TAPAS Classroom. *Journal of the Text Encoding Initiative*, 12. DOI: <https://doi.org/10.4000/jtei.2144>
- Fukushima, K., & Bourrier, K.** (2019). Inside Digital Dinah Craik: Feminist Pedagogy, Cognitive Apprenticeship, and the TEI. *Journal of the Text Encoding Initiative*, 12. DOI: <https://doi.org/10.4000/jtei.2185>
- Hawkins, M.** (2020). Installing your XML editor. *Text and Pointy Brackets*. Retrieved from <https://www.textandpointybrackets.com/foundational-skills-and-knowledge/installing-your-xml-editor> (last accessed: 24th December 2021)
- Ives, M., Del Hierro, V., Kelsey, B., Smith, L. C., & Sumners, C.** (2013). Encoding the Discipline: English Graduate Student Reflections on Working with TEI. *Journal of the Text Encoding Initiative*, 6. DOI: <https://doi.org/10.4000/jtei.882>
- Medieval Manuscripts in Oxford Libraries.** Retrieved from <https://medieval.bodleian.ox.ac.uk/> (last accessed: 24th December 2021)
- Mylonas, E., & Renear, A.** (1999). The Text Encoding Initiative at 10: Not Just an Interchange Format Anymore – But a New Research Community. *Computers and the Humanities*, 33, 1–9. DOI: <https://doi.org/10.1023/A:1001832310939>
- Pierazzo, E.** (2010). *Elena Pierazzo on the Arabic ENRICH Schema*. Retrieved from <http://sabinamessenger.blogspot.com/2010/08/guest-post-elena-pierazzo-on-arabic.html> (last accessed: 24th December 2021)
- Singer, K.** (2013). Digital Close Reading: TEI for Teaching Poetic Vocabularies. *The Journal of Interactive Technology and Pedagogy*, 3. <https://jitp.commons.gc.cuny.edu/digital-close-reading-tei-for-teaching-poetic-vocabularies/>
- Stewart, D.** (2007). The structure of the Fihrist: Ibn Al-Nadim as historian of Islamic legal and theological schools. *International Journal of Middle East Studies*, 39(3), 369. DOI: <https://doi.org/10.1017/S0020743807070511>
- Terras, M., Branden, R., & Vanhoutte, E.** (2009). Teaching TEI: The need for TEI by example. *Literary and Linguistic Computing*, 24, 297–306. DOI: <https://doi.org/10.1093/lc/fqp018>
- Turska, M., Cummings, J., & Rahtz, S.** (2017). Challenging the Myth of Presentation in Digital Editions. *Journal of the Text Encoding Initiative*, 7. DOI: <https://doi.org/10.4000/jtei.1453>

TO CITE THIS ARTICLE:

Faghihi, Y., Holford, M., & Jones, H. (2022). Teaching the Text Encoding Initiative: Context, Community and Collaboration. *Journal of Open Humanities Data*, 8: 15, pp. 1–14. DOI: <https://doi.org/10.5334/johd.72>

Published: 24 May 2022

COPYRIGHT:

© 2022 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.