A World of Difference: Myths and misconceptions about the TEI

James C. Cummings
james.cummings@it.ox.ac.uk
University of Oxford, United Kingdom

Introduction

The Guidelines of the Text Encoding Initiative are generally recognised in the digital humanities as important and foundational standards for many types of research in the field. The Guidelines of the TEI are generalistic, seeking to enable the largest possible user base encoding digital texts for a wide range of purposes. Working on many TEI-based projects, teaching TEI workshops, and advising researchers on data modelling needs, I have encountered many misunderstandings about the TEI. Indeed, one keynote lecture (not at DH) once told me that “the problem with the TEI is it has too many tags and there is no way to change it”. Inspired by myths like this, this paper will detail and expose common misconceptions about the TEI -- all of which have been espoused to me at some point -- but will concentrate on the more technical myths in a hope to increase knowledge about the TEI while dispelling some misconceptions along the way. Some of those to be investigated include:

“The TEI is too big (or complicated)”

While there is some truth to this -- the TEI Guidelines are numerous, consisting of around 565 elements -- no single project needs them all. Indeed, the TEI has mechanisms for customisation and recommends doing so to any project. The Guidelines themselves are modular and not all chapters will be appropriate or necessary to read for all projects.

“There is no way to change the TEI”

Although I have heard even well-respected keynote lecturers (not at DH) espouse this belief, it is patently and demonstrably false. This myth arises from unfamiliarity with the fact that the TEI is a framework entirely based on the concepts of adaptability and modification. Not only does the TEI have a sophisticated literate programming methodology to create meta-schemas which subset, constrain, and extend the vocabulary for any individual encoding project, but it also provides a variety of tools to enable users to do so.

“The TEI is too small (or doesn't have <my:SpecialElement>)”

While seemingly the opposite of #1, a frequent complaint made by those unfamiliar with the customization mechanisms of the TEI is that it does not have the special element needed for a particular encoding project. There is, naturally, a reluctance to add new elements to one’s customization -- and getting more generalized solutions into the TEI Guidelines themselves is indeed a better solution -- however, many new elements are added to the Guidelines through community development across disciplines. Any user is free to add <my:SpecialElement> but generally it is a better idea to get a number of individuals or a special interest group to agree a more detailed proposal.

“The TEI is XML (and XML is broken or dead)”

This idea is usually espoused by those who want to support some other, newer, format. Leaving aside the need some feel to denigrate one format in order to support another, XML is a widely supported format which will be with us for many years to come. However, TEI is **not** XML -- it is currently serialized as such, but previously it has been serialized as SGML, and in the future it may be expressed in another format(s). While there is currently no other widely adopted format which meets the many and varied needs of the TEI’s central format, this does not mean that the TEI cannot be used with many other formats (as input, output, integrated with it).

“XML (and thus TEI) can't handle overlapping hierarchies”

Many people have discussed their concerns of overlapping hierarchies in XML, and while it is true that there are limitations in expressing multiple hierarchies in XML, it also has solutions built into it, such as empty elements to represent one or more alternative hierarchies. Primarily, this misunderstanding is also based on the assumption that all markup is embedded markup. The TEI Guidelines include a chapter on representing non-hierarchical structures, and the TEI framework has many features for representing fragmented element structures, out-of-line and stand-off markup, and the association of additional annotation through URI-based pointing. In addition, many DH text encoding projects only require two hierarchies (e.g. intellectual vs physical representations) and the TEI provides transformation solutions to alternate between these.

“You can't do stand–off markup in XML (or TEI)”

This myth shows a misunderstanding of both XML and the TEI. The former is a language for markup vocabularies and puts no restriction on whether that markup is embedded, out-of-line, or entirely stand-off. The TEI Guidelines provide a number of solutions entirely geared to stand-off markup, and its community is working towards introducing more features in this area. The combination of fine-grained markup, URI-based pointing and/or XPointer schemes, and descriptive markup designed to function this way, means that stand-off markup is supported in the TEI.

“You can't get from TEI to $myPreferredFormat”

One of the benefits of XML is that it is easily processable to other formats. The TEI Consortium provides around 40 conversions to/from other formats, including, for example: bibtex, cocoa, csv, docbook, docx, dtd, epub, html(5), xsl-fo, json, InDesign, latex, markdown, mediawiki, nlm, odd, pdf, rdf, relaxng, slides, txt, wordpress, xlsx, xsd, and many more. There exist RESTful web services like *OxGarage* which can provide a pipeline for these and other conversions.

“There are no tools that understand the TEI”

This is false -- thousands of TEI projects have created many tools which process, mine, convert, and visualize TEI data. While the TEI Wiki lists some of these, one of the problems is that projects do not necessarily advertise and openly share their tools. Much of the software developed by projects is also bespoke and specific -- they are not necessarily generalisable to other projects’ needs. There are also many sophisticated encoding activities (such as stand-off markup) for which there are few general tools, since these are usually implemented in project-specific methods.

“If you create a TEI–based digital edition you must learn other $tech”

While historically it has been the case that to create TEI-based digital editions one must learn, or employ those who know, various technologies, this is increasingly less of an issue. Out of the box software like *eXist-db’s TEI Publisher* and *TEI Boilerplate* mean researchers are able to publish digital editions for themselves. Moreover, the TEI has introduced implementation-agnostic methods for documentation of intended processing models in a TEI customization. This can then be used to generate project-specific code based on changes to the customization, as in the case of the eXist-db implementation of the TEI processing model. This new aspect of the TEI enables developers to write more generalized software which relies on the TEI ODD customization file for information on the processing model.

“TEI is only for Anglo/Western works”

There is much about the TEI Guidelines that is based in Anglo and Western European textual traditions, but the Guidelines also make an effort to enable use in other languages and cultures. The definitions and glosses of elements (etc.) can be viewed in a number of languages (English, German, Spanish, French, Italian, Japanese, Korean, Chinese). There is an entire internationalization framework built into the TEI Guidelines and the TEI Customization language, which means that the schemas can routinely display these internationalized definitions in editors and those creating customisations can have definitions, examples, and attribute value descriptions, in any Unicode-expressable language.

“Interoperability is impossible with the TEI”

Interoperability is a good and laudable goal, but the potential richness of TEI encoding for research and analysis purposes should not be sacrificed for this (depending on the point of the initial encoding project). While interoperability does suffer in a framework that is customizable and extendable (which are necessary for such a generalized system), it is certainly possible. Usually it is a process of crosswalks or some scripted transformation to a lowest common denominator that involves someone knowing both resources. The creation of sub-communities (such as the TEI subset EpiDoc), which agree encoding standards that are tighter than the necessarily general and flexible TEI, can improve this significantly.

“The TEI is only for digital edition(s)”

The TEI may be used for many forms of output, for example camera-ready copy. The primary mistake here is to assume a one-to-one relationship between TEI encoded files and a single particular output. If significant encoding has taken place, a wide variety of outputs are possible. If the format is used to its full potential, many aspects of an edition can be created, as well as supplementary files, indices, introductory material, interactive data visualizations, and more. The use of the TEI can also be used outside of edition-building, for the creation of linguistic corpora, digital facsimiles, and other resources.

Summary

While these are only some of the myths surrounding the TEI, discussing these will be beneficial to the DH audience, and will hopefully lead potential TEI users to question other “received wisdom” about the Guidelines.