

2012-11-17

Creating Digital Editions: An Introduction to the Text Encoding Initiative (TEI)

Welzenbach, Rebecca; Hawkins, Kevin; Schaffner, Paul

<http://hdl.handle.net/2027.42/109403>

Rebecca Welzenbach
How to Read the TEI Guidelines
November 16, 2012

Before we turn you loose in the world, I'd like to take a little bit of time to introduce you to the TEI Guidelines--that is, the big document that everything we've been working on today has been based on. The guidelines, and lots of supporting material, live at:

<http://www.tei-c.org/index.xml>

"The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics."

The website of the TEI contains the full content of each version of the Guidelines that has been published, as well as many other resources that might be useful to you.

The TEI Guidelines were first published in 1994. Today, they are in their fifth formal iteration: TEI P5 was published in 2007, and has been updated every 6 months or so with new corrections and features. Because of this ongoing maintenance, it's best to use the Guidelines right from this site, rather than downloading a PDF or ordering a print copy--both of which you have the option to do!

As we've hinted today, the Guidelines are so massive and flexible that, really, no one expects to memorize the whole thing. A hint: neither is anyone expected to read the whole thing from start to finish. The best way to use the guidelines is as a reference, keeping them close at hand especially while you are designing your project, and but also throughout the encoding process. To get the most out of the guidelines, lets look at how you might find your way into them.

From the TEI homepage, you can make your way to the Guidelines tab and, most probably, the P5 guidelines (this is the most up-to-date version).

The guidelines consist of three major structural chunks:

- prose divided into topical chapters
- technical specifications for each element and attribute
- examples of the element and attribute being used. I

It's important to realize that these are not just different presentations of the same information. Actually, each form contains slightly different information, and to fully understand how any given element works and interacts with the rest of the TEI, you'll want to look at all three. So let's dive in by going to "browse" the TEI guidelines, and navigating to the table of contents:

<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index-toc.html>

Here you'll find several chapters of front matter that will give you even more background. The body of the prose of the Guidelines is divided into chapters, or modules, that treat specific aspects of encoding. This is a good place to start if you are planning a particular project and want to understand better how the TEI can deal with certain phenomena, such as manuscripts and primary sources, names and dates, or dictionaries.

Let's look at Manuscript Description:

<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html>

Each chapter contains several subsections, and all in all, addresses each element that is associated with that module. Each element name is linked to the technical specification, which gives you the specific constraints of that element. As well, examples for each element are given.

The prose gives a more nuanced explanation of what each element is *for* than you'll find elsewhere. Validating against a schema will tell you if your markup is syntactically valid, but the prose of the guidelines will tell you if you're really doing things "right." We use the term "tag abuse" to describe using an element that is legal, but not sensible or appropriate, to meet your needs. In order to be sure that your markup is not just syntactically valid, but that you're avoiding tag abuse, you will want to read the prose where it pertains the elements that you are using.

Let's look next at the element spec. This is a technical summary of each element, revealing what is allowed and forbidden for each element. If you have a specific element in mind and you want to know more about it, the element spec is a good place to start. Let's look at the element "catchwords."

<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-catchwords.html>

(In addition to linking here from the prose, you can arrive at the element spec by doing a search in the site search bar, or by choosing "elements" from the back matter of the Guidelines on the TOC page: <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/REF-ELEMENTS.html>)

At the top we see the element name, a “gloss,” which summarizes the definition of the element, and a link to the section of the prose that gives greater detail and context.

Beneath that is a reference to the “module” that each element belongs to. This is the chapter or section of the Guidelines that discusses this element and how it fits in with other available elements.

In order to understand the next two fields -- “attributes” and “used by”-- you need to know a bit about how the TEI Guidelines are structured. In order to ensure that elements with certain commonalities are all treated consistently, TEI elements are grouped into classes. “Attribute classes” contain elements that use the same attributes. “Model classes” contain elements that appear in the same places. This first field describes which attribute classes an element belongs to, and the next one describes which model classes it belongs to. I’m not going to say too much more about that now, but you can read about it in more detail in Chapter 1 of the Guidelines. To help you interpret this, the attributes that are allowed for each element are listed out here in parentheses, so can see them at a glance.

The next two fields, “Contained By” and “May Contain,” show all the elements that can have this element as a child, or which this element may have as a child.

Below this is the declaration, which is essentially a more compact/technical/machine-friendly way to summarize everything above, as well as what kind of data this element is allowed to contain. When you validate your markup against a schema, it’s checking your use of this element against this declaration.

Finally, we come to the third piece of the Guidelines: examples. Both the prose and the element spec will show at least one example of the element or attribute in question. These examples are designed specifically to illustrate that element or attribute.

<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-catchwords.html>

If you like, you may click “show all” to get an automatically generated list of all examples in the whole Guidelines making use of this element or attribute.

<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/examples-catchwords.html>

However, you should be aware that these examples may not be as useful, since they may have been designed to illustrate other elements and attributes, and may use this one only in passing. The uses will all be legal, but they may not be especially informative.

As you have already noticed, if you're using <Oxygen/> and your document is associated with a schema, you'll be able to see at every moment which elements and attributes are available to you. But it is the guidelines that will give you the fuller picture of how to use them and when it is appropriate to use one or another.

For more help and guidance, it is worthwhile to spend some time digging around the TEI website. There are a number of teaching resources and links to tools, as well as a wiki that contains some supplemental information that may not quite fit into the Guidelines, but is useful all the same. You can also get help from the TEI listserv.

Tools designed to work with the TEI: <http://www.tei-c.org/Tools/>

Support: <http://www.tei-c.org/Support/>