Markup:
XML & TEI

An Introduction to Digital Humanities
University of Washington

with thanks to Till Grallert, Julia Flanders & DHSI

Textual Markup

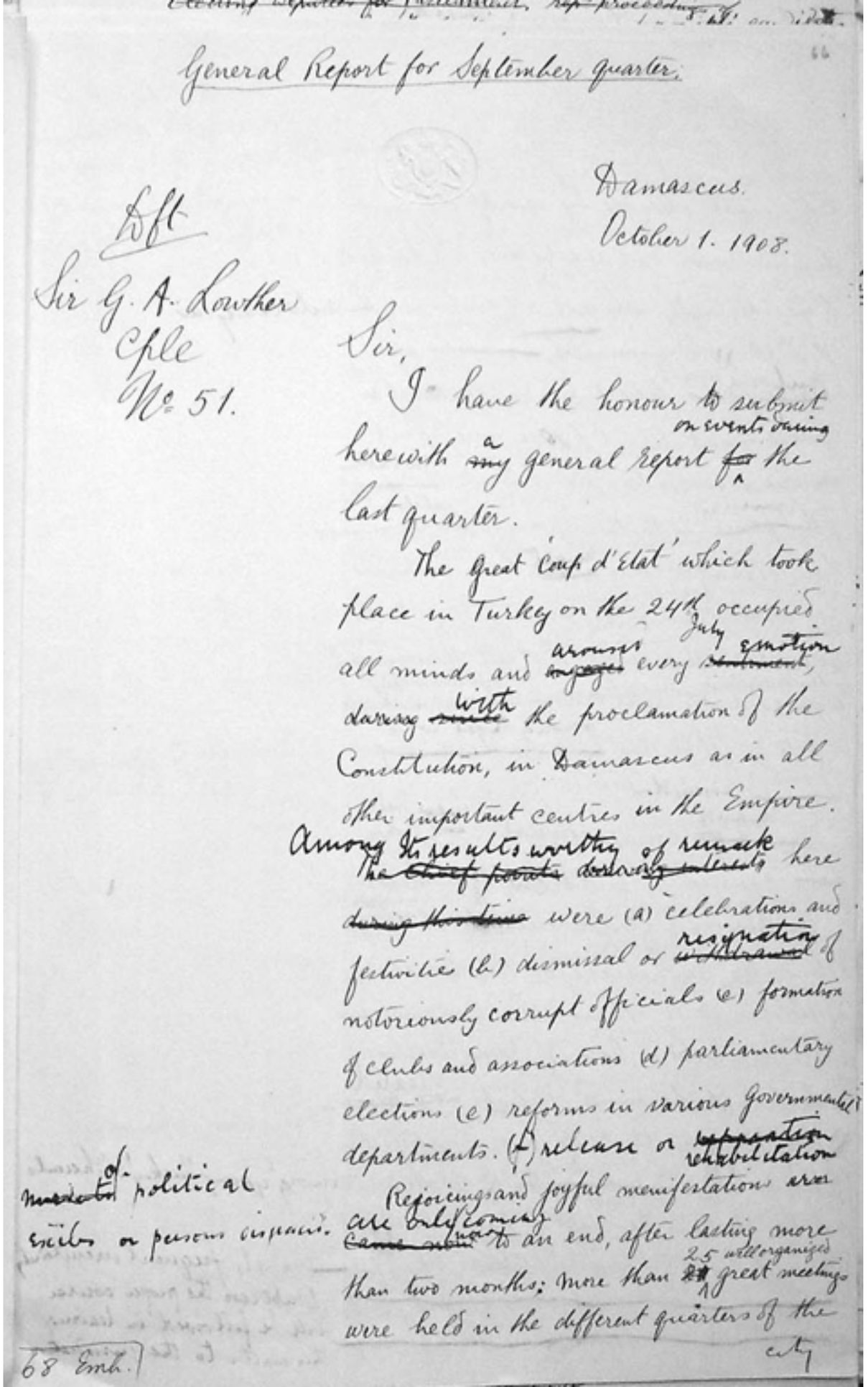
In order to talk about texts, markup and encoding of texts, we need to understand what we mean by these basic concepts.

When we talk about text encoding, what do we mean by a text? What is in a text and which assumptions do we make in reading them?

What is a Text?

Is this text…

Damascus, Quarterly Report,
Devey to Lowther 1 Oct. 1908



General Report for September quarter.

Damascus.
October 1. 1908.

Sir G. A. Lowther
Cple
No. 51.

Sir,

I have the honour to submit herewith ~~my~~ a general report on events during ~~for~~ the last quarter.

The great 'coup d'Etat' which took place in Turkey on the 24th July occupied all minds and aroused every ~~emotion~~, during ~~since~~ with the proclamation of the Constitution, in Damascus as in all other important centres in the Empire.

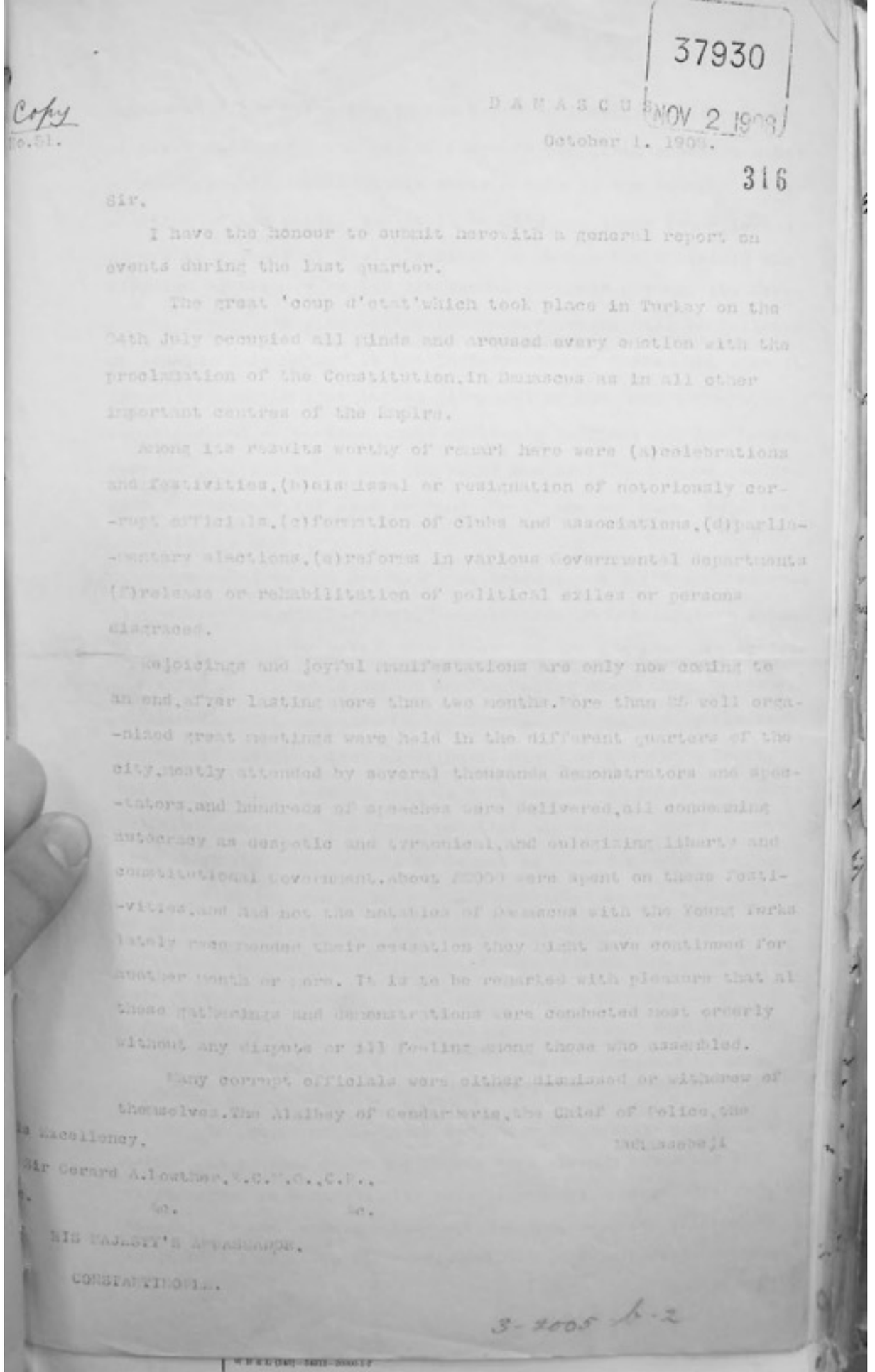
Among the results worthy of remark ~~the chief points during this time~~ were (a) celebrations and festivities (b) dismissal or ~~withdrawal~~ resignation of notoriously corrupt officials (c) formation of clubs and associations (d) parliamentary elections (e) reforms in various Government departments. (f) release or ~~separation~~ exhibitation of political exiles or persons disposed.

Rejoicings and joyful manifestations are ~~being carried~~ coming to an end, after lasting more than two months; more than 25 well organized great meetings were held in the different quarters of the city

68 Emb.

The same as this text…

Damascus, Quarterly
Report, Devey to Lowther 1
Oct. 1908



37930
NOV 2 1908

Copy
No.61.

DAMASCUS
October 1. 1908.

316

Sir,

I have the honour to submit herewith a general report on events during the last quarter.

The great 'coup d'etat'which took place in Turkey on the 24th July occupied all minds and aroused every emotion with the proclamation of the Constitution,in Damascus as in all other important centres of the Empire.

Among its results worthy of remark here were (a)celebrations and festivities,(b)dismissal or resignation of notoriously corrupt officials,(c)formation of clubs and associations,(d)parliamentary elections,(e)reforms in various Governmental departments (f)release or rehabilitation of political exiles or persons distressed.

Rejoicings and joyful manifestations are only now coming to an end,after lasting more than two months.More than 50 well organised great meetings were held in the different quarters of the city,mostly attended by several thousands demonstrators and spectators,and hundreds of speeches were delivered,all denouncing autocracy as despotic and tyrannical,and eulogizing liberty and constitutional government.About £500 were spent on these festivities,and had not the relations of Damascus with the Young Turks lately undergone their cessation they might have continued for another month or more. It is to be regarded with pleasure that all these gatherings and demonstrations were conducted most orderly without any dispute or ill feeling among those who assembled.

Many corrupt officials were either dismissed or withdrew of themselves.The Mudirey of Gendarmerie,the Chief of Police,the

His Excellency.

Sir Gerard A.Lowther,K.C.V.O.,C.B.,

&c. &c.

HIS MAJESTY'S AMBASSADOR.

CONSTANTINOPLE.

3-2005 b-2

…the same as this text…

Damascus, Quarterly Report,
Devey to Lowther 1 Oct. 1908

[37930] No. 1.

Sir G. Lowther to Sir Edward Grey.—(*Received November 2.*)

(No. 697.)

Constantinople, October 24, 1908.

Sir,

I HAVE the honour to forward herewith a despatch from His Majesty's Consul at Damascus reporting generally on events during the last quarter.

I have, &c.

(Signed) GERARD LOWTHER.

Inclosure in No. 1.

Consul Devey to Sir G. Lowther.

(No. 51.)

Damascus, October 1, 1908.

Sir,

I HAVE the honour to submit herewith a general report on events during the last quarter.

The great *coup d'État* which took place in Turkey on the 24th July occupied all minds and aroused every emotion with the proclamation of the Constitution in Damascus, as in all other important centres of the Empire.

Among its results worthy of remark here were—

(*a.*) Celebrations and festivities.
(*b.*) Dismissal or resignation of notoriously corrupt officials.
(*c.*) Formation of clubs and associations.
(*d.*) Parliamentary elections.
(*e.*) Reforms in various Governmental Departments.
(*f.*) Release or rehabilitation of political exiles or persons disgraced.

Rejoicings and joyful manifestations are only now coming to an end, after lasting more than two months. More than twenty-five well-organized great meetings were held in the different quarters of the city, mostly attended by several thousand demonstrators and spectators, and hundreds of speeches were delivered, all condemning autocracy as despotic and tyrannical, and eulogizing liberty and constitutional government. About 2,000*l.* were spent on these festivities, and had not the Notables of Damascus, with the Young Turks, lately recommended their cessation, they might have continued for another month or more. It is to be remarked with pleasure that all these gatherings and demonstrations were conducted most orderly without any dispute or ill-feeling among those who assembled.

Many corrupt officials were either dismissed or withdrew of themselves. The Alaibey of Gendarmerie, the Chief of Police, the Muhassebeji, the Secretary of the Court of Sherá, six Kaïmakams of Kazas, Mohammed Pasha Adm, of Idara Medjliss, with about thirty other minor officials, vacated their posts, mostly to the general satisfaction of the public, and it is debated now among Young Turkey adherents whether it would be right to punish the officials who were led to bribery by the littleness of their pay and its frequent irregularity. Doubtless the proper course will be followed in leaving this matter to the jurisdiction of Parliament when it assembles, and the ignominious dismissal of the more corrupt, rapacious and outrageous among officials suffices for the moment. Moreover, the vacancies are not being easily filled up by really good substitutes, for such persons are at present rare and often unwilling to serve.

Damascus clubs and associations, besides the Union and Progress, now number five, viz., Hurriyet, Ulema, Medical, Commercial, Free Ottoman, besides a Shoemaker Guild. The first two are showing much activity, which is likely to excite emulation. The Ulema, through misapprehension, and with uncalled-for solicitude, are anxious to conserve and uphold Islam against the liberal views of the Young Turks, while the latter have no intention at present to interfere with religious precepts. The first

…the same as this text?

Damascus, Quarterly Report, Devey to Lowther 1 Oct. 1908

example-PROFO618-3.TEIP5.xml

teiCorpus teiHeader fileDesc

1 . General Report for September Quarter

Damascus
October 1. 1908.
Dft
Sir G.A. Lowther
Cple
No. 51

Sir,
I have the honour to submit herwith ~~my~~ a general report ~~for~~ on events during the last quarter.

The great 'coup d'Etat' which took place in Turkey on the 24th July occupied all minds and ~~engaged~~ aroused every emotion with the proclamation of the Constitution, in Damascus as in all other important centres in the Empire. ~~The chief~~ ~~interests~~ Among its results worthy of remark here ~~during this time~~ were

- (a) celebrations and festivities
- (b) dismissal or ~~withdrawal~~ resignation of notoriously corrupt officials
- (c) formation of clubs and associations
- (d) parliamentary elections
- (e) reforms in various governmental departments.
- (f) release or ~~made to~~ rehabilitation of political exiles or persons

Rejoicings and joyful manifestation ~~came now~~ are only coming now to an end, after lasting more than two months: more than 25 well organized great meetings were held int he different quarters of the city, mostly attended by several thousands demonstrators and spectators, and hundreds of speeches were delivered, all condemning autocracy as despotic & tyrannical and eulogizing liberty and constitutional government. About £2000 were spent on the festivities and had not notables of Damascus with the young Turks lately recommended their cessation they might have continued for another month or more. It is to be remarked with pleasure that all these gathering & demonstrations were conducted most orderly

Text Grid Author

A text is not a document

Where is the text?

- in the shape of letters and their layout?
- in the original from which this copy derives?
- in the stories we read into it?
- or in its author's intentions?

TEI's definition:

- A "document" is something that exists in the world, which we can digitize.
- A "text" is an abstraction, created by or for a community of readers, which we can encode.

Encoding of texts

A text is more than a sequence of encoded glyphs or lexical tokens

- It has a structure and a communicative function
- It also has multiple possible readings

Encoding, or markup, is a way of making these things explicit

Only that which is explicit can be reliably found again and displayed

What is the point of markup?

To make explicit (to a machine) what is implicit (to a person)

To add value by supplying multiple annotations

To facilitate re-use of the same material

- in different formats
- in different contexts
- by different users

We don’t have to be limited to the view of one editor or consumer

Styles of markup

In the beginning there was procedural markup:

RED INK ON; print balance; RED INK OFF

Which being generalized became descriptive or semantic markup:

<balance type='overdrawn'>some numbers</balance>

also known as encoding or annotation

descriptive markup allows for easier re-use of data

Some more definitions

- Markup makes explicit the distinctions we want to make when processing a string of bytes

- Markup is a way of naming and characterizing the parts of a text in a formalized way

- It's (usually) more useful to markup what we think things are than what they look like

Separation of form and content

Presentational markup cares more about fonts and layout than meaning

Descriptive markup says what things are, and leaves the rendition of them for a separate step

Separating the form of something from its content makes its re-use more flexible

It also allows easy changes of presentation across a large number of documents

Markup as scholarly activity

The application of markup to a document can be an intellectual activity

In deciding what markup to apply, and how this represents the original, one is undertaking the task of an editor

There is (almost) no such thing as neutral markup – all of it involves interpretation

Markup can assist in answering research questions, and deciding what markup is needed to enable such questions to be answered can be a research activity in itself

Good textual encoding is never as easy or quick as people would believe

Detailed document analysis is needed before encoding for the resulting markup to be useful

XML

Extensible Markup Language (XML) is a simple, very flexible text format derived from SGML (ISO 8879).

Originally designed to meet the challenges of large-scale electronic publishing, XML also now plays an indispensable role in the exchange of a wide variety of data on the Web and elsewhere.

XML: what it is and why you should care

XML is structured data represented as strings of text

XML looks like HTML, except that:

- XML is extensible
- XML must be well-formed
- XML can be validated

XML is application-, platform-, and vendor- independent

XML empowers the content provider and facilitates data integration

XML is an international standard

XML requires use of ISO 10646 (also known as Unicode)

- a 31 bit character repertoire including most human writing systems
- encoded as UTF8 or UTF16

other encodings may be specified at the document level

language may be specified at the element level using @xml:lang

(The @xml:id attribute is another W3C-defined attribute.)

The rules of the XML Game

An XML document represents a (kind of) tree

It has a single root and many nodes

Each node can be

- a subtree
- a single element (possibly bearing some attributes)
- a string of character data

Each element has a name or generic identifier

XML elements and attributes are case sensitive

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader>
         <!-- .... -->
   </teiHeader>
   <text>
         <front>
               <!-- front matter of copy text, if any, goes here -->
         </front>
         <body>
               <!-- body of copy text goes here -->
         </body>
         <back>
               <!-- back matter of copy text, if any, goes here -->
         </back>
   </text>
</TEI>
```

XML syntax: the small print

What does it mean to be **well-formed**?

- There is a single root node containing the whole of an XML document
- Each subtree is properly nested within the root node
- Element/attribute/etc. names are always case sensitive
- Start-tags and end-tags are always mandatory (except with combined start-and-end tags, e.g. <pb/>)
- Attribute values are always “quoted”

A file can be **valid** in addition to being well-formed. This means you obey the rules of a specified schema, such as the TEI.

Representing an XML tree

- An XML document is encoded as a linear string of unicode characters
- It begins with a special *processing instruction*
- Element occurrences are marked by *start-* and *end-tags*
- The characters < and & are magic and must always be “escaped” if you want to use them as themselves, i.e. < and &
- Comments are delimited by <!-- and —>
- Attribute name= value pairs are supplied on the start-tag and may be given in any order, separated by spaces

The XML declaration

An XML document must begin with an XML declaration which does three things:

- specifies that this is an XML document
- specifies which version of the XML standard it follows
- specifies which character encoding the document uses; the default, and recommended, encoding is ‘UTF-8’ (Unicode)

Example:

```
<?xml version=“1.0" encoding="UTF-8"?>
<greetings xmlns="http://www.example.org/greetings">
    <hello type="enthusiastic">hello world!</hello>
</greetings>
```

Declaring namespaces

XML documents can include elements declared in different namespaces.

- a namespace declaration associates a namespace prefix with an external URI-like identifier
- the default namespace may be declared using an xmlns
- other name spaces must all use a specially declared prefix

All TEI documents are declared within the TEI namespace — a way of distinguishing one set of elements from another with the same names (like <p>):

```
<TEI xmlns="http://www.tei-c.org/ns/1.0"> ... </TEI>
```

Example:

```
<?xml version=“1.0" encoding=“UTF-8"?>
   <TEI xmlns="http://www.tei-c.org/ns/1.0">
   <greetings xmlns="http://www.example.org/greetings">
    <hello type="enthusiastic">hello world!</hello>
</greetings>
</TEI>
```

XML Example #1

```
<?xml version=“1.0" encoding="UTF-8"?>
<doc xmlns="http://www.example.org/namespace">
    <p n=“1">this is a paragraph</p>
    <p n=“2">this paragraph mentions
<placeName>Seattle</placeName></p>
</doc>
```

XML Example #2

```
<?xml version=“1.0" encoding="UTF-8"?>
<greetings xmlns="http://www.example.org/greetings">
    <hello type="enthusiastic">hello world!</hello>
</greetings>
```

Test your XML knowledge

Which are correct?

<seg>some text</seg>
<seg> <foo>some</foo> <bar>text</bar> </seg>
<seg> <foo>some <bar></foo> text</bar> </seg>
<seg type="text">some text</seg>
<seg type='text'>some text</seg>
<seg type=text>some text</seg>
<seg type="text"> some text <seg/>
<seg type="text"> some text<gap/> </seg>
<seg type="text">some text</Seg>

The TEI

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.

1987 was a long time ago…

The Text Encoding Initiative was born into a very different world

- the world wide web did not exist
- the tunnel beneath the English Channel was still being built
- a state called the Soviet Union had just launched a space station called Mir
- serious computing was done on mainframes
- most people didn’t have mobile phones

...but also familiar problems

Corpus linguistics and ‘artificial intelligence’ had created a demand for large scale lexical resources in academia and beyond

Advances in text processing were beginning to affect lexicography and document management systems (e.g. TeX, Scribe, tRoff..)

The Internet existed and theories about how to use it ‘hypertextually’ abounded

Books, articles, and even courses in something called “Computing in the Humanities” were becoming commonplace

The birth of the Text Encoding Initiative

Spring 1987: European workshops on standardization of historical data (J.P. Genet, M. Thaller)
Autumn 1987: In the US, the NEH funds an exploratory international workshop on the feasibility of defining “text encoding guidelines”
this resulted in the “Poughkeepsie principles”
Summer 1990: first draft (P1, with the ‘P’ standing for proposal ever since) of guidelines circulated

TEI is old!

So the TEI is very old!

- It comes from a time before the Web, before the DVD, smart mobile phones, cable tv, the iPod, and even XML (which was finalized in 1998)!

- Not much in computing survives 5 years, never mind 25

- Why is it still here, and how has it survived?
  What relevance can it possibly have today?

Why the TEI

The TEI provides

- a language-independent framework for defining markup languages
- a very simple consensus-based way of organizing and structuring textual (and other) resources…
- … which can be enriched and personalized in highly idiosyncratic or specialized ways
- a very rich library of existing specialized components
  an integrated suite of standard stylesheets for delivering
- schemas and documentation in various languages and formats
- a large and active open source style user community

Relevance

Why would you want those things?

because we need to interchange resources

- between people
- (increasingly) between machine

because we need to integrate resources

- of different media types
- from different technical context

because we need to preserve resources

- cryogenics is not the answer!
- we need to preserve metadata as well as data

TEI adopted XML

- In 2002, the TEI consortium published the P4 Guidelines, which were essentially an adaptation of P3 to XML that had been finalized as W3C standard in 1998.

- P5, a complete overhaul of the guidelines, was published in 2008. Updates are regularly published every couple of months ever since. The current version 2.8.0 was released on 6 April 2015.

- The Guidelines are currently maintained as an open source project on the Sourceforge site http://tei.sf.net/, from which released and development versions may be freely downloaded.

TEI XML

- all of XML.

- In addition, TEI XML must be valid against the schema “TEI all”

- TEI all: <?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng" type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"?>

- this processing instruction means that the XML adheres to a specific schema, a set of conventions that tell the computer as well as the human reader about the structure of the element and the data to be expected at any given point in the document.

- this provides interchangeability and a certain degree of interoperability (the latter is often only theoretically applicable)

Note: namespaces vs schemas

a namespace is a way of identifying the provenance of a bunch of elements:
a schema does the same, but it also specifies some rules about how those elements should be used.

a schema allows you to

- ensure that your documents use only predefined elements, attributes, and entities
- enforce structural rules such as ‘every chapter must begin with a heading’ or ‘recipes must include an ingredient list’

a namespace is just a URI;
a schema is a formal specification written in a formal language

Conformance issues

A document is TEI Conformant if and only if it:

- is a well-formed XML document
- can be validated against a TEI Schema, that is, a schema derived from the TEI Guidelines
- conforms to the TEI Abstract Model
- uses the TEI Namespace (and other namespaces where relevant) correctly
- is documented by means of a TEI Conformant ODD file which refers to the TEI Guidelines

or if it can be transformed automatically using some TEI-defined procedures into such a document (it is then considered TEI-conformable).

A useful mental exercise

Imagine you are going to markup several thousand pages of complex material….

- Which features are you going to markup?
- Why are you choosing to markup this feature?
- How reliably and consistently can you do this?

[Now, imagine your budget has been halved. Repeat the exercise!]

Ideas for Markup

- Identification information, page numbers, sources
- Structural information, titles, paragraphs, line breaks
- “chunks” or divisions of text, which may contain a picture, a poem, some prose, or a combination thereof
- within the chunks, we can identify formal units such as
  - a picture,
  - a caption
  - stanzas,
  - lines
  - paragraphs
  - and more…

Check the TEI Guidelines for appropriate tags, and examples of their use.