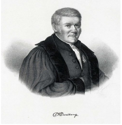
An archival format for TEI

Nicolas Larrousse *CNRS-TGIR Huma-Num (http://huma-num.fr)*

Lou Burnard

There's a lot of TEI-XML out there and we don't want to lose it!

TEI
TEXT ENCODING INITIATIVE
HN
Huma-Num
CINES
Centre Informatique National de l'Enseignement Supérieur



What metadata is needed for the long-term preservation of an XML-TEI file and how should it be supplied?

*The **TGIR Huma-Num** provides a service for TEI users wishing to use CINES facilites for long term preservation*

Verification
Documentation

Storage
Duplication

Restitution
Delivery

Long term preservation of XML resources

What's the problem?

XML data is easy and cheap both to store and to maintain in the long term but...

- XML is just a syntax. To process XML data correctly the meaning and function of the data must be preserved as well
- The processing of an XML data stream is highly application dependent
- Different applications and projects make up their own processing models, and have done so over the long term, so there are competing legacy systems
- Until recently, standards for describing and packaging data have been comparatively immature

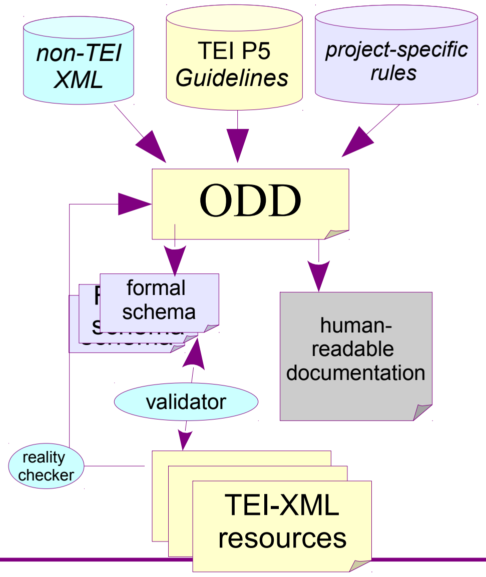
Consequently, although XML data may be easy to store without loss, it may be harder to preserve its semantics, and thus to guarantee its correct processing in the long term.

Is TEI-XML the answer?

YES : it provides an agreed vocabulary, together with a consensus around the meanings and intended use of all the XML elements anyone might ever need, plus ways of evolving that consensus in step with changing priorities

NO : in some places the full TEI has many ways of doing essentially the same thing, in others it is under-specified. This leads to inconsistency and ambiguity when trying to combine TEI-XML resources from different projects, and also suggests that something more is needed to ensure that such resources are processable in the long term.

TEI Conformance and the importance of an ODD



"TEI Conformance" goes beyond simple syntactic validity...

A TEI conformant file must
be valid with respect to TEI All
be accompanied by an ODD documenting its specific TEI usage
respect any additional constraints (e.g. use of elements from non-TEI name spaces)
respect the TEI conceptual model

Suggested procedures:
syntactic validation of each document against the most appropriate TEI schema; for documents containing textual data this would naturally include **TEI All,** but also any **project-supplied XML schema**, and also (for any ODD document supplied) the **standard TEI ODD schema**;
creation of a TEI schema from the supplied ODD and validation of the documents against that in order to validate any project-specific constraints such as attribute values;
comparison of the ODD supplied with an **ODD generated automatically** from the document set;
definition and usage of a set of stylesheets to **convert the resource into a "lowest common denominator"** TEI format.