TEI as an archival format

Lou Burnard, Nicolas Larrousse

TEI as an archival format

Lou Burnard, Oxford University, England

Nicolas Larrousse (nicolas.larrousse@huma-num.fr), Huma-Num CNRS (Centre National de la Recherche Scientifique), France

The adoption of the TEI as a common storage format for digital resources in the Humanities has many consequences for those wishing to interchange, integrate, or process such resources. The TEI community is highly divers, but there is a general feeling that all of its members share an understanding of the best way to use the TEI Guidelines, and that those Guidelines express a common understanding of how text formats should be documented and defined. There is also (usually) a general willingness to make resources encoded according to the TEI Guidelines available in that format, as well as in whatever other publishing or distribution format has been adopted by the project. The question arises as whether such TEI-encoded resources are also suitable for long term preservation purposes : more specifically, if a project wishes to ensure long term preservation of its resources, should it archive them in a TEI format? And if so, what other components (schema files, stylesheets, etc.) should accompany the primary resource files when submitting them for long term preservation in a digital archive? TEI encoded resources typically contain mostly XML-encoded text, possibly with links to files expressed using other commonly encountered web formats for graphics or audio; is there any advantage to be gained in treating them any differently from any other such XML encoded resource?

This is not an entirely theoretical question : as more and more digitization projects seek to go beyond simply archiving digital page images, the quantity of richly encoded TEI XML resources representing primary print or manuscript sources continues to increase. In France alone, we may cite projects such as the ATILF, OpenEditions, BVH, BFM, Obvil and many more for all of which the TEI format is likely to be seen as the basic storage format, enabling the project to represent a usefully organised structural representation of the texts, either to complement the digital page images, or even to replace them for such purposes as the production of open online editions. When such resources are deposited in a digital archive, how should the archivist ensure that they are valid TEI and will continue to be usable ? One possibility might be to require that such resources are first converted to some other commonly recognised display format such as PDF or XHTML; and indeed for projects where the TEI form is considered only as a means to the end of displaying the texts, this may well be adequate. But since TEI to HTML or TEI to PDF are lossy transformations, in which the added value constituted by TEI structural annotation is systematically removed this seems to us in general a less than desirable solution. We would like to be able to preserve our digital resources without loss of information, so as to facilitate future use of that information by means of technologies not yet in existence. Such data-independence was, after all, one of the promises XML (and before it SGML) offered.

The data archivist needs to be able to test the coherence and correctness of the resources entering the archive, and also to monitor their continued usability. For an

XML-based format, this is a relatively simple exercise. An XML file must be expressed using one of a small number of standard character encodings, and must use a tagging system the syntactic rules of which can be written on the back of a not particularly large envelope. The algorithm by which an XML document can be shown to be syntactically correct, (”well formed”) is expressible within the same scope and producing a piece of software able to determe that correctness is consequently equally trivial. The XML Recommendation adds a layer of “syntactic validation” to this, according to which the use of XML tags within a set of documents can be strictly controlled by means of an additional document known as a schema, defining for example the names of all permitted XML elements and attributes, together with contextual rules about their valid deployment. Syntactic validation of an XML resource against its schema is also a comparatively simple and automatic procedure, rerquiring only access to the schema and an appropriate piece of software. (Given the dominant position enjoyed by XML as a data format, the current wide availability of reliable open-source validators for it seems unlikely to change, even in the long term)

However, the notion of “TEI Conformance” as it is defined in the current Guidelines goes considerably beyond the simple notion of syntactic validity. An archivist concerned to ensure the coherence and correctness of a new resource at this further level needs several additional tools and procedures, and a goal of our project is to determine to what extent the goal of ensuring such conformance is quixotic or impractical. In particular, we will investigate the usefulness of the TEI’s ODD documentation format as a means of extending the scope of what is possible in this respect when using a conventional XML schema language such as RELAX NG or ISO Schematron.

Our initial recommended approach for ingest of a conformant TEI resource might include :

- 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

Such an approach suggests that the “submission information package” for a TEI resource will contain a number of ancillary documents or references to documents, notably to a TEI P5-conformant ODD from which a tailored set of syntactic and semantic validators can be generated using standard transformations. We hope to report on this and on the results of our initial experiments with some major French-language resources at the Conference.