Long term preservation of TEI Corpora

Nicolas Larrousse, Michel Jacobson

Long term preservation of TEI Corpora

Nicolas Larrousse
Huma-Num / CNRS
Paris, France
Nicolas.Larrousse@huma-num.fr

Michel Jacobson
Huma-Num / CNRS
Paris, France
Michel.Jacobson@huma-num.fr

ABSTRACT

This paper will present the implementation of TEI as an archival format done by the French research infrastructure dedicated to Social Science and Humanities, Huma-Num, liaising the TEI community and their practices with the staff of the preservation center[5] and their technical and archival needs.

CCS CONCEPTS

• **Digital Humanities**;

KEYWORDS

Long term preservation, Text Encoding Initiative,

ACM Reference Format:
Nicolas Larrousse and Michel Jacobson. 2020. Long term preservation of TEI Corpora. In *Digital Tools & Uses Congress (DTUC '20), October 15–17, 2020, Hammamet, Tunisia.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3240117.3240128

1 INTRODUCTION

Many of the digital resources produced by research communities, or at least from the Humanities, use the TEI format. For instance in France, the quantity of richly encoded TEI XML resources representing manuscripts, tapuscripts or other native sources is increasing regularly: some examples are resources such as books and journals on the "OpenEdition" platform (See https://www.openedition.org), Renaissance manuscripts in the "Bibliothèques Virtuelles Humanistes" textbase[7] (See http://www.bvh.univ-tours.fr), transcriptions and annotations of texts in the "Consortium CAHIER" corpus (https://cahier.hypotheses.org), transcriptions of oral recordings in the "ORTOLANG" (See https://www.ortolang.fr/) and "COCOON" (See https://cocoon.huma-num.fr/) repositories, etc. Considering the huge amount of work required to create these resources, there is a need to think about their long-term preservation[8] in order to make them reusable in the future, more than 20 years from now, by someone who was not involved in their creation. What are the prerequisites for reusability? There are many, but put briefly, the main goal is to be able to understand the documents both on the technical (readability) and semantic (understandability) level.

DTUC '20, October 15–17, 2020, Hammamet, Tunisia

ACM ISBN 978-1-4503-6451-5/18/10…$15.00
https://doi.org/10.1145/3240117.3240128

Huma-Num (https://www.huma-num.fr/), the French research infrastructure dedicated to Social Science and Humanities, aims at supporting research communities by providing a set of platforms and tools for the processing, conservation, dissemination and long-term preservation of digital research data throughout their lifecycle (https://documentation.huma-num.fr/1080). For long-term preservation[6], Huma-Num provides a service based on the CINES (https://www.cines.fr/en/long-term-preservation/) facility.

2 GOALS AND PROPOSED SOLUTIONS

In the case of TEI[3], technical understanding is easy enough to define as encoded resources typically contain XML-encoded text, possibly with links to files expressed using other commonly used formats for graphics, audio, music, etc. While many tools to check the syntactic consistency of an XML file have long been available, we wished to provide more than a simple validation as the purpose of the TEI is to express the meaning of a text rather than its form.

On the semantic level, although the TEI consortium publishes extended guidelines both for the documentation and definition of corpora, there is a great diversity of use within the TEI community due to the very diverse types of objects dealt with and the objectives pursued. This abundance is illustrated by the evocation of what Lou Burnard[2] has called the "TEI cornucopia".

So in order to be able to reuse these data in the future we need to perform multi-level checks:

- Syntactic validation specifically designed for TEI and associated files The main objective is to define and express "TEI conformance": data archivists at CINES are responsible for ensuring that the digital format will be readable in the future, which means firstly that they need to verify its conformance to the format when they receive the digital resource. As TEI encoded resources use the TEI XML schema, we decided to assume that all TEI documents intended to be archived should first be checked against the TEI schema. This means that we do not accept other contents (e.g. for instance external "name spaces", in the XML sense). Although it is technically possible to encapsulate objects directly in TEI documents, we forbid it. So to add an image expressed in SVG, it will be considered as an "external" object accessible from the TEI document via a reference (e.g. "TEI url attribute"). Likewise, this file should be checked against the SVG schema to ensure its readability.
- Semantic validation based on the corpus creation process In most cases, TEI producers do not use all the available tags; instead, they pick some for their own purpose and make choices about how they will use them for their specific needs to describe their scientific objects. The right way to document this process is to use an ODD[1] (One Document Does it all) document in which the producer lists all the tags

used in the corpus and thus prevent the use of other tags. It is also possible to add some precise constraints (e.g. "this tag is required or must contain a date"). And last but not least it is possible to document everything in the ODD file itself. It should be noted that the ODD document could represent, if necessary, a good base to migrate the format of TEI files. The ODD file is expressed in TEI so we can also check it against the TEI schema.
Here is an excerpt from an ODD file:

```
<elementSpec ident="idno" module="header"
mode="change">
<attList>
<attDef ident="type" mode="change">
<desc>
Indicates that the number "idno" should have
a type "IRHT-medium" or "IRHT-ark" and
that the default value is "IRHT-ark"
</desc>
<defaultVal>IRHT-ark</defaultVal>
<valList type="closed" mode="replace">
<valItem ident="IRHT-medium">
<desc>Identifier from Medium database</desc>
</valItem>
<valItem ident="IRHT-ark">
<desc>
ARK number from database
"Bibliothèque Virtuelle des Manuscrits Médiévaux"
(BVMM, https://bvmm.irht.cnrs.fr/)
</desc>
</valItem>
...
```

A specific schema (e.g. in RelaxNG), which materializes the constraints expressed for the corpus, will be generated from the ODD file in order to check the corpus against it.
To summarize, with an ODD, we can build a customized (restricted) schema, impose constraints and also provide some human-readable documentation about the scientific choices made to build the corpus.
Therefore, in our specifications a specific ODD is required for a TEI corpus to be "archivable".

Beside the technical and semantic validation, we propose to add some "environmental" documentation to provide information about the production context of the corpus. By environmental documentation we mean a general description of the scientific project, images of the original document (e.g. facsimile) and also different representations or renditions of the corpus based on the TEI documents (e.g. pdf, HTML) with their associated stylesheets.

The improvement of data quality to meet the needs of sustainability is accomplished by improving the level of requirements in several respects – technical, scientific (modeling) and documentary.

3 IMPLEMENTATION

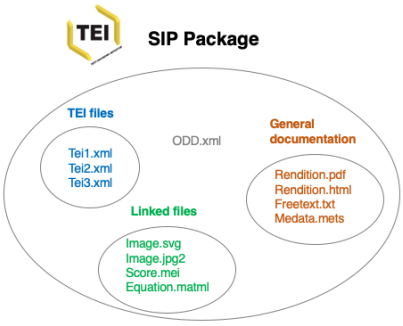
The CINES archival service is based on the OAIS[4] model (Open Archival Information System) which provides a general organizational framework (e.g. people, system, etc.) in order to manage the preservation of information for the long term.

In short, OAIS defines different entities (e.g. Producers, Consumers, Manager) that communicate by means of Information Packages in the course of the whole process. In the OAIS model, the SIP (submission information package) should be built by the data producer in conformance with recommendations made by CINES.

After some exchanges with representatives of the TEI community and staff at CINES, we agreed on the general structure for the package (SIP) to be archived. It should contain:

- TEI files valid against the "TEI All" schema and against the schema generated from the ODD provided
- Linked files expressed in a format accepted by CINES (See https://facile.cines.fr/)
- An ODD documenting the specific TEI usage for the corpus valid against the "TEI All" schema
- General documentation as described previously

Figure 1: Description of TEI SIP Package



When a TEI SIP package is sent at CINES, the following operations are carried out:

- Check the existence of an ODD file
- Validate all TEI XML files (including ODD) against the TEI All schema
- Validate all TEI XML files (excluding ODD) against the schema generated from ODD
- Check the existence of linked files mentioned as references in the TEI files
- Validate linked files classically (e.g. validate SVG files against the SVG schema)
- Validate documentation files classically (e.g. validate PDF files with a tool such as JOVE)

In order to be able to carry out these operations over the long term, the CINES maintains a copy of the TEI documentation (e.g. schemas) which will be enriched over time to take into account the evolution of the TEI. In the same vein, the CINES keeps an extensive documentation describing the production context and specific management metadata associated with each project.

4 FUTURE WORK

The technical implementation at CINES was completed in 2019, and after some polishing (e.g. choice of a specific TEI schema),

the system is considered stable. Currently, two projects are underway at CINES: books expressed in TEI from the "OpenEdition Books" platform and medieval manuscripts from "IRHT" (Institut de Recherche et d'Histoire des Textes - https://www.irht.cnrs.fr/). Additionally, some tests are in progress with linguistic resources from ORTOLANG.

The next step is to provide extended documentation for TEI producers based on the experience gained with existing and future projects. Another objective is to refine the validation process implemented at CINES to better integrate the different needs of TEI producers.

There are also some issues to be discussed with the TEI council regarding existing tools (e.g. Roma, ODD editor) and their future.

REFERENCES

[1] Syd Buman and Julia Flanders. 2004. Odd Customizations. In *Extreme Markup Languages proceedings*. Extreme Markup Languages, Montreal, Canada. http://conferences.idealliance.org/extreme/html/2004/Bauman01/EML2004Bauman01.html

[2] Lou Burnard. 2014. *What is the Text Encoding Initiative?* OpenEditionPress, Marseille. https://doi.org/10.4000/books.oep.426

[3] Lou Burnard and Nicolas Larrousse. 2013. TEI as an archival format. In *TEI - Text Encoding in the Web Conference proceedings*. TEI Consortium, Rome, Italy. https://hal.archives-ouvertes.fr/hal-02153026

[4] CCSDS. 2012. Reference Model For An Open Archival Information System (OAIS). (2012).

[5] CINES. 2004. Centre Informatique National de l'Enseignement Supérieur (2019), A digital archiving solution for long term preservation. (2004).

[6] Michel Jacobson, Nicolas Larrousse, and Marion Massol. 2014. La question de l'archivage des données de la recherche en SHS (Sciences Humaines et Sociales. In *Archives et données de la recherche proceedings 2014*. ICA, Paris, France. https://halshs.archives-ouvertes.fr/halshs-01025106

[7] Nicolas Larrousse, Christophe Jacobs, Michel Jacobson, Gilles Kagan, Joël Marchand, and Cyril Masset. 2019. Un Manuscrit Naturellement Rescuing a library buried in digital sand. In *DH 2019 proceedings*. ADHO, Utrecht, Netherlands. https://hal.archives-ouvertes.fr/hal-02153003

[8] Claudia Loebbecke and Manfred Thaller. 2005. Preserving Europe's Cultural Heritage in the Digital World. In *ESIS proceedings*. ESIS, Regensburg, Germany. https://aisel.aisnet.org/ecis2005/31