Analysis of Text Encoding Approaches: A Case Study

Aja Teehan and John G. Keating

An Foras Feasa: The Institute for Research in Irish Historical & Cultural Traditions
National University of Ireland Maynooth
Maynooth, Co. Kildare, Ireland
Motivation

• Text Encoding is a task commonly undertaken during the production of Digital Humanities resources, and is used to support metadata creation as well as full-text encoding of documents. Consequently, there are a large number of text-encoders active within humanities research communities, and within the TEI community in particular.

• We know of the many ways in which mark-up may be developed (see, for instance, the lively SIGs), some of the processes of development (Ide and Veronis: 1998) and the many features to which it can be applied (Burnard: 1999).

• There have been discussions of the suitability of mark-up for certain documents (Bradley: 2005, Lavagnino: 2006), and the implications of mark-up for textuality (McGann and Buzzetti: 2006) ... but despite all of this we know relatively little of how it is applied in practice.
Members of the Wider Text Encoding Community

- An Foras Feasa is active in the production of Digital Humanities resources and is involved in document-encoding.

- We are a multi-disciplinary team drawn from Computer Science and Humanities disciplines.

- Our approach is informed by our disciplines, and has been developed based upon well-established software engineering methodologies and developing theories of e-textuality.

- This approach has been successful for us, allowing us to produce Human Usable digital resources, e.g. Alcalá Account Book Project, Amharc Eireann Newsreels Collaborative Writing Environment and the Irish in Europe Project.
Members of the TEI Community

- We do not use TEI - it is not compatible with our current approach.
- TEI has been successfully used by many other organisations and projects.
- What is this other approach(es)?
- We hope that by reflecting upon current practices within the whole text-encoding community we can successfully develop strategies, methodologies, theories, and eventually tools, to support text encoders.
- In this way we can identify ways in which we, as a community, can build upon existing expertise to improve the digital resources we deliver.
Overview of Our Approach

- Software engineering methodologies: Use Case analysis, including Primary and Secondary Use Cases.

- Theories of digital textuality: Logical, Physical and Interaction Classes.

- These are only some of a suite of tools that we use to help us to develop our digital resources e.g. the iterative design paradigm, rapid application programming, participatory design, object-oriented theory, and existing and developing theories of textuality.
Overview of Our Approach

• We situate Use Case analysis within Activity Theory.

• Activity Theory tells us that our tool (encoding process and product) must adapt to meet the needs of the subject (user) and the object (document), in order to provide the desired outcome.

• We are interested in producing Human Usable documents, which require a different kind of encoding approach to embodied in TEI; our activities are different to those encapsulated in TEI (tools), but are similar to others in the whole text encoding community.
Overview of Our Approach

- Use Cases are analogous to outcomes. Therefore, the tool must adapt to each group of Use Cases.

- Primary and Secondary Use Cases are derived from a combination of the original information encoded and communicated within the document, and the needs of the individual researcher, and the research community.
Digital Alcalá Account Book Manuscript Classes

• Logical model: the notion and semantics of the Alcalá Account Book is described, indicating what it is, what it can be used for, and how to use it.

• Physical model is added, corresponding to the Alcalá Account Book manuscript, what it can be used for, and how to use it.

• Interaction model, provides a description of the features (attributes) of the digital edition of the Alcalá Account Book manuscript, of the user-requirements (methods) and how to implement them.

• Define the classes, instantiate the object; the digital edition of the Alcalá Account Book manuscript.

• This object includes all of the attributes of all of the classes, that is, it encapsulates what tasks (methods) it can perform and it knows how to perform them; without methods, it is useless.

All of the concrete classes shown implement the enclosed classes' attributes and methods (inheritance). Ref multiple hierarchies, <document>, <facsimile>, <text>.
A Document Encoding Activity: Theory of e-Textuality

Physical

Logical

Interaction

Scholar → Encoding → Model Physicality

Query Physicality

Encoding

Model Document Logic

Query Document

Use Document

Scholar → Encoding → Model Interaction
Whole System Approach

• We adopt a whole system approach (holistic) first using social theories (Activity Theory, Conversational Framework, GOMS - Goals, Operators, Methods, and Selection, SFL) to examine the activities of the communities. We then use the UML, specifically Use Cases, to specify and develop appropriate tools to support the activities.

• We develop text encoding systems that support human usable texts - we use the same Use Cases to develop the encoding schemes (in XML) AND the accompanying software to develop the tools that produce the expected outcomes.

• The identification, selection and realisation of appropriate Use Cases is necessary for document encoding and the production of human usable texts.
Dissemination

- Presentations (international and national)
  - DH2009, European Researchers’ Night, HEAnet
- Journal Articles
  - Jarbuch fur Computerphilologie
  - LLC (invited)
- Workshops
  - Cost Action A32, Digital Critical Editions (upcoming)
- Internal and External Courses
  - Undergraduate and Post-graduate
  - DHO summer school
- TEI MM!
• AT can be used (traditionally) to analyse the activities of individual scholars who disseminate their research for peer evaluation; have personalised tools, objectives, etc.

• These researchers add to the level of knowledge within their communities by contributing tools, outcomes and processes. A meta-theory (like AT) is necessary to theorise about best practices and artefacts.

• There must be other approaches: TEI is one, but there is probably more than one approach within TEI. With so many successes, for varied outcomes, the tools must have changed!

• What the text encoding community do (TEI + wider), and why do they do it?
The Case Study

• 24th of August: posted a call to the Humanist Mailing List; respondents would be provided with five images from a guestbook and asked to encode them as examples of their own approaches to text encoding.

• Well-received: following the receipt of 14 (+2) expressions of interest, a full description of the case study was issued by email on the 14th of September.

• The respondents were sent a description of the study along with five sample images, non-authoritative transcriptions and imagemaps.

• Our source document was a guestbook from the Castlehyde Estate House, Co. Cork. This estate house is historically significant as the lands once belonged to the family of Douglas Hyde, the first president of Ireland.
The Document

- Our source is a guestbook from the Castlehyde Estate House. This is now in a private collection and has not been catalogued.

- The guestbook seems to have originated as a collection of hand-tabulated loose leaf paper pages in 1927, the last recorded entry was in 1987. These pages appear to have been bound into a printed guestbook around 1935.

- The book is leather bound and measures roughly 27cm x 20cm x 3cm, the word "Castlehyde" is embossed on the centre of the front cover.

- 149 of pages have text on them, 51 are blank pages at the back of the book, and the cover.

- Both hand-tabulated pages and pre-printed pages.
Below you will find links to five downloadable zipped folders. Each folder corresponds to a single image; this equates to 2.5 pages of our document. A non-authoritative text transcription of the image is also included; we have indicated where problems with the transcription occurred by using square brackets around these areas. There may also be other problem areas. Each folder also contains a low-density (JPG) image of the source. Finally, we include an imagemap providing coordinates for some segments of the text we thought might be of interest. Please feel free to use, or not use, each of the items as you see fit.

We have undertaken an encoding of these images ourselves and we will supply this encoding to you all along with all of your other, anonymised, versions. In this case study you will create an encoding to support a Digital Critical Edition of the guestbook to reside within An Foras Feasa's archive, which is Fedora Commons based and runs on a Mac server. Your encodings should be replicable across the guestbook in order to create this Digital Critical Edition. Dublin Core and MODS metadata will be used for the resource. You can see An Foras Feasa's Digital Critical Edition of the Alcalá Account Book manuscript here: http://archives.forasfeasa.ie/

We are conducting a case study to compare approaches to full-text encoding of a document, for a Digital Critical Edition to reside within a repository. Ideally, the case study will provide four documents, detailed as follows:

Full-Text Encoding
- We would be obliged if you would provide a full-text encoding of the images in the folders. While metadata provision would be, of course, welcome, the text must also be machine readable. Any XML-based encoding language can be used and the finished product may be provided in any number of files.
- Your encoding approach should reflect the way you normally perform text encoding.
- It should also be replicable by an intelligent, industrious individual, within eight 40-hour weeks, across all of the Guestbook (we all have limited resources in the real-world!).

Description of Process
- In addition, we would be obliged if you could supply to us a brief written description of the encoding process addressing such issues as time spent encoding, the use cases and the rationale behind your approach.

Versions of Encoding
- We would appreciate copies of earlier versions if you have produced some that differ significantly from the end-product.

Questionnaire
- After we have received the encodings, and any other relevant documentation, we will send you a short questionnaire to complete. This will be used as part of a reflective and reflexive process.

We intend to publish our findings and to disseminate the anonymous encodings to the community. Please indicate the level of permission you grant us in relation to this using the attached document. You can, of course, withdraw from the process at any stage.

All the encodings, along with permission document, should be received by Monday the 28th of September, two week’s time. We understand that this places a demand on your time and appreciate your participation. If at any stage you cannot complete the case study, for any reason, please send us the work you have completed - especially the actual encoding; any and all input is valuable. If you have any further questions please do not hesitate to contact us.
Imagemap and Non-Authoritative Transcription

• We did not provide explicit Use Case descriptions.

• We did request some documentation describing the rationale and process of encoding.

• We wanted to see what Use Cases would be supported by the encodings, as this would provide us with one mechanism for separating and analysing the various approaches.

• This is the non-authoritative transcription and a sample of the imagemaps provided.

• The image on the next slide illustrates some of the difficulties for encoders.

<table>
<thead>
<tr>
<th>NAME</th>
<th>ADDRESS</th>
<th>DATE OF ARRIVAL</th>
<th>DATE OF DEPARTURE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colin Bell</td>
<td>The Acres</td>
<td></td>
<td>1st June</td>
</tr>
<tr>
<td>Miss Martille</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E A Montgomery</td>
<td>Wales</td>
<td></td>
<td>14th June</td>
</tr>
<tr>
<td>Capt R Conway V</td>
<td>Spitak Rhuddlan W.Wales</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Valet (nurse)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E A Montgomery</td>
<td>Glandalane</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ruth J. Hallian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mrs Hyde-Thomson</td>
<td>192 Queens Gate London</td>
<td>15th-July</td>
<td>21st-July</td>
</tr>
<tr>
<td>Hume R. Jones</td>
<td>Portrush</td>
<td>18th-July</td>
<td>19th-July</td>
</tr>
<tr>
<td>Betty Jones</td>
<td>do</td>
<td>do</td>
<td>do</td>
</tr>
<tr>
<td>Mary Stokes</td>
<td>Monerakelle, Moneragalla, Ceylon</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ken Doncaster</td>
<td>Byway, &quot; &quot;</td>
<td></td>
<td>&quot; &quot;</td>
</tr>
<tr>
<td>E. d. Bloomfield</td>
<td>Hines Park Waterford</td>
<td>May 1st</td>
<td>&quot; &quot;</td>
</tr>
<tr>
<td>D.S.C. Bloomfield</td>
<td>&quot; &quot;</td>
<td>&quot; &quot;</td>
<td>&quot; &quot;</td>
</tr>
<tr>
<td>Mrs. Stokes</td>
<td>Dublin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mr &amp; Mrs H.J.D. Stokes</td>
<td>Ceylon</td>
<td></td>
<td>Aug.3</td>
</tr>
<tr>
<td>Name</td>
<td>Address</td>
<td>Date of Arrival</td>
<td>Date of Departure</td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td>----------------</td>
<td>------------------</td>
</tr>
<tr>
<td>C. G. Ross</td>
<td>123 Ross Ave</td>
<td>13 July 1934</td>
<td>21 July 1934</td>
</tr>
<tr>
<td>Dr. J. Smith</td>
<td>321 Smith St</td>
<td>18 July 1934</td>
<td>29 July 1934</td>
</tr>
<tr>
<td>Mr. &amp; Mrs. Johnson</td>
<td>456 Johnson Ln</td>
<td>23 Sept. 1934</td>
<td>5 Oct. 1934</td>
</tr>
</tbody>
</table>

Data outside cell boundaries:

- Group of two people, one entry, two rows
- Two related people, two entries, two rows
- Group of three, two entries, two rows, one entry containing two names

Two types of page layout.
Results

• Strong initial response rate, but ultimately only 4 submissions.

• To what do we attribute this? Time pressure, also some reluctance to expose encodings.

• Without a larger sample, we can really only talk about overall structure and the features that were captured.

• ...And ask for more participants!
So ... who did the best one?

<table>
<thead>
<tr>
<th></th>
<th>Encoding 1</th>
<th>Encoding 2</th>
<th>Encoding 3</th>
<th>Encoding 4</th>
<th>Encoding 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tabular Structure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantic Structure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Problematic Overlaps of Physical and Logical Classes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Searchable Date</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Searchable Address</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Encoded Individuals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rendering Features</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mechanism for &quot; or &quot;ditto&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type of Page</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text/image Linking</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

But that is not surprising; these are (mostly) our use cases.
To What Do We Attribute Differences?

• The differences can arise because of various complications: the amount of time it took (no-one gets paid!), which is directly related to the level of difficulty, and how ‘invested’ the participants were.

• The biggest differences must have arisen because of different Use Case scenarios envisaged for the document.

• People considered their Use Cases, though some considered them in relation to their encoding practice, rather than the Users. What drives the encoding?

• From a User’s perspective, some of the implied Use Cases were inconsistent, e.g., marking the rendering of the ink, but not the type of page. If one of the interaction class use cases was “present the text version as it appeared on the document” then this is problematic, if it was not, then why mark the ink colour? Is there another use case?

“my view is that what's important here is to record as accurately as possible the semantic content. Each row records someone, or some group of someones, registering at this hotel on some date, and (sometimes) when they left. What matters (I am guessing) is who they were, where they came from, how long they stayed. “

“In general, my markup scheme privileged the semantic structure of the document over its physical structure. Examples of this would be those sections in which material that clearly belonged to one cell spilled over into the next row as indicated by the ruling on the paper. These were treated as a single cell with an internal line break. However, clear violations of the table structure by its creators are preserved – as, for instance, when a few writers record information about their address in the date column.”

“Depending on the needs of the project, there are many ways to handle this encoding, but since we are not privy to this information, we elected to leave the encoding this way. Further expansion of this encoding could be used, depending on the scope and the post-encoding needs of the project.”
Formal Analysis Methodologies

• There are theoretical and computational frameworks of analysis.

• Simple analysis: validating, errors, well-formed, etc.

  • The Personal Software Process: users record the process they perform when developing software. Metrics are produced. Not applicable in this scenario, but there is a framework for analysis produced based upon entry criteria, planning, development, postmortem and exit criteria. Can also be linked to Use Case analysis. (Putz)

• Complex Analysis: many frameworks for software, few for XML.

  • There is a fuzzy linguistic model that can be used to evaluate the information quality of XML documents, evaluation uses only Users’ perceptions and is thus User centred. Links human evaluation scheme to computer rated quality of XML. (Herrera-Viedma et al., 2007)

  • We can use a Weighted Element Tree Model to measure the structural similarity of XML Documents (Wang et al., 2009).
Future Plans - Please Participate!

• We want to perform formal analysis, but we must have sufficient numbers.

• We cordially invite you to participate. All, even partial, participation is welcome.

• We have to delay the questionnaire and corpus until we have sufficient numbers.

• Happy to link to the results from the TEI website to serve as example material for individuals; they will be able to review the different approaches taken in relation to one document.

• We wish to encourage people to engage in data modeling prior to encoding. It is not solely the document that decides the encoding, it is also the approach, based upon Use Cases - problematic for Access TEI!
Questions, Comments?

Aja Teehan
John Keating

firstname.lastname@nuim.ie