

TEI Members' Meeting, 2009, Ann Arbor

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



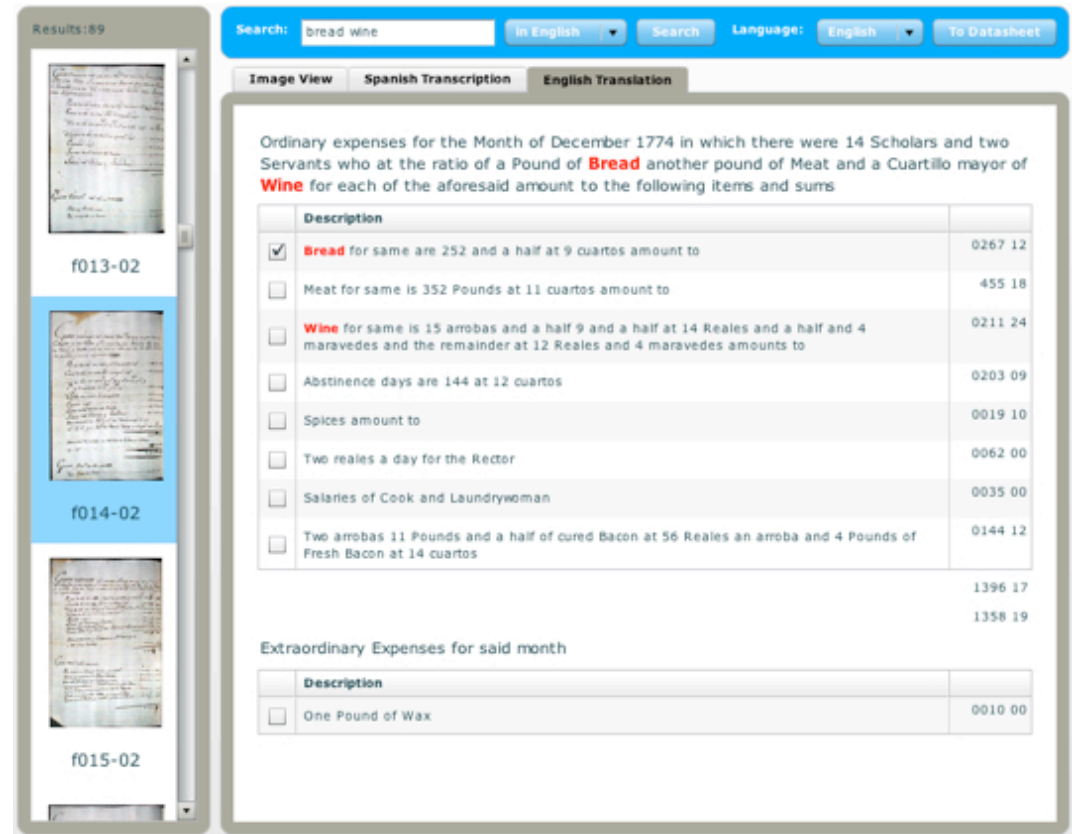
NUI MAYNOOTH
Ollscoil na hÉireann Má Nuad



- 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.



Results: 89

Search: bread wine In English Search Language: English To Datasheet

Image View Spanish Transcription English Translation

Ordinary expenses for the Month of December 1774 in which there were 14 Scholars and two Servants who at the ratio of a Pound of **Bread** another pound of Meat and a Cuartillo mayor of **Wine** for each of the aforesaid amount to the following items and sums

Description	
<input checked="" type="checkbox"/> Bread for same are 252 and a half at 9 cuartos amount to	0267 12
<input type="checkbox"/> Meat for same is 352 Pounds at 11 cuartos amount to	455 18
<input type="checkbox"/> Wine for same is 15 arrobas and a half 9 and a half at 14 Reales and a half and 4 maravedes and the remainder at 12 Reales and 4 maravedes amounts to	0211 24
<input type="checkbox"/> Abstinence days are 144 at 12 cuartos	0203 09
<input type="checkbox"/> Spices amount to	0019 10
<input type="checkbox"/> Two reales a day for the Rector	0062 00
<input type="checkbox"/> Salaries of Cook and Laundrywoman	0035 00
<input type="checkbox"/> Two arrobas 11 Pounds and a half of cured Bacon at 56 Reales an arroba and 4 Pounds of Fresh Bacon at 14 cuartos	0144 12
	1396 17
	1358 19

Extraordinary Expenses for said month

Description	
<input type="checkbox"/> One Pound of Wax	0010 00

f013-02

f014-02

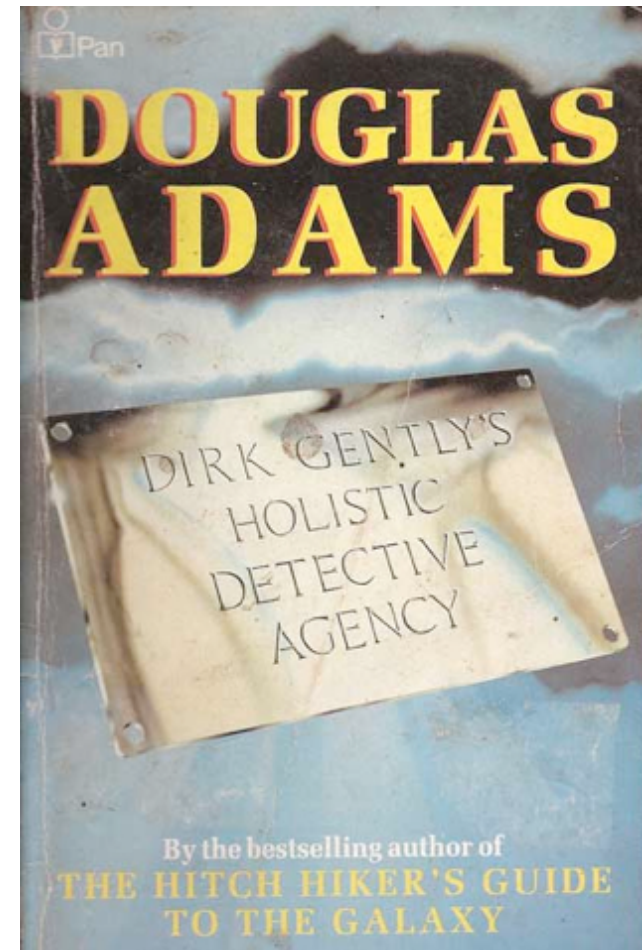
f015-02

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.

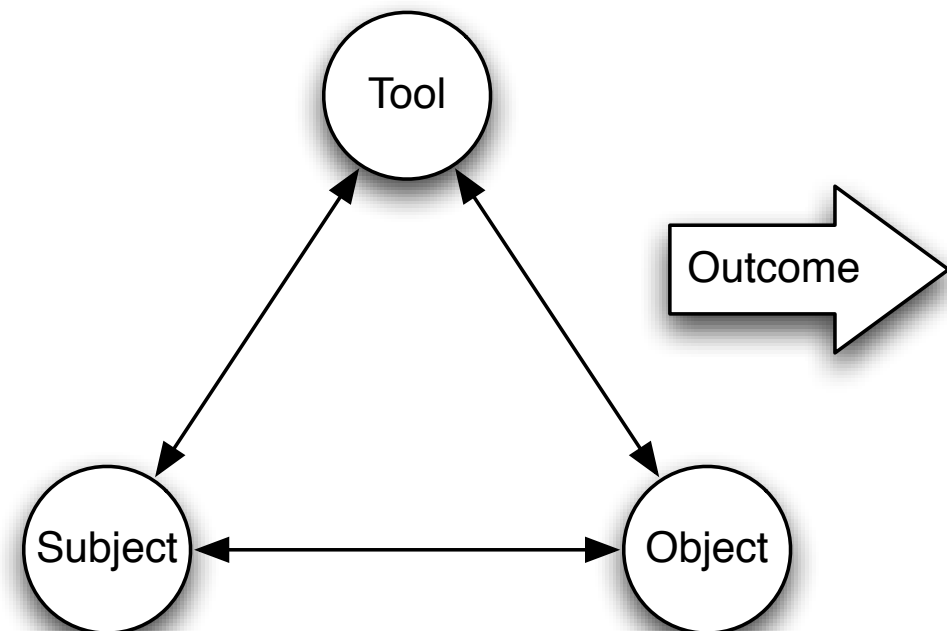
Overview of Our Approach

- 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.



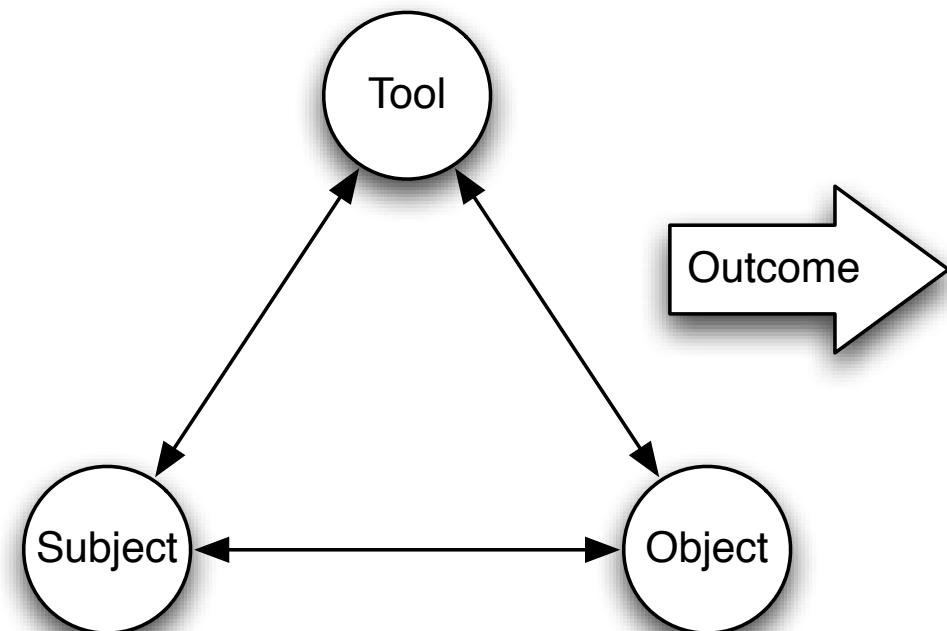
Overview of Our Approach

- 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.



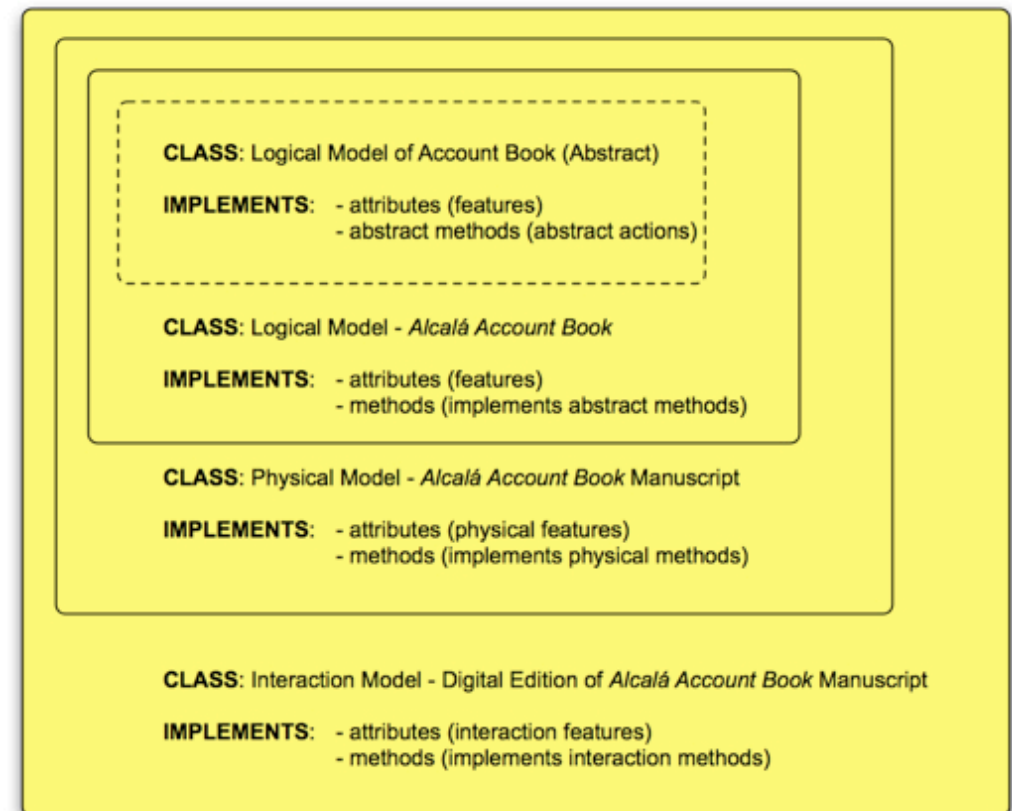
Overview of Our Approach

- 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.



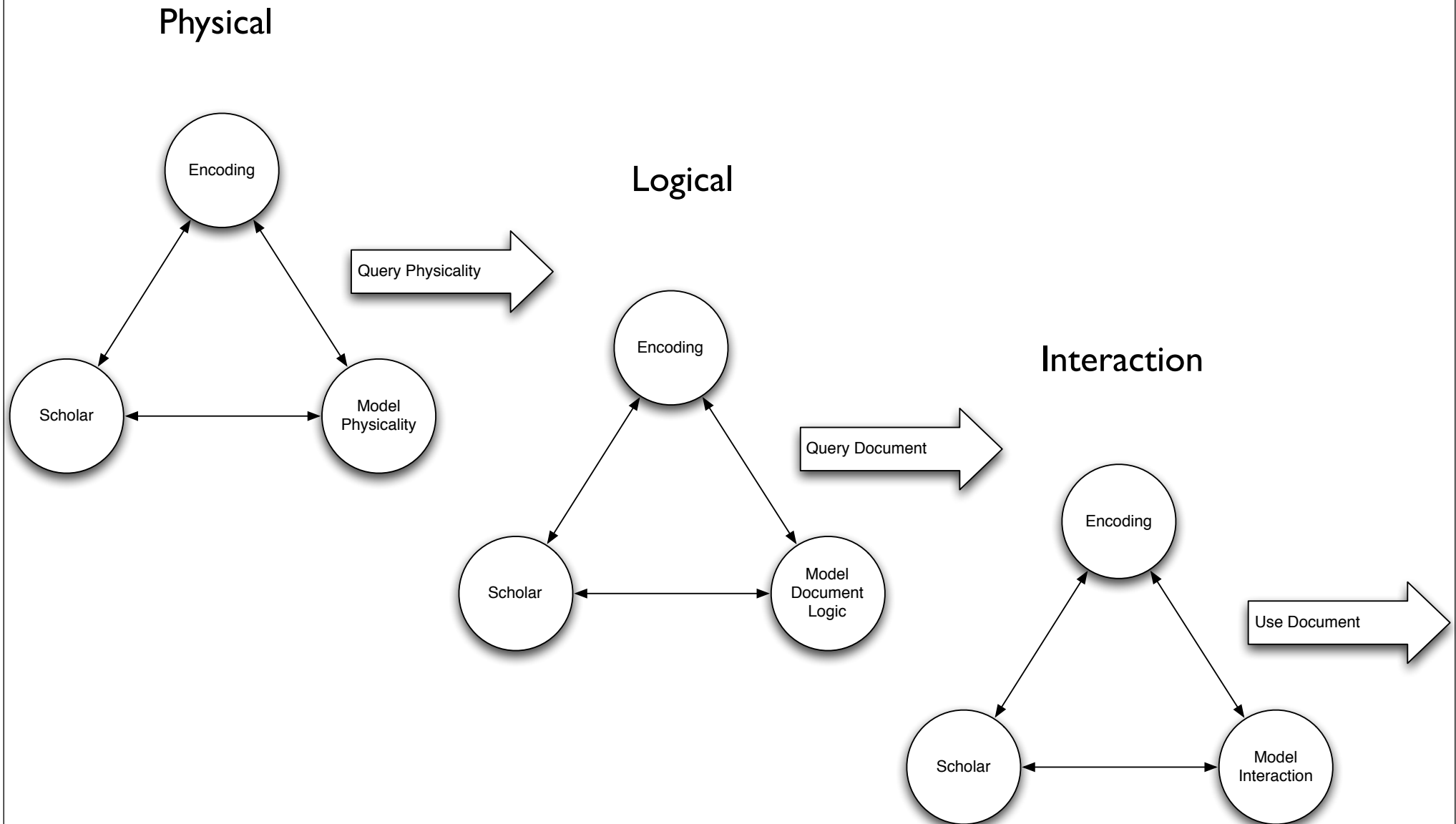
Digital Alcalá Account Book Manuscript Classes

- 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.



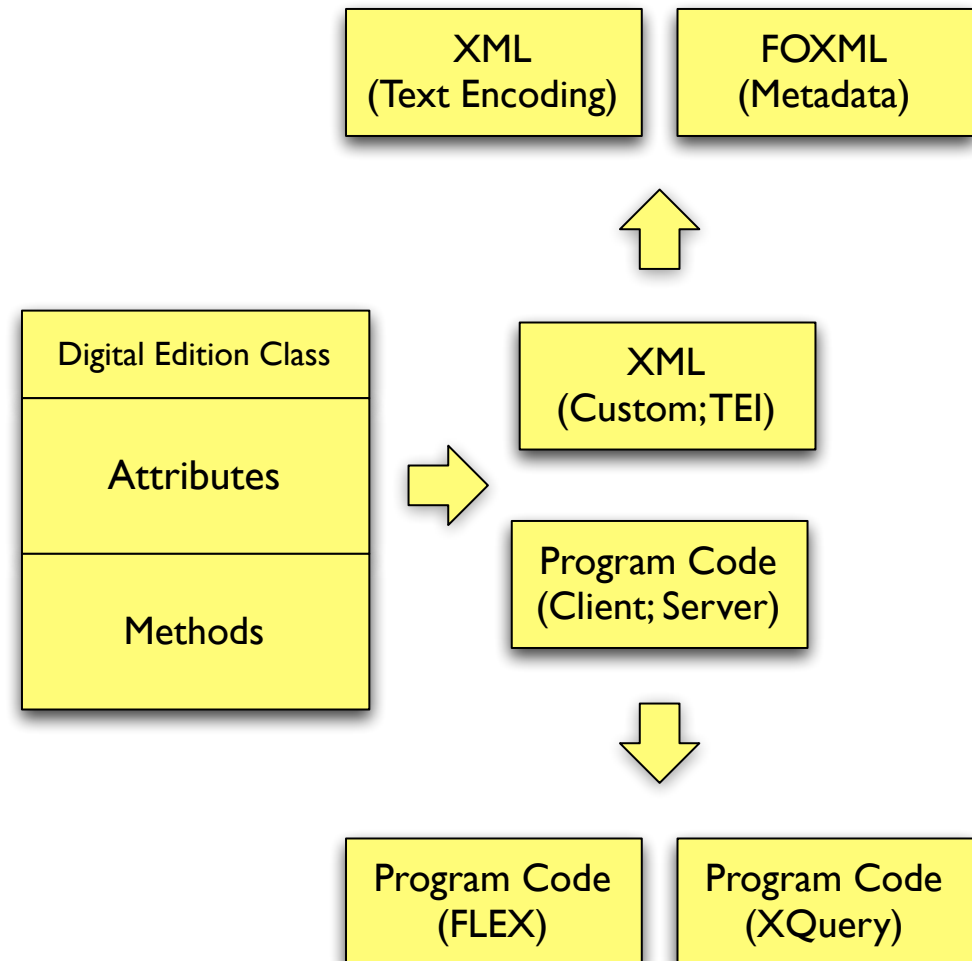
All of the concrete classes shown implement the enclosed classes' attributes and methods (inheritance). Ref multiple hierarchies, <document>, <facsimile>, <text>.

A Document Encoding Activity: Theory of e-Textuality



Whole System Approach

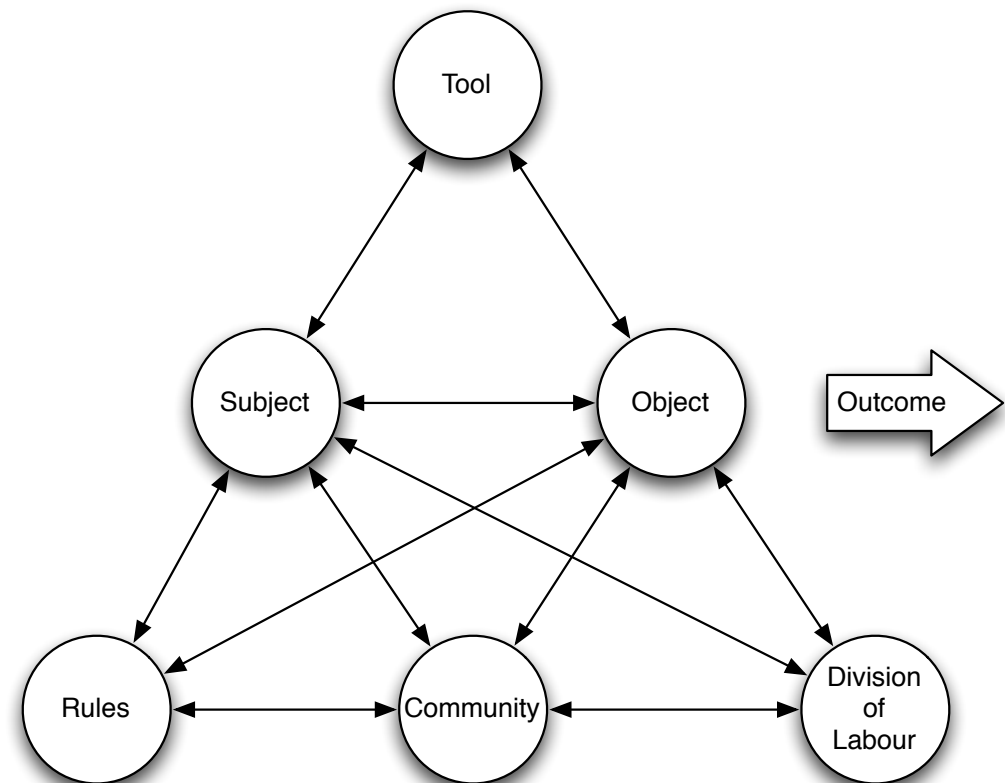
- We adopt a whole system approach (holistic) first using social theories (Activity Theory, Conversational Framework, GOMS - Goals, Operators, Methods, and Selection, SFL) to examine the activities of the communities. We then use the UML, specifically Use Cases, to specify and develop appropriate tools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



- 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!

Document Encoding as a Community Activity

- 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.

From: Humanist Discussion Group <willard.mccarty@mccarty.org.uk>
Subject: [Humanist] 23.250 Calling all text encoders!
Date: 25 August 2009 08:24:08 IST
To: humanist@lists.digitalhumanities.org
Reply-To: Online seminar for digital humanities <humanist@lists.digitalhumanities.org>

Humanist Discussion Group, Vol. 23, No. 250.
Centre for Computing in the Humanities, King's College London
www.digitalhumanities.org/humanist
Submit to: humanist@lists.digitalhumanities.org

Date: Mon, 24 Aug 2009 16:51:58 +0000
From: "John G. Keating" <John.Keating@nuim.ie>
Subject: Calling all Text Encoders!

Colleagues,

As part of our ongoing efforts to understand text encoding, we are embarking on a small case study that will investigate different approaches to encoding a single text (a guestbook from a historically significant estate house located in Ireland). We would like to invite text encoders to participate in the study, by volunteering to encode 5 sample pages from the text. All encoders will receive the same sample.

Participants will have the freedom to encode the text using any XML- based approach that they consider to be appropriate. We are happy to answer any questions about the text to help you with your encoding. Following submission of your encoding, you will be invited to complete a short questionnaire. All encodings and questionnaire responses will be fully anonymised.

Ideally we would like to have participants return their encodings prior to 15 September 2009. We plan to disseminate the results of the study to the text encoding community as soon as possible after this date.

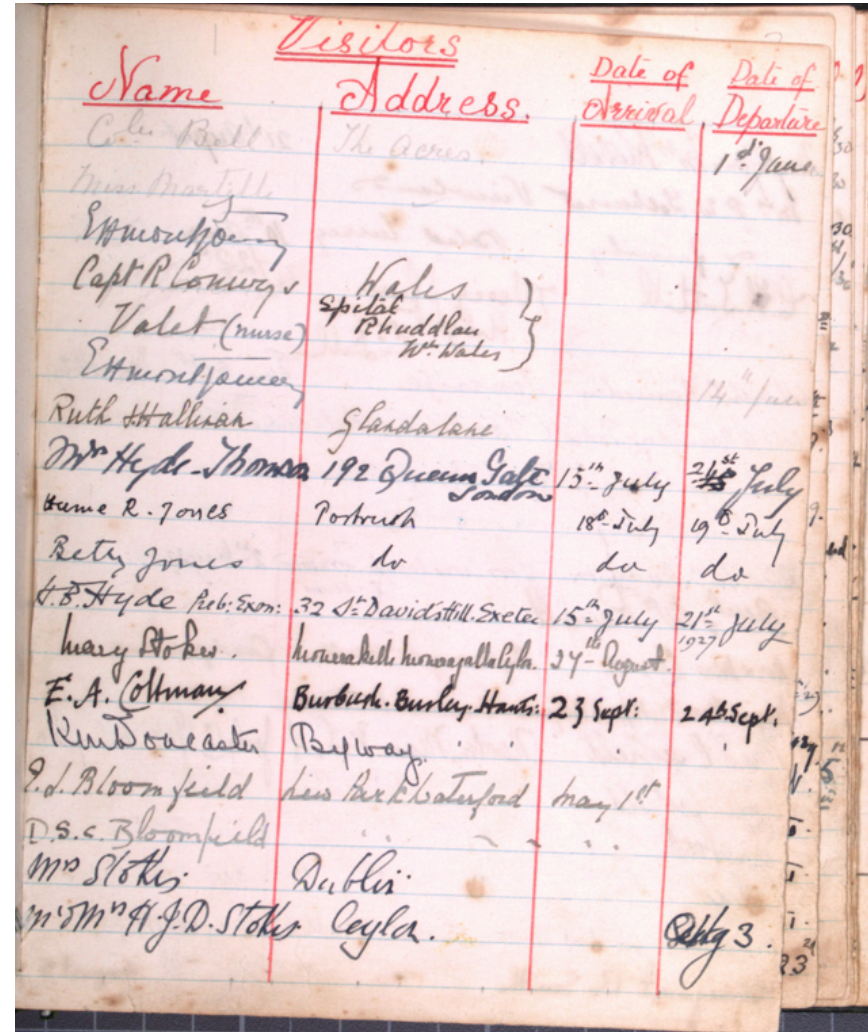
Please email myself (john.keating@nuim.ie) or Aja Teehan (aja.teehan@nuim.ie) for a preview page if you are interested in participating. If you know someone who would like to participate, please forward this message.

Best wishes,

Aja Teehan and John Keating.

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.



<u>Name</u>	<u>Address</u>	<u>Date of Arrival</u>	<u>Date of Departure</u>
Col. Ballin	The Acres		1 st June
Miss Mary F. L.			
Spencer			
Capt R. Conway	Wales		
Valet (nurse)	Spital Road, Huddellau, W. Wales		
Edmund Jones			14 th June
Ruth Hallinan	Glandatane		
Mr Hyde-Thomson	192 Queen's Gate, London	15 th July	21 st July
James R. Jones	Portrush	18 th July	19 th July
Betty Jones	do	do	do
G. B. Hyde	32 St. David's Hill, Exeter	15 th July	21 st July
Mary Stokes	Marine Hill, Marazion, Cornwall	24 th August	1927
E. A. Colman	Burbuck, Burley, Hants	23 Sept	24 th Sept
Kindred Carter	Bayway		
P. d. Bloomfield	115 Park Road, Watford	May 1 st	
P. S. C. Bloomfield			
Mr Stokes	Dublin		
Mr & Mrs J. D. Stokes	Ceylon		

Detailed Description (based on email)

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.

PAGE 1 FRONT NAME

ADDRESS

DATE OF ARRIVAL

DATE OF DEPARTURE

Coln Bell	The Acres		1 st June
Miss Martille			
E A Montgomery			
Capt R Conway V	Wales		
Valet (nurse)	Spitak Rhuddlan W.Wales		
E A Montgomery			14 th June
Ruth J. Hallian	Gandalane		
Mrs Hyde-Thomson	192 Queens Gate London	15 th -July	21 st July
Hume R. Jones	Portrush	18 th -July	19 th -July
Betty Jones	do	do	do
H.B. Hyde Pub:Exon:	32 St. David's Hill. Exeter	15 th -July	21 st -July 1927
Mary Stokes	Monerakelle, Moneragalla, Ceylon.	27 th -August	
E.A. Coltmany	Burbank. Burley. Hants:	23 Sept:	24 th Sept:
Ken Doncaster	Byway, " "	" "	" "
E. d. Bloomfield	Hines Park Waterford	May 1 st	
D.S.C. Bloomfield	" " " "	" "	
Mrs. Stokes	Dublin		
Mr & Mrs H.J.D. Stokes	Ceylon		Aug.3

```
<area shape="poly" alt="row ruth" coords="128,1476,128,1644,1260,1784,2084,1720,2092,1544" nohref title="row ruth" />
<area shape="poly" alt="row thomas" coords="128,1648,128,1844,472,1932,2084,1972,2728,2024,2944,2036,3068,1936,3072,1756" nohref title="row thomas" />
```


Castlehyde Estate Guestbook

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

<u>Visitors</u>					
<u>Name</u>	<u>Address.</u>	<u>Date of</u> <u>Arrival</u>	<u>Date of</u> <u>Departure</u>	<u>DATE OF</u> <u>DEPARTURE</u>	
C. B. Bell	The Acres		1 st June		
Miss Mary Fille					
Wm. J. Conroy	Hales				
Capt R. Conroy	Spital				
Valet (nurse)	Rhuddlan				
E. J. Conroy	W. Hales				
Ruth Hallinan	Glandatane				
Mr Hyde-Thomson	192 Queens Gate	15 th July	21 st July		
James R. Jones	Portrush	18 th July	19 th July		
Betty Jones	do	do	do		
H. B. Hyde	32 St. David's Hill, Exeter	15 th July	21 st July		
Mary Stokes	harrowhill, harrow, pallisley	24 th August	1927		
E. A. Colman	Burbuck, Burley, Hants.	23 Sept.	24 Sept.		
Kind O'neaster	Bayway				
P. d. Bloomfield	Leas Park, Waterford	May 1 st			
P. S. C. Bloomfield					
Mr Stokes	Dublin				
Mr M. J. D. Stokes	Coyler.				

Data outside cell boundaries

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?

	Encoding 1	Encoding 2	Encoding 3	Encoding 4	Encoding 5
TEI	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Tabular Structure	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Semantic Structure	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Problematic Overlaps of Physical and Logical Classes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Searchable Date	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Searchable Address	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Encoded Individuals	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Rendering Features	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Mechanism for \" or \"ditto\"	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Type of Page	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Text/image Linking	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

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.**”

- 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!

Thank you!

Questions, Comments?

Aja Teehan
John Keating

firstname.lastname@nuim.ie