Elli Bleeker

Addressing Ancient Promises : Text Modeling and Alexandria

Vortrag
Digital Humanities Kolloquium der Berlin-Brandenburgischen Akademie der Wissenschaften ;
(Berlin) : 2018.11.02

Persistent Identifier: urn:nbn:de:kobv:b4-opus-29325

Addressing Ancient Promises

Text Modeling and Alexandria

Elli Bleeker
Research and Development - Humanities Cluster, Royal Academy of Arts and Sciences

DH-Kolloquim, Berlin-Brandenburgische Akademie der Wissenschaften
2 November 2018

NOTE

For my presentation, I made use of the RISE extension of Jupyter Notebook, that allows you to create a reveal.js-based presentation. You can download Jupyter notebook here (https://jupyter.org/install) and an introduction of RISE here (https://rise.readthedocs.io/en/stable/index.html).

You can download my entire talk here (https://github.com/bleekere/TAG/blob/master/talk-bbaw-2018/talk-bbaw_20181102.ipynb). After installing RISE in the same folder as you have downloaded this presentation, you can see it as a slideshow.

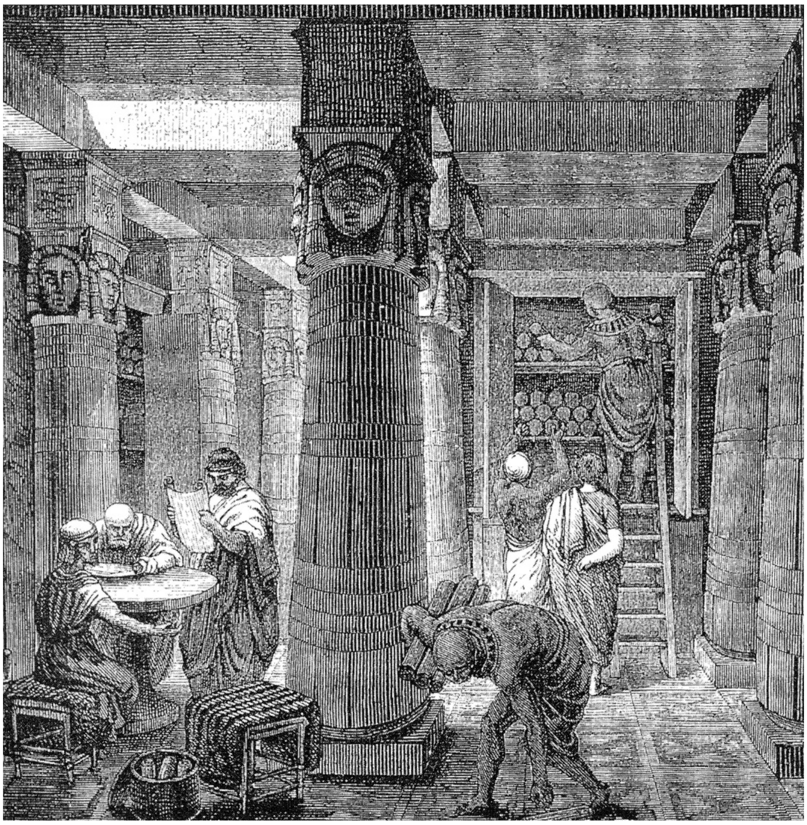
The following is a downloaded version of my talk, that reads as one long text. The weird formatting is because of the conversions of the RISE-slides into text.

In *The Call of the Cthulhu* (1926), science fiction writer H.P. Lovecraft wrote:

"The most merciful thing in the world, I think, is the inability of the human mind to correlate all its contents. We live on a placid island of ignorance in the midst of black seas of infinity, and it was not meant that we should voyage far. The sciences, each straining in its own direction, have hitherto harmed us little; but some day the piecing together of dissociated knowledge will open up such terrifying vistas of reality, and of our frightful position therein, that we shall either go mad from the revelation or flee from the deadly light into the peace and safety of a new dark age."

H.P. Lovecraft, 1928

Despite the ominous picture that is painted here, humankind has had the desire to "piece together dissociated knowledge" for centuries. Around 290 BC, King Ptolemy the First (or his son, king Ptolemy the Second; the record's not clear on that) charged the founding of a research institute which they called the "Mouseion". Part of this research institute was the library of Alexandria, famous to this day, which had for objective to be a "universal library". The goal was to "collect all the books in the world" and, notably, "the writings of all men *as far as they were worth any attention*".



The librarians were collectors of information and developed a methodology of examining manuscript copies and establishing "the best" text. This methodology forms the basis of textual criticism as we know it today. They created what can be seen as the first centre of knowledge production.

The goal of the Mouseion and Alexandria was

to unite all scholarly knowledge

There was undeniably a political reasoning behind this objective: the ancient Egyptian rulers certainly understood the truth behind the idiom "knowledge is power" and were pretty rigorous in their methods of aquisition and their dissemination policy. But, politics aside, the combination of a vast library collection and an international research center facilitated the production of more knowledge. Scholars from all over the civilised world came to work in the Mouseion. They had considerable academic freedom and made avid use of the libary's collection which resulted in high-end scholarship and led to advances in the fields of natural sciences, geography, mathematics, literature, and philosophy. I think it safe to assume that today's digital libraries and archives still have this noble ambition.

But what is that, exactly: *unite all scholarly knowledge*? Arguably, the least problematic word in that phrase is "scholarly", although what is deemed scholarly and what not is also a matter of debate. The other words, however, are even trickier. What does it mean, to "unite"? Does it suffice to bring together all books in one physical place? Books may stand side-by-side on a shelve, but that doesn't mean that their contents are united, let alone synthesised. And what is knowledge, actually? Is it even possible to have it *all*?

I'm tempted to respond to that last question with "no, it isn't". That is incidentally also the principle the Open World Assumption of the Semantic Web:

"No single agent or observer has complete knowledge... We admit that our knowledge of the world is incomplete."

Open Semantic Framework Wiki

It basically means that the absence of a piece of information doesn't make it false: something can be true even if we cannot prove it's true and the other way around.

Phrases like "unite scholarly knowledge" also remind us of visions of the early digital textual scholars, who saw in the digital medium the way to make information accessible for everyone to look at from different perspectives, across disciplinary divides. They believed it possible to create every expanding knowledge sites that transcend individual project "silos" following the Linked Open Data principles. Those of you working in digital textual scholarship will know that, despite our best attentions, we have not yet reached that goal.

In fact, such closed data silos are still being developed. In a recent "review-turned-reflection" on the Mirador image viewer, my colleague Joris van Zundert identifies at least two causes: the limited financial resources (intra-institutional collaboration is more costly, both in time and in money, than a local solution) and the convenience of the developers who incidentally *also* have to deal with limited time and money. In short, even though *in principle* most scholars would agree that

"... keeping tools and data locked in one place behind a one-size-fits-all interface would be in stark contrast to the heterogeneous demands and requirements scholars had for document resources that were spread, quite literally, across the planet..."

Joris van Zundert, 2018

it's still pretty much our reality.

In this talk, I describe how digital text modeling can, potentially, fulfil the objective of the great research libraries of the past: unite and disseminate scholarly knowledge. I identify three requirements (or let's call them "desiderata", as a salute to Peter Robinson):

Desiderata

1. an **inclusive data model** that

   1.1. natively models advanced textual features

   1.2. provides a way to formalise the meaning of markup;

1. **support for multiple perspectives on a source text**;

1. an **editing and version management tool** that fits within a **distributed architecture**.

The discussion of these desiderata form primarily the background for a presentation of *Alexandria*, a text repository system that is also the reference implementation of a new data model called TAG (Text-as-Graph). Both are under development at the R&D group of the HuC.

I'll go in broad sweeps over concepts like information modeling, with which I'm sure all of you are familiar, but it gives me the chance to abuse computers a bit while citing computer scientists.

Because these requirements are related to the principles of the Semantic Web, the Linked Open Data (LOD) framework, and the ideals of a distributed architecture, I will touch upon those topics as well, but only in passing: the main goal of my talk is to establish an informed discussion about information modeling, data structures and digital textual scholarship. And, of course, to present and promote our own *Alexandria*.

Definitions

First, some definitions (which I'll probably use interchangeably anyway).

Information

Information scholar Michael Buckland distinguishes three types of information:

1. Information-as-process
2. Information-as-knowledge
3. Information-as-thing

1. **information-as-process**: the process of acquiring information, which may lead to knowledge.
2. **information-as-knowledge**: intangible. It needs to be expressed in a physical way, as a signal or communication. The expression makes the third form:
3. **information-as-thing**: tangible. Books, databases, bits, bytes, files, etc. Data. Text. Can be touched or measured. Representations of knowledge (because knowledge itself cannot be made tangible). Defined by Buckland as "evidence": it is related to understanding but in a passive way, as inferences can be drawn from it. This is the main focus of our work today.

Data

"There is a tendency to use "data" to denote numerical information and to use "text" to denote natural language in any medium. Some humanist scholars are even pointedly against considering texts as data, like Stephen Marche (2012)

"*Literary texts cannot be reduced to data because they are too complex, very messy and never complete*"

Stephen Marche, 2012

In a similar vain, computer scientist William Kent said that

"*Information in its 'real' essence is probably too amorphous, too ambiguous, too subjective, too slippery and elusive, to ever be pinned down precisely by the objective and deterministic processes embodied in a computer.*"

William Kent, 1978

Still, I'd argue that this comment rather protests what "data" has come to mean in DH: the reduction of literary texts to quantifiable objects. Personally, I see no objection to the term: before they can be processed and analysed, literary texts are transformed into data with a certain structure. There's no question that this transformation entails a reduction and a loss of information, but that's rather the nature of modelling.

Modeling

By definition,

a model is always a selection of the original

"We always pay a price for theoretical knowledge: reality is infinitely rich; concepts are abstract, are poor. But it’s precisely this 'poverty' that makes it possible to handle them, and therefore to know."

Franco Moretti, 2000

said Franco Moretti. Still, even though we understand the limits of models, it is considered

"the holy grail of computer science to algorithmically express the natural world"

Teresa Nakra, 2006

At least, according to professor Teresa Nakra, who's a computer scientist and musicologist. Why is this a holy grail? Is it really so hard?

Because the natural world is ambiguous and complex; its products open to many different interpretations.

The computer, on the other hand, is rather dumb. Simplistic, if you will.

William Kent described computer programming as

"... the art of getting and imbecile to play bridge or to fill out his tax returns by himself. It can be done, provided you know how to exploit the imbeciles limited talents, and are willing to have enormous patience with his inability to make the most trivial common sense decisions on his own."

"The first step toward understanding computers is an appreciation of their simplicity, not their complexity."

Kent, 1978

Despite this utter stupidity we're apparently dealing with on a day to day basis, we can already do lots of cool stuff with a computer when it comes to modeling complex documents. Achievements that, if we follow Kent's argument, can be wholly and uniquely attributed to our human intelligence! So, unfaltering, we continue with our quest.

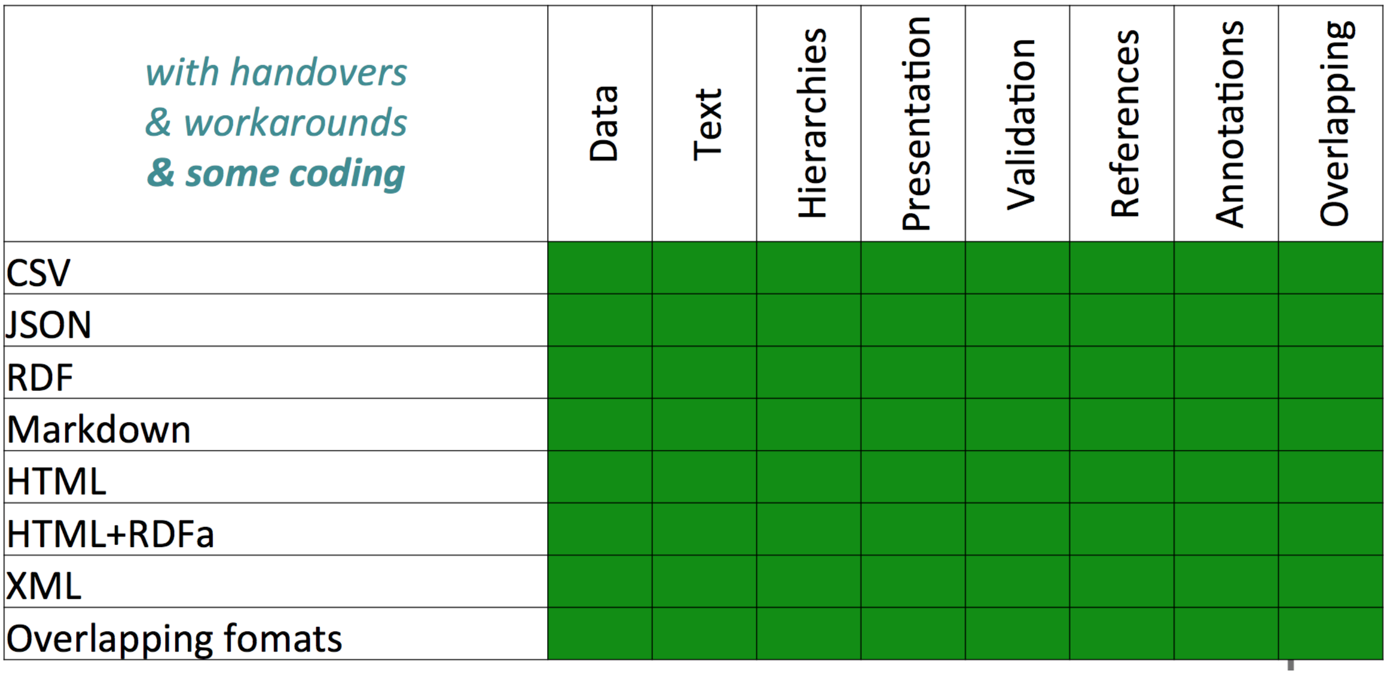
This brings me to the first point, the first "desideratum" if you will:

1. A Flexible and Inclusive Data Model

We have at our disposition a number of different data models to express and represent text. I'll give a brief outline of the most common or extraordinary ones. Keep in mind the difference between a data model and a syntax! A syntax is a serialisation of a data model, e.g., XML is a serialisation (= a linearly ordered expression) of the OHCO model, but SGML as well.

Data model	Serialisation
String	plain text (ranges)
OHCO	SGML, XML
Graph	Turtle, N-Triple, RDF/XML...

Each of these formats has disadvantages but also their own merits. As Fabio Vitali argued a few years back, with the help of some coding and hacks you can express almost everything in every data model (though it ain't always pretty).



*with handovers & workarounds **& some coding***	Data	Text	Hierarchies	Presentation	Validation	References	Annotations	Overlapping
CSV	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
JSON	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
RDF	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Markdown	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
HTML	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
HTML+RDFa	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
XML	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Overlapping fomats	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩

Image by Vitali (2016)

Making use of what is already there has a lot of benefits: it is more or less stable, people know and understand it, there's usually a community, tutorials and tools made to work with that format. The downsides:

1. Models influence the way we think and argue about text.

It's a sneaky one: if we use certain models long enough, they can influence the way we think and argue about text. They can even - very subtly - encourage us to ignore certain features that are not represented in that particular model. Patrick Sahle noted, with regard to TEI,

"TEI konzentriert sich for allen auf den Text als Werk(-Struktur), als sprachliche Äußerung und als kanonisierte, definierte oder auch variante Fassung. Der Text als intentionale Mitteilung, als semantischer Inhalt, aber auch der Text als physisches Object, als Dokument, wird nur am Rande unterstützt. Der Text als komplexes Zeichen, als semiotic Entität, spielt bei der TEI keine Rolle."

Patrick Sahle, 2013

1. Workarounds and local "solutions" hinder interoperability and reusability

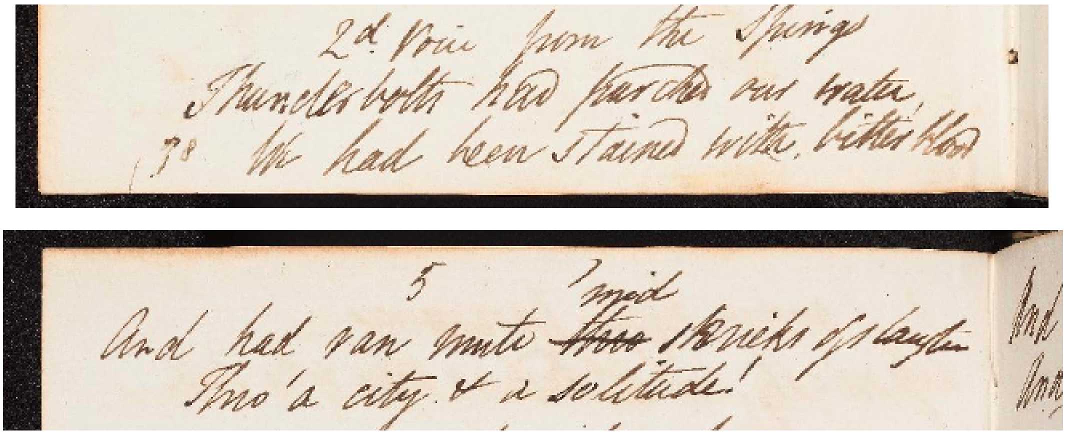
The more complex your texts (or what you want to do with it), the more coding, hacking and workarounds. I am not saying that modeling complex textual features in XML or RDF is impossible. It may lead to a reduced human-readability of the file (which is more important than you may think, especially when it concerns humanities scholars), but you may argue that the file is not intended to be read by humans, as long as machines understand it. However, it also hinders interoperability and delegates a lot of responsibility and complexity to the application level that processes the data.

Let's take a closer look at what I mean, exactly, with "complex textual features" that are hard to express in existing models. I'll show an example from a modern literary manuscript, but you can find these structures anywhere.

Difficult textual features

- Overlapping structures
- Discontinuous elements
- Non-linear elements

Overlapping structures



2d Voice from the Springs
Thunderbolts had parched our water
78 We had been stained with bitter blood

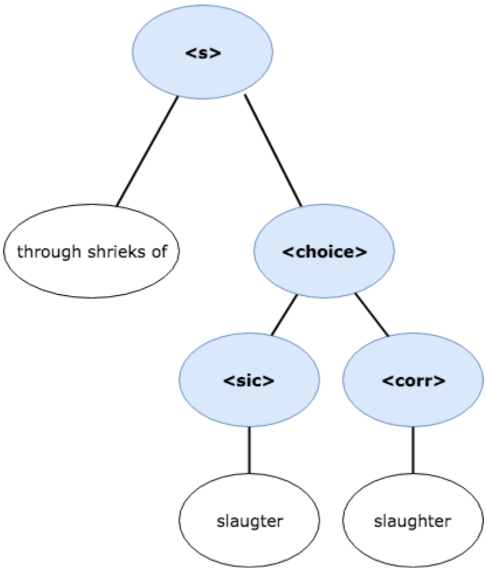
5 'mid
And had ran mute ~~thro~~ shrieks of slaugter
Thro' a city & a solitude!

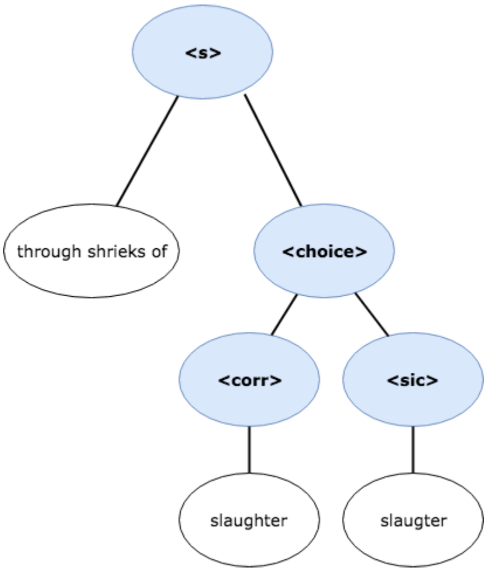
Fragment from the authorial manuscript "Prometheus Unbound" by Percy Bysshe Shelley (pages 21v and 22v; retrieved from http://shelleygodwinarchive.org/sc/oxford/prometheus_unbound/act/i/#/p8 (http://shelleygodwinarchive.org/sc/oxford/prometheus_unbound/act/i/#/p8))

Non-linear structures

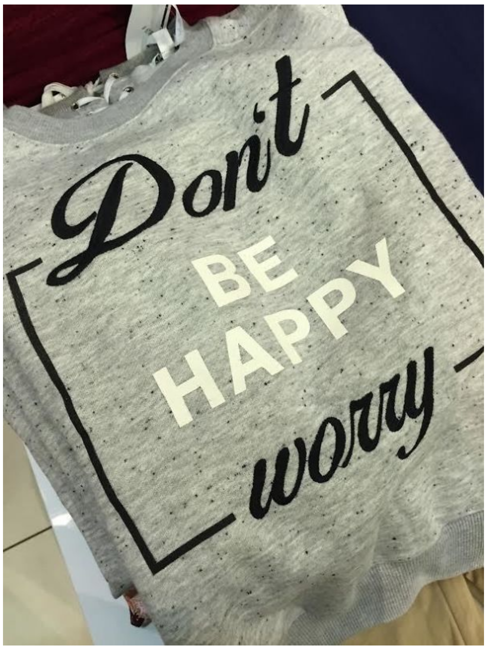
```
<s> through shrieks of
  <choice>
    <sic>slaugter</sic>
    <corr>slaughter</corr>
  </choice>
</s>
```

```
<s> through shrieks of
  <choice>
    <corr>slaughter</corr>
    <sic>slaugter</sic>
  </choice>
</s>
```





Discontinuous elements



If you are interested in such phenomena, and if you're modelling text I can hardly imagine you're *not* interested in them, then there's value in using a data model that is close to your understanding of text. A model that can deal with these features natively. With this in mind, we developed

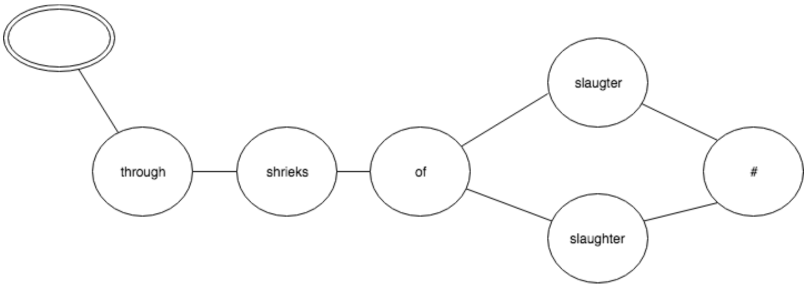
a hypergraph model for text

Data model		Serialisation	
String		plain text (ranges)	
OHCO		SGML, XML	
Graph		Turtle, N-Triple, RDF/XML, **hypergraph for text**	

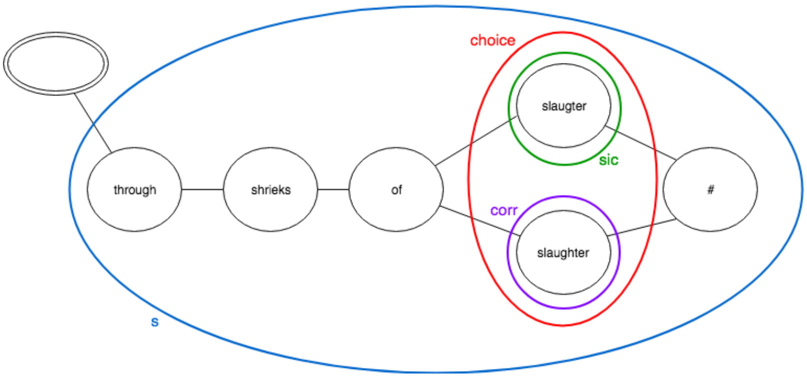
Like the graphs we're used to, a hypergraph consists of nodes and edges, but it also has hyperedges that can connect more than two nodes. To have the model support our understanding of text, we needed to define it as precisely as possible. In our definition,

Text is a multilayered, nonlinear object containing information that is at times ordered, partially ordered and unordered.

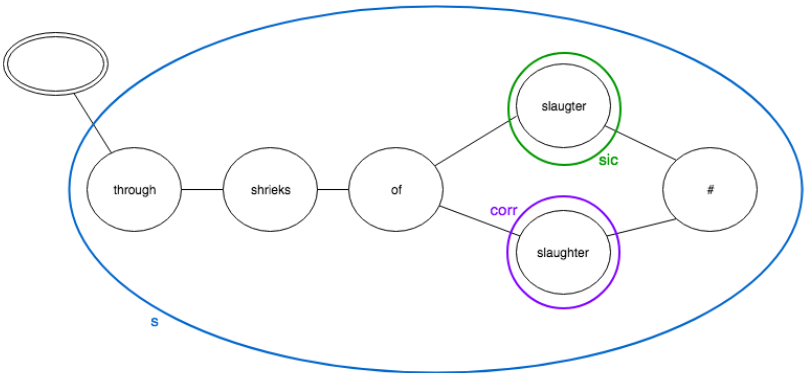
Nonlinearity in TAG



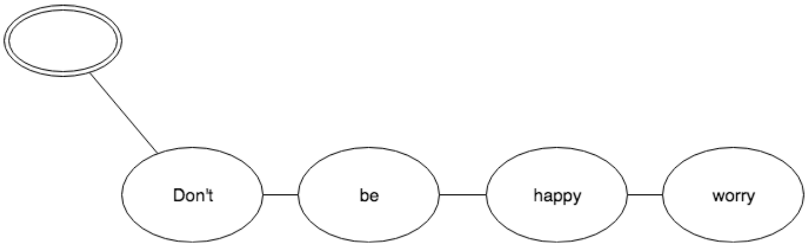
Nonlinearity in TAG



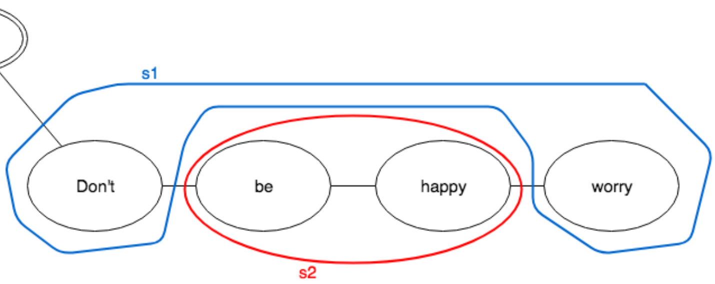
Nonlinearity in TAG



Discontinuity in TAG



Discontinuity in TAG



If I describe how the TAG hypergraph model deals with complex texts, we're here for the rest of the evening. In addition to the data model, we've designed a serialisation (a markup language) called TAGML, which requires a grammar and a schema and that poses challenges of its own. If you want, we can get back to it later. For now, let's move on to the second desideratum:

2. Multiple Perspectives

What do I mean with that? Well, remember the Open World Assumption that we can never be entirely sure we know everthing? You can also put it this way:

Scholars disagree on everything.

This is very much okay, because opposing views are the driving force behind research. Also: views do not always oppose one another but happily coexist. In digital text modeling, these coexisting perspectives are often a cause for overlap. So typically editors either choose one dominating perspective. I know of only one example of an edition project that actually encoded both perspectives: the Faust Edition.

In the TAG model, overlap is not an issue anymore. But how can we deal with these coexisting yet different views on the same data? Our solution is

Layers

Layers are, in fact, TAG's solution to many issues.

Layers classify a set of markup nodes.

These nodes can be classified as belonging to a certain research perspective (like nodes expressing the materiality of a document, or nodes expressing linguistic information) or they identify which user has added the markup (like all markup added by Elli, or all markup added by Frederike).

Layers are hierarchical: the nodes within one layer are hierarchically ordered.

Layers may share markup nodes and textual nodes.

In other words: layers may overlap, but within one layer there can be no overlapping markup. This feature touches upon the discussion of containment and dominance, but we'll get to that when we have time.

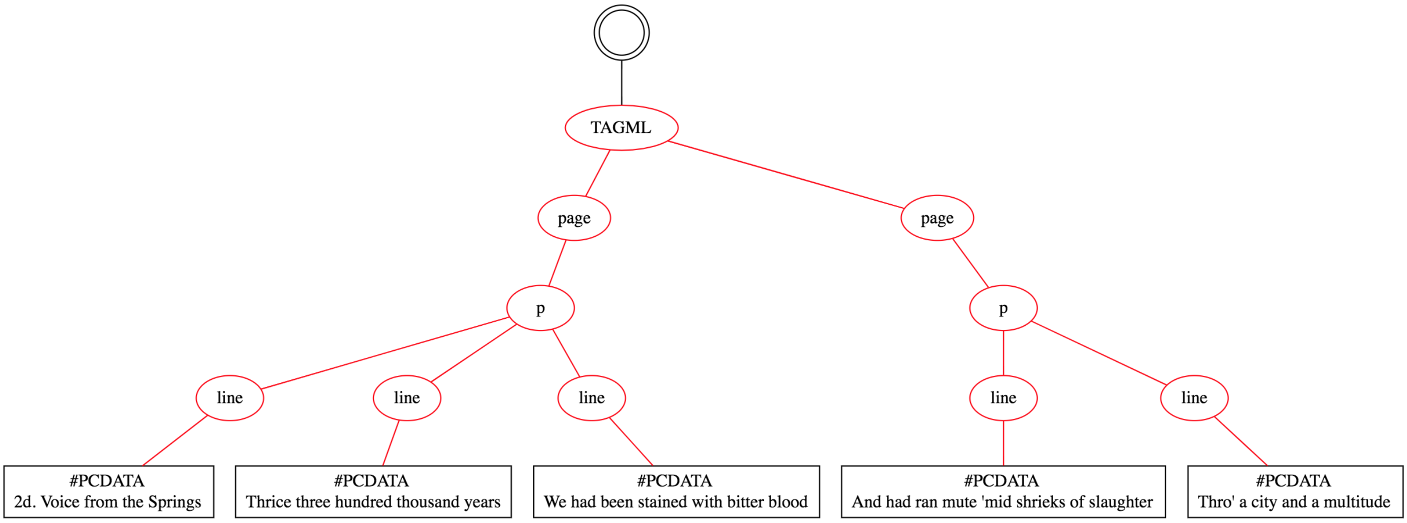
Layers can start locally, at any point in the document

In the second part of this lecture, I'll give a demo to show how layers work, but here you can already see some examples in fragments of a TAGML transcription:

Astrid

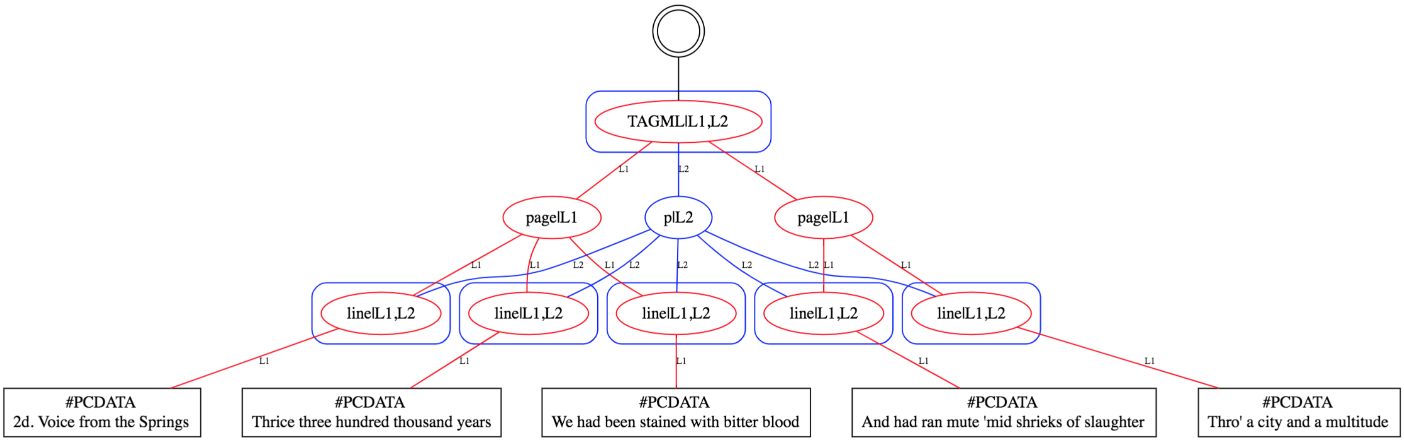
```
[TAGML>
[page>
[p>
[line>2d. Voice from the Springs<line]
[line>Thrice three hundred thousand years<line]
[line>We had been stained with bitter blood<line]
<p]
<page]
[page>
[p>
[line>And had ran mute 'mid shrieks of slaughter<line]
[line>Thro' a city and a multitude<line]
<p]
<page]
<TAGML]
```

The layers in TAG are Multi-Colored Trees (MCT).


TAGML
page
page
p
p
line
line
line
line
line
#PCDATA
2d. Voice from the Springs
#PCDATA
Thrice three hundred thousand years
#PCDATA
We had been stained with bitter blood
#PCDATA
And had ran mute 'mid shrieks of slaughter
#PCDATA
Thro' a city and a multitude

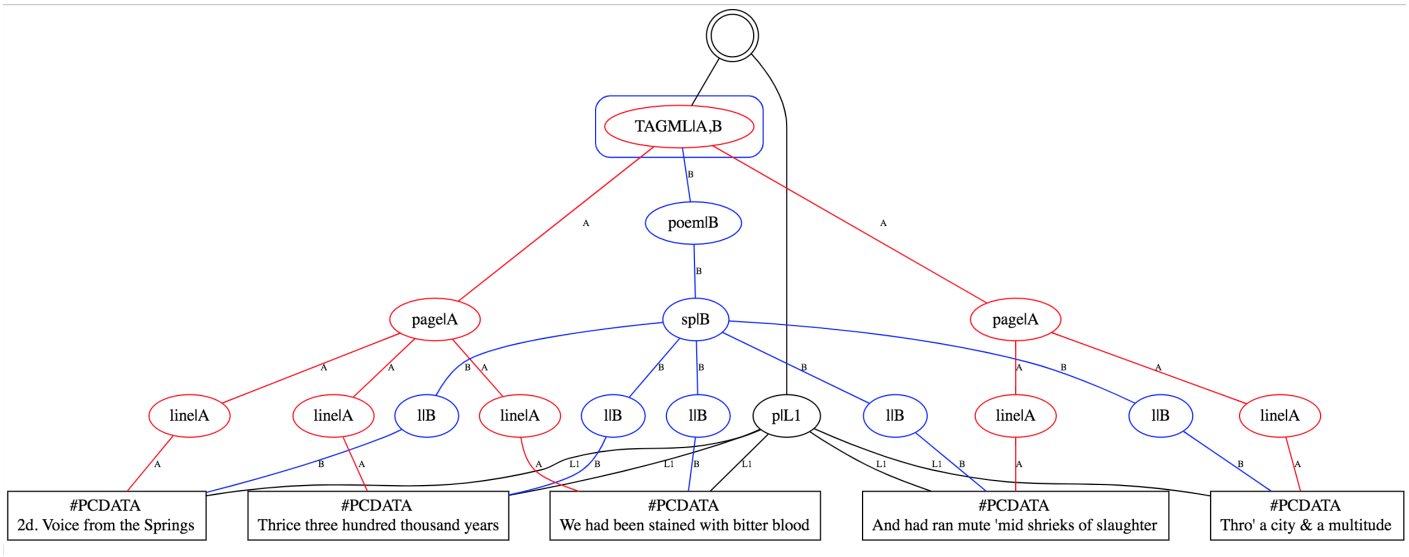
Astrid

```
[TAGML>
[page|+L1>
[p|+L2>
[line>2d. Voice from the Springs<line]
[line>Thrice three hundred thousand years<line]
[line>We had been stained with bitter blood<line]
<page|L1]
[page|L1>
[line>And had ran mute 'mid shrieks of slaughter<line]
[line>Thro' a city and a multitude<line]
<p|L2]
<page|L2]
<TAGML]
```


TAGML|L1,L2
L1
L2
L1
page|L1
p|L2
page|L1
L1
L1
L2
L1
L2
L2
L2
L2
L1
L1
line|L1,L2
line|L1,L2
line|L1,L2
line|L1,L2
line|L1,L2
L1
L1
L1
L1
L1
#PCDATA
2d. Voice from the Springs
#PCDATA
Thrice three hundred thousand years
#PCDATA
We had been stained with bitter blood
#PCDATA
And had ran mute 'mid shrieks of slaughter
#PCDATA
Thro' a city and a multitude

Bram

```
[TAGML>
[page|+A,+L1>
[poem|+B>
[p|A,+L2>
[sp|B>[l|B>[line|A>2d. Voice from the Springs<l]<line]
[line|A>Thrice three hundred thousand years<l]<line]
[line|A>We had been stained with bitter blood<l]<line]
<page|L1]
[page|A,L1>
[line|A>[l|B>And had ran mute 'mid shrieks of slaughter<l]<line]
[line|A>[l|B>Thro' a city & a multitude<l]<line]
<p|L2]
<sp]
<page]
<TAGML]
```


TAGML|A,B
poem|B
page|A
sp|B
page|A
line|A
line|A
l|B
line|A
l|B
l|B
p|L1
l|B
line|A
l|B
line|A
#PCDATA
2d. Voice from the Springs
#PCDATA
Thrice three hundred thousand years
#PCDATA
We had been stained with bitter blood
#PCDATA
And had ran mute 'mid shrieks of slaughter
#PCDATA
Thro' a city & a multitude
A
B
L1

Working with layers has a number of important implications:

Implications

- abandoning a shared conception of digital editing in favor of multiple perspectives
- coexisting views, not one view leading

- increased readability *and* reusability
- documentation: clearly communicate what information is in a file

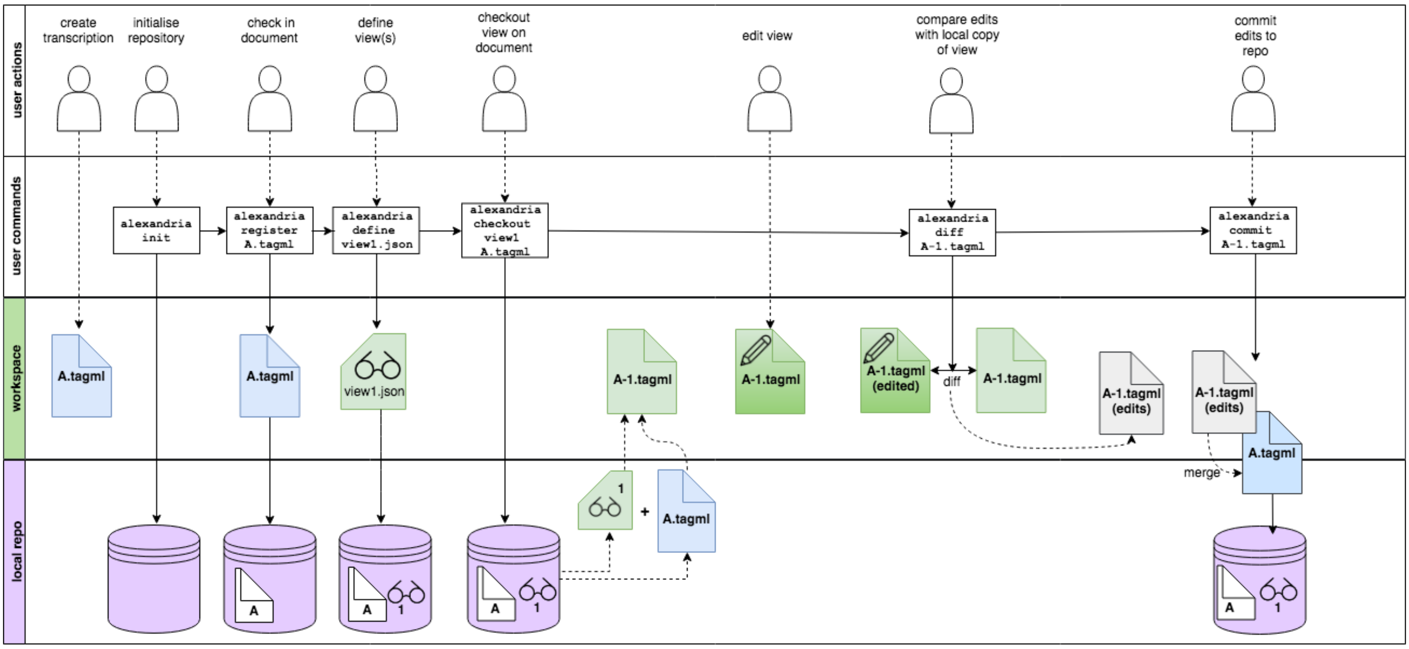
With regard to the readability and reusability: this is the case when filtering out one or more layers, e.g., "show me only the markup with the layer ID [A]". With regard to the documentation: we're working on generating documentation to a certain extent.

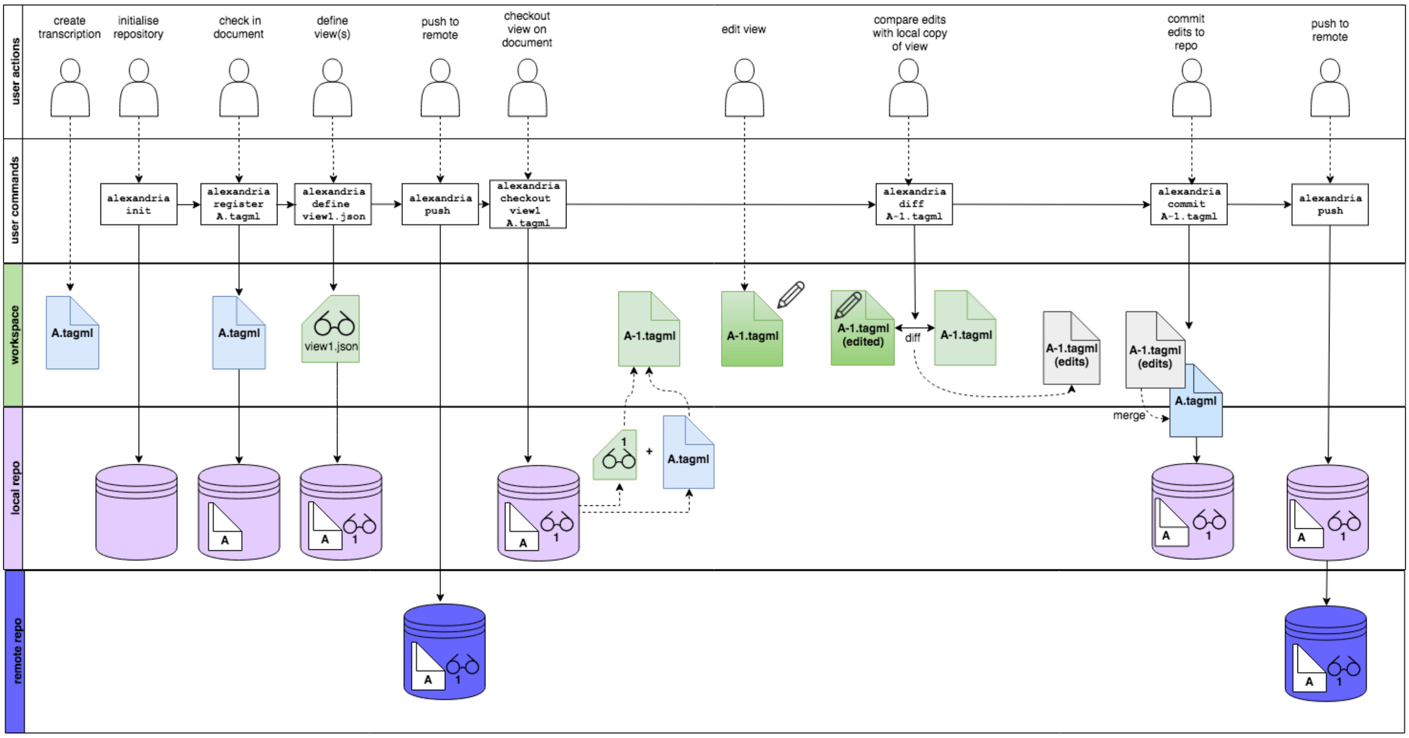
We have come to the third and final "desideratum" or required component: a platform-independent system to work with TAG files containing multiple layers.

3. Distributed System

First, let's talk about the system.

Alexandria works similar to git:





[demo]

You may imagine (or not, but then let me point it out to you) that a workflow with layers implies a comparison between two marked-up files, finding changes in markup, text or both.

We have developed a versioning tool that is able to do that (although it requires more testing) so the challenge is sooner conceptual (or philosophical) than technical.

It comes down to questions like:

What defines "change"? How do you identify a new version?

Do you want to label the changes as additions (keeping the first version) or as substitutions (disgarding the first version)? If I change Astrid's `[line>` markup to `[s>`, should we keep the `[line>` information as well?

Is a perspective only markup, or also text?

Intuitively we'd say that a perspective consists of text *and* markup: in one perspective, a transcriber can identift a typo with `[sic> [corr>` while another transcriber may choose to disgard the typo and silently correct it. As a result, one perspective contains both `slaugter` and `slaughter` while the other perspective only contains `slaughter`.

Still, there are no ready answers to these and similar questions. On the contrary, there are possibly opposing views, but as I've just argued that opposing views can lead to innovation that seems fine.

We just need more testing and more people to work with TAG and profound philosophical discussions about markup, modeling and text.

Remember the Holy Grail of computer science? According to Joris van Zundert, there's also a holy grail of digital textual scholarship:

"*The ultimate transcription environment has become something of a holy grail within digital textual scholarship.*"

And, in line with the metaphor, all attempts to find that grail have been in vain:

"*There is also a graveyard somewhere for scholarly transcription environments... In the case of transcription tools, a defining trope would be that the tool was built as an integrated transcription environment*"

Van Zundert, 2018

It's a truism that if you want a large community, that has developed its own set of preferred methods, tools and convictions, to drop all that and adopt your does-it-all tool, you're gonna fail.

Alexandria is therefore implemented as a command-line tool without an interface. It's a skeleton, a basic structure for which you can developed your own interface, if you please, or which you can integrate in an existing environment of choice. It is not dependent on a platform or an operating system and fits within a distributed architecture.

With Alexandria, in short, we do *not* provide an integrated transcription environment that will, in due time, find it's way to its friends on a graveyard, but we provide a flexible tool that can be integrated in an editorial workflow.

Not coincidentally, *Alexandria* is created by two developers who were also part of the *Interedition* project (2007-2013) that can be accredited with tools like CollateX and StemmaWeb among others. *Alexandria* is an open source software component that allows for customisation.

In conclusion

With the TAG data model and the reference implementation *Alexandria* we've taken some significant steps towards that holy grail of computer science: expressing the natural world algorithmically so that the computer - that imbecile - can understand and be an even better instrument for scholarly research. A while ago, Wendell Piez said of his LMNL data model that

"*both the analysis and the representation of literary texts are enabled in new ways by being able to overlap freely in the model*"

Wendell Piez, 2014

We can paraphrase that, saying

Both the analysis and representation of, and the collaboration on literary texts are enabled in new ways by being able to natively express complex textual features in the model.

Using *Alexandria* and the TAG data model does have some epistemological consequences for our understanding of a version, of text, of information produced by scholars, etc. I'll just mention some, without necessarily answering them. Food for thought and a basis for the discussion.

- Can we represent information about text so that others can meaningfully interact with it?

- What does it mean when our textual models are *conceptually* no longer limited by a particular format or structure?

- Where do we stand regarding ambiguity in our description of information?

The layer-feature of TAG allows you to express ambiguous information about a text. On the other hand, there's significant value in disambiguating the tags used for each layer, for instance by linking them to a sort of ontology that formalises the semantics of your tag set. This facilitates processing, querying and reusing. Now that I mention "ontology" we may also have to discuss semantics in realition to TAG. If there's time...

I want to emphasise here that we shouldn't pass lightly over the value of existing models and data structures. RDF and XML have much going for them:

Don't throw the baby out with the bath water!

- It's easy to start right away with XML; there are tutorials all over the web, there's a strong and active community
- There's never enough time and money; if you're on a tight budget it can make sense to cut technical training of editors or to stay clear of experimenting
- Not everyone has revolutionary ambitions; sometimes all you want to do is publish fast and cheap.

These are solid reasons and we don't argue against them. All of it is possible in TAG, and perhaps even easier: we can follow the TEI encoding standard but avoid the complicated workarounds when you encounter forms of overlap.

Furthermore, the TAG model entails a strict separation between responsibilities (encoding - schema - semantics), outsourcing certain responsibilities to the application layer which makes it easier to process.

Again, I cite Piez:

"The primary goal of text encoding in the humanities should not be to conform to standards ... Rather, we encode texts and represent them digitally in order to present, examine, study, and reflect on the rich heritage of knowledge and expression presented to us in our cultural legacy"

Piez, 2014

The combination of TAG and *Alexandria* provides us with a powerful modelling tool:

TAG and *Alexandria*

- multiple coexisting views
- inclusive data model
- modular, open source software

Even though we are in the midst of development, these features already influence how we model, think about, proceas and analyse text. The abstract objective of the ancient Museion, "to unite all scholarly knowledge" has become a concrete and even attainable goal.

Credits

Research & Development team involved in TAG: Ronald Haentjens Dekker, Bram Buitendijk, Astrid Kulsdom, Elli Bleeker.

See Alexandria (https://github.com/HuygensING/alexandria-markup-server/blob/master/README.md) and TAG (https://github.com/HuygensING/TAG) on Github

References

- The BECHAMEL project, Illinois School of Information Science, see https://ischool.illinois.edu/research/projects/bechamel-markup-semantics-project (https://ischool.illinois.edu/research/projects/bechamel-markup-semantics-project).
- Bleeker, Elli, Bram Buitendijk, Ronald Haentjens Dekker, and Astrid Kulsdom. "Including XML markup in the automated collation of literary texts." Presented at XML Prague 2018, Prague, Czech Republic, February 8–10, 2018. In *XML Prague 2018 - Conference Proceedings, pp. 77–95*. http://archive.xmlprague.cz/2018/files/xmlprague-2018-proceedings.pdf (http://archive.xmlprague.cz/2018/files/xmlprague-2018-proceedings.pdf)
- Buckland, Michael K. 1991. "Information as Thing". In *Journal of the American Society for Information Science* vol. 42, no. 5, pp. 351-350.
- Ciotti, Fabio and Francesca Tomasi. 2016. "Formal Ontologies, Linked Data, and TEI Semantics". In *Journal of the Text Encoding Initiative*, issue 9, September 2016 - December 2017.
- Davidson, Justin. 2006. 'Measure for Measure: Exploring the Mysteries of Conducting.' The New Yorker, pp. 60-69
- Jagadish, H. V., Laks VS Lakshmanan, Monica Scannapieco, Divesh Srivastava, and Nuwee Wiwatwattana. 2004. "Colorful XML: one hierarchy isn't enough." In *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, pp. 251-262. ACM, 2004.
- Haentjens Dekker, Ronald and David J. Birnbaum. 2017. "It's more than just overlap: Text As Graph." Presented at Balisage: The Markup Conference 2017, Washington, DC, August 1 - 4, 2017. In *Proceedings of Balisage: The Markup Conference 2017. Balisage Series on Markup Technologies, vol. 19*. DOI: https://doi.org/10.4242/BalisageVol19.Dekker01 (https://doi.org/10.4242/BalisageVol19.Dekker01).
- Haentjens Dekker, Ronald, Elli Bleeker, Bram Buitendijk, Astrid Kulsdom and David J. Birnbaum. "TAGML: A markup language of many dimensions." Presented at Balisage: The Markup Conference 2018, Washington, DC, July 31 - August 3, 2018. In *Proceedings of Balisage: The Markup Conference 2018. Balisage Series on Markup Technologies, vol. 21*. DOI: https://doi.org/10.4242/BalisageVol21.HaentjensDekker01-XML-Prague-paper (https://doi.org/10.4242/BalisageVol21.HaentjensDekker01-XML-Prague-paper)
- Kent, William. 1978. *Data and Reality: Basic Assumptions in Data Processing Reconsidered*. North-Holland Publishing Company.
- Lovecraft, H.P. 1928. *The Call of the Cthulhu*. Available online at http://www.hplovecraft.com/writings/texts/fiction/cc.aspx (http://www.hplovecraft.com/writings/texts/fiction/cc.aspx)
- Marche, Stephen. 2012. "Literature Is not Data: Against Digital Humanities". In *Los Angeles Review of Books*. Online: https://lareviewofbooks.org/article/literature-is-not-data-against-digital-humanities/# (https://lareviewofbooks.org/article/literature-is-not-data-against-digital-humanities/#)
- Moretti, Franco. 2000. "Conjectures on World Literature". In *New Left Review* vol.1. Online available at https://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature (https://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature)
- Ore, Christian-Emil and Oyvind Eide. 2009. "TEI and Cultural Heritage Ontologies: Exchange of Information?" In *Literary and Linguistic Computing* vol. 24, no. 2, pp. 161-172.
- Peroni, Silvio, Aldo Gangemi, and Fabio Vitali. 2011. "Dealing with Markup Semantics". In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics 2011)*, edited by Clara Ghidini *et al*, New York: ACM, pp. 111-118.
- Piez, Wendell. 2014. "TEI in LMNL: Implications for Modeling". In *Journal of the Text Encoding Initiative*, vol.8.
- Robinson, Peter. "Desiderata for Digital Editions/Digital Humanists should get out of textual scholarship". Presented at Digital Humanities Conference 2013, Lincoln, Nebraska, 19 July 2013. Slides at https://www.slideshare.net/PeterRobinson10/peter-robinson-24420126 (https://www.slideshare.net/PeterRobinson10/peter-robinson-24420126)
- Sahle, Patrick. 2013. Digitale Editionsformen-Teil 3: Textbegriffe Und Recodierung. Norderstedt: Books on Demand. http://kups.ub.uni-koeln.de/5353/ (http://kups.ub.uni-koeln.de/5353/)
- Tennison, Jeni. 2006. "Overlap, containment and dominance". *Jeni's musings*, 2008-12-06. http://www.jenitennison.com/2008/12/06/overlap-containment-and-dominance.html (http://www.jenitennison.com/2008/12/06/overlap-containment-and-dominance.html)
- Vitali, Fabio. 2016. "The Expressive Power of Digital Formats". Presented at DiXiT Convention II: "Academia, Cultural Heritage, Society", Köln, March 14-16, 2016. Available at http://dixit.uni-koeln.de/wp-content/uploads/Vitali_Digital-formats.pdf (http://dixit.uni-koeln.de/wp-content/uploads/Vitali_Digital-formats.pdf)
- Van Zundert, Joris. 2018. "On Not Writing a Review about Mirador: Mirador, IIIF, and the Epistemological Gains of Distributed Digital Scholarly Resources". In *Digital Medievalist vol. 11, no.1*. DOI: http://doi.org/10.16995/dm.78 (http://doi.org/10.16995/dm.78)

Extra Slides

[In case there's time, or to better answer any questions]

Implication for Archival Studies

- the content of archives is always curated; a process of selection and exclusion
- narratives in archives are a double-edged sword; they can both highlight *and* obscure objects
- the utter arbitrariness of curators; a collection or exposition can convey a false sense of scholarly accuracy and completeness

Textual Awareness

Including humanist scholars in the process of text modeling is unavoidable; excluding them undesirable. Dirk van Hulle coined the term "textual awareness" to describe the understanding of how a text is made (from a textual genetics point of view); I think it apt to expand that definition by including an awareness of the process of data transformations.

- the understanding of how a text is made
- an awareness of the process of data transformations

This necessitates a different focus in the education of humanities scholars, which should at least include information modeling and background knowledge of different data structures. As William Kent wrote:

"*There is a flow of ideas from mind to mind; there are translations along the way, from concept to natural language to formal language and back again...*"

Kent, 1978

We should appreciate the productivity of scholarly discourse and, instead of striving towards objective descriptions of our ideas, we should formalise our interpretations.

TAG and *Alexandria* provide us with new, unprecedented ways to do just that.

Semantics

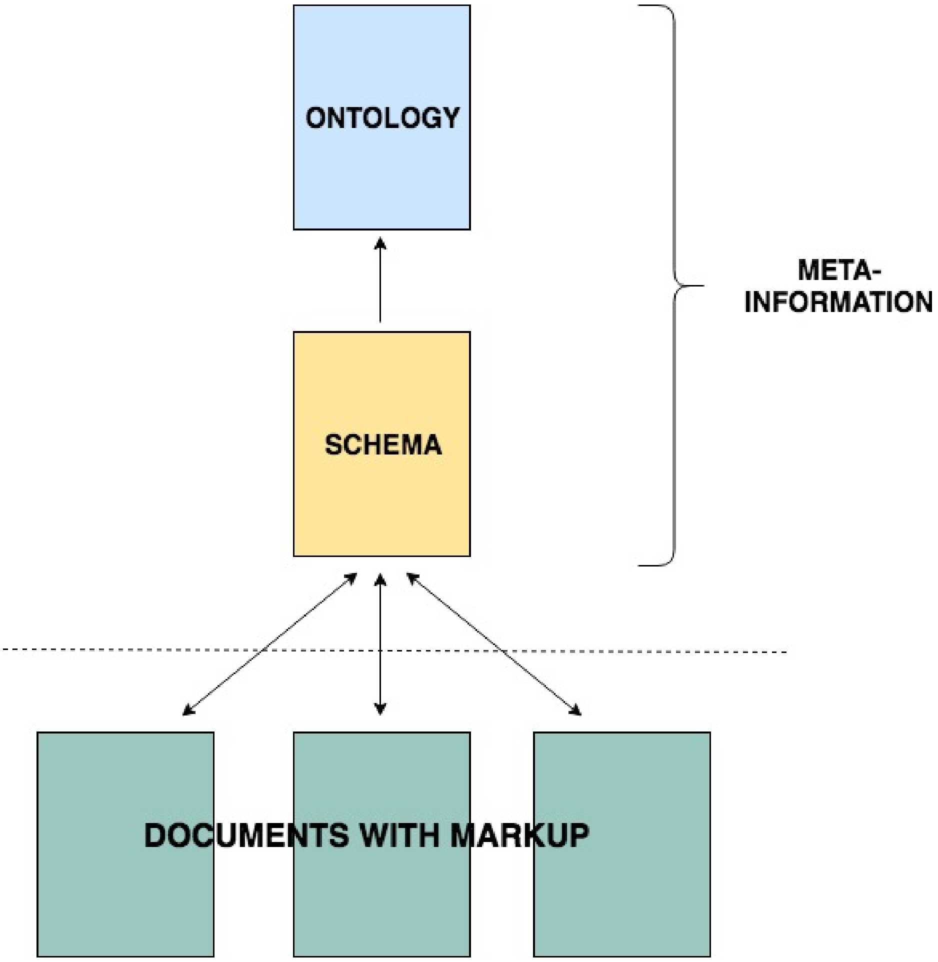
In order for the system to properly manage and process the information you give it, it needs constraints:

- Naming conventions
- Semantic properties of and relationships between categories
- Pre-definition of information (data type X needs to be interpreted as Y)

TEI is primarily written for humans, it provides documentation rather than a formal description of the semantic of tags.

This would be our ideal separation of concerns:

Text characters → markup → schema → ontology



Promising work in the area of TEI, formal ontologies and semantic markup:

- TEI Abstract Model by the TEI Consortium
- TEI and CIDOC-CRM by Christian-Emil Ore and Oyvind Eide (2009)
- LA EARMARK (Peroni *et al* 2011) presents a TEI Ontological Model: "an ontological extension of the TEI framework to partially formalise the semantics of the markup constructs it provides" that uses OWL formal ontologies (Ciotti and Tomasi, 2016).
- BECHAMEL project by David Dubin, Michael Sperberg-McQueen, Claus Huitfeldt and Allan Renear

Data typing

In XML, all annotations by default behave like strings. Other data types (like `<date="2018-11-02"/>`) need to be specified in the schema.

TAGML has data typing: users can specify whether a specific annotation is a List, a Number, a String, a Boolean, etc.

```
[lecture date={year=2018 weekday="Friday" month=11 almostWeekend=True}> some text here <lecture]
```

Containment versus Dominance

"Containment is a happenstance relationship ... while dominance is one that has a meaningful semantic. A page may happen to contain a stanza, but a poem dominates the stanzas that it contains."

In other words: dominance is meaningful; containment is not.

While XML doesn't distinguish between dominance and containment (an element both contains its descendants and dominates them); TAGML does. Again, this is possible by means of layers. If we explicitly want to express dominance, we make sure that the elements are in the same layer (e.g., `[page>` dominates `[line>` and are both in the same layer). The "happenstance relationship" between elements is expressed when layers overlap.