XML-Free Digital Scholarly Editions

Desmond Schmidt

University of Queensland

March 15, 2016

How this talk is organised

1. Why no XML?
2. Outline of technical solution
3. The benefits of not using XML
4. Demos
5. Question and answer

Theory versus practice

'There is nothing so practical as a good theory'

(Balisage/Extreme Markup conference motto)

This book is not a guide to producing a digital edition . . . as it does not provide explanations of how to use specific tools and techniques. Rather it investigates the changes and the methodological implications of the application of computational methods to all stages of the editorial workflow. . .

Elena Pierazzo, *Digital Scholarly Editing*

The fault is that there is not enough theory; the person concerned ought to have learnt from experience. What he learnt from experience might well be true theory, even if he were unable to impart it to others

Immanuel Kant *Theory and Practice*

Three theories about XML in the humanities

1. The OHCO theory: text is fundamentally hierarchical
2. The belief/theory that TEI-XML is interoperable
3. The theory of reuse: one input, many outputs

1. The OHCO theory

- The OHCO Theory (ordered hierarchy of content objects) stated that texts were 'fundamentally hierarchical' in spite of some exceptions. It forms basis for the XML model of text
- Its proponents (Renear, Mylonas and Durand) withdrew it shortly afterwards: 'hierarchies are the exception, not the rule'
- Most people continued to believe it was true – it was too convenient as a justification for the perceived utility of XML.
- But when they applied the theory to the practice of making digital scholarly editions they discovered problems.

1. The overlap problem

SGML/XML forces the encoder to impose on the text a strict hierarchical structure that often runs counter to the perceived structure of the text.

1. Overlap of features within the text (e.g. a quotation running over a poetic line, overlapping formats) – not too serious, not that common.
2. Overlap between different perspectives of the text (e.g. text-to image links, part of speech versus formatting structure etc.) – common, serious.
3. Overlap between internal changes to a document (e.g. changes to structure, lines/paragraphs that join up) – pervasive phenomena in manuscripts, serious.

Consensus: Overlap is a serious and unsolvable problem for XML.

1. Overlap within/between versions

Shakespeare, *King Lear*

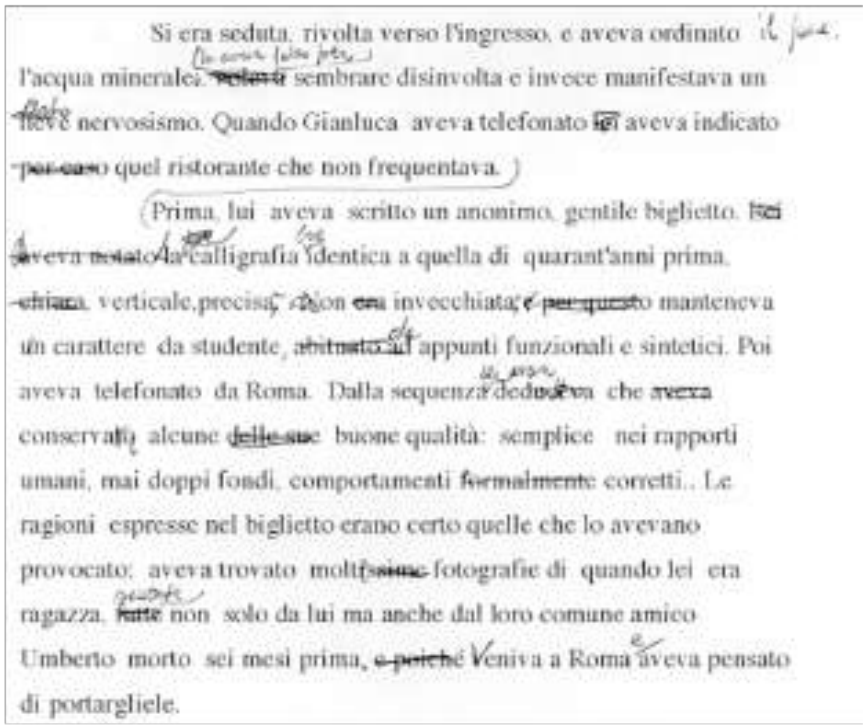
First Quarto, Act 1, Scene 1
Lear. The bow is bent & drawen make from the shaft.
Kent. Let it fall rather,
Though the forke inuade the region of my heart,
Be Kent vnmannerly when Lear is man,
What wilt thou doe ould man, think'st thou that dutie
Shall haue dread to speake, when power to flatterie bowes,

First Folio, ibid.
Le. The bow is bent & drawne, make from the shaft.
Kent. Let it fall rather, though the forke inuade
The region of my heart, be Kent vnmannerly,
When Lear is mad, what wouldest thou do old man?
Think'st thou that dutie shall haue dread to speake,
When power to flattery bowes?

Hack:

'<app><rdg wit="Q1"></l><l></rdg></app>'
is not XML.
You need: <app><rdg wit="Q1"><rdg><lb/></rdg>
</app>', but you lose l-element and attributes on <l>.

Francesca Sanvitale, 'Orient Express'



Si era seduta, rivolta verso l'ingresso, e aveva ordinato l'acqua minerale. ~~Voleva~~ sembrare disinvolta e invece manifestava un lieve nervosismo. Quando Gianluca aveva telefonato ~~le~~ aveva indicato ~~per caso~~ quel ristorante che non frequentava.

Prima lui aveva scritto un anonimo, gentile biglietto. ~~Lei~~ aveva notato la calligrafia identica a quella di quarant'anni prima, ~~chiara~~, verticale, precisa, non ~~era~~ invecchiata, ~~e per questo~~ manteneva un carattere da studente, ~~abituato~~ agli appunti funzionali e sintetici. Poi aveva telefonato da Roma. Dalla sequenza deduceva che aveva conservato alcune ~~delle sue~~ buone qualità: semplice nei rapporti umani, mai doppi fondi, comportamenti ~~formalmente~~ corretti.. Le ragioni espresse nel biglietto erano certo quelle che lo avevano provocato: aveva trovato molte ~~sicure~~ fotografie di quando lei era ragazza, ~~tutte~~ non solo da lui ma anche dal loro comune amico Umberto morto sei mesi prima, ~~e perché~~ Veniva a Roma e aveva pensato di portargliele.

Hack:

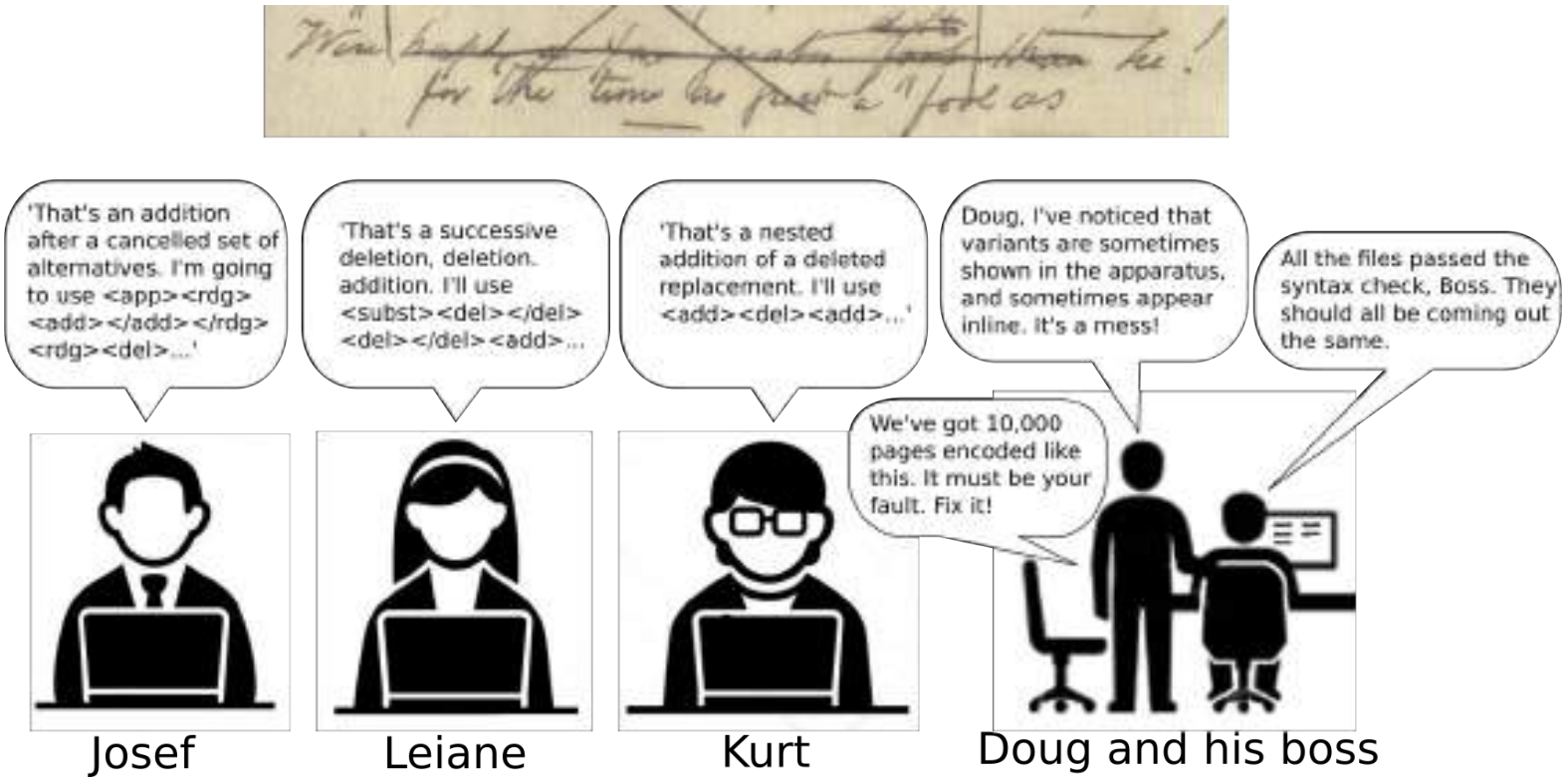
'<del> </del>' is not XML
You need: <del> para 1 para 2 </del>
<add> copy of para 1, copy of para 2 </add>

2. Interoperability and TEI-XML

- Interoperability is the property of data that allows it to be loaded *unmodified* and *fully utilised* in *several* applications.
- Interoperability is *essential* for collaboration, crowdsourcing, archiving, sharing and reuse.
- The TEI don't claim their schema is interoperable. All their experts say that it *isn't.* But most people believe that it *is* (Dee 2014)
- 'Interchange' is often mistaken for 'interoperability' (Mueller 2011)
- Interchange is 'hard' (Bauman 2011), involves remodelling, the writing of software converters, and data damage

2. Markup variability

Different encoders on different days will encode the same features differently (Durusau 2006), even when trained on a restricted set of tags.



2. Markup variability in searching

Q. "Find all quotations of Shakespeare in letters."
XPath: /div[type=letter]/quote/ref[contains(.,'Shakespeare')]

1st level: , <div1 type="letter">, ...
2nd level: <q>, <quote>, <cit><quote>...
3rd level: <ref>, <bibl>, <quote></quote><ref></ref>
4th level: 'Shakespear', 'Shake-speare', 'Shakespeare'

A. two-thirds of matches are missed on each level. Out of 2,142 potential matches only 40 will be found. Recall = 1.8%.

"machine-exploitable extraction of document components such as 'retrieve all letters of the document collection' or 'display all quotations in a chapter' pose an enormous problem ... This problem has been generally recognised."

(Geyken et al., 2012)

3. Reuse, or 'one input, many outputs'

- SGML/XML is based on the 'descriptive/procedural' distinction.
- The *theory* states that the structure of a document can be kept separate from its end-use

"the development and use of software-independent markup schemes like SGML and the TEI Guidelines will prove more important, in the long run, to the success of electronic scholarly editions, ... than any single piece of software can."

Michael Sperberg-McQueen 1994

3. Barriers to reuse

Failure to observe the descriptive/procedural distinction prevents document reuse.

- Whenever you encode something in the source to obtain a specific end-result you are violating the 'descriptive/procedural' distinction
- Information specific for one task can't be reused for a different task, and removing it may damage the text
- In practice, texts are encoded for one type of output, and can't easily be reused for another (Hillesund 2005)
- TEI Boilerplate (as used in TAPAS) abandons the descriptive/procedural distinction (Walsh and Simpson 2013)

Examples from the TEI Guidelines

```
<pb n="1" facs="page1.png"/>
```

Assumes: file type is 'png', location is current directory, external file 'page1.png' exists

```
<zone ulx="28" uly="75" lrx="175" lry="178">
  <line>Les cloches ont quasi fi-</line>...
</zone>
```

Assumes: the text is in an appropriate font and size to fit in the zone; the zone describes an image of a specific resolution

```
<cRefPattern matchPattern="(.+) (.+):(.+)"
  replacementPattern="#xpath(//div[n='$1']/
  div[n='$2']/div[n='$3]')">
```

Assumes: the existence of external programs able to process the regular and XPath expressions

What the experts say

I am most decidedly not of the persuasion that angle-brackets are good for you. They are not good for you. I can see no advantage at all in learning how to type angle-brackets and quotation marks and all that. . . . it's not actually going to improve your life or improve your editions if you can do it.

And we have been really badly let down in the last 20 years in the digital humanities because the tools have not come. It is actually more difficult now to make a digital edition than it was 20 years ago.

(Peter Robinson 2015 – author of 20% of TEI Guidelines)

What the experts say

The next few years will be crucial for the survival and expansion of the TEI: in order to survive and overcome the new challenges that come with the fast-evolving world of data representation it will have to part ways with XML as a sole technological implementation

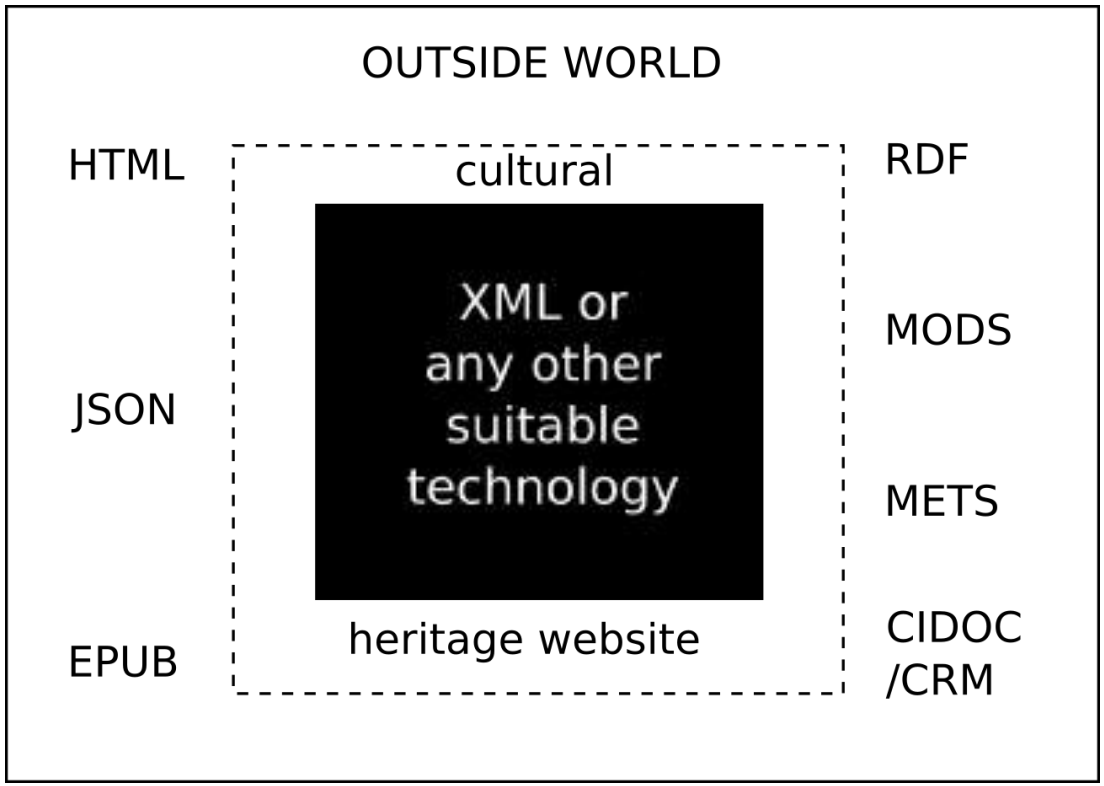
(Elena Pierazzo – chair of TEI Board 2012–2015)

Why embedded markup isn't the answer

- Any markup language would necessarily:
  1. If computable be context-free, and hence hierarchical
  2. If embedded by humans be subject to variability, so not interoperable
  3. If used for a specific purpose be not reusable for another
- So no 'new funky language' can ever solve these problems

How can we ever do without XML and TEI?

- In practice virtually no one shares or reuses TEI-XML
- Standards only matter *outside* an application
- Increasingly the outside world is abandoning XML



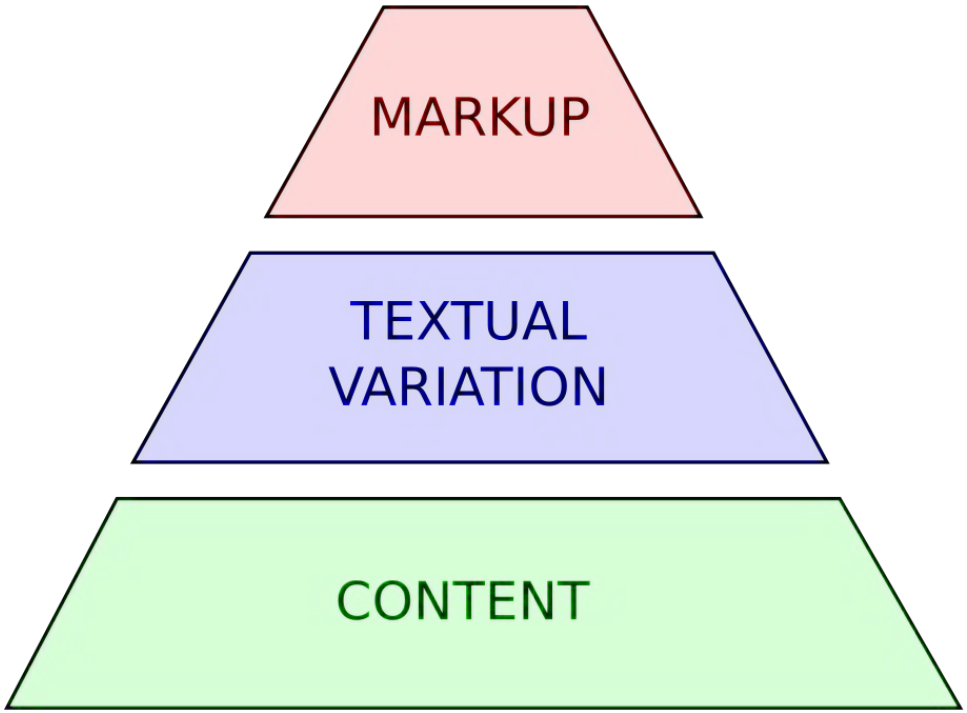
Ecdosis: a true model of historical texts

'An "OHCO structure" is not a model of the text, but a possible model of its expression.'

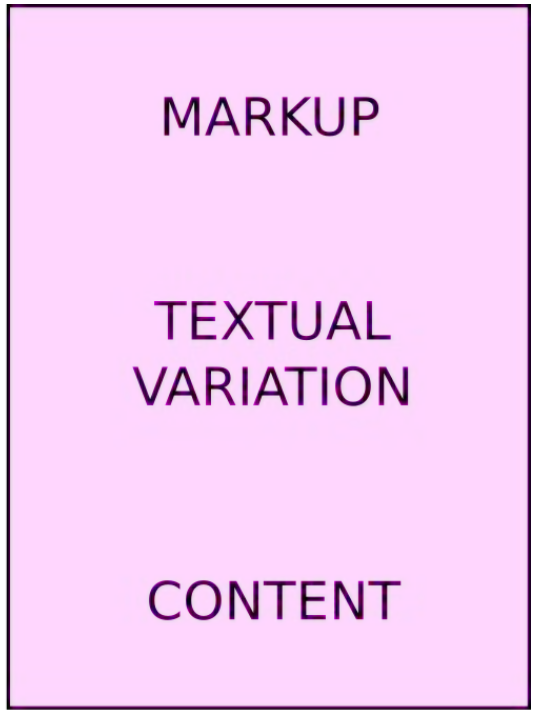
(Dino Buzzetti 2002)

- *Ecdosis* (Greek 'edition') is a general editing system for producing digital scholarly editions
- It is based on two technologies (*not* markup languages):
  1. Multi-version documents
  2. Standoff properties

How transcription data is organised



ECDOSIS

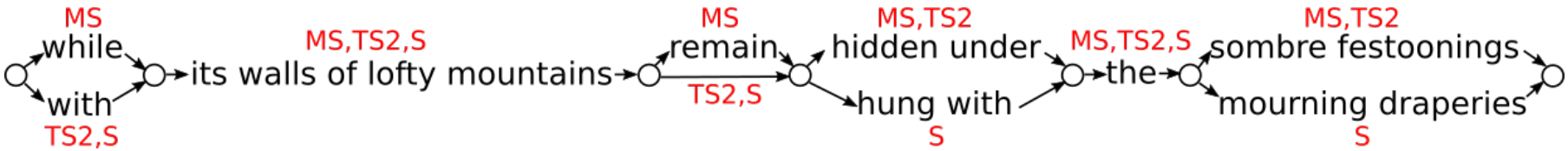


TEI-XML

Multi-version documents (1)

XML assumes that text is linear. But this doesn't take account of the multi-dimensionality of historical texts. This is true even of printed books:

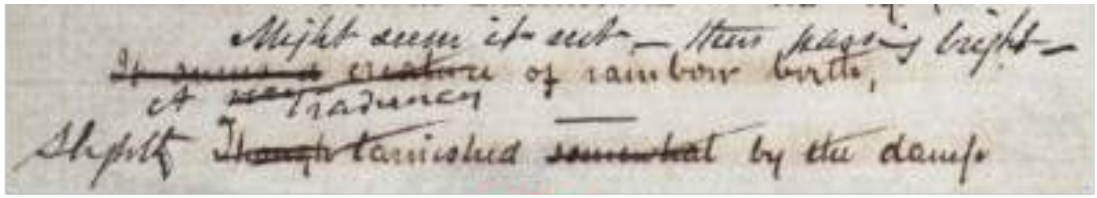
MS: while its walls of lofty mountains remain hidden under the sombre festoonings
TS2: with its walls of lofty mountains hidden under the sombre festoonings
S: with its walls of lofty mountains hung with the mourning draperies

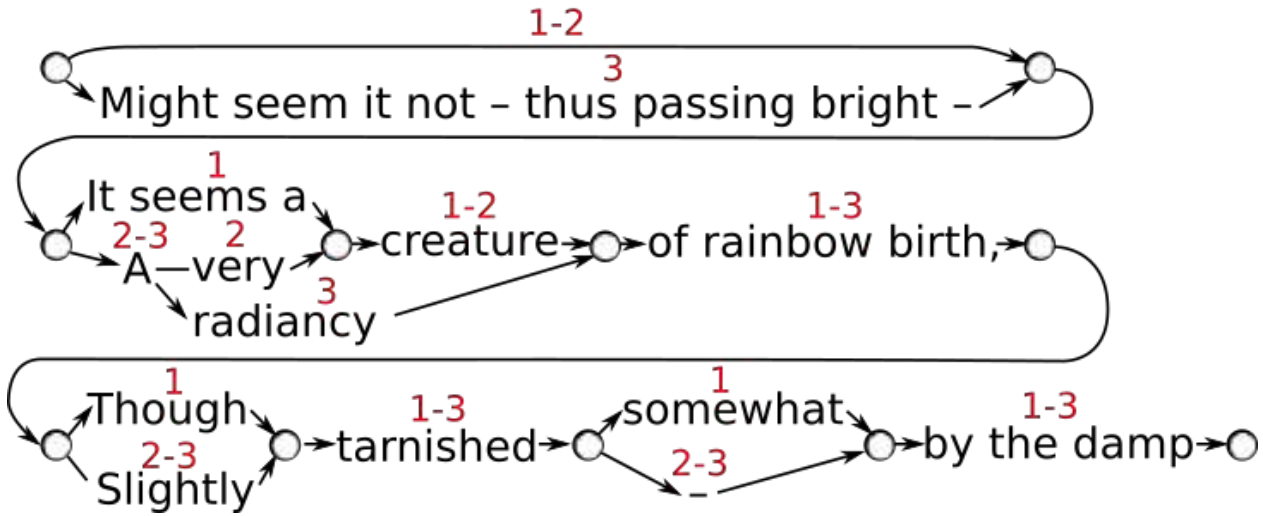


Three separate versions of Conrad's *Nostromo*

Multi-version documents (2)

In manuscripts the non-linear structure of text is clearly visible:





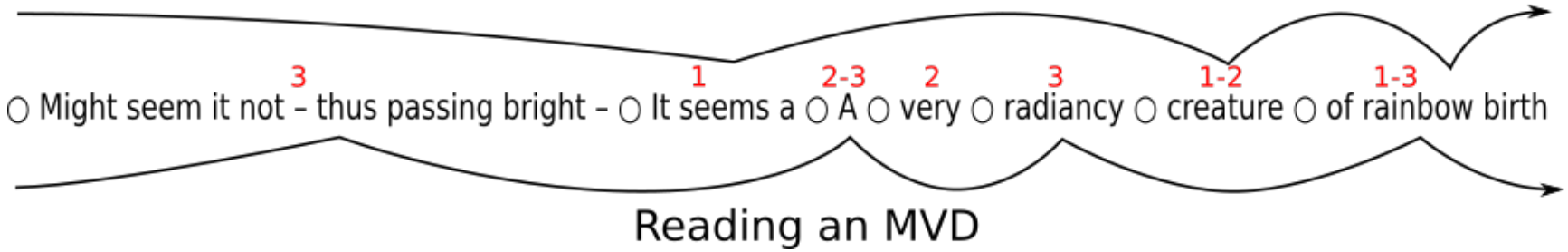
Internal variation in Charles Harpur's 'Kangaroo Hunt'

Multi-version documents (3)

A multi-version document (MVD) is simply a variant graph written out as a list of text fragments and their versions.

MVD

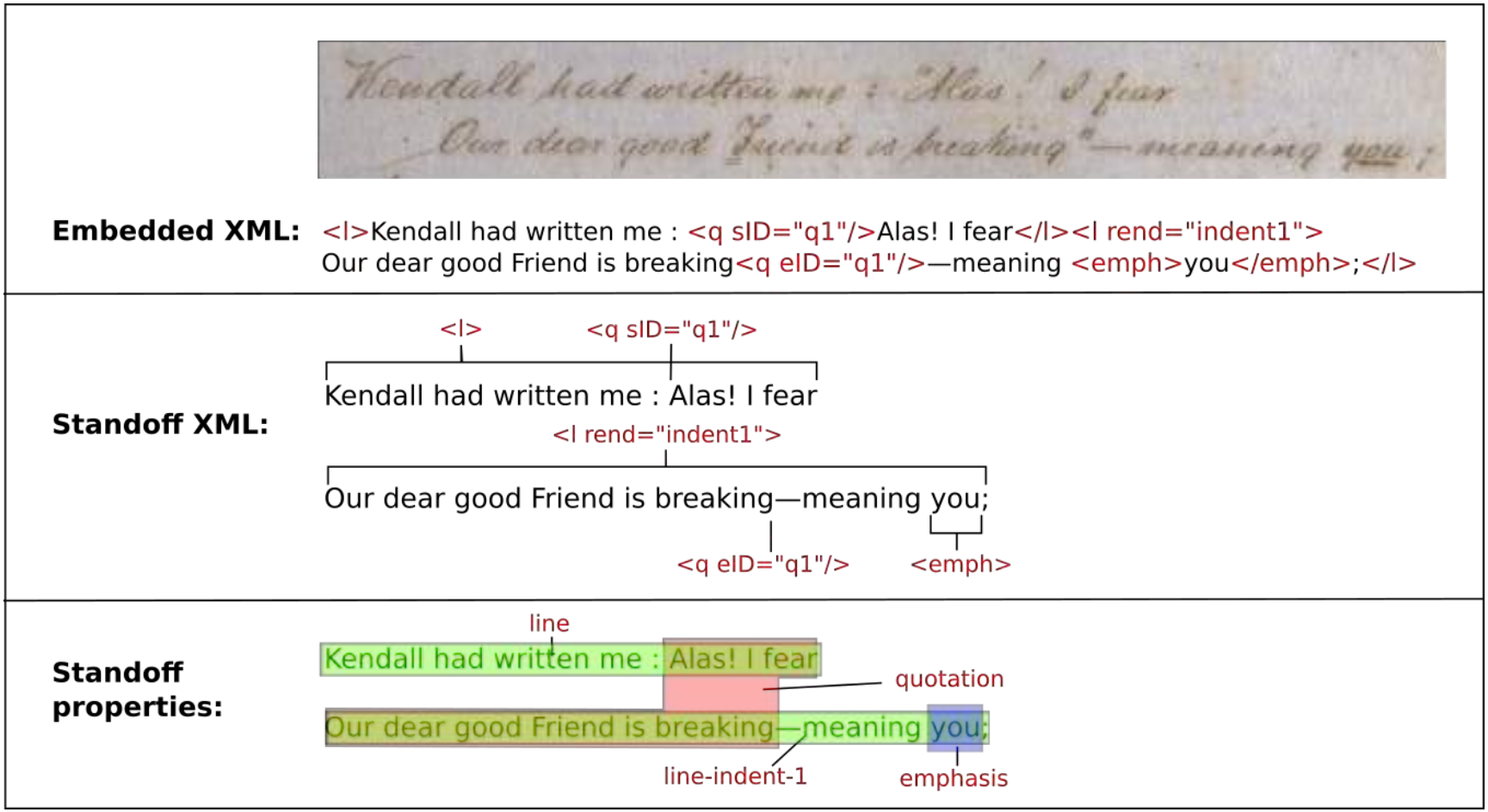
[1-2] [3] Might seem it not – thus passing bright –
[1] It seems a [2-3] A [2] very [3] radiancy [1-2] creature
[1-3] of rainbow birth [1] Though [2-3] Slightly
[1-3] tarnished [1] somewhat [2-3] – [1-3] – by the damp



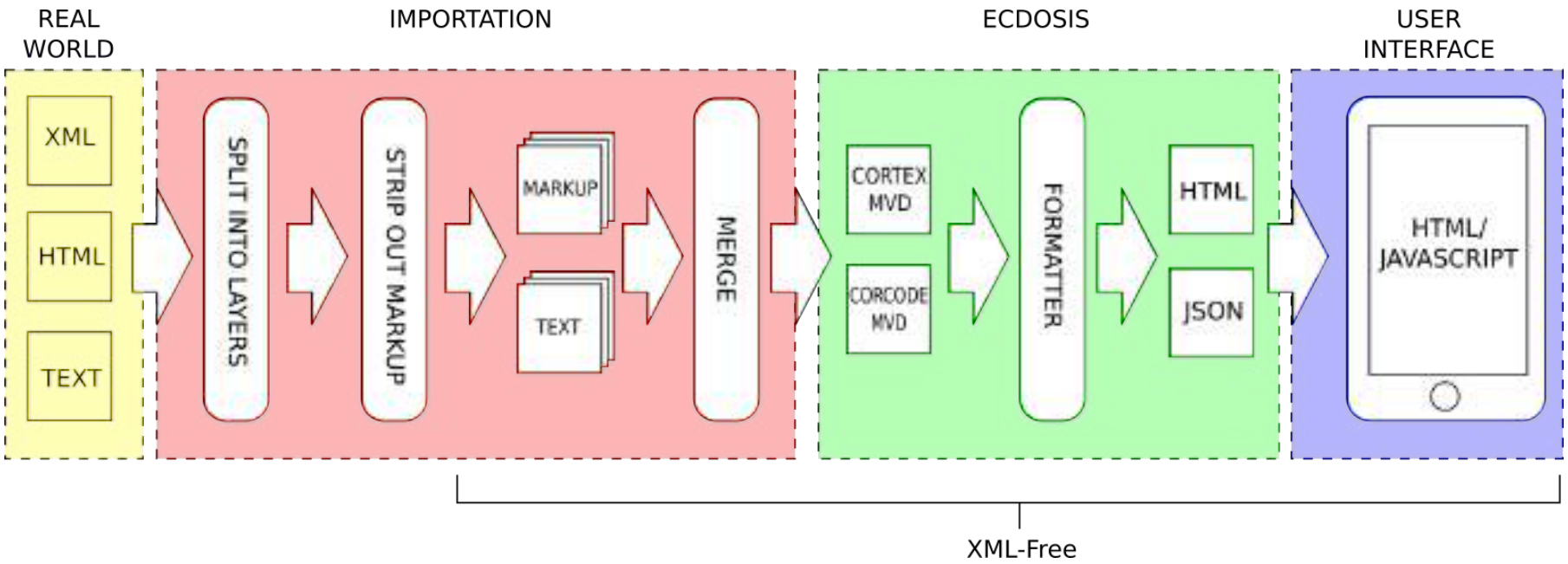
Reading an MVD

Standoff properties (1)

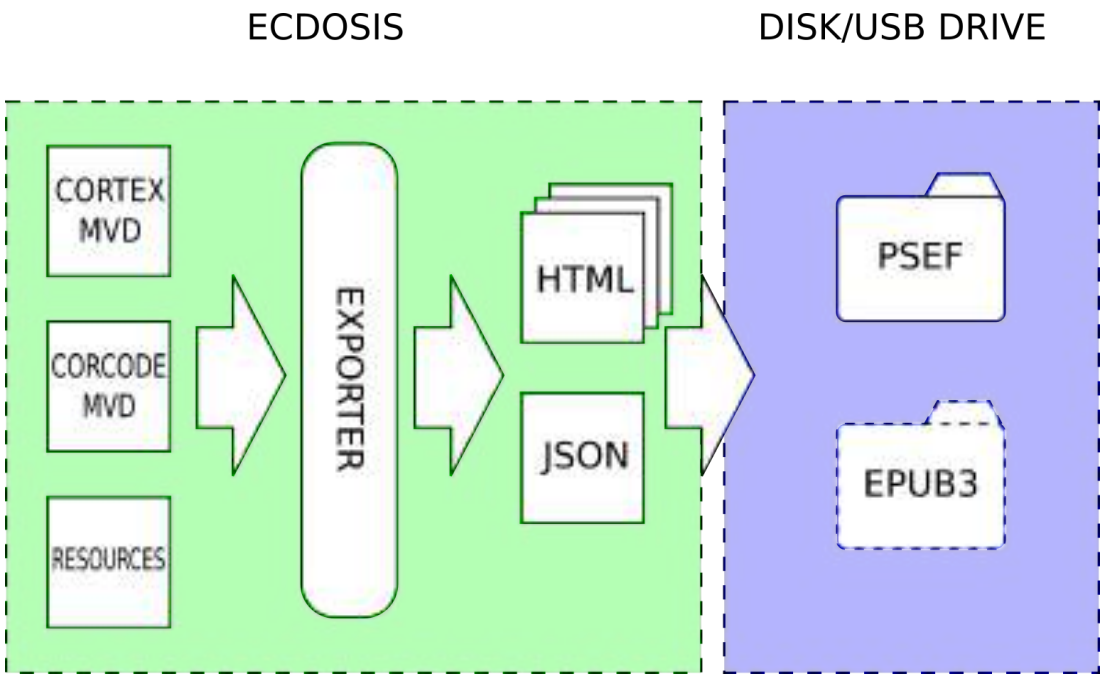
To allow text reuse and to ensure interoperability the markup must be taken out of the text.



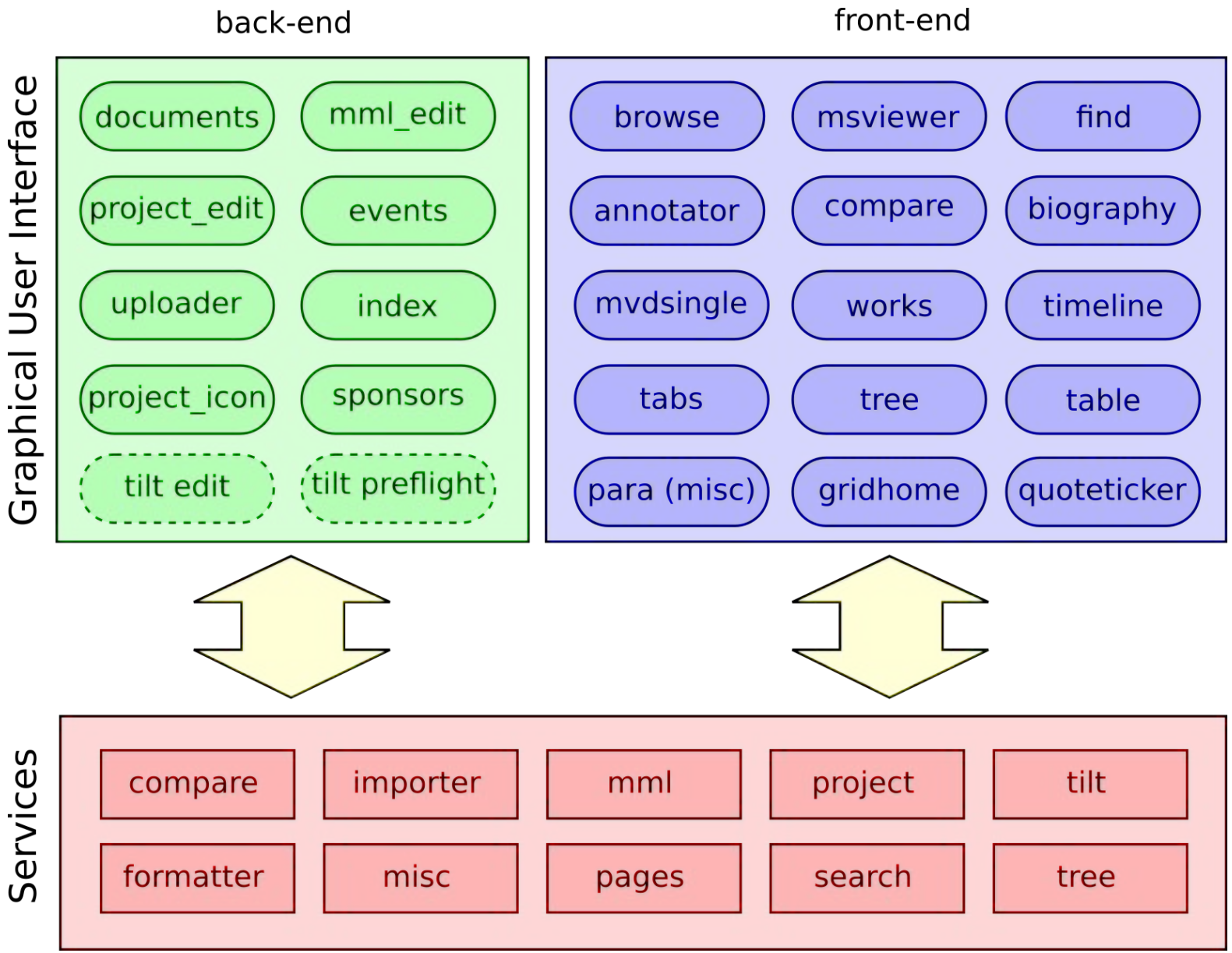
Importing and viewing



Exporting


ECDOSIS
DISK/USB DRIVE
CORTEX MVD
CORCODE MVD
RESOURCES
EXPORTER
HTML
JSON
PSEF
EPUB3

Ecdosis Overview



Demos

1. Browse
2. Single view
3. Find
4. Compare
5. Tree
6. Table
7. Biography
8. Timeline
9. TILT preflight

A closing thought

The digital edition is not about

- Markup
- Turning editors into programmers
- Bottom up ‘modelling’ of the text, then thinking about how to present it to the user
- Static data as in a book, transferred to the screen
- Not just for scholars

A closing thought

The digital edition is not about

- Markup
- Turning editors into programmers
- Bottom up 'modelling' of the text, then thinking about how to present it to the user
- Static data as in a book, transferred to the screen
- Not just for scholars

The digital scholarly editions is about:

- *Visual* interfaces
- Interactive content just as much as text
- Creating tools that model scholarly and user processes