Proposed Changes to the DESpecs (version 1.1.2 to version 1.2)

Wolfgang Schmidle

Max Planck Institute for the History of Science, Berlin, Germany

18th November 2008

Contents

1 General Remarks **2**
- 1.1 The Results From China . . . 2
- 1.2 Additional Remarks . . . 2

2 Necessary Changes to Existing Topics **2**
- 2.1 Pages . . . 2
- 2.2 Tables . . . 3
- 2.3 Anchored Marginal Notes . . . 3
- 2.4 Figures . . . 3
- 2.5 Latin Alphabet . . . 3
- 2.6 Italics . . . 3
- 2.7 Vertical Text . . . 4
- 2.8 Latin Ligatures . . . 4
- 2.9 Greek Alphabet . . . 4
- 2.10 Greek Numbers . . . 4
- 2.11 Greek Ligatures . . . 5
- 2.12 Mathematics . . . 5
- 2.13 Table of All Tags . . . 5

3 New Topics That are Needed for the Books in Batch 3 **5**
- 3.1 Special Instructions . . . 5
- 3.2 Fraktur . . . 6
- 3.3 Indexes and TOCs . . . 6

4 Material That Cannot Easily be Included **6**
- 4.1 Text Flows . . . 6
- 4.2 Anchored Comments . . . 7

5 Additional New Topics **7**
- 5.1 Chinese . . . 7

6 Points That Need to be Checked **7**

7 Sources for This Document **8**

1 General Remarks

1.1 The Results From China

Everyone is invited to take a closer look at the first results from China. They are on the wiki.

Currently, thirteen books are being typed. Eleven books are typed completely. We have a sample of the first 100 pages of the Conimbricensis and we need to decide whether the results are good enough for giving an ok for the rest of the text. The problem is that our special instructions seem to fail (see section 4). We can either produce a better version of the special instruction (I don't see how; maybe more examples?), or remind them of the rules (probably rather pointless), or give our ok and correct it afterwards.

We still have no working sample of the book that contains Greek text (book 12: Simon Stevin).

1.2 Additional Remarks

Starting with changes to version 1.1.2 of the Specs, we have to keep track of the changes. In addition to a new version, we will need to provide Formax with a list of changes.

2 Necessary Changes to Existing Topics

Some parts of the Specs need to be changed in reaction to the questions and results from China or in anticipation of batch three. Completely new topic for batch three are discussed in section 3.

We will *not* introduce a rule for library stamps. The rule is: Type everything that is printed. (We might add this as an explicit rule, though; The Chinese typed one library stamp and tagged another stamp as figure.)

2.1 Pages

Check whether they have ignored pages that are cut off in Mersenne (1635). If not, we need to give explicit rules for handling badly digitised texts.

What do we do with badly digitised texts such as Archimedes (1543), Aristolteles (1573), and the text from Google (Mersenne 1644)?

2.2 Tables

Section Tables (2.3.2): Include the additional rules from the “Special Instructions for Tables”, i.e. rules for (1) table elements that (horizontally) span more than one cell and (2) multi-line text within a table cell. Add a clarification that the `<tb>` should be typed on a separate line.

A new section structure be necessary, namely more than one main rule in a single section: Rule 1, example 1, rule 2, example 2, etc. Examples for weird tables and weird non-tables (as notes?).

Add a rule for table elements that (vertically) span more than one cell, or wait until a real example occurs?

A	E1	B
C	E2	D

```
<tb>
A # E1 \\ E2 # B
C # # D
</tb>
```

2.3 Anchored Marginal Notes

Ghetaldi (1603) contains a variant of anchored marginal notes: The anchor in the main text is always `*`, and it is not repeated in the note itself. The rule (or rather a note) would be to mark the asterisk as `<n *>`. However, judging by our experience with anchored comments in the Conimbricensis (see section 4.2), I don’t think it makes sense to introduce this rule as they will probably get it wrong anayway.

2.4 Figures

For the time being, we have decided not to add an explicit rule: Tag figures even if they are purely ornamental.

2.5 Latin Alphabet

Real typos are very rare in Latin text, but some letters are mixed up repeatedly. Add a list of common mistakes, such as b and h in italics? Probably it won’t have much (positive or negative) effect. Malcolm prefers a confusion matrix for post-processing.

Add real example for bold face, subscript, superscript: One example for superscript is the ^{9} ligature, i.e. `{us}`.

2.6 Italics

In the table of all all tags, allow `it` in footnotes and marginal notes. See also section 2.8.

2.7 Vertical Text

Make the rule that vertical text in captions should be typed as normal text a general rule. This makes sense because it also occurs in tables. (On the other hand, they seem to ignore verticalness anyway.)

2.8 Latin Ligatures

Check whether simple ligatures in italics such as `_{$p}_` in Euclid (1607) have been typed correctly. If not, add some of them to the list of ligatures. (Applies also to Ghetaldi 1607 in batch 3.) Add `_{ij}_`. Add the character `<002>` from the list of unknown characters (it doesn't make sense to add `<001>`.) Add the `{pro}` from Benedetti.

tan{quam} in Agricola (1561).

2.9 Greek Alphabet

The whole Greek section needs an overhaul before we send the first proper Greek text.

Problem of using escape sequences for typing greek letters with breathing and accent (still not completely decided whether we accept this or not). They have ignored the list of Greek ligatures. In general, the small portions of Greek text that we have seen so far didn't work well, but in these small samples even the typesetters got it wrong massively. For a detailed discussion see the "Analysis of the First Samples From China". It may make sense to introduce a `<greek>` tag for the sole reason to remind them to look into the Greek module in the specs. Or alphabet tags, see section 3.2. We may add an explanation about the position of the breathing, but this can only work with texts where the typesetter gets it right, and a similar clarification about the accents was ignored.

We still wait for the first 50 pages of Pappus Alexandrinus, but the Greek in this book is also badly typeset. MS: Send 10-20 pages of Greek text that is easily readable and see what happens. For example, use the Euclid lat/gr Book (1607?).

Find a better monospaced font, or type the examples with a non-monospaced font!

2.10 Greek Numbers

Move the note about the Greek number 6 at the end of the Greek whole-page example to a new section "Greek Numbers"? (Thus, the section itself would be new, but it would consist of material that is already there.) Problem is: The information given in the note is probably the only information that we need to give for Greek numbers, as all other numbers can be typed without knowing that they are numbers.

2.11 Greek Ligatures

We will see what the Chinese make out of the first 50 pages of Simon Stevin. Add the additional ligatures in Archimedes (1544). Use the Rgreekl2 font as the main source of ligature images. Eternal question: Resolve very simple ligatures silently, or would the distinction between very easy and not-very-easy be confusing?

2.12 Mathematics

Rule for mathematical notation in Archimedes (1544). Since it is somewhat idiosyncratic, it might be better to put it in a Special Instruction (see section 3.1).

2.13 Table of All Tags

Add `it`, see section 2.6. Also add new tags such as `\\` to the table.

3 New Topics That are Needed for the Books in Batch 3

3.1 Special Instructions

We may need to provide Formax with (very short) Special Instructions for some of the books in batch 3:

- Agricola: Specific question mark
- Aristoteles (1573), Simplicius (1551): Rule for types the two pages per image (problem of the digitalisation).
- Perhaps the mathematical notation in Archimedes (1544), e.g. the notation for lines.
- Simplicius (1551): Specific type of block quotations.

General problem with special instructions: They tend to be ignored more easily than the rules in the main Specs. The reason may be the separate document, or simply that the rules in the Special Instructions are very specific, difficult and/or apply rarely. Thus, we may want to use Special Instructions only where it cannot be avoided.

3.2 Fraktur

Fraktur is needed for the dictionary at the end of Agricola (1561), i.e. only for a short part of the book. Provide an alphabet, but make them type latin characters. Provide a list of ligatures, for example “der”. Introduce a rule for Sperrung (<sp>, probably in the Fraktur section, or alternatively somewhere in the section ”General Markup”?).

(We might also explain Fraktur in a Special Instruction for Agricola now and add it to the Specs only later. But since the rules will not be very book-specific, I think it makes more sense to put them in the Specs directly.)

(E-mail to Malcolm about “Lichtenberg’s physikalische und mathematische Schriften (1803-1806)” from 2008-10-29: Fraktur ligatures; how to type Fraktur, especially umlaut, hyphens; Also: Jupiter symbol; alphabet tags <fr>, <la>, <gr>; more examples of weird tables; smaller font in a subheading; caption within a figure)

3.3 Indexes and TOCs

- Casati (1686) in batch 2 already contains an index, but it is easy to type.
- Difficult indexes in Agricola (1561), Heron “Gli Artificiosi” (1589), Simplicius (1551).
- Table of contents in Piccolomini (1558). (The index in Simplicius is similar to the toc in Piccolomini.)

The underlying problem is left- and right-justified text with empty regions inbetween. We need either a general rule or (probably better) treat indexes and tables of contentns separately.

4 Material That Cannot Easily be Included

4.1 Text Flows

Text flows are defined in the “Special Instructions for Conimbricenses” (which should in fact be “Conimbricensis”). They are specific to this book and work only partially: The first text flow gets recognised, but the second text flow is only recognised if the text does not span the whole line width. In other words, they have no way of deciding whether a seemingly normal paragraph is part of the second text flow or not. They have difficulties ignoring the catchword of a text flow.

Since the rules are formulated in book-specific way using italics as distinguishing mark, we cannot easily incorporate this in the normal Specs anyway.

4.2 Anchored Comments

Anchored comments are also defined in the “Special Instructions for Conimbricenses”. They do not seem to work at all: The first anchored note is recognised and tagged correctly (!), but after that no anchored comment is recognised. A likely problem is that comment and anchor in the main text are not on the same page, contrary to footnotes and anchored marginal notes. I don’t see an easy way of successfully incorporating them into the normal Specs.

5 Additional New Topics

These topics might be added after version 1.2.

5.1 Chinese

MH and MS will compile a list of points where the general rules need to be adapted to Chinese text. One example is the markup of a different font that is used only in prefaces.

If we want to have bills from Formax for Chinese texts before 12th December, we need to work fast.

6 Points That Need to be Checked

The first result from China was: 50 pages Conimbricensis, 50 pages Benedetti, 50 pages Euclid. The second result from China was pages 51-100 of Conimbricensis.

- In result 1, the Chinese got the `<red>` wrong. We have told them about it, but in result 2 there is no red text, so we don’t know whether they will type it correctly in the future.
- See whether the Special Instructions for tables work properly.
- Some tags have not been tested yet: `<q>`, footers, `<col>`, `<tb>` (although they have asked for clarifications regarding tables), `<bf>`, `<_>`
- Mathematical symbols or fractions have not been tested yet.
- Are there any marginal notes or footnotes that consist of more than one paragraph?
- Do they get old-style numerals right?

7 Sources for This Document

- Text: “Analysis of the First Samples From China”
- Remaining points from the second part of the DESpecs appendix
- Text: “Special Instructions for Conimbricenses”
- Text: “Special Instructions for Tables”
- Notes from the DESpecs meetings
- Wiki: Provisional list
- Wiki: First evaluation
- Some e-mails