Spreadsheet annotation spec

Henry S. Thompson
$Id: annotate.xml,v 67:b66bcfae65d6 2017/06/12 15:20:14 ht $

1. Introduction

This is a first pass at defining an annotation menu structure for spreadsheets. The assumption is that we'll have an 'Annotate' entry in the Excel right-button menu for selected regions, which will pop up region-appropriate menus.

2. Top-level menus

Right-clicking 'Annotate' when over a selected range will create a new defined name of the form `_Xnnn`, where `X` is one of `R`, `C` or `M` for **r**ows (horizontal range selection), **c**olumns (vertical range selection) or **m**atrix (for two-dimensional range selection) respectively, and `nnn` is a serial number for the relevant selection type.

If the selection is a single cell I guess we try popping up a selection type menu, with choices 'Row', 'Column', 'Matrix' and 'None' (the latter resulting in `_Nnnn`).

The comment field (attribute in the XML) of the defined name should contain a feature-value dictionary, represented in JSON/Python style, that is, using the following BNF

```
fvd := '{' ( fvp ( ',' fvp )* )? '}'
fvp := key ':' value
key := string
value := string | number | fvp | array
string := '"' char* '"'
array := '[' ( value ( ',' value )* )? ']'
```

with whitespace ignored, 'number' being the usual integer or decimal representation and 'char' being ASCII-only (?).

If possible, the selected range should appear as the value of the new name *without* single-quotes.

Some features can and should be computed, others require annotator decision. Some features and/or feature values are unique to a particular selection type, others are shared across all or some types.

Accordingly, in order for the annotator to supply the required information, a form should pop up with all the features appropriate to the selection type. Literal or array-valued form fields will just require a value menu (allowing multiple selection in the array-valued case), but features with dictionary values will require cascading sub-forms.

The next two sections document the annotator-supplied and software-supplied features. Except for 'comment', whose value is free text, allowed values are tabulated.

3. Annotator-supplied features

3.1. All types

comment string: unconstrained. By its nature difficult to exploit, really should only be used to document a problem with the available feature&value vocabulary or structure.

3.2. Both one-dimensional types

type string: `"data"|"key"|"label"`

"key" is my preferred word for what Dresden call "attribute". In the simpler cases, think of it as what you find as the first row/column of the 2nd argument to an HLOOKUP/VLOOKUP call.

content fvd:
- **type** string: `"currency"|"date"|"datetime"|"integer"|"float"|"key"|"label"|"string"|"ti`

The "key" and "label" content types are for use (as in the Dresden paper example) where compound keys/labels are indicated by row or column spans.

3.3. Matrices

type string: `"table"|"data"|"label"|"condition"`
content fvd:
- **type** string: `"rows"|"columns"|"cells"`

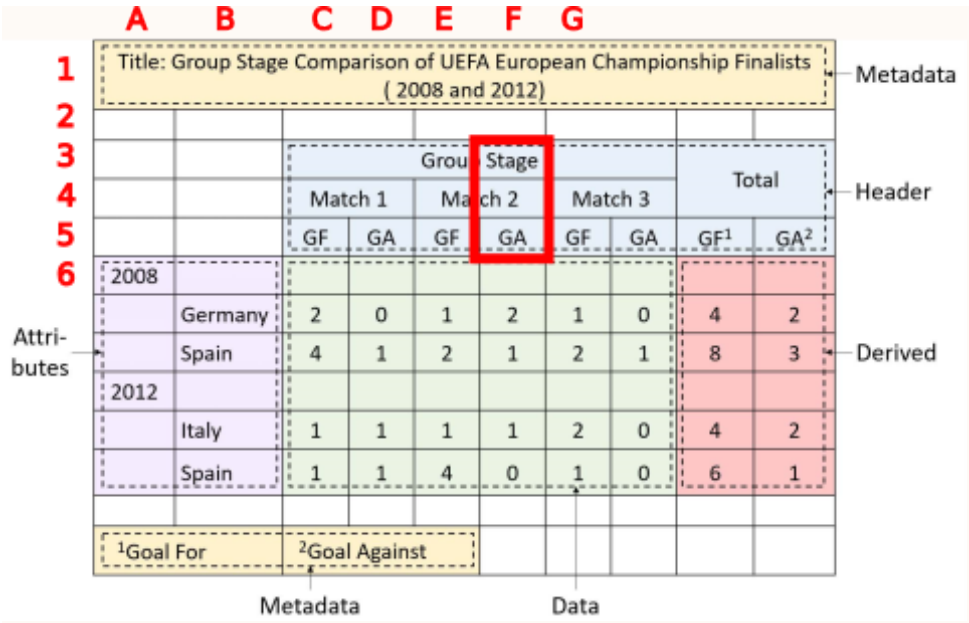
When a form for a matrix is completed, if `type` is 'data' a pop-up should offer to auto-fill based on `content/type`. If chosen, this fills the matrix with named ranges of the appropriate orientation (rows, columns or, in the case of `cells`, both). If it's not too hard, it would be good to go on to pop up the form for each generated range in turn, either having asked in advance for appropriate features whose values are the same for all the ranges, or carrying forward values from one to the next as defaults.

4. Software-supplied features

5. Issues

5.1. Compound labels and keys

There's a problem with defining the structure I want for compound labels and keys, in that you can't for example select the 6th column of rows 3 through 5 in the Dresden example, to denote the "Group stage/Match 2/GA" column label:



Excel would allow you to define a name for F3:F5 in that spreadsheet, but I don't *think* you can select that range with the mouse.

5.2. Metadata

Nothing in the above proposal provides a way to annotate what Dresden call 'Metadata'. We could simply provide another 1-D type, e.g. 'meta', I suppose, or just allow uninteresting regions to remain unannotated. There is a difference between on the one hand informative prose such as occurs in the Dresden example with the Metadata label, and regions whose type is just not obvious (as e.g. lots in the Kenneth Lay sheet from the Enron dataset...