Using git for version control of Excel spreadsheets - part 2 of 3

The nitty gritty

Excel files, or .xlsx files are a zipped archive of .xml files. It is therefore easy to unzip the file and then parse the individual .xml files to extract information that is required.

That is where the ease ends however, as there are some rather tricky aspects to converting the xml to text in a robust way.

XML schema for .xlsx files

Within the unzipped folder structure, the individual worksheets are stored in .xml files in the xl/worksheets/ folder.

The top of the file contains various pieces of header information, which I'll ignore for now. The interesting bits for the purposes of comparing text outputs from a conversion script are how the values, text and formulas are stored and how to get at these.

New rows are specified by

```
<row>
<\row>
```

Within the row header, attributes `r=""` and `spans=""` specify the row number and number of columns in the row respectively.

```
<row r="6" spans="1:9">
</row>
```

Populated cells result in a `<c><\c>` header which also contains various attributes. These attributes depend upon the content of the cell - which defines the cell type.

Cell types attribute `t=""` include

- `s` : string
- <blank> : normal formula or a value only
- `array` : array formula
- `shared` : shared formula

Values

The simplest cell definition contains only a `<v><\v>` header.

```
<row r="3" spans="1:9">
        <c r="A3">
             <v>9.48</v>
        </c>
</row>
```

The value of the cell is stored in the `<v><\v>` header.

Strings

String are stored in a separate XML file called `sharedStrings.xml` and referred to by an index which is stored in the `<v><\v>` header.

```
<row r="1" spans="1:9">
        <c r="A1" t="s">
            <v>0</v>
        </c>
</row>
```

The `t="s"` flag indicates that a lookup is required.

Formulas

There are three types of ways for storing formulas.

Stored directly

In this case, the text equivalent (minus the = sign) is stored in the `<f><\f>` value. The `ca=""` attribute refers to the location within the calcChain.xml file of this cell.

```
<c r="E7">
        <f ca="1">E6+$C$6</f>
        <v>4.45</v>
</c>
```

The result of the formula calculation is stored in the `<v><\v>` value.

Shared

Shared formulas are created in Excel when copying using fill command or equivalent. The formula is stored once in the first cell in which it is defined and assigned a shared index using the `si=""` attribute in the `<f><\f>` header.

```
<c r="G8">
        <f t="shared" ref="G8:G10" si="1">F8+G7</f>
        <v>6</v>
</c>
```

In subsequent cells that share this formula, the formula is referenced with the same shared index `si=""`.

```
<c r="G9">
        <f t="shared" si="1"/>
        <v>10</v>
</c>
```

One issue with this is that Excel must recompute the references dynamically upon opening, and so any text representation of shared formulas must also be recomputed before accurate text based differences could be computed.

Again, the computed result of the formula is stored in the `<v><\v>` value.

Array

Array based formulas are stored directly in the value of the `<f><\f>` header.

```
<c r="I9">
        <f t="array" ref="I9">H9</f>
        <v>14</v>
</c>
```

The computed result of the formula is stored in the <v><\v> value.