Adding USFM transformations to Proskomma pre-1.0

Desired behaviour

- Split a USFM document into two parts:
  - the original document minus some markup
  - the extracted markup
- Merge markup back into a USFM document, which may have been modified
- Merge translation notes into a USFM document as footnotes

Constraints

- Proskomma pre-1.0 only supports USFM/USX import (so no TSV or markdown)
- Proskomma pre-1.0 expects all documents to be Scripture-like (so no specific handling of stand-off markup)
- Because there is no document type system, there is no way to create new tags or attributes specifically for non-Scripture documents. (This is a limitation on USFM/USX import – it is possible to create new classes of scope once the data has been imported.)

Workarounds

- Represent stand-off markup as a secondary sequence of the original document, connected to the main sequence via a block graft
- Structure stand-off sequences so that they could be parsed as (quirky) Scripture markup
- Convert TSV files into USFM for import

Mutations

```
extractDocumentScope(
  documentId, # The id of the document to be split
  scopeName,  # The name of the scope to be extracted (eg ‘charTag/wj’)
  alignUnit,  # The unit of alignment (chapterVerse|chapter|block)
  label       # The tag for the stand-off sequence (discard if not present)
)

mergeDocumentScope(
  documentId,
  label
)

mergeInlineGrafts(
  documentId,
  mergeDocumentId, # The id of, eg, translation notes imported as USFM
  graftSubType     # eg ‘footnote’
)
```

Structure of Stand-off Sequence

Common Features

Store the label of the stand-off sequence as a tag

Alignment units are specified using scopes:

```
alignUnit/chapterVerse/2:4
alignUnit/chapter/2
alignUnit/block/47
```

Alignment is specified using scopes:

```
alignTo/payload/grace/1/4
alignTo/lemma/χαρις/1/4
alignTo/x-align/χαριτι/1/4
alignTo/strong/G877/1/4
```

The insertion position is specified using scopes, the default being immediately around the token:

```
nestAlign/inside/w
nestAlign/outside/zaln
```

Document Scope Features

- Sequence type = ‘alignScope’
- One block per alignment unit (verse, chapter or block)
- block scope contains the alignment unit
- block content contains one or more alignment:
    - one or more *alignTo* start scope
    - zero or one *nestAlign* start scope
    - one or more start scope to be added to the USFM
    - matching close scopes

Document Inline Graft Features

Sequence type = ‘alignInlineGrafts’

One block per graft

block scope contains the alignment unit

- block content contains one alignment with graft content:
    - exactly one *alignTo* start scope
    - zero or one *nestAlign* start scope
    - block content (items)
    - matching close scopes

Inline Grafts as USFM

- One *\p* paragraph per graft
- One alignUnit wrapped by *\w word |x-type="alignUnit"*
- One alignment wrapped by *\w word|x-type="alignment"*
- Rest of paragraph is the content of the graft, eg *\f* … for a footnote.

Proskomma will parse the USFM, producing a separate sequence for each set of graft content, the id of which can then be used to add the footnote to the target sequence.