GROBID table retrieval

In an exploratory manner, we used the software package GROBID (GeneRation Of BIbliographic Data; Github) to determine whether tables could be automatically identified with out-of-the-box open-source software. The added benefit of automated retrieval is that it increases the scale of data extraction from tables at a later stage in the pipeline. The drawback is that any automated process will not be 100% and might require fine-tuning. This exploratory testing is to roughly estimate the baseline of automated retrieval.

```
# This is a dependency for grobid
# Depending on your system you might need to add others
sudo apt-get install libxml2

# Get the latest stable release of GROBID
wget https://github.com/kermitt2/grobid/archive/grobid-parent-0.4.1.zip
unzip grobid-parent-0.4.1.zip

# Build grobid
cd grobid-parent-0.4.1/
mvn clean install
cd ../

# find the jar file and test whether it runs
java -jar grobid-grobid-parent-0.4.1/grobid-core/target/grobid-core-0.4.1.one-jar.jar

# Input directory to read PDFs from
DIR=""
# Input directory to save to (if it doesn't exist, be sure to create it)
SAVE=""
java -jar grobid-grobid-parent-0.4.1/grobid-core/target/grobid-core-0.4.1.one-jar.jar
-gH grobid-grobid-parent-0.4.1/grobid-home/
-dIn $DIR -dOut $SAVE -exe processFullText
```

GROBID restructures PDF documents (amongst others) into structured and encoded text (in the Text Encoding Initiative, TEI, format). It does this via machine learning algorithms and is specifically aimed at retrieving bibliographic information (e.g., title, author, date, affiliation, and abstract). However, it also restructures elements of a PDF such as the location of a table. This results in the following structured document. The TEI XML file as extracted by GROBID indicates coordinates of the content (for PMR: how does it map onto the svg file?) and can separate the table header from its contents.

```
<figure type="table" xml:id="tab_1" validated="true" coords="5,48.30,136.35,233.73,497.72">
  <figDesc coords="5,48.30,136.35,215.63,8.40;5,48.30,147.35,56.42,8.40">
    Table 1 Baseline demographic and sexual behaviour characteristics
  </figDesc>
  <table coords="5,48.30,161.86,233.73,472.21">
    Percentages are of group total unless specified Control group (%) Intervention group (%) Gender Male 31/101 (30.69) 29/99 (29.29) Female 70/101 (69.31) 70/99 (70.71) Age Mean (SD) 20.60 (2.39) 20.39 (2.42) 16-19 years 33/101 (32.67) 36/99 (36.36) 20-24 years 68/101 (67.33) 63/99 (63.64) Ethnicity White 55/101 (54.46) 59/99 (59.60) Black 32/101 (31.68) 21/99 (21.21) Asian 0/101 (0.0) 2/99 (2.0) Chinese 0/101 (0.0) 0/99 (0.0) Other 14/101 (13.86) 17/99 (17.17) Refused/missing 0/101 (0.0) 0/99 (0.0) Sexual orientation Heterosexual 83/101 (82.18) 88/99 (88.89) Gay or lesbian 5/101 (4.95) 3/99 (3.03) Bisexual 10/101 (9.90) 5/99 (5.05) Refused/missing 3/101 (2.97) 3/99 (3.03) STI infection at baseline No infection 53/101 (52.48) 58/99 (58.59) Chlamydia-positive 42/101 (41.58) 35/99 (35.35) Gonorrhoea/NSU 5/101 (4.95) 5/99 (5.05) Chlamydia/gonorrhoea/ NSU diagnosis 1/101 (0.99) 1/99 (1.01) Sexual behaviour Condom use at last sex 35/101 (34.65) 32/99 (32.32) Condom use at last sex with someone new 52/101 (51.49) 48/99 (48.48) Testing prior to last sex with someone new 37/101 (36.63) 32/99 (32.32) Partner testing at last sex with someone new* 12/101 (11.88) 11/99 (11.11) Number of sexual partners in last 12 months 0 0/101 (0.0) 0/99 (0.0) 1 9/101 (8.91) 6/99 (6.06) 2+ 92/101 (91.09) 93/99 (93.94) *Participants were asked " The last time you had sex with someone new, did they get tested for sexually transmitted infections before you had sex? " .
  </table>
</figure>
```

Figure 1: Excerpt of TEI XML code as extracted by GROBID

Retrieval rate

After running GROBID on an initial, closed-access corpus, we retrieved the number of identified tables using the following bash script

```
find corpus-grobid -type f -name '*.tei.xml' -print0 | while IFS= read -r -d '' file; do
    printf "%s,%s\n" "$file" $(grep 'type="table"' "$file" | wc -l);
done > data/grobid-tables.csv
```

The information about the number of tables was then conjoined with the manual coding of the number of tables in each paper. As a result, as many tables were extracted from a paper when the difference between manual and GROBID identification was zero; too few tables were extracted if the difference was >0 and too many tables were extracted if the difference was <0. The tables identified were not matched manually to tables in the content; the above serves as a heuristic of evaluating the extracted tables. This might result in overestimating the performance of GROBID.

```
x <- read.csv("../data/nr-tables.csv")

fn <- sum(x$diff[x$diff > 0])
fp <- abs(sum(x$diff[x$diff < 0]))
tp <- sum(x$manual[x$diff == 0])

precision <- tp / (tp + fp)
recall <- tp / (tp + fn)
```

	No table	Table	Total
'No table'	-	24 (0.209)	24 (0.209)
'Table'	21 (0.183)	70 (0.609)	91 (0.791)
Total	21 (0.183)	94 (0.817)	115 (1)

The precision of using GROBID to extract tables was 0.769 whereas the recall was 0.745. The table above depicts the classification problem, with columns indicating the true situation and the rows indicating the result from GROBID. There were 21 false positives and 70 true positives, resulting in a estimated Positive Predictive Value (PPV) of 0.769. As such, it seems like GROBID has reasonable performance to extract tables automatically.

Considering the limited corpus at the moment and the limited scope of the current project, this exploration is meant mainly as an illustration for the potential of GROBID or other automated retrieval of tables in future projects. As such, these out-of-the-box numbers are promising for effective retrieval of tables. For the current project, manual table extraction is preferred to keep the project on track. If there is remaining time at the end of the project, it is possible to pick up automated retrieval of tables.

Future possibilities

GROBID is a machine learning library, which would allow a training set to be created to improve table detection (or other aspects of a paper). The documentation for the GROBID package is unclear (to CHJH) how this would work exactly and would require considerable additional work. Nonetheless, for scalability, this could prove crucial. Nonetheless, investing in using GROBID to this end would also flow back into future endeavours to train for example figure retrieval. Additionally, the GROBID library is used by for example CERN and HAL Research Archive, which would create additional value for any contributions made.