ContentMine

Text and Data Mining Services

CM-UCL
Systematic rapid evidence assessment II

Prepared by: Dr Jo Brook

Prepared for: Prof James Thomas

Date: 25th January 2018

Deliverable 3: CM-UCLII Content Enhancement Report

Contents

Introduction

The primary objectives of this project are:

- to lessen the burden of current systematic reviewers by increasing their throughput and accuracy
- to promote the value of automatic analysis of the scholarly literature

In targeting these objectives, the current project builds on the approach and software developed for the project **CM-UCL Systematic rapid evidence assessment I,** April 2017 **(CM-UCLI).** That project demonstrated that data held in each cell of a table grid could be extracted for tables conforming to the 'APA 'style commonly used in scholarly publications.

The specific aims of the CM-UCLII project were to investigate two main areas of enhancement:

- Enhancing content beyond extracting grid cells by row and column coordinates
- Extending the range of table styles which could be processed by the system to include tables not conforming to the general APA style (in particular continuation tables without headers, tables rotated 90 degrees and tables with formatting using vertical lines, blocks of colour or similar.

This report gives and overview of the software developed to process and investigate these areas and the results focused on a small corpus of representative papers.

For a more extensive account of the technical and analytical work on the project is available at the project's Open Notebook and Github repository.

Corpus

After discussion of scope with ContentMine a corpus of 25 papers was supplied by researchers at UCL working on the Human Behaviour Change Project.

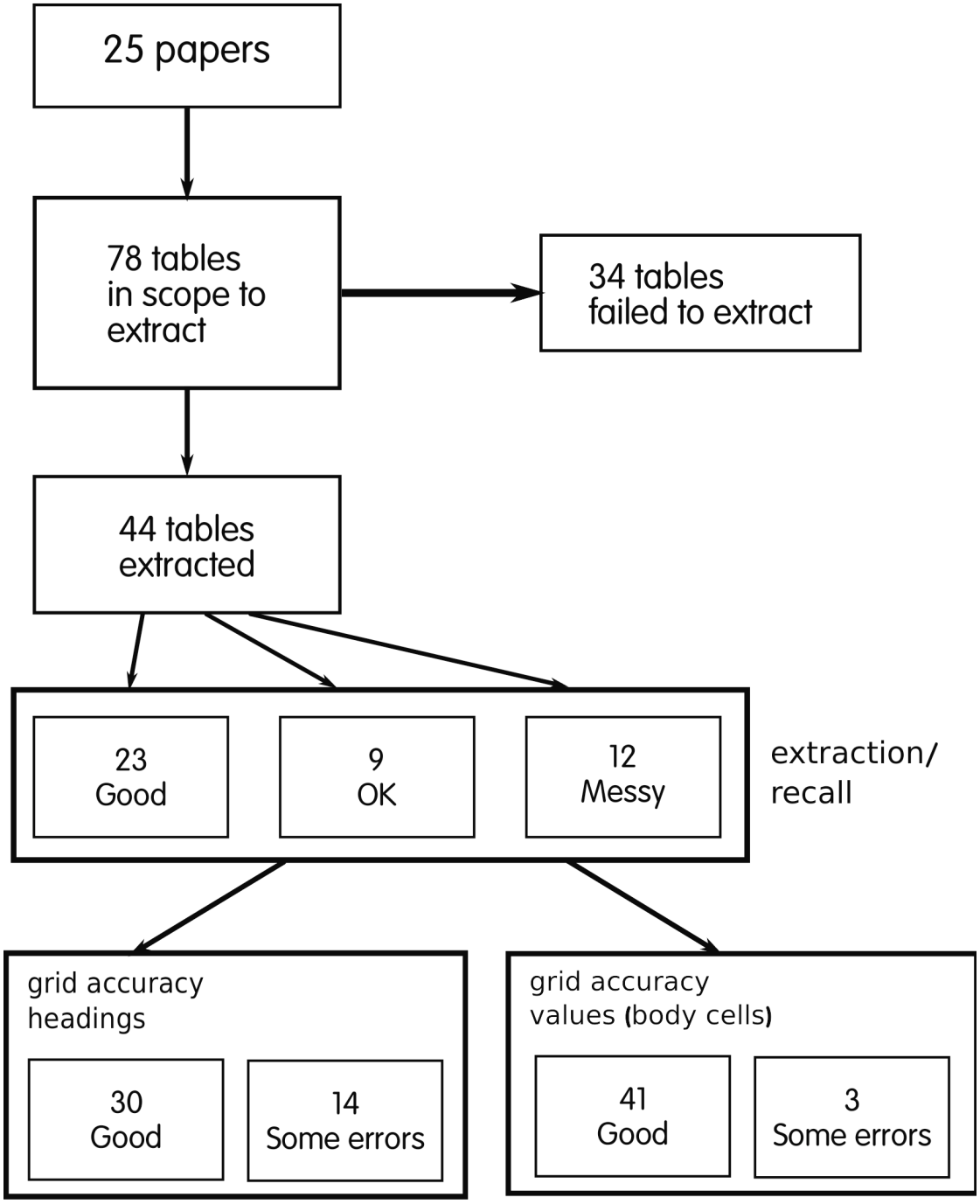
At ContentMine's request this corpus included examples from a range of different publishers and publication styles ('pubstyles'). It also consisted of both Open Access and non-Open Access publications. The project scope required that all papers were published in 2000 or after and were digitally originated PDFs (documents requiring OCR were out of scope).

The focus of interest for this project was on **baseline** and **outcomes** tables specifically – other tables were not of interest. So within each paper, tables other than these were considered out of scope. In particular these papers often featured a Table 1 simply presenting a verbal description of the experimental design in a tabular format. A vocabulary of synonyms for 'baseline' and 'outcome' was provided by UCL (e.g., 'baseline' may be 'demographic'). An inclusive policy was taken in selecting a corpus of tables to process, including tables containing numerical values and excluding any containing only text.

	Open Access	**Non-Open Access**	**Total**
Paper Type	16	9	25
Baseline/outcomes tables	52	26	78

The Open Access part of the corpus is avaliable on GitHub here .

Overview of results



In extraction (or recall), **Good** tables were ones whose grid was accurately extracted for both headings and values, and in which semantic restructuring (subtable identification, supercolumn headings, splitting column values) was successful.

OK Individual data and header cell contents were extracted and output in a usable format. These differ from the 'Good' tables in that either a small numbers of errors were observed in accuracy of the grid, or semantic restructuring was not successful, including false positives and negative in subtable detection. So the table overall was less accurately processed than in the 'Good' cases.

Messy tables were those whose grid structure was not accurately extracted using the row/column by the x,y coordinate approach as well as any in which individual numerical values were incorrect or ambiguous due to unknown characters.

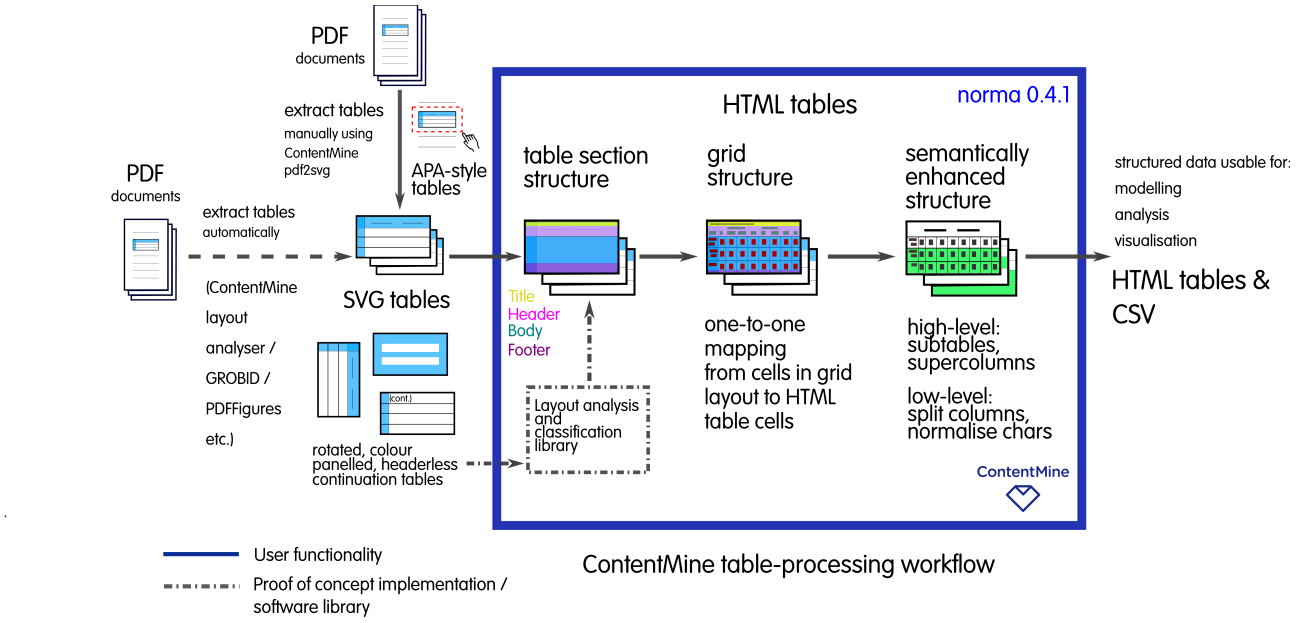
Accuracy of headers compared to data cells

Tables which were Good/OK/Messy the relative grid accuracy of headers and data cells could be compared. Data cells were more often correctly extracted than headers. This reflects the greater variation in header styles, row headers containing descriptive text broken over several lines, whereas numerical cells more rarely contained linebreaks.

In a machine-assisted document analysis system, the ability to render the values as usable data, respecting the row and column structure and semantics is a positive outcome and one which would apply to larger tables than the ones included in this project. It and is presented as a tool to reduce time-consuming and error-prone manual re-keying of tabulated data.

Overview of Features and Workflow Structure

ContentMine Table Processing Workflow



In the previous project CM-UCLI software was developed which resolved APA tables to a **grid structure** using common x and y coordinates, This project CM-UCLII takes this grid structure as a starting point for extracting usable data in the form of individual numerical values and also identifying semantic structures (subtables, hierarchical supercolumn headers, compound columns, table headers and footers) and outputting these in a semantically structured format.

The development was divided into two areas of investigation – **Content Enhancement and Semantic Structure** improving accuracy and usability and **Extending recall** to non-APA tables.

Content Enhancement and Semantic Structure

The content enhancement work in the CM-UCL-II project uses the **grid** resolved in CM-UCL-I project as its starting point. For suitable tables this is defined by heuristic analysis of x and y coordinates:

- columns – shared or similar x coordinate
- rows – shared or similar y coordinate

After these values have been identified, the following structure can be resolved:

- value cells – arranged as a grid according to columns and rows identified as above

- headers

As seen in CM-UCL-I this approach works well on APA-like tables whose structure can be automatically decomposed into **T**itle / **H**eader / **B**ody / **F**ooter.

The CM-UCL-II project has added a number of content enhancement features to the ContentMine pipeline in two main areas working with both the macro and micro-structure of the table:

1. **Semantic Restructuring**

- Macro level: capturing subtables, supercolumn tree headers,
- Micro level: splitting compound column content into supplemental columns

2. **Content Normalisation**

- Character normalisation – e.g., dashes used in place of minus signs

Data representation and output format: HTML has proved to be a good intermediate format format for data transformations. The ContentMine pipeline uses an XML-compliant subset of HTML5 to represent table structure including header, footer, and subtables.

A CSV file containing only the numeric values and basic headers is also output, suitable for use in other analysis or display tools.

Semantic restructuring

Subtables are detected based on layout using indentation. The project focused on the most common variant which uses leading whitespace to indicate rows within a subtable. Subtables whose rows used initial dashes or colour contrasts to indicate a subtable were not investigated in this project.

The subtable-finding process was applied to all tables in the workflow. Subtables were accurately found in 14 cases. It was also observed that the policy of interpreting whitespace indentation to indicate the presence of a subtable led to a number of false positives. Future development of the system will make subtable detection a user-configurable feature as part of a machine-assisted data-extraction workflow.

Column-header trees With a few exceptions, column-header trees (hierarchical or nested headings) are indicated by a layout including short horizontal lines between the header rows which span a number of lower-level headers and indicate a hierarchical relationship. A typical scenario would be variables such as mean, SD and CI reported at different stages of a longitudinal study, under super-column headings of 1 month, 6 months etc.

Splitting compound columns The use cases for these within the corpus fall into three main types:

- mean (SD) / mean (SE)
- Confidence Interval (CI)
- count or proportion of population

The project set out to utilise these common types of data in two ways:

- to decompose the contents of the columns according to the pattern (a regex or regular expression)
- to title new columns formed by these decompositions respective to the parts of the column header

In practice however, the wording and syntax of the column headers showed too much variation and often relied for its interpretation on context either within the table (e.g., in the header or footer) or even in the running text.

The system therefore takes the approach of detecting columns containing any cells with **compound** content, defined as multiple numerical values plus any other punctuation and whitespace. For each numerical value a new column is created with a separate cell for each numerical value. In this way the individual values become usable, and are available for import into a spreadsheet or database for instance.

The headers of the newly formed columns are given names indicating their relationship to the original table grid – the original column is suffixed with a numerical index, e.g., 'mean (SD) :0', 'mean (SD) :1'. Rather than restructure the original table, the new columns are appended as supplemental to the main body of the table, sharing the original row labels.

Content normalisation Authors and publishers are not consistent in the use of punctuation and symbols in numerical data intended for print and human (visual) reading only. In producing semantic and usable data, the system needs to ensure that visually similar characters are output as the characters or symbols which have the semantics appropriate for the context. A common example is the variety of different Unicode and non-standard dash characters which are used to represent minus signs. To ensure usable and accurate numerical data, it is important to identify and normalise these to a standard ASCII minus and this was the focus of the normalisation in the project.

The system successfully produced semantic minus signs, except in one case in which the published paper had used dashes and minuses in an inconsistent way across the column. A natural extension to the current work would be to apply data consistency checking and data cleansing processes to numerical content. This could be based on a combination of conventions for the domain and statistical tables and also the publication style (or 'pubstyle'), derived from analysis of the style of published papers.

Extending recall beyond APA tables

The second strand of development investigated extending the types of tables processed beyond APA style.

APA-style table is one whose large-scale structure is indicated by horizontal lines only.

Rows and columns are indicated by positioning and separated only by whitespace, not by additional horizontal and vertical lines or coloured panels.

The system segments these tables into **T**itle, **H**eader, **B**ody and **F**ooter. This structure is used as the basis for the semantic representation of the tables internally, and the aim is to convert other table formats into this overall structure for onward processing.

Below is an example of an APA-style table and its decomposition into THBF structure:

Title
Header
Body
Footer

Table 1. Baseline characteristics as a percentage of the sample, unless indicated otherwise

	Intervention ($n = 456$)	Control ($n = 451$)	Total sample ($n = 907$)	Significant difference I and C
Male	60.3	60.1	60.2	n.s.
Age, mean (SD)	20.9 (1.7)	20.8 (1.7)	20.8 (1.7)	n.s.
Education				n.s.
Higher professional education	26.5	25.9	26.2	
University	73.2	73.8	73.5	
Moderators				n.s.
Contemplation stage[a]	20.4	22.4	21.4	
Problem drinking[b]	38.8	39.2	39.0	
Freshmen	23.3	19.3	21.3	
Fraternity or sorority membership	50.4	52.3	51.4	
Carnival participation T1	43.9	45.5	44.7	
Measures				n.s.
Heavy drinking[c]	82.2	82.3	82.2	
Frequency of binge drinking	81.8	81.6	81.7	
Weekly alcohol consumption, mean (SD)	22.0 (15.9)	21.6 (16.0)	21.8 (15.9)	

Note. All differences between conditions were non-significant ($P > 0.05$). SD, standard deviation; T1, 1-month follow-up.
[a]Readiness to change alcohol consumption was assessed through one item asking the participants which statement applied best to them. Participants selecting 'I want to reduce drinking alcohol within the upcoming 6 months' or 'I want to reduce drinking alcohol within the upcoming month' were considered to be in the contemplation stage of change, meaning that they were willing to reduce their alcohol consumption in the near future.
[b]Assessed with the AUDIT and dichotomized into 0 = "no problem drinking" (AUDIT score of ≤15) and 1 = "problem drinking" (AUDIT score of ≥16).
[c]Drinking > 14 or 21 (female/male) glasses of standard units of alcohol per week and/or drinking five or more glasses of standard alcohol units per occasion at least once per week (=binge drinking).

The work on extended recall coverage focused on the following types of tables:

- rotated 90 degrees on the page
- continuations of previous tables without their own headers
- tables whose internal structure is indicated using coloured panels, line-grids or other non-APA layouts

Overall structure and schema

The Baseline and Outcomes tables were analysed in the context of a "schema" which would be generic to systematic reviewing and "understand" the common semantics and vocabulary.

The major schema features are often well-defined implicitly: (1) the table sections (2) inter-section semantics a) superscripts b) alignment of columns (3) super-columns (4) hypertables (5) common vocabulary and semantics. Following the 2017-09-25 EPPI-CM

meeting this analysis will appear in the final report, giving a qualitative idea of the frequency of well-defined hypertables, and be available for EPPI-CM formalisation.

Section analysis

The formulation (**T**itle, **H**eader, **B**ody, **F**ooter) was universal with the main orderings THBF?, HBT?F. The continuation and rotated tables are orthogonal to this and conform after stitching and/or rotation. Our model is based on interpreting repeated row-like objects (text, whitespace, rules, panels). The main challenge is the lack of explicit inter-section separators. Sometimes rules were omitted, or added into H and B making formal parsing difficult. Sometimes panels overlapped boundaries or were used inconsistently. The current approach is to identify all syntactic events and later to use several methods to predict the best separators.

The syntactic row-events are:
(i) horizontal rules, both “full width” and short
(ii) whitespace between text-text or text-other
(iii) text style changes (font size, weight, style, font family, etc.)
(iv) panels

Every table generates a signature which can then be used to find separators.

The identification of the sections involves:
(a) content (e.g. “Table 2”) or leading subscripts (in footer)
(b) horizontal spacing (header has horizontal spaces unlike title)
(c) short rules (normally only header)
(d) order

In the project time we did not find a unique method which found all separators. We propose 4 possible methods for future development beyond this project:

1) heuristics (the current approach), extended to regexes on the signatures
2) Machine Learning on the signatures
3) per-publisher pub-styles (maybe required for BMJ and Lancet)
4) human intervention on a per-table basis, but storing the results as a training set.

User Guide

The ContentMine document processing system is known as norma. This document introduces the commands used to transform tables to semantic format.

System requirements

- Java 1.8 runtime (Oracle or OpenJDK)
- OS: Linux, MacOS X

The following instructions assume a command-line shell such as Linux or MacOS X.

Windows is not currently supported for this version.

Version

This document is intended for use with the current snapshot release norma-0.4.1-SNAPSHOT

Running norma on an existing corpus

To get the binary and the current corpus (the Open Access subset of the CM-UCL-II corpus), either:

If you have git installed, do:

```
git clone https://github.com/ContentMine/cm-uclii/
```

or download the following and and unzip it:

https://github.com/ContentMine/cm-uclii/archive/master.zip

This repo has sub-directories:

- **`corpus-oa-uclii-01`** -- a directory in CProject format, containing the tables manually extracted from the OA papers and converted into ContentMine SVG
- **`norma-20171212`**

and also contains bash shell scripts:

- **`transform.sh`**
- **`cleancproject.sh`**

`norma-20171212/bin/norma` is the latest norma binary (version norma-0.4.1-SNAPSHOT)

Usage

Converting tables from SVG to HTML

The core functionality for the CM-UCL project is to convert the tables which have been converted to SVG into HTML.

To run this conversion process on the Open Access papers from the CM-UCLII corpus, open a terminal. Go to the the directory `cm-uclii` and use the following command line:

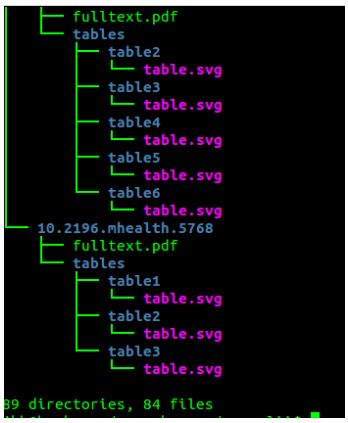
```
./bin/norma --project corpus-oa-uclii-01 --fileFilter "^.*tables/table(\\d+)/table(_\\d+)?\\.svg"  --outputDir corpus-oa-uclii-01 --transform svgtable2html
```

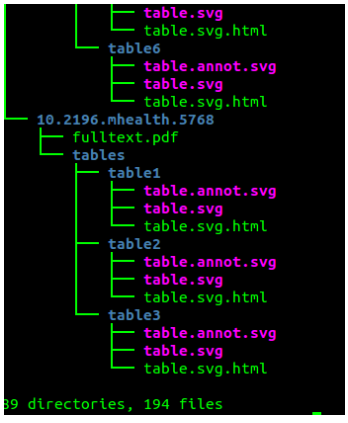
This will generate HTML5 files stored under the `corpus-oa-uclii-01` directory tree.

The name of each document is output to the command line as norma runs. More detailed output from the run including any errors are logged to `norma.log` in the current directory. Each day's log is rolled over into a dated log file so that the current day's log does not grow too large.

Expected behaviour – details

A successful run will output resulting HTML files (`table.svg.html`) and some intermediate files capturing the system's analysis of the table (e.g., `table.annot.svg`, a version of the original `table.svg` with annotations indicating its structure and other results of layout analysis).





Creating an HTML demo page to inspect results for the whole corpus

Additional bash scripts are provided for convenience to create a browsable HTML page to view the results across the whole corpus.

Before using the scripts, set the NORMA environment variable to point to the full absolute path to the location of the main norma executable, e.g., `/home/abc/work/cm-uclii/norma-20171212/bin/norma`

To create a browsable HTML page showing styled tables, run:

```
transform.sh corpus-oa-uclii-01
```

To view the browsable page open a browser (e.g., Firefox or Chrome and enter the following file URL, replacing /home/abc/work with the path to the directory containing of the cm-uclii directory):

```
file:///home/abc/work/cm-uclii/corpus-oa-uclii-01/tableViewList.html
```

Example outputs

A example of a table whose structure has been resolved to identify subtables, super-column headers and columns with compound contents split into individual numerical values. Note that the colours and styling for this demo HTML is provided by using an external CSS3 stylesheet which references the purely semantic markup in the HTML5 file.

For DOI: 10.1093.alcalc.ags133 Table 4:

10.1093.alcalc.ags133

table1 table2 table3 table4 table5

Table 4. Moderating effects on heavy drinking and frequency of binge drinking using logistic regression (intention-to-treat analysis) at the 1- and 6-month follow-up

	1-month follow-up			6-month follow-up			1-month follow-up		6-month follow-up	
	OR	95% CI	P	OR	95% CI	P	95% CI :0	95% CI :1	95% CI :0	95% CI :1
Heavy drinking by										
Gender	0.72	[0.35–1.44]	0.35	1.39	[0.78–2.49]	0.27	0.35	1.44	0.78	2.49
Readiness to change T0	0.61	[0.26–1.48]	0.28	1.28	[0.64–2.55]	0.49	0.26	1.48	0.64	2.55
Problem drinking T0	0.75	[0.35–1.65]	0.48	0.74	[0.41–1.33]	0.31	0.35	1.65	0.41	1.33
Freshmen T0	1.04	[0.45–2.38]	0.93	1.15	[0.57–2.33]	0.69	0.45	2.38	0.57	2.33
Fraternity or sorority membership T0	1.07	[0.53–2.18]	0.85	0.87	[0.48–1.56]	0.63	0.53	2.18	0.48	1.56
Carnival participation T1	0.78	[0.31–1.94]	0.59	0.99	[0.55–1.80]	0.99	0.31	1.94	0.55	1.80
Frequency of binge drinking by										
Gender	0.75	[0.37–1.52]	0.43	1.38	[0.78–2.44]	0.27	0.37	1.52	0.78	2.44
Readiness to change T0	0.67	[0.28–1.61]	0.37	1.30	[0.65–2.59]	0.45	0.28	1.61	0.65	2.59
Problem drinking T0	0.70	[0.32–1.49]	0.35	0.75	[0.41–1.36]	0.33	0.32	1.49	0.41	1.36
Freshmen T0	1.00	[0.45–2.24]	0.99	1.14	[0.57–2.32]	0.71	0.45	2.24	0.57	2.32
Fraternity or sorority membership T0	1.08	[0.54–2.18]	0.83	0.90	[0.50–1.61]	0.72	0.54	2.18	0.50	1.61
Carnival participation T1	0.83	[0.33–2.09]	0.69	1.06	[0.59–1.91]	0.84	0.33	2.09	0.59	1.91

Note. T0, baseline assessment; T1, 1-month follow-up.

corpus-oa-uclii-01/10.1093.alcalc.ags133/tables/table4/table.svg.html

Subtables are indicated by dark green headings and pale green background.
Additional columns generated by splitting compound column contents are coloured yellow, or yellow-green for values within subtables. Row headings (observation labels) are orange within

subtables, red for top-level rows.

Appendix: Example Workflow

The following example shows how a table in PDF

This table is from Voogt et al. (2013), DOI: 10.1093.alcalc.ags133, Table 2:

Original PDF layout:

This is an APA-style table containing subtables, supercolumn headers and compound columns.

Table 2. Percentage of heavy drinking and frequency of binge drinking at the 1- and 6-month follow-up by condition (WDYD intervention versus control): intention-to-treat (multiple imputation) and completers-only analyses

	Intervention		Control				
	n	%	*n*	%	OR	95% CI	*P*
Heavy drinking							
1-month follow-up							
Intention-to-treat	456	81.5	451	82.8	0.92	[0.64–1.31]	0.63
Completers-only	412	81.6	409	83.1	0.90	[0.62–1.29]	0.55
6-month follow-up							
Intention-to-treat	456	68.0	451	66.0	1.10	[0.83–1.46]	0.52
Completers-only	412	67.5	409	65.5	1.09	[0.82–1.46]	0.55
Frequency of binge drinking							
1-month follow-up							
Intention-to-treat	456	80.2	451	82.3	0.88	[0.61–1.25]	0.46
Completers-only	412	80.6	409	82.9	0.86	[0.60–1.22]	0.39
6-month follow-up							
Intention-to-treat	456	67.0	451	65.2	1.09	[0.82–1.44]	0.56
Completers-only	412	66.7	409	65.0	1.08	[0.81–1.44]	0.61

The original document makes use of indentation to indicate subtables and columns (n, %) are grouped under super-headers of **Intervention** and **Control.**

The basic grid resulting from layout analysis by column and row coordinates has this form:

Table 2. Percentage of heavy drinking and frequency of binge drinking at the 1- and 6-month follow-up by condition (WDYD intervention versus control): intention-to-treat (multiple imputation) and completers-only analyses

	Intervention		**Control**				
	n	**%**	**n**	**%**	**OR**	**95% CI**	**P**
Heavy drinking							
1-month follow-up							
Intention-to-treat	456	81.5	451	82.8	0.92	[0.64–1.31]	0.63
Completers-only	412	81.6	409	83.1	0.90	[0.62–1.29]	0.55
6-month follow-up							
Intention-to-treat	456	68.0	451	66.0	1.10	[0.83–1.46]	0.52
Completers-only	412	67.5	409	65.5	1.09	[0.82–1.46]	0.55
Frequency of binge drinking							
1-month follow-up							
Intention-to-treat	456	80.2	451	82.3	0.88	[0.61–1.25]	0.46
Completers-only	412	80.6	409	82.9	0.86	[0.60–1.22]	0.39
6-month follow-up							
Intention-to-treat	456	67.0	451	65.2	1.09	[0.82–1.44]	0.56
Completers-only	412	66.7	409	65.0	1.08	[0.81–1.44]	0.61

This grid is represented using basic HTML table structure (`td` and `tr`) as an intermediate format.

Note that this simple grid extraction does not reflect the indentation within the left-most

column, nor the subtable structure this conveys. Column super-headers Intervention and Control also do not 'span' the underlying n and % columns. This corresponds to the output from the pipeline developed for CM-UCLI.

Result

The following shows the results of running the table through the current CM-UCLII system. The semantic restructuring, content enhancement and normalisation processes have been applied and the following features have been identified and represented in the semantically structured output.

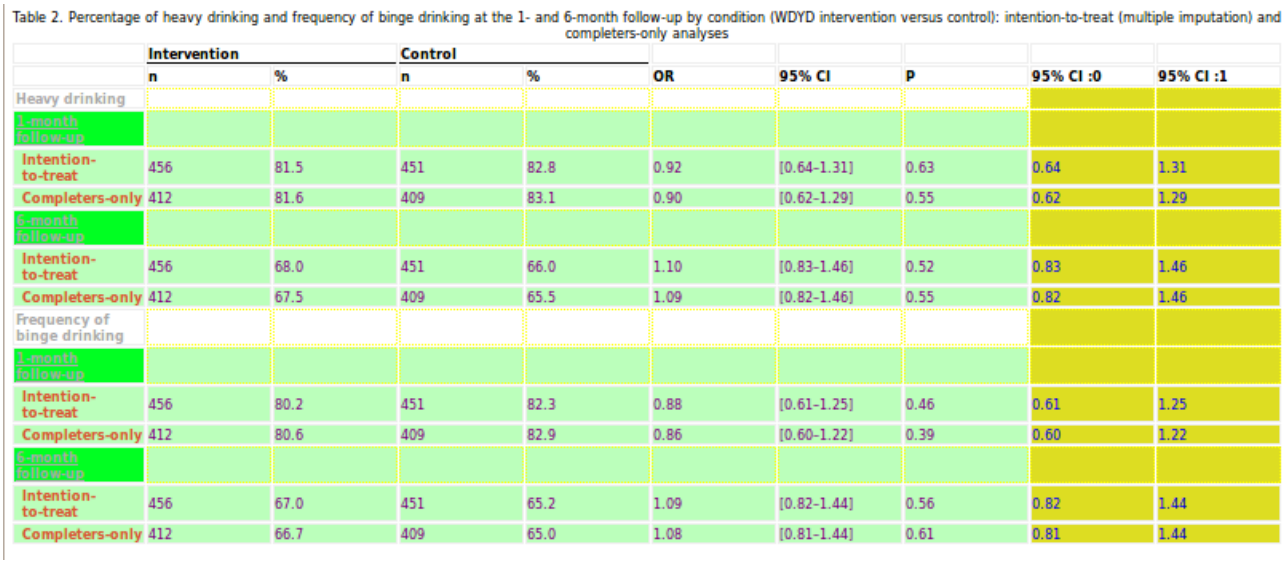


Table 2. Percentage of heavy drinking and frequency of binge drinking at the 1- and 6-month follow-up by condition (WDYD intervention versus control): intention-to-treat (multiple imputation) and completers-only analyses

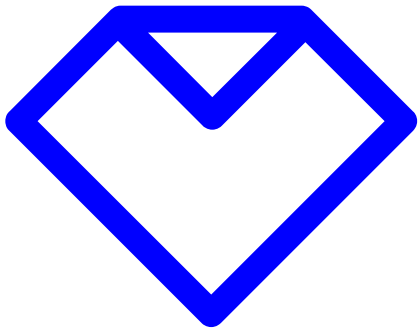
	Intervention		Control						
	n	%	n	%	OR	95% CI	P	95% CI :0	95% CI :1
Heavy drinking									
1-month follow-up									
Intention-to-treat	456	81.5	451	82.8	0.92	[0.64-1.31]	0.63	0.64	1.31
Completers-only	412	81.6	409	83.1	0.90	[0.62-1.29]	0.55	0.62	1.29
6-month follow-up									
Intention-to-treat	456	68.0	451	66.0	1.10	[0.83-1.46]	0.52	0.83	1.46
Completers-only	412	67.5	409	65.5	1.09	[0.82-1.46]	0.55	0.82	1.46
Frequency of binge drinking									
1-month follow-up									
Intention-to-treat	456	80.2	451	82.3	0.88	[0.61-1.25]	0.46	0.61	1.25
Completers-only	412	80.6	409	82.9	0.86	[0.60-1.22]	0.39	0.60	1.22
6-month follow-up									
Intention-to-treat	456	67.0	451	65.2	1.09	[0.82-1.44]	0.56	0.82	1.44
Completers-only	412	66.7	409	65.0	1.08	[0.81-1.44]	0.61	0.81	1.44

The output is in the form of semantically structured HTML5. For presentation purposes, CSS3 stylesheet has been applied referencing and highlighting these semantic structures. Style and presentation are kept separate.

In the above example, subtables have been identified (styled in green). **Intervention** and **Control** have been identified as supercolumn headers spanning **n** and **%** columns.

A CSV file containing only the numeric values and basic headers is also output, suitable for use in other analysis or display tools.

ContentMine



ContentMine Ltd
16 Mill Lane
Cambridge
CB2 1RX

+44 (0)1223 324379

info@contentmine.org | www.contentmine.org