Data qubes from Workbook

PhuseSubTeamAnalysisResults@example.org

2015-01-11

Contents

Setup

First load the package.

```
library(rrdfqbcrnd0)
```

Generating RDF data cube from specification in a spreadsheet and export as turtle file

The generation of RDF data cube can be specified in a spreadsheet. The outputs below shows the meta data for generation of the DM RDF data cube.

```
RDFCubeWorkbook<- system.file("extdata/sample-cfg", "RDFCubeWorkbook.xlsx", package="rrdfqbcrnd0")
cubeMetadata <- read.xlsx(RDFCubeWorkbook,
                          sheetName=paste0("DM-Components"),
                          stringsAsFactors=FALSE)
knitr::kable(
  cubeMetadata[ cubeMetadata$compType %in% c("dimension", "attribute", "measure"),
               c("codeType", "compName","nciDomainValue", "compLabel")]
  )
```

codeType	compName	nciDomainValue	compLabel
DATA	trt01a	NA	Treatment Arm
SDTM	sex	C66731	Sex (Gender)
DATA	saffl	NA	Safety Population Flag

codeType	compName	nciDomainValue	compLabel
DATA	procedure	NA	Statistical Procedure
DATA	factor	NA	Type of procedure (quantity, proportion…)
SDTM	race	C74457	Race
NA	measure	NA	Value of the statistical measure
NA	unit	NA	Unit of measure
NA	denominator	NA	Denominator for a proportion (oskr) subset on which a statistic is based

```
knitr::kable(cubeMetadata[ cubeMetadata$compType=="metadata",c("compName","compLabel")])
```

	compName	compLabel
10	obsURL	https://phuse-scripts.googlecode.com/svn/trunk/scriptathon2014/data/adsl.xpt
11	obsFileName	dm.AR.csv
12	dataCubeFileName	DC-DM-R-V
13	cubeVersion	0.5.2
14	createdBy	Tim Williams
15	description	Cube with 6 Dimensions (factor, procedure, race, saffl, sex, trt01a), 2 Attributes (denominato
16	providedBy	PhUSE Results Metadata Working Group
17	comment	Example Demographics data supplied by Ian Fleming via R.. All dimensions have a Codelist
18	title	Demographics Analysis Results
19	label	Demographics results data set.
20	wasDerivedFrom	demog.AR.csv
21	domainName	DM
22	obsFileNameDirectory	!example
23	dataCubeOutDirectory	!temporary

The next statements demonstrates how to create two RDF data cubes according to the specfications in the excel spreadsheet. Note the contents of the RDF data cube is read from the csv file given by obsFileName in directory given by obsFileNameDirectory. The value “‘!example´´´ specifies that the file should be read from sample data in the package. The dataCubeOutDirectory give the directory name for the generated RDF data cube.

```
dm.cube.fn<- BuildCubeFromWorkbook(RDFCubeWorkbook, "DM" )
cat("DM cube stored as ", dm.cube.fn, "\n")
```

```
## DM cube stored as  /tmp/RtmpiHiTyi/DC-DM-R-V-0-5-2.TTL
```

```
ae.cube.fn<- BuildCubeFromWorkbook(RDFCubeWorkbook, "AE" )
cat("AE cube stored as ", ae.cube.fn, "\n")
```

```
## AE cube stored as  /tmp/RtmpiHiTyi/DC-AE-R-V-0-5-2.TTL
```

Notes

In the read.xlsx, if all cells in a column is missing, then the input fails.

Future version may replace the use of the DomainName, eg. DM and AE in the examples above, with another way of deriving identification of the table

The attribute denominator may be changed to a dimension to handle more complex situations. For example if there are percentages for TRT01A and SEX using respectively TRT01A and SEX as denominator. This will be represented by two observations with by definition the same dimensions but different value for the attribute denominator. However, this will violate the intergrity constraints for a RDF Data Cube (TODO: Add IC name).

Input the generated turtle file

Now look at the generated cubes by loading the turle files. Note: by specifying prefix the output contains is shown using the prefixes. Note for future: This may be a disadvantage if the value of the prefix, say ds, changes.

The rest of the code only depends on the value of dataCubeFile.

```
## [1] "Number of triples: 1039"
```

First set values for accessing the cube.

The next statement shows the first 10 triples in the cube.

s	p	o
dccs:saffl	qb:dimension	prop:saffl
dccs:saffl	rdfs:label	Safety Population Flag
dccs:saffl	rdf:type	qb:ComponentSpecification
ds:obs2	prop:saffl	code:saffl-Y
ds:obs2	prop:unit	*NULL*
ds:obs2	rdfs:label	2

The next statement shows the first 10 triples in the cube, where the subject is a qb:Observation.

s	p	o
ds:obs35	rdfs:label	35
ds:obs35	prop:race	code:race-AMERICAN_INDIAN_OR_ALASKA_NATIVE
ds:obs35	rdf:type	qb:Observation
ds:obs35	prop:denominator	RACE
ds:obs35	prop:measure	0
ds:obs35	prop:procedure	code:procedure-percent

s	p	o
ds:obs35	prop:factor	code:factor-proportion
ds:obs35	prop:trt01a	code:trt01a-Placebo
ds:obs35	prop:saffl	code:saffl-Y
ds:obs35	qb:dataSet	ds:dataset-DM

The cube components are shown in the next output.

```
##          vn                                       label
## 1    factor Type of procedure (quantity, proportion...)
## 2 procedure                       Statistical Procedure
## 3      race                                        Race
## 4     saffl                      Safety Population Flag
## 5       sex                                Sex (Gender)
## 6    trt01a                               Treatment Arm
```

vn	label
factor	Type of procedure (quantity, proportion…)
procedure	Statistical Procedure
race	Race
saffl	Safety Population Flag
sex	Sex (Gender)
trt01a	Treatment Arm

The codelists are shown in the next output.

```
##           vn                                            clc
## 1     factor                                     factor-AGE
## 2     factor                              factor-proportion
## 3     factor                                factor-quantity
## 4  procedure                                procedure-count
## 5  procedure                                  procedure-max
## 6  procedure                                 procedure-mean
## 7  procedure                               procedure-median
## 8  procedure                                  procedure-min
## 9  procedure                              procedure-percent
## 10 procedure                                procedure-stdev
## 11      race          race-AMERICAN_INDIAN_OR_ALASKA_NATIVE
## 12      race                                     race-ASIAN
## 13      race                 race-BLACK_OR_AFRICAN_AMERICAN
## 14      race race-NATIVE_HAWAIIAN_OR_OTHER_PACIFIC_ISLANDER
## 15      race                                     race-WHITE
## 16      race                                     race-_ALL_
## 17     saffl                                        saffl-Y
## 18       sex                                          sex-F
## 19       sex                                          sex-M
## 20       sex                                          sex-U
## 21       sex                                         sex-UN
## 22       sex                                      sex-_ALL_
## 23    trt01a                                 trt01a-Placebo
## 24    trt01a                    trt01a-Xanomeline_High_Dose
## 25    trt01a                     trt01a-Xanomeline_Low_Dose
## 26    trt01a                                   trt01a-_ALL_
##                                     prefLabel
## 1                                         AGE
## 2                                  proportion
## 3                                    quantity
## 4                                       count
## 5                                         max
## 6                                        mean
## 7                                      median
## 8                                         min
## 9                                     percent
## 10                                      stdev
## 11          AMERICAN INDIAN OR ALASKA NATIVE
## 12                                      ASIAN
## 13                  BLACK OR AFRICAN AMERICAN
## 14 NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER
## 15                                      WHITE
## 16                                      _ALL_
## 17                                          Y
## 18                                          F
## 19                                          M
## 20                                          U
## 21                                         UN
## 22                                      _ALL_
## 23                                    Placebo
## 24                       Xanomeline High Dose
## 25                        Xanomeline Low Dose
## 26                                      _ALL_
```

vn	clc	prefLabel
factor	factor-AGE	AGE
factor	factor-proportion	proportion
factor	factor-quantity	quantity
procedure	procedure-count	count
procedure	procedure-max	max
procedure	procedure-mean	mean
procedure	procedure-median	median
procedure	procedure-min	min
procedure	procedure-percent	percent
procedure	procedure-stdev	stdev
race	race-AMERICAN_INDIAN_OR_ALASKA_NATIVE	AMERICAN INDIAN OR ALASKA NA

vn	clc	prefLabel
race	race-ASIAN	ASIAN
race	race-BLACK_OR_AFRICAN_AMERICAN	BLACK OR AFRICAN AMERICAN
race	race-NATIVE_HAWAIIAN_OR_OTHER_PACIFIC_ISLANDER	NATIVE HAWAIIAN OR OTHER PAC
race	race-WHITE	WHITE
race	race-*ALL*	*ALL*
saffl	saffl-Y	Y
sex	sex-F	F
sex	sex-M	M
sex	sex-U	U
sex	sex-UN	UN
sex	sex-*ALL*	*ALL*
trt01a	trt01a-Placebo	Placebo
trt01a	trt01a-Xanomeline_High_Dose	Xanomeline High Dose
trt01a	trt01a-Xanomeline_Low_Dose	Xanomeline Low Dose
trt01a	trt01a-*ALL*	*ALL*

Notes

instead of using gsub the codelist values should be obtained in a more straightforward way

this involves a new version of the ph.recode function

the rrdf package could be extended to expand the URI using the Jena expandPrefix method

---

The dimensions are shown in the next output.

p

prop:trt01a
prop:race
prop:factor
prop:procedure prop:sex
prop:saffl

Then the attributes as shown in the next output.

p

prop:unit
prop:denominator

And finally the SPARQL query for observations, showing only the first 10 observations.

```
## prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
## prefix skos: <http://www.w3.org/2004/02/skos/core#>
## prefix prov: <http://www.w3.org/ns/prov#>
## prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
## prefix dcat: <http://www.w3.org/ns/dcat#>
## prefix owl: <http://www.w3.org/2002/07/owl#>
## prefix xsd: <http://www.w3.org/2001/XMLSchema#>
## prefix qb: <http://purl.org/linked-data/cube#>
## prefix pav: <http://purl.org/pav>
## prefix dct: <http://purl.org/dc/terms/>
## prefix mms: <http://rdf.cdisc.org/mms#>
## prefix cts: <http://rdf.cdisc.org/ct/schema#>
## prefix rrdfqbcrnd0: <http://www.example.org/rrdfqbcrnd0/>
## prefix code: <http://www.example.org/dc/code/>
## prefix prop: <http://www.example.org/dc/dm/prop/>
## prefix dccs: <http://www.example.org/dc/dm/dccs/>
## prefix ds: <http://www.example.org/dc/dm/ds/>
##  select * where { ?s a qb:Observation  ;
##        qb:dataSet ds:dataset-DM  ;
##  prop:trt01a ?trt01a;
## prop:race ?race;
## prop:factor ?factor;
## prop:procedure ?procedure;
## prop:sex ?sex;
## prop:saffl ?saffl; prop:unit ?unit;
## prop:denominator ?denominator; prop:measure      ?measure ;
##  optional{ ?trt01a skos:prefLabel ?trt01avalue . }
## optional{ ?race skos:prefLabel ?racevalue . }
## optional{ ?factor skos:prefLabel ?factorvalue . }
## optional{ ?procedure skos:prefLabel ?procedurevalue . }
## optional{ ?sex skos:prefLabel ?sexvalue . }
## optional{ ?saffl skos:prefLabel ?safflvalue . }
##  }
```

trt01avalue	racevalue	factorvalue	procedurevalue	sexvalue	safflval
Placebo	AMERICAN INDIAN OR ALASKA NATIVE	proportion	percent	*ALL*	Y
Xanomeline High Dose	AMERICAN INDIAN OR ALASKA NATIVE	proportion	percent	*ALL*	Y
Xanomeline High Dose	BLACK OR AFRICAN AMERICAN	proportion	percent	*ALL*	Y
Placebo	*ALL*	quantity	count	F	Y
Placebo	*ALL*	quantity	count	*ALL*	Y
Xanomeline Low Dose	*ALL*	AGE	stdev	*ALL*	Y
Xanomeline High Dose	*ALL*	AGE	mean	*ALL*	Y
Xanomeline Low Dose	*ALL*	proportion	percent	M	Y
Placebo	*ALL*	AGE	mean	*ALL*	Y
Placebo	WHITE	quantity	count	*ALL*	Y

Here is how to re-produce the metadata for the workbook. First get the dimensions, measure and attribute

compType	compName	codeType	nciDomainValue
dimension	prop:trt01a	NA	NA
dimension	prop:race	NA	C74457
dimension	prop:factor	NA	NA
dimension	prop:procedure	NA	NA
dimension	prop:sex	NA	C66731
dimension	prop:saffl	NA	NA
attribute	prop:unit	NA	NA
attribute	prop:denominator	NA	NA
measure	prop:measure	NA	NA

Secondly, get the metadata for the workbook. To get the metadata element “cubeVersion” a workaround is needed. The cubeversion is not directly available but from dcat:distribution derived as the result of paste0(“DC-”, domainName,“-R-V-”,cubeVersion,“.TTL”).

	compType	compName	compLabel
	metadata	title	Demographics Analysis Results
	metadata	distribution	DC-DM-R-V-0-5-2.TTL
	metadata	comment	Example Demographics data supplied by Ian Fleming via R.. All dimensions hav
	metadata	label	Demographics results data set.
	metadata	description	Cube with 6 Dimensions (factor, procedure, race, saffl, sex, trt01a), 2 Attributes
	metadata	obsFileName	dm.AR.csv
compLabel	metadata	cubeVersion	0.5.2

For comparison, see the meta data from the excel workbook in the beginning of the document.