Create Integrity Contraints SPARQL Queries from RDF data cube definition

mja@statgroup.dk

2015-03-13

Contents

Preliminaries

When developing, the script is intented to run from the package root after the setup for development as defined in the README.md.

```
knit(input="inst/data-raw/create-qb-IC-dataset.Rmd",
     output="inst/data-raw/create-qb-IC-dataset.md")
```

R-code

IC-19 is two queries, so it is split into IC-19a and IC-19b: For IC-20 and IC-21 special handling are needed. The queries are templates and the value of p should be inserted as $p in the template.

```
library(RCurl)
library(XML)
library(devtools)

qbURL<-"http://www.w3.org/TR/2014/REC-vocab-data-cube-20140116/"
if (! url.exists(qbURL) ) {
  stop(paste0("Can not access URL ",qbURL))
}


# Acknowledgement: I got the approach from
# http://stackoverflow.com/questions/1395528/scraping-html-tables-into-r-data-frames-using-the-xml-pack

webpage <- getURL(qbURL)
# The following two lines is suggested in the stackoverflow post
# Apparantly not needed here
## Process escape characters
## webpage <- readLines(tc <- textConnection(webpage)); close(tc)

# Parse the html tree, ignoring errors on the page
pagetree <- htmlTreeParse(webpage, error=function(...){}, useInternalNodes = TRUE)
```

```
# appears that integrity checks starte with h3 and then a table with class bordered-table
# so that's what we look for
both<- getNodeSet(pagetree,"//*/h3[@id]|//*/table[@class='bordered-table']/tbody/tr/td/pre")

       irq20<- "
SELECT ?p WHERE {
    ?hierarchy a qb:HierarchicalCodeList ;
                qb:parentChildProperty ?p .
    FILTER ( isIRI(?p) )
}
"

irq21<-"
SELECT ?p WHERE {
    ?hierarchy a qb:HierarchicalCodeList;
                qb:parentChildProperty ?pcp .
    FILTER( isBlank(?pcp) )
    ?pcp  owl:inverseOf ?p .
    FILTER( isIRI(?p) )
}
"

storeIC<-function(ictitle,instantiationRq,rq) {
   return( list(
     ictitel= ictitle,
     HasInstantiation= nchar(instantiationRq)>0,
     instantiationRq= instantiationRq,
     rq= rq) )
  }

qbIClist<- list()
for (i in 1:(length(both)-1)) {
  icname<- xmlGetAttr(both[[i]],"id",default="none")
  if (grepl('ic-[1-9]([0-9])*', icname ) ) {
   ictitle<- unlist(xmlValue(xmlChildren(both[[i]])$text ))
   rq<- xmlValue(xmlChildren(both[[i+1]])$text)
#   print(paste0( "Node ", i, ", IC name ", icname, " - ", ictitle ))
   if (icname %in% "ic-19") {
     ## XXX change list to vection - use unlist ??
###     print(i)
      rq<- paste0(unlist(strsplit(xmlValue(xmlChildren(both[[i+1]])$text),"\n"))[1:8], collapse="\n")
      qbIClist[["ic-19a"]]<- storeIC(gsub("IC-19", "IC-19a", ictitle), "", rq)
      rq<- paste0(unlist(strsplit(xmlValue(xmlChildren(both[[i+1]])$text),"\n"))[10:17], collapse="\n")
      qbIClist[["ic-19b"]]<- storeIC(gsub("IC-19", "IC-19b", ictitle), "", rq)
     } else if ( icname == "ic-20" ) {
       qbIClist[[icname]]<- storeIC( ictitle, irq20, rq)
     } else if ( icname == "ic-21" ) {
       qbIClist[[icname]]<- storeIC( ictitle, irq21, rq)
     } else {
       qbIClist[[icname]]<- storeIC( ictitle, "", rq)
   }
  }
  }
```

Here are the integrity constraints:

```
for (icname in names(qbIClist)) {
  ##    fileConn<-file(paste0(icname, ".rq"))
  icall<- qbIClist[[icname]]
  cat( paste(names(icall),icall,collapse="\n",sep="\n"),"\n")
##    close(fileConn)
  }
```

Data are stored in the data directory, following R packages by Hadley Wickham and Writing R Extensions.

knit runs the script in the data-raw directory, so it would be expected to use pkg="../.." to store the qbIClist in the data directory However, it did not work - hence the setwd below-

```
devtools::use_data(qbIClist,pgk="../..",overwrite=TRUE)
```

```
# This stores the qbIClist in the data directory
# Consider making it internal

devtools::use_data(qbIClist,overwrite=TRUE)

print("Done")
```