rols: an R interface to the Ontology Lookup Service

Laurent Gatto[*]
Computational Proteomics Unit[†]
University of Cambridge, UK

October 7, 2014

Contents

---

[*] `lg390@cam.ac.uk`

[†] `http://cpu.sysbiol.cam.ac.uk`

1 Introduction

The Ontology Lookup Service[1] (OLS) [1, 2] is a spin-off of the PRoteomics IDEntifications database (PRIDE) service, located at the EBI. OLS provides a unified interface to 93 ontologies (see below).

rols makes use of the SOAP service at the EBI to post XML requests. The SOAP XML responses are then parsed and returned in an *R* friendly data structure. This is achieved using Duncan Temple Lang's *SSOAP* package [3].

2 Brief rols overview

2.1 Ontologies

There are 93 ontologies available in the OLS, listed in the table 1 below. Their name is to be use to defined which ontology to query.

Name	Description
AAO	Amphibian Gross Anatomy
APO	Yeast phenotypes
ATO	Amphibian Taxonomy
BFO	Basic Formal Ontology
BS	Biosapiens Annotations
BSPO	Spatial Reference Ontology
BTO	BRENDA tissue / enzyme source
CARO	Common Anatomy Reference Ontology
CCO	Cell Cycle Ontology
CHEBI	Chemical Entities of Biological Interest
CL	Cell Type
CMPO	Cell Microscopy Phenotype Ontology
DDANAT	Dictyostelium discoideum Anatomy
DOID	Human Disease
DPO	Drosophila Phenotype Ontology
ECO	Evidence Codes
EDAM	EMBRACE Data and Methods
EFO	Experimental Factor Ontology
EHDA	Human Developmental Anatomy, timed version
EMAP	Mouse Gross Anatomy and Development, timed
EMAPA	Mouse Gross Anatomy and Development, abstract
ENA	European Nucleotide Archive Submission Ontology
ENVO	Environmental Ontology
EO	Plant Environmental Conditions

---

[1] http://www.ebi.ac.uk/ontology-lookup/

EV	eVOC (Expressed Sequence Annotation for Humans)
FAO	Fungal Gross Anatomy
FBbi	Biological Imaging Methods
FBbt	Drosophila Gross Anatomy
FBcv	Flybase Controlled Vocabulary
FBdv	Drosophila Development
FBsp	Fly taxonomy
FIX	Physico-Chemical Methods and Properties
FMA	Foundational Model of Anatomy Ontology
FYPO	Fission Yeast Phenotype Ontology
GAZ	Gezetteer ontology
GO	Gene Ontology
GRO	Cereal Plant Development
HAO	Hymenoptera Anatomy Ontology
HOM	Homology Ontology
HP	Human phenotype ontology
IDO	Infectious Disease Ontology
IEV	Event (INOH)
IMR	Molecule Role (INOH)
LSM	Leukocyte Surface Markers
MA	Mouse Adult Gross Anatomy
MAT	Minimal Information About Anatomy ontology
MFO	Medaka Fish Anatomy and Development
MI	Molecular Interaction (PSI MI 2.5)
MIAA	Minimal Information About Anatomy ontology
MIRO	Mosquito Insecticide Resistance
MOD	Protein Modifications (PSI-MOD)
MP	Mammalian Phenotype
MPATH	Mouse Pathology
MS	PSI Mass Spectrometry Ontology
NEWT	NEWT UniProt Taxonomy Database
NPO	NanoParticle Ontology
OBA	Ontology of bio-attributes
OBO_REL	OBO Relationship Types
OPL	Ontology of Parasite Life Cycle
Orphanet	Orphanet
PAR	Protein Affinity Reagents
PATO	Phenotypic qualities (properties)
PM	Phenotypic manifestation (genetic context)
PO	Plant Ontology (Structure, Growth and Developmental Stage)
PRIDE	PRIDE Controlled Vocabulary
PRO	Protein Ontology
PW	Pathway Ontology
REX	Physico-Chemical Process
RO	Multiple Alignment

SBO	Systems Biology Ontology
SEP	Separation Methods
SO	Sequence Types and Features
SPD	Spider Comparative Biology Ontology
SYMP	Symptom Ontology
TADS	Tick Gross Anatomy
TAIR	Arabidopsis Development
TAO	Teleost Anatomy and Development Ontology
TAXRANK	Taxonomic rank vocabulary
TGMA	Mosquito Gross Anatomy
TO	Cereal Plant Trait
TRANS	Pathogen transmission
TTO	Teleost taxonomy
UBERON	Uber anatomy ontology
UO	Unit Ontology
VariO	Variation Ontology
WBPhenotype	C. elegans phenotype
WBbt	C. elegans gross anatomy
WBls	C. elegans Development
XAO	Xenopus anatomy and development
ZDB	Zebrafish Anatomy and Development
ZEA	Maize Gross Anatomy
ZFA	Zebrafish Anatomy and Development
qcML	Quality Control Ontology

Table 1: Available ontologies.

2.2 Interface

Table 2 summarised the common interface available for the 93 ontologies of table 1. More information is provided in the respective manual pages.

3 Use case

A researcher might be interested in the trans-Golgi network and interested in knowing in which ontologies his favourite organelle is referenced. This can be done by querying all ontologies with a relevant pattern. The code below describes how to achieve this.

```
> library("rols")
> alltgns <- olsQuery("trans-golgi network")
```

As shown below, 4 different ontologies have matched the query string.

Function	Description
olsVersion	Returns the OLS version
ontologies	Returns all available ontologies
ontologyNames	Returns all ontologyNames
ontologyLoadDate	Returns the ontology load date
isIdObsolete	Is the ontology id obsolete
term	Returns the term of a given identifier
termMetadata	Retuns an identifier's metadata
termXrefs	Returns the idenifier's ontology cross references
rootId	Retuns the root identifiers of an ontology
allIds	Returns all identifiers and terms of an ontology
olsQuery	Returns matching identifiers
parents	Returns the parent(s) of a term.
childrenRelations	Returns the children relation type(s).

Table 2: Functions available to query the ontologies.

```
> alltgns
                                                                CCO:C0000738 
                  "CCO:trans-Golgi network transport vesicle membrane" 
                                                                  GO:0005802 
                                                 "GO:trans-Golgi network" 
                                                                CCO:C0000984 
                           "CCO:trans-Golgi network transport vesicle" 
                                                                CCO:C0000975 
                        "CCO:clathrin coat of trans-Golgi network vesicle" 
                                                                CCO:C0001468 
                                          "CCO:trans-Golgi network membrane" 
                                                                   FMA:61756 
                                                  "FMA:Trans Golgi network" 
                                                                CCO:C0000381 
                                                  "CCO:trans-Golgi network" 
                                                                PR:000016299 
                   "PRO:trans-Golgi network integral membrane protein 2" 
                                                                PR:000016925 
                   "PRO:trans-Golgi network integral membrane protein 1" 
                                                                  GO:0032588 
                                           "GO:trans-Golgi network membrane" 
                                                                  GO:0044795 
                "GO:trans-Golgi network to recycling endosome transport" 
                                                                  GO:0012510 
                    "GO:trans-Golgi network transport vesicle membrane" 
                                                                  GO:0098540 
    "GO:lumenal side of trans-Golgi network transport vesicle membrane" 
```

```
                                                             GO:0098541 
"GO:cytoplasmic side of trans-Golgi network transport vesicle membrane" 
                                                             GO:0098564 
                       "GO:trans-Golgi network transport vesicle lumen" 
                                                             GO:0030130 
                      "GO:clathrin coat of trans-Golgi network vesicle" 
                                                             GO:0030140 
                             "GO:trans-Golgi network transport vesicle" 
```

```
> allonts <- sapply(strsplit(alltgns, ":"), "[", 1)
> onto.tab <- table(allonts)
> onto.tab
```

```
allonts
CCO FMA  GO PRO 
  5   1   9   2 
```

The description of the 4 ontologies of interest can then be used to subset the ontology description:

```
> ontologies()[names(onto.tab),]
```

```
    Name                            Description
CCO  CCO                    Cell Cycle Ontology
FMA  FMA Foundational Model of Anatomy Ontology
GO    GO                          Gene Ontology
PRO  PRO                       Protein Ontology
```

To restrict the search to a specific ontology of interest, one can specify the ontolgy name as a parameter to `olsQuery`.

```
> gotgns <- olsQuery("trans-golgi network", "GO")
> gotgns
```

```
                                                          GO:0005802 
                                               "trans-Golgi network" 
                                                          GO:0098540 
    "lumenal side of trans-Golgi network transport vesicle membrane" 
                                                          GO:0098541 
"cytoplasmic side of trans-Golgi network transport vesicle membrane" 
                                                          GO:0098564 
                       "trans-Golgi network transport vesicle lumen" 
                                                          GO:0030130 
                      "clathrin coat of trans-Golgi network vesicle" 
                                                          GO:0032588 
                                      "trans-Golgi network membrane" 
                                                          GO:0030140 
                             "trans-Golgi network transport vesicle" 
                                                          GO:0044795 
```

```
               "trans-Golgi network to recycling endosome transport"
                                                          GO:0012510
                    "trans-Golgi network transport vesicle membrane"
```

Details about relevant terms can be retrieved with the `term` and `termMetadata` functions. This functionality provides on-line access to the same data that is available in the *GO.db*, and can be extended to any of the 93 available ontologies.

```
> term("GO:0005802", "GO")

[1] "trans-Golgi network"

> mtd <- termMetadata("GO:0005802", "GO")
> names(mtd)

[1] "related_synonym_3" "related_synonym_2" "related_synonym_1"
[4] "definition"        "related_synonym_4" "exact_synonym_1"
[7] "comment"           "exact_synonym_2"

> mtd

related_synonym_3: Golgi trans face
related_synonym_2: Golgi trans-face
related_synonym_1: late Golgi
definition: The network of interconnected tubular and cisternal
  structures located within the Golgi apparatus on the side distal to
  the endoplasmic reticulum, from which secretory vesicles emerge. The
  trans-Golgi network is important in the later stages of protein
  secretion where it is thought to play a key role in the sorting and
  targeting of secreted proteins to the correct destination.
related_synonym_4: maturing face
exact_synonym_1: trans Golgi network
comment: There are different opinions about whether the TGN should be
  considered part of the Golgi apparatus or not.  We follow Alberts et
  al, 1994 (ISBN:0815316194), who consider it to be a part.
exact_synonym_2: TGN

> strwrap(mtd["comment"])

[1] "There are different opinions about whether the TGN should be considered"
[2] "part of the Golgi apparatus or not.  We follow Alberts et al, 1994"
[3] "(ISBN:0815316194), who consider it to be a part."

> strwrap(mtd["definition"])

[1] "The network of interconnected tubular and cisternal structures located"
[2] "within the Golgi apparatus on the side distal to the endoplasmic"
[3] "reticulum, from which secretory vesicles emerge. The trans-Golgi"
[4] "network is important in the later stages of protein secretion where it"
```

```
[5] "is thought to play a key role in the sorting and targeting of secreted"
[6] "proteins to the correct destination."
```

Below, we execute the same query using the *GO.db* package.

```
> GOTERM[["GO:0005802"]]

GOID: GO:0005802
Term: trans-Golgi network
Ontology: CC
Definition: The network of interconnected tubular and cisternal
    structures located at the side of the Golgi apparatus distal to the
    endoplasmic reticulum, from which secretory vesicles emerge. The
    trans-Golgi network is important in the later stages of protein
    secretion where it is thought to play a key role in the sorting and
    targeting of secreted proteins to the correct destination.
Synonym: Golgi trans face
Synonym: Golgi trans-face
Synonym: late Golgi
Synonym: maturing face
Synonym: TGN
Synonym: trans Golgi network
```

4 On-line vs. off-line data

It is possible to observe different results with *rols* and *GO.db* [4], as a result of the different ways they access the data. *rols* or *biomaRt* [5] perform direct online queries, while *GO.db* and other annotation packages use database snapshot that are updated every release.

Both approaches have advantages. While online queries allow to obtain the latest up-to-date information, such approaches rely on network availability and quality. If reproducibility is a major issue, the version of the database to be queried can easily be controlled with off-line approaches. In the case of *rols*, altough the load date of a specific ontology can be queried with `olsVersion`, it is not possible to query a specific version of an ontology.

5 Session information

- R version 3.1.1 (2014-07-10), `x86_64-unknown-linux-gnu`
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, utils
- Other packages: AnnotationDbi 1.26.1, Biobase 2.24.0, BiocGenerics 0.10.0, DBI 0.3.1, GO.db 2.14.0, GenomeInfoDb 1.0.2, RSQLite 0.11.4, knitr 1.6, rols 1.6.1, xtable 1.7-4

- Loaded via a namespace (and not attached): BiocStyle 1.2.0, IRanges 1.22.10, RCurl 1.95-4.3, SSOAP 0.8-0, XML 3.98-1.1, XMLSchema 0.7-2, codetools 0.2-9, evaluate 0.5.5, formatR 1.0, highr 0.3, stats4 3.1.1, stringr 0.6.2, tools 3.1.1

References

[1] Richard G Côté, Philip Jones, Rolf Apweiler, and Henning Hermjakob. The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, 7:97, 2006. `doi:10.1186/1471-2105-7-97`.

[2] Richard G Côté, Philip Jones, Lennart Martens, Rolf Apweiler, and Henning Hermjakob. The ontology lookup service: more data and better tools for controlled vocabulary queries. *Nucleic Acids Res.*, 36(Web Server issue):372–376, 2008. `doi:10.1093/nar/gkn252`.

[3] Duncan Temple Lang. *SSOAP: Client-side SOAP access for S*, 2011. R package version 0.8-1. URL: `http://www.omegahat.org/SSOAP,http://www.omegahat.org,http://www.omegahat.org/bugs`.

[4] Marc Carlson, Seth Falcon, Herve Pages, and Nianhua Li. *GO.db: A set of annotation maps describing the entire Gene Ontology*. R package version 2.7.1.

[5] S Durinck, Y Moreau, A Kasprzyk, S Davis, B De Moor, A Brazma, and W Huber. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–40, Aug 2005. `doi:10.1093/bioinformatics/bti525`.