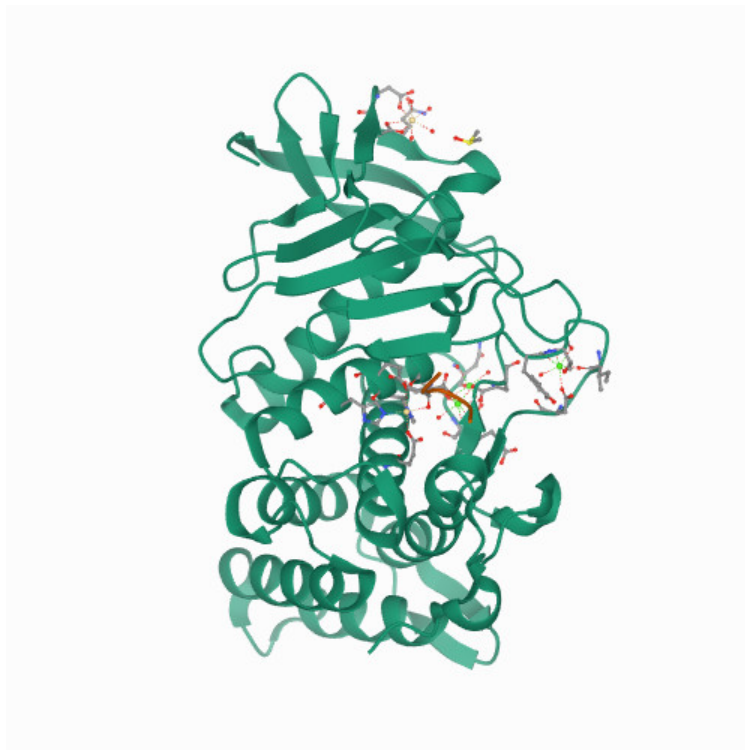


# Creation, publication, and semantic querying of biomedical data

---

Exploitation and semantic organization of data

**Francisco Mellado Martínez**



MSc in Bioinformatics  
Faculty of Biology  
University of Murcia

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Ontology</b>	<b>2</b>
2.1	Creation . . . . .	2
2.2	Querying . . . . .	2
2.3	Trifid publication . . . . .	5
<b>3</b>	<b>Conclusions</b>	<b>5</b>

# 1 Introduction

Dementia is a broad category of neurodegenerative pathologies, whose main symptom is a decline in cognitive ability severe enough to interfere with activities of daily living. Among them, Alzheimer Disease (AD) is the most common type, accounting for 60% and up to 80% of the total Dementia cases.

AD is one of the fastest rising diseases among the leading causes of death. About 10% of people of age 65 and older suffer from AD. Epidemiologic data in the last decade highlighted a sharp increase of AD incidence: According to the estimations, the number of AD subjects will increase from the 26.6 million worldwide in 2006 up to 107 million by 2050, with 16.5 in Europe. To note, 68% of the increase would be localized in the low- and middle-income countries [1].

Rather than single genes, a better approach would be investigating AD as an event related to alterations affecting entire biological pathways. A plethora of mechanisms, including neuroinflammation, defects in mitochondrial dynamics and function, dopamine metabolism as well as autophagy malfunction pathways impairments in the brain. Even if the processes were to be discussed separately, they are strictly linked with each other and are often synergistic in its damage to the Central Nervous System. This is reason enough to make an ontology from the ground up, so we can see how this pathways interacts with each other [2].

I will use data for these biological pathways we named. After that, we are to include the data in Blazegraph repository to create five queries in SPARQL. Then, we will implement an R script to execute these queries through SPARQL package. Finally, we will publish the dataset by using trifid tool. All the files needed to realize this task are stored in the directory `/home/alumno14/semántica`.

## 2 Ontology

### 2.1 Creation

To create the ontolog, we used Protege, an OWL ontology development environment. Once we created It, we exported the code as RDF/XML, and we used W3 RDF validator <https://www.w3.org/RDF/Validator/> to check wether our code was working as intended. However, since the ontology was too large this tool couldn't produce a png representation of our graph.

### 2.2 Querying

The querying itself was done using the Query tab of Protege. We also upluaded our results to Blazegraph, then we uploaded an R file containing the queries to our workspace; this script that allows to execute the queries. We will now describe and elaborate on each query:

Our first query was the following; using SELECT, it shows how many reactions take place in each cell compartment. Additionally, we ordered it using the clause *Order by*.

```
SELECT ?localization (count(distinct ?reaction) as ?count)
WHERE{
?reaction gr_ont:hasLocation ?localization} GROUP BY ?localization
Order by ?count
```

localization	count
endoplasmatic_reticulum_membrane	"1" <sup>AA</sup> <http://www.w3.org/2001/XMLSchema#integer>
mitochondria	"5" <sup>AA</sup> <http://www.w3.org/2001/XMLSchema#integer>
cell_membrane	"5" <sup>AA</sup> <http://www.w3.org/2001/XMLSchema#integer>
cytoplasm	"8" <sup>AA</sup> <http://www.w3.org/2001/XMLSchema#integer>

Figure 1: First query

For our second query, we do something similar. This time we count each gene, and how many of them are encoded in each Chromosome. This time we used an additional restriction; every chromosome that doesn't have at least 2 genes doesn't appear in the result.

```
SELECT ?Chromosome (count(distinct ?Gene) as ?count)
Where{
?Gene gr_ont:locatedIn ?Chromosome.
}
Group by ?Chromosome
HAVING (?count >1)
ORDER BY ?count
```

Chromosome	count
XX	"4" <sup>AA</sup> <http://www.w3.org/2001/XMLSchema#integer>
XVII	"3" <sup>AA</sup> <http://www.w3.org/2001/XMLSchema#integer>
I	"2" <sup>AA</sup> <http://www.w3.org/2001/XMLSchema#integer>
VI	"4" <sup>AA</sup> <http://www.w3.org/2001/XMLSchema#integer>

Figure 2: Second query

For the third query, we did the following. We extracted each cytokine (a subtype of protein) and if it had a UniProt ID and an additional comment.

```
SELECT ?cytokine ?UniProt ?Definicion
WHERE {
?cytokine rdfs:subClassOf gr_ont:cytokine .
?cytokine uni: ?UniProt .
?cytokine rdfs:comment ?Definicion
}
```

cytokine	UniProt	Definicion
CCL2	P13500	"C-C motif chemokine 2 (HC11) (Monoc"
IL6	P08887	"Interleukin-6 (IL-6) (B-cell stimulatory fa
IL4	P05112	"Interleukin-4 (IL-4) (B-cell stimulatory fa
IL1	P01583	"Interleukin-1 alpha (also hematopoietin
TNFB	P01374	"Lymphotoxin-alpha (LT-alpha) (TNF-bel
TNFA	P01375	"Tumor necrosis factor (Cachectin) (TNF
IL1B	P01584	"Interleukin-1 beta (IL-1 beta) (Catabolin

Figure 3: Third query

For the fourth, we asked for anything that has molar mass in kD. It shows everything that has at least more than 30kD of mass using the clause FILTER.

```

SELECT ?s ?mass
WHERE {
?s gr_ont:molecularWeight ?mass .
FILTER (?mass >= 30)
}
ORDER BY ?mass

```

s	mass
AmyloidB	"31" <sup>^^</sup> <http://www.w3.org/2001/XMLSchema#int>
CASP6	"34" <sup>^^</sup> <http://www.w3.org/2001/XMLSchema#int>
ATG10	"34" <sup>^^</sup> <http://www.w3.org/2001/XMLSchema#int>
ATG3	"35" <sup>^^</sup> <http://www.w3.org/2001/XMLSchema#int>
ADH	"39" <sup>^^</sup> <http://www.w3.org/2001/XMLSchema#int>
ATG7	"40" <sup>^^</sup> <http://www.w3.org/2001/XMLSchema#int>
ALDH	"41" <sup>^^</sup> <http://www.w3.org/2001/XMLSchema#int>
MAO	"55" <sup>^^</sup> <http://www.w3.org/2001/XMLSchema#int>
ATG5	"55" <sup>^^</sup> <http://www.w3.org/2001/XMLSchema#int>
MTOR	"300" <sup>^^</sup> <http://www.w3.org/2001/XMLSchema#int>

Figure 4: Fourth query

In the fifth query, we used another subtype of SPARQL query. This time we used CONSTRUCT, that returns a single RDF graph specified by a graph template. In this case, it displays a graph in which we can see the protein and the gene that encodes it. However, we couldn't implement this query in the R script.

```

CONSTRUCT { ?s gr_ont:encodedBy ?Gen}
WHERE
{
{ ?s rdf:type gr_ont:Protein } UNION { ?s gr_ont:encodedBy ?Gen }
}

```

Subject	Predicate	Object
ATG3	encodedBy	ATG3_gene
IL4	encodedBy	IL4_gene
IL1	encodedBy	IL1_gene
TNFB	encodedBy	TNFB_gene
CASP2	encodedBy	CASP2_gene
ATG10	encodedBy	ATG10_gene
TNFA	encodedBy	TNFA_gene
ATG7	encodedBy	ATG7_gene
ADH	encodedBy	ADH_gene
CASP1	encodedBy	CASP1_gene
CCL2	encodedBy	CCL2_gene
CASP6	encodedBy	CASP6_gene
ATG5	encodedBy	ATG5_gene
MTOR	encodedBy	MTOR_gene
CASP3	encodedBy	CASP3_gene
ALDH	encodedBy	ALDH_gene
IL6	encodedBy	IL6_gene
IL1B	encodedBy	IL1B_gene

Figure 5: Fifth query

Finally, our last query is a simple ASK query. This return a boolean TRUE or FALSE whether the condition has a solution. No information is returned about the possible query solutions. In this particular example, we ask if reaction17 is mediated by Toll Like Receptors (TLR):

```
ASK
WHERE { gr_ont:reaction17 gr_ont:mediatedBy gr_ont:TLR}
```

Result
True

Figure 6: Sixth and last query

All the queries were implemented in a R script that can also be found on the given directory.

## 2.3 Trifid publication

The data should be published following FAIR principles to make them reproducible and more accessible. One of the most important parts of this process is the metadata which includes two types of information. On one hand, there are Metadata about the content. On the otherhand, we have technical metadata for example the URI of the dataset, the type of license, or the URI of the graphs, or who has generated these data.

The other remarkable point of publishing the data is that we have to establish a connection between blazegraph and trifid for being able to make SPARQL queries. So I had to change the configuration of blazegraph by specifying the SPARQL endpoint URL which in my case was:

*<http://155.54.239.183:3041/blazegraph/namespace/Alzheimers/sparql>.*

Then, I loaded the metadata and the dataset in an n-Quads format which is a linebased RDF syntax with a similar flavor as N-triplets to the repository. I just had to convert my RDF file into Ntriplets, and then I added the ontology iri using a bash script. Finally, we can run trifid to prove that the dataset is published.

zazukoTrifid

<http://dayhoff.inf.um.es:8181/paco/ontologies/2022/4/Alzheimers#ALDH>

<http://dayhoff.inf.um.es:8181/paco/ontologies/2022/4/Alzheimers>

type	Dataset
type	Ontology
License	<a href="http://creativecommons.org/licenses/MIT/">http://creativecommons.org/licenses/MIT/</a>
label	Alzheimer Disease
distribution	sparql
namedGraph	graph
primaryTopic	Common pathways of Alzheimer's

[json-ld](#) | [turtle](#) | [n3](#) [Back to top](#)


 zazuko

Figure 7: Trifid

## 3 Conclusions

This practical assignment has clarified the way to manipulate data to create ontologies or datasets related to biological processes based on existent ontologies such as Uniprot or Gene Ontology. More important, I have learned to make queries to make a better use of these ontologies. Finally, I found very interesting the FAIR principles which allow us to describe how research outputs should be organized so they can be more easily accessed, understood, exchanged, and reused.

However, I found the last part of the assignment particularly hard to do. I found myself lost quite often considering that no one in our class this year came from a computer science background we had to work extremely hard to come off with a solution. I hope it can be better explained next year.

## References

- [1] Marco Calabrò et al. “The biological pathways of Alzheimer disease: a review”. In: *AIMS Neuroscience* 8.1 (Dec. 2020), pp. 86–132. ISSN: 2373-8006. DOI: [10.3934/Neuroscience.2021005](https://doi.org/10.3934/Neuroscience.2021005). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7815481/> (visited on 05/23/2022).
- [2] Hugo YK Lam et al. “AlzPharm: integration of neurodegeneration data using RDF”. In: *BMC Bioinformatics* 8.3 (May 2007), S4. ISSN: 1471-2105. DOI: [10.1186/1471-2105-8-S3-S4](https://doi.org/10.1186/1471-2105-8-S3-S4). URL: <https://doi.org/10.1186/1471-2105-8-S3-S4> (visited on 05/23/2022).