A graph-based taxonomy RESTful data service

Leyla Garcia[1*], Leonardo Gonzales[1*], Maria Martin[1]

[1] Protein Function Development Group, EMBL-EBI, Wellcome Genome Campus, CB10 1 SD, Hinxton, UK

* These authors have equally contributed to this work

{ljgarcia, lgonzales, martin}@ebi.ac.uk

Abstract. In Life Sciences, taxonomy refers to a standard nomenclature and classification system for living organisms. In early days, organisms were clustered together mainly based on shared morphological characteristics. Nowadays, evolutionary patterns linking organisms have become more relevant when classifying organisms. Applications and tools serving taxonomical data aim to facilitate navigation. Here we present a graph-based RESTful data service built on top of Neo4J. Our service supports common queries as well as advanced queries including common ancestor between two or more taxonomic identifiers, a path to ancestors or descendants within a specified level, and the shortest path between two taxa. Availability: http://www.ebi.ac.uk/proteins/api/doc/index.html

Keywords: Graph-based databases, taxonomy, web data services.

1 Background

Taxonomical data services aim to facilitate navigation through the hierarchy. They commonly support queries based on taxon identifier as well as scientific and common names. Responses include taxon identifier, scientific and common names, taxonomical rank, e.g., genus, order or species, and lineage. For instance, the National Center for Biotechnology Information (NCBI) provides taxonomy data services via e-utils tools, supporting queries based on one or multiple taxonomical identifiers as well as scientific and common names. Both Extended Mark-up Language (XML) and JavaScript Object Notation (JSON) formats are supported as well as File Transfer Protocol (FTP) downloads and end-user interfaces.

The European Nucleotide Archive (ENA) provides a JSON-only taxonomy service, supporting queries by taxon identifier and name. The name-based query can be user to retrieve a list of suggested taxa based on the first characters. The Universal Protein Resource (UniProt) serves taxonomical data as part of its Resource Description Framework (RDF) distribution, accessible via its SPARQL Protocol and RDF Query Language (SPARQL) endpoint. SPARQL flexibility together with SPARQL property paths make it possible to create queries such as getting a path forwards or backwards from a starting node. Both ENA and UniProt support FTP downloads and end-user interfaces.

Desirable features when traversing a taxonomical tree such as finding the shortest path from one taxa to another or the common ancestor for multiple taxa are not commonly supported by traditional taxonomical data services. Here we present a graph-based taxonomy RESTful data service supporting not only common queries but also those related to taxonomical paths.

2 Our Neo4J-based RESTful taxonomy Service

Here we present a graph-based RESTful data service built on top of Neo4J. Our service supports common queries –those related to taxonomic identifiers and scientific or common names, as well as advanced queries. In addition to the basic data, i.e., taxon identifier, scientific names, common names and lineage, we also provide links to the related taxa. Advanced queries include the common ancestor between two or more taxonomic identifiers, a path to ancestors or descendants within a specified level, and the shortest path between two taxa. Such advanced queries might also be accomplished by using SPARQL; however, they can turn out to be complex and long queries. In our service, complexity is hidden to service users. Additionally, stress and load tests have been performed on our service to ensure its reliability and robustness. Both XML and JSON responses are supported. The full documentation is available at http://www.ebi.ac.uk/proteins/api/doc/swagger/index.html?moduleId=taxonomyApi. Some query samples are provided here:

- Full data by taxon identifier, e.g., http://www.ebi.ac.uk/proteins/api/taxonomy/id/9606
- Partial data by taxon identifier, i.e., basic information, children, parent or siblings, e.g., http://www.ebi.ac.uk/proteins/api/taxonomy/id/9606/node
- Lineage by taxon identifier, e.g., http://www.ebi.ac.uk/proteins/api/taxonomy/lineage/9606
- Full data by scientific, common and mnemonic names, e.g., http://www.ebi.ac.uk/proteins/api/taxonomy/name/homo?searchType=STARTSWITH&fieldName=SCIENTIFICNAME, useful for autocomplete
- All taxa related with the queried taxon identifier in a specific direction and depth level, e.g., http://www.ebi.ac.uk/proteins/api/taxonomy/path?id=8782&depth=3&direction=BOTTOM
- Path between two taxa showing their relationship, e.g., http://www.ebi.ac.uk/proteins/api/taxonomy/relationship?from=8782&to=9606
- Lowest common ancestor of two or more taxa, e.g., http://www.ebi.ac.uk/proteins/api/taxonomy/ancestor/8782,9606

Our graph-based RESTful taxonomy data service is a modern and robust service supporting traditional queries as well as some new features that make it easier to find relationships across taxa. This service is part of the set of protein data web services delivered by the Protein Function Development Group at the European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI). We are exploring the creation of a library of web components using and showcasing our web services.