Linked Data and Life Sciences

Ismael Navas-Delgado, Maria del Mar Roldán-García, José F. Aldana-Montes

A living organism is an open system that keeps a continuous interchange of chemical compounds, energy and information with its environment. This interchange involves a high number of elements (molecules) related among them in a dynamic hierarchical and modular way. The development of new techniques has enabled the discovery of the information related with the elements. As a result the number of biological data has an exponential growth that makes its analysis unfeasible.

Currently, the publication of these data as Linked Data has enable the access to this growing information resource for opening new scientific studies. In this sense, there are many Linked Data repositories for biological data with a high degree of linking. These resources make use of different vocabularies, sharing in many cases the use of core ontologies such as Gene Ontology, which provides a large thesaurus that can be used to annotate and classify available data. In parallel, most important databases are starting to publish part of their data, but in those cases more specialized ontologies are use to describe the data.

Thus, the Life Sciences is a use case that needs solutions for most of the key problems in Linked Data community:

- Computation, storage, and management aspects for large sets of data (billions of triples);
- The use of rich semantics for taking advantage of these data requires inference algorithms able to deal with ABox and TBox reasoning at large scale;
- The available information required to get more holistic views of Molecular Biology problems increases daily, due to a large amount of data provided by Genome Projects and high-throughput technologies. The integration of the available data requires to facilitate the discovery of new links for the flood of data coming from high throughput experiments;
- This scenario requires the development of specialized applications that must hide the complexity of the data discovery, access and analysis;
- Another worth mentioning disadvantage of the present trends and tools is the continuous overlapping of new information strata that frequently lead to cover up the previous information.