



HAL
open science

AskOmics, a web tool to integrate and query biological data using semantic web technologies

Xavier Garnier, Anthony Bretaudeau, Olivier Filangi, Fabrice Legeai, Anne Siegel, Olivier Dameron

► To cite this version:

Xavier Garnier, Anthony Bretaudeau, Olivier Filangi, Fabrice Legeai, Anne Siegel, et al.. AskOmics, a web tool to integrate and query biological data using semantic web technologies. JOBIM 2017 - Journées Ouvertes en Biologie, Informatique et Mathématiques, Jul 2017, Lille, France. pp.1. hal-01577425

HAL Id: hal-01577425

<https://inria.hal.science/hal-01577425v1>

Submitted on 25 Aug 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

AskOmics, a web tool to integrate and query biological data using semantic web technologies

Xavier GARNIER^{1,2}, Anthony BRETAUDEAU^{1,3}, Olivier FILANGI^{1,3}, Fabrice LEGEAI^{1,4}, Anne SIEGEL² and Olivier DAMERON²

¹ Institut de Génétique, Environnement et Protection des Plantes (IGEPP) - Institut national de la recherche agronomique (INRA) : UMR1349, Agrocampus Ouest - Agrocampus Ouest, UMR1349 IGEPP, F-35 042 Rennes, France

² DYLISS (INRIA - IRISA) - INRIA, Université de Rennes 1, CNRS : UMR6074 - Campus de Beaulieu, F-35 042 Rennes Cedex, France

³ Plateforme bioinformatique GenOuest - Université de Rennes 1, Biogenouest - France

⁴ GENSCALE (INRIA - IRISA) - Université de Rennes 1, CNRS : UMR6074, INRIA - Campus de Beaulieu, F-35 042 Rennes Cedex, France

Corresponding author: xavier.garnier@irisa.fr

Research programs involving genetics, genomics and epigenetics are quickly growing. Lot of experiments producing large amount of data are now feasible in the frame of a laboratory. As well, the tools analyzing the data generated by these experiments are often available. Results of these experiments can be stored and structured into files and loaded into databases, and lots of these databases are public. However, these bases are usually implemented according to schemes and techniques that do not allow their interoperability in an easy manner.

The technologies from the Semantic Web, especially RDF and SPARQL are one of the key elements for combining databases, which has led to the emergence of linked data. It is based on triples (subject, predicate and objects) describing the relationships between elements stored into interoperable triplestores allowing distributed querying. Because of its flexibility, versatility and ontology-awareness, numerous biological databases, such as UniProt, PubChem or ChEBI and Reactome at EBI, give access to their data via a SPARQL endpoint.

We present AskOmics, a new software, that uses the Semantic web technologies, which helps to integrate multiple format of data and query them through a user-friendly interface. AskOmics is a free and open-source software (AGPL licence) available on GitHub (<https://github.com/askomics/askomics>).

AskOmics supports both intuitive data integration and querying while shielding a non-expert user from most of the technical difficulties underlying the web semantic technologies. Because large and heterogeneous biological datasets are often difficult to integrate, AskOmics users can provide simple tabulation-separated files (TSV), that are transformed automatically into RDF triples, then stored into a triplestore. Finally, for data querying, AskOmics provides a visually intuitive interface to obtain a comprehensive view of the biological study.

During data integration, user provides input files in common formats (currently TSV and GFF) to be converted into RDF triples. AskOmics generates triples corresponding to the data (the content), and also triples which describe the data (the abstraction). Triples are loaded in a triplestore in order to persist data and optimize queries.

The query interface is composed of a dynamic graph at the left and a right view for filtering attributes. On the graph, each node represents an entity. Entities are linked between them with arrows. Attributes of the selected entities are displayed on the right view. To build the graph, AskOmics query the abstraction. Users build their queries by starting from a node of interest and sequentially select its neighbors and filter on attributes, creating a path on the abstraction. This path is converted into a SPARQL query and sent to the triplestore. Finally, the results are displayed as a table and can be downloaded as a TSV file.

AskOmics has been applied successfully to the analysis of large scale datasets on the aphid embryogenesis, on the variability of Brassicaceae in response to clubroot disease, and to the analysis of biological pathways.