BioWes – From Design of Experiment, through Protocol to Repository, Control, Standardization and Back-Tracking

Antonín Bárta, Petr Císař, Dmytro Soloviov, Pavel Souček, Dalibor Štys, Štěpán Papáček, Aliaksandr Pautsina, Renata Rychtáriková, and Jan Urban

Institute of Complex Systems, South Bohemian Research Center of Aquaculture and Biodiversity of Hydrocenoses, Faculty of Fisheries and Protection of Waters, University of South Bohemia in České Budějovice, Zámek 136, Nové Hrady, 373 33, Czech Republic
abarta@frov.jcu.cz

Abstract. The amount of data produced by current experiments in systems biology is enormous. Some database and software support for biology experiments exist but usually deal just with some part of data and metadata management. Primary data are only occasionally analyzed in-depth and shared. Studies showed that cost of data sharing is cheaper than experimental work. Experimental costs are several orders higher than data. Therefore, being up to date is a worldwide problem for all users of biology applications. In practice, the problem extends to general fields such as knowledge mining, experiment quality and repeatability, and the philosophical epistemology of biological problems.

The BioWes project is inspired by several similar projects that try to solve a substantial contemporary problem of sharing big amount of experimental data. There are several projects that offer the solution for data sharing (for different types of data). The problem is that the amount data produced by experimentalists is constantly increasing and the speed of internet will always be a step behind. The effective and easier way how to share experimental data between researchers is to share metadata. Metadata means the overall knowledge about the experiment that consist of complex information of experimental procedure and knowledge that can be extracted from data automatically or manually by post-processing. The data itself is meaningless without any additional knowledge concerning the experiment. There is no project that can offer the whole concept of experimental data sharing and data processing based on the sharing of knowledge.

Keywords: Database, Repository, Metadata, Data management, Protocols, Experiment setup, Design of experiment.

1 Introduction

The project BioWes is inspired by several similar projects that try to solve a substantial contemporary problem of sharing big amount of experimental data. There are several projects that offer the solution for data sharing (for different types of data).

F. Ortuño and I. Rojas (Eds.): IWBBIO 2015, Part II, LNCS 9044, pp. 426–430, 2015.

The problem is that the amount of data produced by experimentalist is constantly increasing and the speed of internet will always be a step behind [1]. The effective and easier way how to share experimental data between researchers is to share the metadata. Metadata means the overall knowledge about the experiment that consist of complex information of experimental procedure and knowledge that can be extracted from data automatically or manually by post-processing. Data itself is meaningless without any additional knowledge concerning the experiment. There is no project that can offer the whole concept of experimental data sharing and data processing based on the sharing of knowledge (see Fig.1).

The main reason of sharing metadata data is to save money and time necessary for experimentation and to compare the results between different experimenters. Data sharing and especially metadata sharing can be understood as the advertisement of the experiments of a particular experimenter. Experimental data sharing and comparison can help to improve experimental procedures and defining of standards in this area [1, 2].

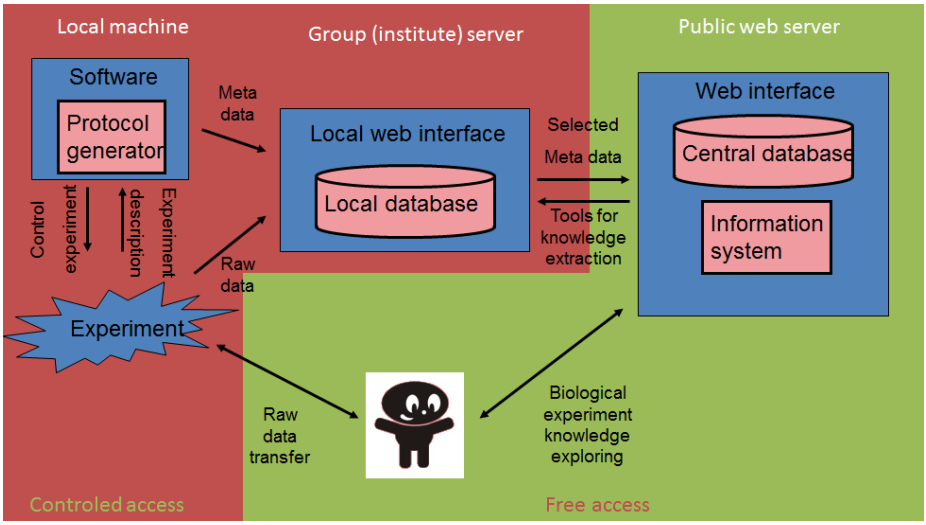

Fig. 1. Scheme of the sharing and usage of data and metadata database on the example of Bio-Wes system

2 Protocol Generator

Scheme of the usage of BioWes system are shown on Fig. 1. Software interface, which is in the direct contact with a user (experimenter) is called Protocol Generator. It is a standalone application that should ensure the repeatability and correctness of the biological experiments. The tool is designed to lead the experimenter through the particular type of experiment as a supervisor and to help him. Protocol generator has two purposes: the first one is to check that the procedure of the experiment has been

done precisely and the second one is to produce all important settings that are part of the experiment in the form of report on the experiment. The method to ensure precise realization of experiment is to check if all the necessary parameters and steps of the experiment have been set and done. The list of necessary parameters and steps for the particular type of experiment comes up from the analysis of biological experiments from different research institutes.

The description of the experiment can be created by the user for specific experiment. Graphical user interface Protocol designer has been implemented for this purpose. The protocol template can be created by any BioWes user who can define all the important conditions of the experiment. The user can use 10 basic components for definition of the protocol. The protocol of experiment can be shared among the people who realize the experiment instead of students to ensure the repeatability of the experiment.

The template can be later modified for new experiment to speed up the process. Main advantage of the electronic protocol is that there is a direct link between the protocol and experimental data. Both are stored in the central database and can be used for obtaining future data.

Protocol generator supports external plugins for mining information about setting of devices from external files. The plugins can read the information about some parameters of experiment form files produced by measurement device (magnification of microscope) and fill it into the protocol. Plugins are using open interface and therefore it can be created by the users for specific devices.

3 Local Database

The main purpose of local data management tool is to organize and store the raw experimental data directly on the site of the institution (experimenter). The local data management tool provide the functionality of data storing, searching, filtration and reporting. The tool is connected to the Protocol generator to support the reporting of experiments on the higher level of metadata. Local data management is realized as a specialized database that will be optimized to the type of experimental data produced by particular institution. The database with uniform interface is modified according to the needs of the particular experimenter to reach the aims of different experiments. The local management tool provide the communication module (interface) to global data management tool (web based data sharing). The global data management tool is used for metadata sharing between different institutions and the public. The process of communication between local and global data management tools will be under full control of the institution. Therefore only metadata can be shared with the rest of scientific community or the public. The advantage of this approach is the direct control of „what I am sharing with the others“. [3]

4 Sharing and Management

The central data storage is realized as combination of local data storage (located at the institution) for raw data and one central data storage selected metadata. The data structures in central database will be defined generally to cover all the different metadata types and to upgrade the structures in the future. [4]

The central data storage serve as the first option for searching the experimental data through metadata and allow the user to find the proper experiments and results. All the metadata are available using the XLM data structures exchange.The BioWes experimental data and metadata management allows:– fast information search – the usage of data structures indexing and advanced algorithms of query execution plans minimize the time response of data storage – standardized system control – the commercial database ensure the safe and secured operation of the database.

The central database will be in cooperation with the information system. The information system provide the parameterization of the central storage, user accounts control and policy. Data from central database are presented using the web presentations to the different kind of the scientific and general audience controlled by the access rights.

The interface is used as an interpreter between central data storage, local data storage and visualization framework.

The user friendliness of the central database is supported by the visualization framework. The visualization framework will be implemented as several software modules and interface for extensions of information system. The visualization framework allow simple and intelligible visualization and comparison of the metadata and results of searches. It is based on the mix of existing modules and third party modules. The third party modules could be plugged into the central data storage for the user of the system. The modules are focused on raw data processing, data mining and aggregation of the metadata. Standard interface of the central data storage will allow the user to upload the extracted information back into the central storage, describe it and integrate it into current structures.

One part of the web solution will be the offer of tools for raw data post-processing. These tools are highly specialized for experimental data processing. They can be used by anyone to produce metadata from raw experimental data and share metadata using BioWes web solution. The list of the tools can be extended by any third party tool for biological data processing. Several tools we are working on directly under the project:

– Cell time lapse image processing and representation
– LC-MS measurements filtration and analysis
– Software for behavior analysis of aquatic organisms

Acknowledgement. The study was financially supported by TACR project TA01010214 BioWes, by the Ministry of Education, Youth and Sports of the Czech Republic - projects 'CENAKVA' (No. CZ.1.05/2.1.00/01.0024), 'CENAKVA II' (No. LO1205 under the NPU I program); and by the Postdok JU (CZ. 1.07/2.3.00/30.0006).

References

[1] Haug, K., et al.: MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. NAR, gks1004 (2012)
[2] Freire, J., Bonnet, P., Shasha, D.: Computational reproducibility: State-of-the-art, challenges, and database research opportunities. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. ACM (2012)
[3] Mayer-Schönberger, V., Cukier, K.: Big data: A revolution that will transform how we live, work, and think. HMH (2013)
[4] Borgman, C.L.: The conundrum of sharing research data. JASIST, 1059–1078 (2012)