Bioenergy Metabolomics DB

Marcelo Soares Souza[1], Vanessa de A. Soares[2], Lucas Soares de Brito[3], Patrícia Verardi Abdelnur[4], Eduardo Fernandes Formighieri[5]

Abstract

In the last years, with the evolution of the mass spectrometry, the metabolites databases passed through a large information increase. This growth generated a demand for computational systems to store this information, which requires several additional tools, instances and a well-prepared infrastructure (e.g., servers, storages and hardware maintenance). The objective of this work is to present the current stage of the Bioenergy Metabolomics DB system development, and its main planned functionalities. To build this system we used free software and tools, like the Python language, Django framework, PostgreSQL database and the Git version control. The adopted development method was based on Scrum, with the adoption of sprints and frequent meetings. The hardware and software infrastructure is installed, configured and in production. The "Base system", called GBP, which comprises the main functionalities of the system, is already functional and almost complete, and we are working into the Metabolomics related modules.

Introduction

Metabolites are small molecules chemically transformed during metabolism and, in this way, provide a functional reading of the cellular state. The set of metabolites of an organism is called metabolome (BAKER, 2011). With the mass spectrometry evolution (e.g. instrumentation, software products and bioinformatics), it is possible to measure thousands of metabolites simultaneously from minimal sample amounts, resulting in large information increase to the metabolites databases in the last years (PATTI et al., 2012).

[1] Informático, graduado em Informática pela Universidade Católica do Salvador, marcelo@libertais.org
[2] Engenheira de Software, graduada em Engenharia de Software, Universidade de Brasília, vanessa.soares@colaborador.embrapa.br
[3] Engenheiro de Redes de Comunicações, mestre em Engenharia de Sistemas Eletrônicos e de Automação, Universidade de Brasília, lucas.brito@colaborador.embrapa.br
[4] Química, doutora em Química Orgânica, pesquisadora da Embrapa Agroenergia, patricia.abdelnur@embrapa.br
[5] Engenheiro-agrônomo, doutor em Biologia Funcional e Molecular, pesquisador da Embrapa Agroenergia, eduardo.formighieri@embrapa.br

Through the metabolomics, it has been identified alternative paths in metabolic pathways, in many cases related to specific phenotypes. Despite the technology great advances, the results are still lists of molecular formulas, being necessary a contextual analysis to define the most likely compounds in each case. The complexity and the large amount of data generated, added to the large amount of resources invested, such as equipment and protocols development, have generated demand for the development of computational systems, to enable and optimize the use of data. This system must to store the information in a safe, centralized and organized way, allowing easy access to the information and reports, with specific access control for the different types of information.

The system development involves several tools and instances, including: Data Base Management System (DBMS), for an organized, safe and quick information access storage; Programming language, base for the system development; Framework, to speed up the development and facilitate its maintenance; and documentation and development best practices, aiming at both the development process improvement and the future maintenance possibility, increasing the system lifetime.

Additionally, an appropriate infrastructure is necessary for the system availability, including server racks with real and virtual servers, backup system, access control, storage and constant hardware and software monitoring and maintenance.

This work presents the current development state of *Bioenergy Metabolomics DB*, that aims to store metabolomics analyses raw data in a safe and organized manner, allow advanced searches, reports generation and, further, external access to selected data. This work was developed by Bioinformatics Research Group at the Bioinformatics and Bioenergy Laboratory – LBB (lbb.cnpae.embrapa.br), in collaboration with the Metabolomics Research Group ("client"). All used software products are free to use and run in Linux OS.

Material and methods

Material

Only Free or Open Code software have been selected. All of them are consolidated, with large amount and variety of documentation and with good global growth prospects of continuity.

It was chosen the multi-paradigm programming language Python (https://www.python.org), widely used for systems development for the Web and with growing adoption by the Bioinformatics area researchers and professionals. To enhance the results, facilitate the collaborative development and increase the system life cycle, it was adopted the Django framework (https://www.djangoproject.com), a library that provides a set of functionalities commonly used for the web platform development, based on the model-template-view (MTV) standard.

For the data storage it was adopted the Object-Relational Data Base Management System (ORDBMS) PostgreSQL (https://www.postgresql.org), a robust and scalable solution for data storage, handling and availability for computational systems, and that uses the research language called Structured Query Language (SQL). PostgreSQL also can be used by other bioinformatics tools, allowing an easy data interchange.

All the development process is being done using the "best practices" of code development and maintenance, assisted by the Git version control tool (https://git-scm.com/) and supported in the GitLab platform (https://about.gitlab.com), which provides a simpler interface, allowing a better monitoring of the activities as well as tracking changes made throughout the development process.

Methods

Several meetings were made for the knowledge of the features needed so that the system meets the demand the best way possible (requirements gathering). From these information, it was made an initial system planning, needed for tooling establishment and suitable methods.

The architecture chosen is based on the MTV model, base of the Django framework. This model consists on the separation of the system in related layers, providing a better logical separation of the system subparts and a better maintainability, making it safer and more effective.

The MTV standard divides the platform into three layers, namely: i) *Template*, directly accountable for the interaction with the user; ii) *Model*, responsible for the persistence (storage) of the data in the Data Base and for the setting of business rules for the data processing; and iii) *View*, responsible to intermediate the relation between the interaction with the user layer (through the Templates) and the data.

The data modeling is intrinsically related to the business rules modeling in the Model layer, and it has not been made a traditional modeling based on Model Entity Relationship.

It was adopted a method of fast development based partially on the Scrum model (http://scrummethodology.com), an iterative and incremental development framework used in the project management mainly in software development. It was defined a general scope through *Brainstorming* meetings, where it was defined the essential requisites and more important business rules. It is held weekly meetings of development monitoring of the tasks in LBB, in which is presented and discussed what was made and the next actions are defined (*sprints* – www.scrumguides.org/scrum-guide.html). These tasks are constantly revised in order to ensure the most possible proper delivery to the updated client demand.

The development system is composed by a base and by subsystems. The base is the GBP, a registration, edition and visualization system of the Projects data made at LBB. In GBP are registered, for example, the Users (Team), Project general Data, Specific Objectives, Expected Results, Components, Plan of Action, Activities and Tasks. This detailing level is necessary to allow detailed access control based on the projects information, and is being developed in a flexible and comprehensive way in order to accommodate different specific modules to be developed at LBB.

The subsystem Bioenergy Metabolomics DB was planned to be an application of support to activities of Metabolomics, which extends the functionalities of the GBP and offers an interface for the management of experiments (storing results), equipment, methods and techniques used, providing a database that will be the basis both for the external availability of selected information and for the creation of a decision support System (definition of metabolites through the molecular formulas), through the last planned module.

Results and discussion

In terms of hardware and software infrastructure, the servers operational systems (Ubuntu - www.ubuntu.com and Debian – https://www.debian.org) were updated in the real servers (last Long Term Stable version), virtual machines were installed and configured on KVM (www.linux-kvm.org), including, among others, one for backup (Bacula – www.bacula.org), one for PostgreSQL DBMS and one for infrastructure management tool (Zabbix – www.zabbix.com). All the needed tools to the development were installed, configured and tested.

GBP Base

After completion of the infrastructure, the next step was the development of GBP (Projects Basic Management), which first version is nearing completion, already including the registration of projects and of all correlated data. To finished it, we are working into navigability and documentation aspects, but it already has a stable basis to enable the development of the subsystem *Bioenergy Metabolomics DB*. The Figure 1 shows one of the interfaces of GBP.

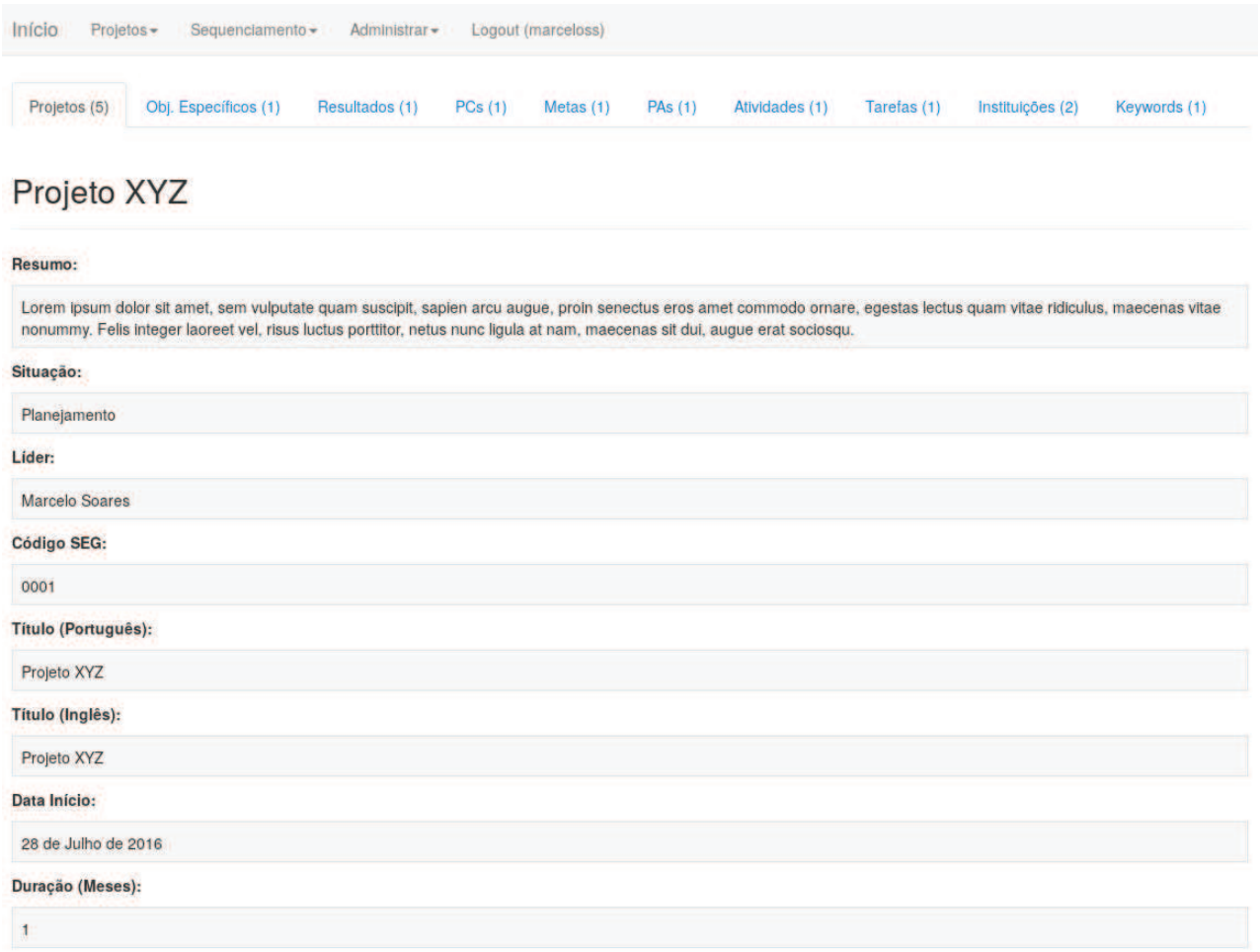

Figure 1. GBP beta interface example – top of the project general data CRUD (Create, Read, Update, Delete) page.

The Figure 2 presents the initial development of a visualization/navigation interface, as a simple way to find data to posterior detailed visualization.

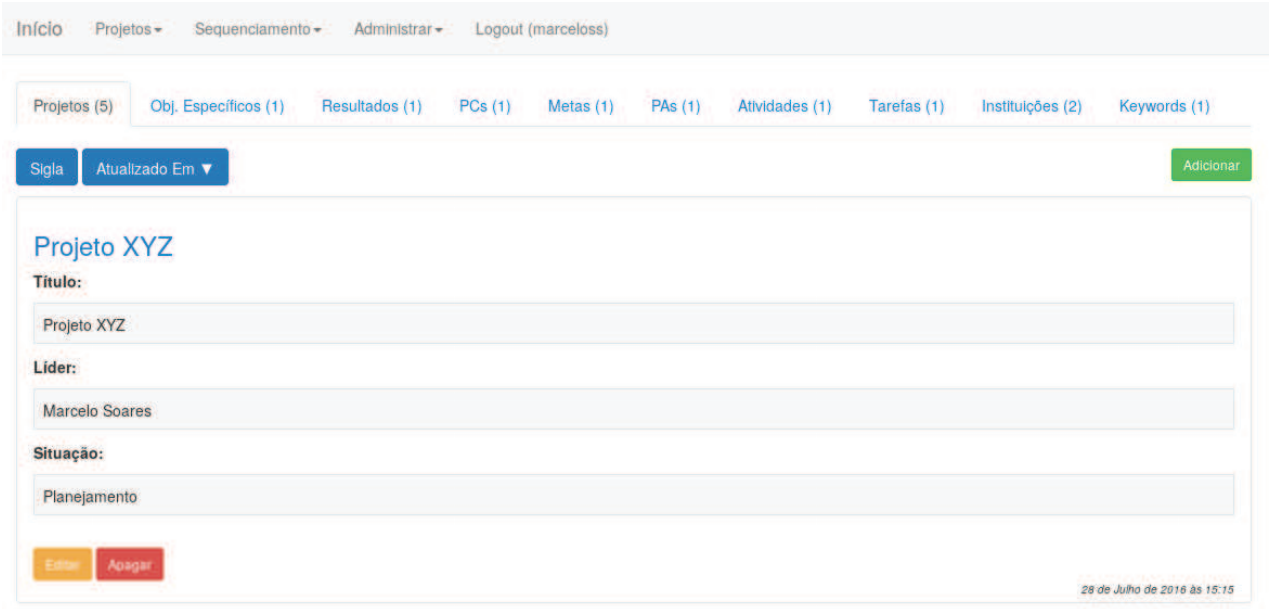

Figure 2. GBP beta interface example – top of the projects main information visualization page.

Bioenergy metabolomics DB subsystem

The *Bioenergy Metabolomics DB* is in the final stages of requisites analysis, with face-to-face meetings and defined scope for the first module, which will include the equipment and processes management to be made by the researchers, and the subsystem specific authorization profiles.

Initial module (internal)

Seeks an organized storage of the pre-existing metabolomics results of the client, including advanced searches, download of raw and compiled data files, and organized results presentation. It is under advanced phase of planning, and will be a quick and simple solution for initial organization of the present results.

Raw-data DB module (internal)

This module aims at the complete organization of the client metabolomics analyses, including registration and edition of all the equipment, types of analyses and its parameters. Also for all the results (linked to people, projects, biological material, equipment, analysis, parameters etc). It is in advanced stage of requirements gathering and planning.

Metabolomics DB module (external)

It seeks the external and controlled availability of selected information. It will be the public interface for research relevant results availability, in principle by previous registration. In phase of requisites gathering and initial planning.

Decision support module (internal)

Aims to help the decision-making on the metabolites setting from the chemical formulas generated by the metabolomics analyses. In stage of requirements gathering, it will be better evaluated in the final development phase.

Conclusions

These methods showed to be essential to a safer and more efficient evolution of the development, and are indispensable to reduce the inherent risks to any web based system, like unavailability and data loss. The chosen tooling showed to be suitable and efficient, and will be used in the next subsystems to be developed at LBB.

The accurate definition of the demand is an essential stage, and the meetings for demand definition (project and development scopes) generated an additional learning for both parts involved, and became gradually more efficient.

Financial support

This study was supported by the Embrapa (Project: Functional genomics, transcriptomics and metabolomics, of xylose-fermenting yeast for increasing efficiency in the production of second generation ethanol - YEASTOMICS. SEG: 02.12.01.006.00.00).

References

BAKER, M. Metabolomics: from small molecules to big ideas. **Nature Methods**, London, v. 8, n. 2, p. 117-121, 2011.

PATTI, G. J.; YANES, O.; SIUZDAK, G. Metabolomics: the apogee of the omics trilogy. **Nature Reviews Molecular Cell Biology**, London, v. 3, n. 4, p. 263-269, 2012.