System architecture for data acquisition, extraction and analysis for experiments with weblabs

Marcelo Queiroz Leite[1] ,Leandro Halle Najm[1] ,Pedro Luiz Pizzigatti Corrêa[1], Ayrton Vollet Neto[2], Vera Lucia Imperatriz Fonseca[3]

[1]Electrical Engineering – Digital Systems – Engineering School of USP
[2]Biological Science – FFCLRP-USP
[3]Biology Institute – IB-USP

{marcelo.queiroz, leandro.najm, pedro.correa}@poli.usp.br, ayrtonvollet@gmail.com, vlifonse@ib.usp.br

ABSTRACT

The concern about understanding the effects of climate change on the environment, such as the decline of pollinators found in nature, the need to support researchers to run experiments efficiently and to share those results, makes it necessary to study and develop virtual laboratories through the web, in order to gain greater insight on the behavior of bees, which is the main pollinator of plants. The hardware structure and software architecture for the implementation of virtual labs, also known as weblabs are presented. In addition to the information technology involved, sensor networks were used for data acquisition, and tools for extraction and analysis of data collected transforming them into information that can be easily used by researchers.

KEYWORDS: weblab, bees, lonworks, wireless sensor networks, Kepler.

1 Introduction

The importance of environmental services, such as air purification, stabilization of climate and pollination of crops, is increasingly evident [1]. The need for research in these areas has increased, generating complex experiments involving a large volume of data, such as the bee laboratory on the web.

The application of new information technologies is necessary to understand the effect of climate change on the environment. One way to help researches to obtain results faster is the development of laboratories available via the Internet. The main idea of these laboratories, called weblabs is to allow: the implementation, monitoring and analysis of experiments without the presence of a researcher in the laboratory, extending its use for everyone with access to a computer connected to the Internet.

Performing experiments in laboratories is essential in many activities; however, the costs of acquiring and assembling the physical structure are sometimes an inhibiting factor for the creation of research laboratories, in which the solutions do not advance at the same speed those problems. Another disadvantage is the limitation of access to the laboratory due to specific operation times and the need for physical space [1].

The virtualization of laboratories, with simple and objective web interfaces will allow researchers to conduct experiments using appropriate equipment and services, without the need to maintain their own facility. In these types of environments, users can control the equipment, schedule actions, enter data, share findings and interact with other researchers after the end of the experiment.

With the physical infrastructure installed and configured correctly, it is necessary to configure the software for data collection and analysis tools to process that data collected by transforming them into information for the researchers to use.

The steps to collect, filter, integrate, analyze data and generate results in a research using weblab technologies and specific tools for biodiversity analysis are described here. To demonstrate this scenario, a reference architecture is used to integrate the data from information collection to analysis of these data and generation of results.

2 Objective

A proposal of system architecture for extracting data acquired by sensor networks is presented and discussed, aiming to carry out experiments with weblabs. This proposal led to trial implementation for the validation of resources used, the instrumentation – the sensors, and other components so that they could be used by researchers in a reliable and simplified way.

Moreover, we sought to define the mechanisms for data acquisition and types of storage, extraction of information in databases in order to provide information quality in weblab, which was analyzed by means of specific tools. The equipment of interest is defined, and a survey is conducted on services needed for an experiment to be controlled via the Internet, including startup, monitoring, analysis and sharing of results. So it is proposed a Web laboratory for researchers from anywhere with Internet access that can perform various experiments in real time and with real results. The experiment was programmed based on the steps showed in Figure 1:

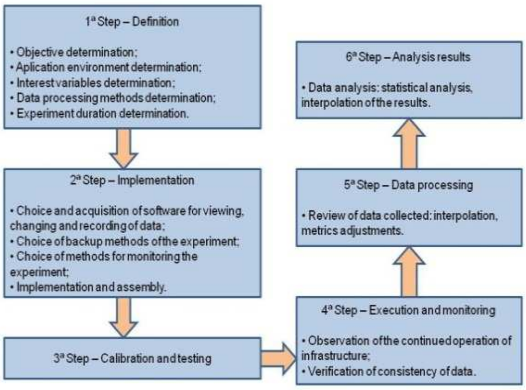

Figure 1. Steps of experiment realization

3 Methods and Materials

This section discusses the technologies used in the experiment, describing how individual characteristics from them, which together have enabled the implementation of these activities.

Kyatera is the first optics platform of education and research in the Southern Hemisphere. It is an Internet 2.0 network, very high speed, dedicated exclusively to scientific and technological research.

The KyaTera – Tupi (ancient Indian tribe of Brazil) kya (fishing net) and the Greek tera (giant) - is already connecting six laboratories of Campinas and Sao Paulo, unifying research at UNICAMP, PUC-Campinas, at USP, and the Foundation contributing authors. But the first phase of the project provides for the interconnection of 50 laboratories.

The Kyatera project provides the infrastructure for fiber-optic network that includes metropolitan area networks (phone) and the Advanced Internet links between the various centers of teaching and research. The infrastructure of the experiment can be observed in Figure 2:

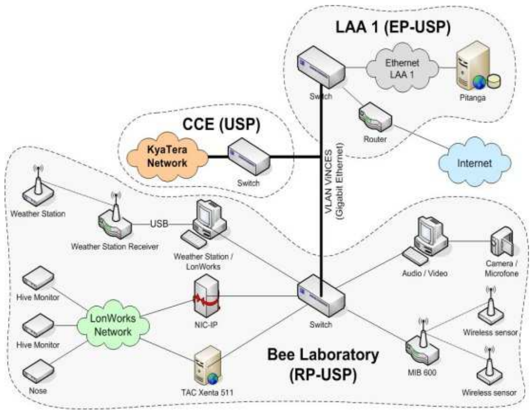

Figure 2. VINCES network architecture

The sensor structure are installed in a beehive of stingless bees specie of *Scaptotrigona depilis* (*Hymenoptera, Apidae, Meliponini*) at Ribeirão Preto, São Paulo state Brazil, in a Bee Laboratory of Biology department of USP (University of São Paulo), integraded by Vinces network with LAA (Laboratório de Automação Agrícola) in Polytecnical School of USP.

3.1 LonWorks

Lonworks is a standard network protocol for the performance and reliability of control applications. The platform is built on a protocol for low bandwidth created by Echelon Corporation to run on twisted pair, fiber optics, data transmission over the grid and radio frequency [2]. The Figure 3 shows the lonworks and wireless sensors, installed in the beehive.

Figure 3. Sensors in beehive

3.2 Wireless Sensor Networks

Device networks capable of sense the environment, execute local processing and wireless communication. The wireless sensors used are form the company Crossbow [3] MicaZ family, where after spread on the environment automatically communicate to the network gateway, model MIB600, which communicates to the TCP / IP, allowing the collected data to be accessed by a computer connected to a network. For the connection of sensors to the TCP / IP was used MIB600 that acts as a gateway.

3.3 Weblabs

Laboratories to perform experiments are essential in many science activities. Among the resources used by researchers, weblab has become a tool of assistance to researchers with a good acceptance from the scientific field, because through this technology, experiments can be shared and evaluated by various scientists, since they are allowed to join the team responsible for running the experiment. A weblab laboratory or remote access is an integrated hardware and software that allows control and monitoring of a programmable electronic device via the Web.

3.4 Bioabelha

The project called Bioabelha is a weblab for study of pollinators. At this weblab are included: images data with high definition and videos for taxonomic study and bee behavior; real time weblab monitoring of beehives, tools and other materials for distance learning, that can be saw in Figure 4.

The data with patterns defined of videos and images in high definition of bees was development for extend the data base available in site WebBee [www.webbee.org.br]. The WebBee is a system of information on pollinators, more specifically on brazilian native bees and was created to organize and centralize your specific information [4]. Its main objective is to be used in researches and study about the biology of bees, the behavior and systematic. Your original database it restricted to images of low definition, since supposed was be accessed a variety of users with several types of internet connections.

Figure 4. Weblab Bioabelha

3.5 Metadata for biodiversity

Metadata is data that fully describes the data represented, allowing the user to decide on the use of such data as best as possible [5]. To facilitate this work, the use of metadata is recommended, because it is the data about the data itself, or reports of significance of the data contained. The data itself lose their meaning without a context, or an unknown one, a value '20' without context is just a number, when it is introduced a context of time, the meaning is reduce to seconds, minutes, years, centuries,etc., yet there is still a gap to be filled. With the use of metadata, the context is known, stating that '20' is actually the age in years of something, for example. In Figure 5, it is showed some examples of metadata in daily life.

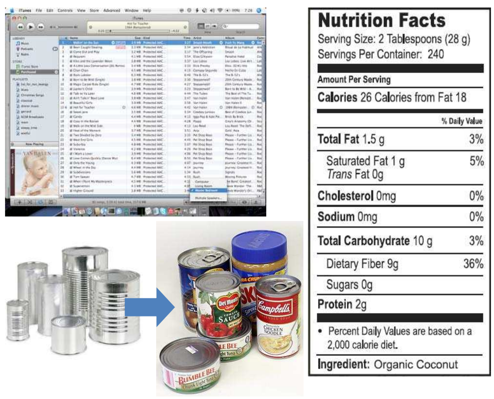

Figure 5. Daily use of metadata

To allow the sharing of data between heterogeneous systems, were developed models of data schemas designed to standardize the data to be exchanged, these standards for biodiversity schemas have been proposed by various initiatives such as the Biodiversity Information Standards [6] and Global Biodiversity Information Facility [7]. Among the standards developed, include the Darwin

Core, ABCD, Plinian Core and EML, the latter being employed in the proposed work. The following is a description of this metadata:

- DARWIN CORE [8] – designed to facilitate the Exchange of information on the geographical distribution of organisms and the physical existence of biotic samples in collections. The Darwin Core standard aims to facilitate the exchange of information between robust distributed databases [9];

- ABCD [10] – it is a data specification for biological collection units, including living and preserved specimens. It is intended initially to support the exchange and integration of primary data collection and detailed observation data;

- PLINIAN CORE [11] – Considered a set of concepts that define the basic attributes needed to integrate and retrieve information on species. It also provides a natural support and structure the relationship between species;

- EML [12] – Used to catalog information on ecological and environmental data, its specification is based on a series of XML documents that could be modular and extensible to document ecological data;

3.6 Extraction of data

With the use of different technologies to data acquisition, it is expected that the data is not stored in the same databases or in the same format. The wireless sensor network, with the xbow sensors, stored in postgresl DBMS and the Lonworks sensors stored in MySQL. Those data were extracted using Java Web technology, it was used Hibernate (https://www.hibernate.org) to connect into the databases and map the fields of the schemas of each technology to load into the main database in postgresql accessed by the weblab.

3.7 SEEK Project

The SEEK Project [13] (Science Environment for Ecological Knowledge) is a system designed not only to enable the acquisition and storage of data, but also to provide the integration, processing, analyzing and synthesizing biodiversity data that by the time of its development were considered intractable. The virtual infrastructure SEEK encompasses three integrated systems, according to Figure 6.

- Ecogrid – Layer capable of allowing the integration of data from heterogeneous sources of information that will be used for analysis in the upper layers;

- Semantic Mediation System – This layer mediates between conceptual differences of the various sources of data made available by Ecogrid. Thus, data from different sources can be grouped and made available, facilitating the construction of models and subsequent analysis.

- Analysis and Modeling System – Na automated visual environment [13] where ecologists can create change and incorporate analysis that make up the workflows and new models. In this layer the tool used to support your browser is a software Kepler [14], a visual interface to develop analysis based on ecological data, in the following sections will be discussed more detailed about that tool.

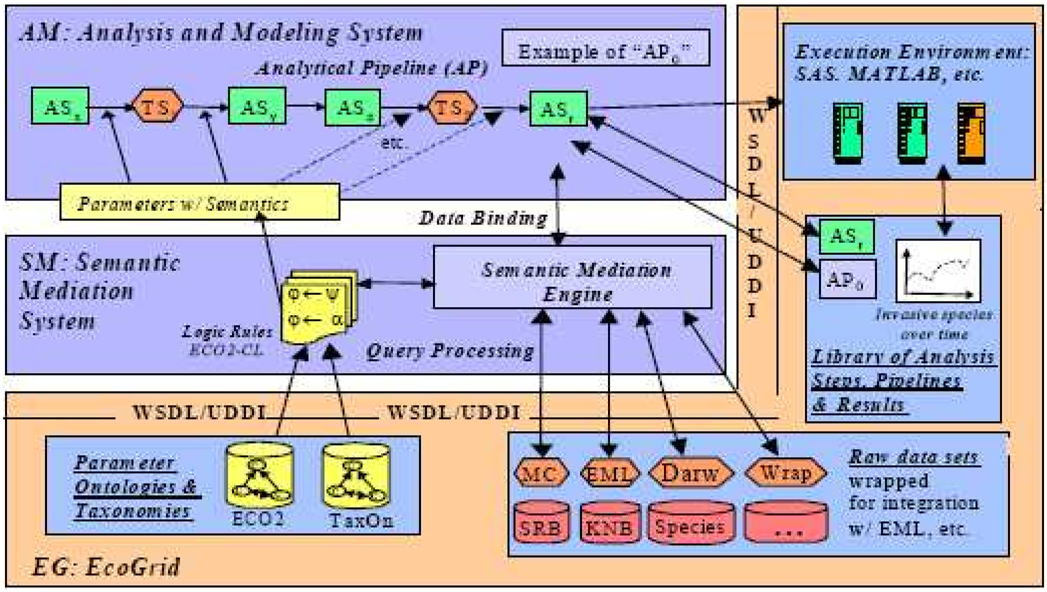

Figure 6. SEEK Architecture

3.8 Analysis Tool Kepler

These tools provide resources that aims to facilitate the work of preparation and implementation of tests specified by the researcher.

Kepler is a Java-based application that performs analysis involving collections and ecological data, allowing the modeling of the sequence of processes used in scientific analysis [14]. Based on the system Ptolemy [15], is a robust platform to support various types of computer suites suitable for different types of analysis (data processing device, or integral equations).

Provides to the user a graphical interface and a runtime environment that can execute workflows within the graphical interface as from the command line. Thanks to its flexibility it can be used in other areas of knowledge beyond the Ecoinformatics. In [16] are examples of other beneficiated areas from the use of the tool, for example, Biology, Geology, Oceanography and Chemistry.

4 Results

The experiment evaluated the functional and nonfunctional requirements for weblab portal. It was observed the web services and systems necessary for the development of the portal, as well as items that require optimization to obtain valid and useful data to users. The necessity of automatic extraction and analysis of data was observed by the complexity of manual work that researchers should do, like binary and hexadecimal conversions, queries in databases, manual metadata insert, etc.

This complexity confirms the necessity of software architecture of monitor that allows the management of information for weblab experiments. The Figure 7 shows the proposed architecture of such monitor.

With the monitor extracting the data and storing into a centralized DBMS, the analysis of such data already became automatic to the researcher, eliminating manual processes that could cause errors or misunderstandings.

At follow sections are described using this architecture, by collecting information in the field through bee experiments realized with weblab of bees.

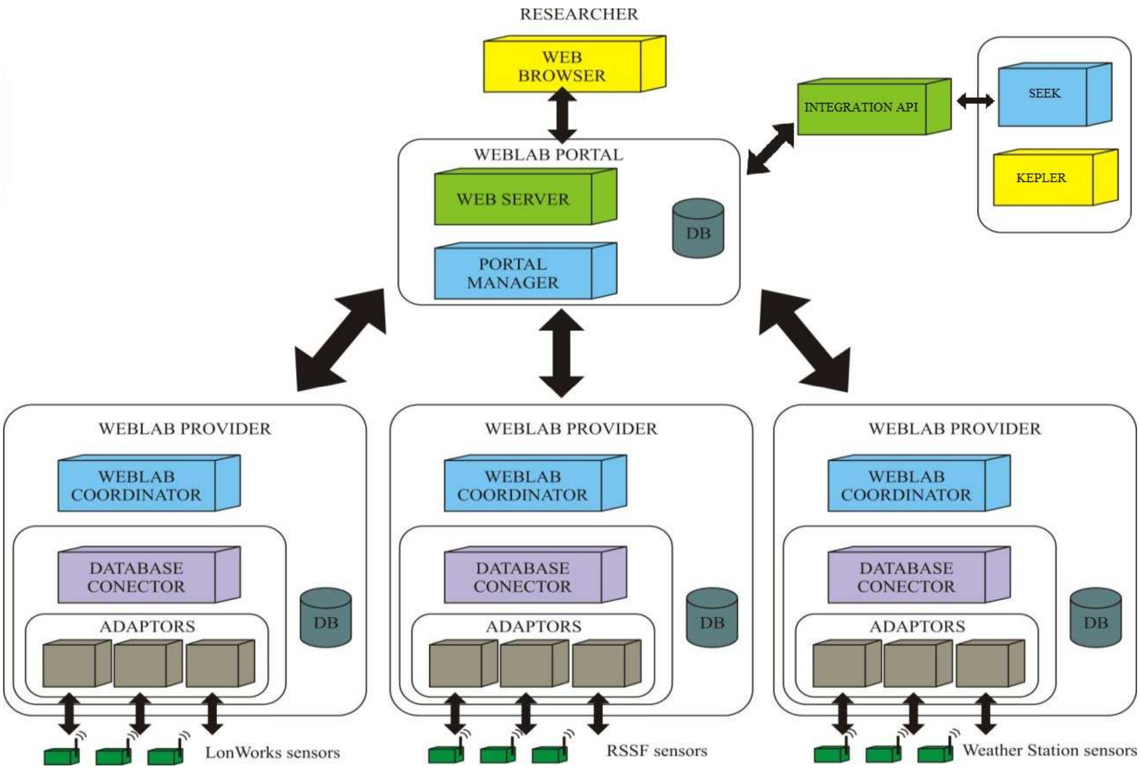

Figure 7. Monitor software architecture for weblabs experiments.

4.1 Temperature Data

Looking at the location of beehives two factors influencing the internal conditions: the temperature inside the chamber and the external environment, which is of significant importance, because the environment presents no barrier and borders, allowing the passage of air, facilitating the exchange of heat and humidity. Therefore, the comparison of data is compromised, as shown in Figure 8, the temperature at the entrance is always smaller than the room, this does not represent the temperature.

In the Figure 8 can also be observed that the temperature of the creation region is always higher than in the room, entry, and the pots of food, suggesting that S. depilis performs the thermoregulation of the region.

The input data represent a hive(?) of new data in the literature of this type of study, representing a good indicator for the study of ventilation and temperature control. However, due to the location of the hive inside the room is unable to make the comparison of data with the room temperature. We can infer that the outside ambient temperature was colder than the room through these data.

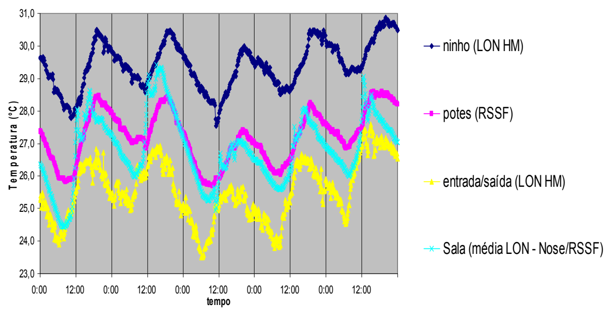

Figure 8. Temperature every 10 minutes in different parts of the beehive

As in other studies of thermoregulation in stingless bees, the highest temperatures were found in the region of the brood cells, and the highest average (29,4 ºC ± 0,73) (Table 1), slightly lower than the temperature found by [17] working with a species of the same gender, *Scaptotrigona postica*, who found an average of 32 ± 3 ºC this region.

	Nidus	Pots	Input/Output	Room
Average Temperatures (ºC)	29,4(SD 0,73)	27,2 (SD 0,8)	25,6 (SD 0,91)	26,8 (SD 1,06)
Average RH (%)	80 (SD 1,5)	72 (SD 2,2)	64 (SD 5,3)	54 (SD 5,6)

Table 1. Temperature and relative humidity (RH) average in the beehive

4.2 Relative humidity data

Continued monitoring of relative humidity (RH) in various parts of a colony is a major achievement and very significant pioneer in research with bees, since they are found in such literature. It is estimated that bees do not collect water directly, but the moisture present in the nest comes from the dehydration of nectar and metabolism of bees.

As shown in Figure 9, the HR in the region is held up higher and more constant than in other parts studied. This may be the result of maintaining the RH of active workers and the nest structures, such as housing, or just a byproduct of the presence of housing for other purposes.

The workers of meliponin ventilate the nest through the beating of wings at the entrance, to control the internal climate. There are no studies linking this with the humidity, but observing Figure 9, it is assumed that the high variation of RH at the entrance of the colony could indicate ventilation, and you can also make inferences about this behavior in future work.

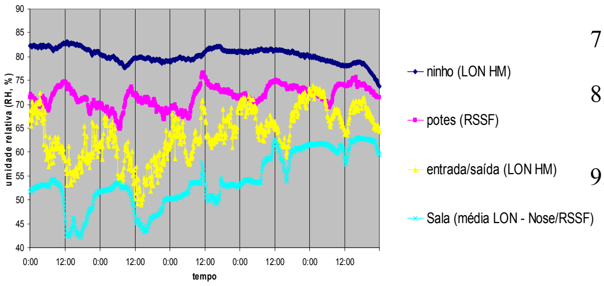

Figure 9. Relative Humidity (RH) in different parts of the beehive

Conclusions

Traditionally the most attention has been placed to the weblab and resources for collect data, but, no reference model to better integrate this infra-structure with software and hardware are demonstrated.

This paper presents a reference model to integrate one or many data collectors controlled by one weblab. Its use, do not must be viewed with a pattern, but the start point to create an others integration architectures, depending on the experiment performed.

Actually the validation of this model is being done through researches executed by specialists in biodiversity, the researchers of laboratory the Biology Institute of University of São Paulo (USP) and laboratory of bees in the Faculty of Philosophy, Sciences and Letters from the campus of USP in Ribeirão Preto.

Acknowledgements

Weblab Bioabelha has been part of the project Virtual Network System of Ecosystem Services, and this project is maintained by Polytechnic School of USP.

References

1. Ferreira, M. S. J. Distribuited systems architecture for envonmental services weblabs. 83 p. Dissertation (Masters) – Polytechnical School, USP. São Paulo, 2007.
2. Echelon. LonWorks Technology Overview. 2009.
3. Crossbow Technology. Crossbow Wireless Sensor Networks. 2009.
4. SARAIVA et al., WebBee – a Web-based Information Network on Bees – Journal of computer Engeneering and Digital Systems, n.1, v.1, nov. 2003
5. Almeida, L.F.B. Methodology Dissemination of Geographic Information and Metadata. Thesis. Center of Mathematical Sciences and Nature – UFRJ, 1999.
6. Biodiversity Information Standards BIS. Available at http://www.tdwg.org/activities/darwincore
7. Global Biodiversity Information Facility GBIF <http://www.gbif.org>.
8. DWC – DarwinCore Group – DwC. Available in < http://www.tdwg.org/activities/darwincore/>. Acessed in 10 jan. 2010.
9. Kakodkar A.P., Kerkar S.S., Varghese N.S., Kavlekar D.P. & C.T. Achuthankutty Darwin Core Based Data Streamlining with Digimus 2.0. Bioinformatics Centre, National Institute of Oceanography, 2009
10. ABCD - ABCD Schema 2.06 – ratified TDWG Standard. Available at http://bgbm3.bgbm.fu-berlin.de/TDWG/CODATA/Schema/default.htm. [12 dec. 2009]
11. PLINIANCORE – Plinian Core Project. Available at http://www.pliniancore.org/en/inicio.htm. [10 dec. 2009]
12. EML – Ecological Metadata Language. Available at http://knb.ecoinformatics.org/software/eml [12 dec. 2009]
13. SEEK – Science Environment for Ecological Knowledge. Available at http://seek.ecoinformatics.org>. [22 dec. 2009]
14. Altintas I., Berkley C, Jaeger E., Jones M., Ludascher B., Mock, S. Kepler: An Extensible System for Design and Execution of Scientific Workflows. National Center for Ecological Analysis and Synthesis, 2004.
15. PTOLEMY – The Ptolemy Project. Available at <http://ptolemy.eecs.berkeley.edu>. Acessed 17 dec. 2009.
16. KEPLER – Kepler: An Extensible System for Scientific Workflows. Available at http://kepler.ecoinformatics.org [17 dec. 2009]