SOCIENTIZE

SOCiety as Infrastructure for E-Science via technology, innovation and creativity

Deliverable no.	D3.3.1
Deliverable name	Report on the deployment, administration and use of the infrastructure
Dissemination level	PU
WP no.	3
WP name	Infrastructure Operation and Deployment
Date	10/09/2013
Date of delivery	04/10/2013
Actual date of delivery	
Status	Final
Author(s)	Cândida G. Silva, Rui M. M. Brito, F. Sanz
Reviewer (s)	Eduardo Lostal, Francisco Brasileiro

SOCIENTIZE is supported by the European Commission under Contract Number: RI-312902

Change log

Version	Date	Author/Editor	Reason for change / issue
1	10/09/2013	Cândida G. Silva	Creation
2	01/10/2013	Francisco Sanz	Added statistics
3	02/10/2013	Cândida G. Silva	Review
4	0210/2013	Eduardo Lostal	Review
5	02/10/2013	Francisco Brasileiro	Review
6	04/10/2013	Francisco Sanz	Final

Table of Contents

1. SUMMARY

This deliverable reports on the deployment, administration and use of the infrastructure under development in the SOCIENTIZE project. This deliverable is under the responsibility of the WP3 leader and includes the contributions of all the other partners involved in WP3.

Basic aspects of the Infrastructure Operation and Deployment (WP3) related with the setup and operation of the hardware and software infrastructure are described. Additionally, the usage of the infrastructure in the implementation of the cell image analysis (*Cell Spotting*) and the semantics maps (*Mind Paths*) applications is also summarized.

2. INTRODUCTION

The main objective of WP3 is the setup and operation of the hardware infrastructure. This includes internal servers and external resources.

Our first step was to analyze existing technologies and resources. After the selection of the technological components, these were deployed under the hardware infrastructure described below. On one hand, we maintain a production branch and two testing branches of the citizen science infrastructure. In one of the testing branch, we test new features while in the other WP4 deploys new experiments. On the other hand, SOCIENTIZE website and the whole CMS used is maintained in a production branch with a testing branch for new features before moving to production.

Although technological components were selected to start experiments' deployment, we keep testing and evaluating all possible technologies susceptible to be used under SOCIENTIZE. This technology surveillance is continuously shared among all the partners of the project, thus we also provide some tools to support such communication.

We also need a way to describe the infrastructure and provide mechanisms for the connection with current and foreseen external resources. This will be achieved by publishing APIs for each element in the infrastructure, making use of standards always that is possible.

The remainder of the document is structured as follows. In Section 3, Infrastructure description, we revise the software and hardware supporting the project development. In Section 4, Maintenance and application porting policies, we present the procedures that need to be followed to update the system infrastructure as well as the application porting process. Next, in Section 5, we address some interoperability issues. In Section 6, entitled *Infrastructure Usage*, we summarize how the infrastructure is used to support the applications being developed, and in Section 7, we present some aggregated statistics. Finally, Section 8 presents the outcomes of the technical events.

3. INFRASTRUCTURE DESCRIPTION

3.1 Hardware

BIFI-UNIZAR provides most of the hardware infrastructure (Figure 1) of the project, although other partners, mainly UC and UFCG, provide their own hardware to install and test software related to the project.

We are providing OpenVZ[1] virtual machines to deploy different software components. OpenVZ is a container-based virtualization for Linux. It creates multiple secure, isolated Linux containers on a single physical server enabling better server usage and ensuring that applications do not conflict among them. Each container performance and execution is exactly like a stand-alone server. A container can be rebooted independently and has root access, users, IP addresses, memory, processes, files, applications, system libraries and configuration files. These virtual machines are hosted in four physical nodes that are described in Table 1.

Table 1. Description of the physical nodes supporting SOCIENTIZE infrastructure.

Name	CPU	Mem	HD	OS
srv1.ibercivis.es	Intel(R) Xeon(R) CPU E5520 @ 2.27GHz (x16)	24GB	1TB (Raid 1)	Debian 6.0
srv2.ibercivis.es	Intel(R) Xeon(R) CPU E5520 @ 2.27GHz (x16)	48GB	1TB (Raid 1)	Debian 6.0
srv3.ibercivis.es	Intel(R) Xeon(R) CPU E5520 @ 2.27GHz (x16)	48GB	1TB (Raid 1)	Debian 6.0
srv4.ibercivis.es	Intel(R) Xeon(R) CPU E5520 @ 2.27GHz (x16)	24GB	1TB (Raid 1)	Debian 6.0

3.2 Virtual Hosts

On top of the physical nodes, several virtual machines can be deployed as needed. One of the advantages of OpenVZ is that it allows moving the virtual hosts across the different physical servers. This allows us a great flexibility to do, for example, maintenance tasks. Nineteen virtual hosts, not all of them related to SOCIENTIZE project, are currently running in the physical nodes described above.

1 http://openvz.org/

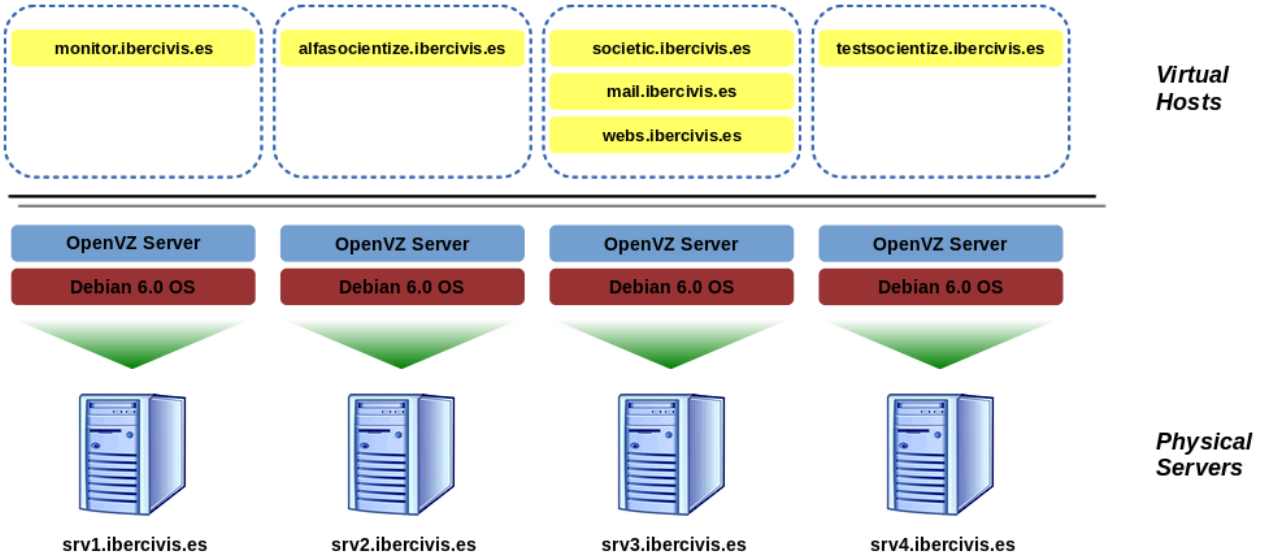


Figure 1. Diagram of the servers supporting the infrastructure of SOCIENTIZE project.

3.2.1 Virtual Hosts related with SOCIENTIZE

Most relevant virtual hosts (Figure 1) related to SOCIENTIZE project are:

- **monitor.ibercivis.es:** This host runs in srv1.ibercivis.es. It is responsible for the daily incremental backups and weekly full backups of the other hosts. This is performed through a software called BackupPC[2] under a 2TB file system mounted using RAID5. Additionally, a weekly snapshot of the virtual host is performed using the vzdump tool which is stored in the same file system.
- **alfasocientize.ibercivis.es:** This host is used to develop PyBossa[3] apps. This server maintains the same configuration as the one present in societic.ibercivis.es, the production PyBossa infrastructure. This server is hosted by srv2.ibercivis.es.
- **societic.ibercivis.es:** Hosted under srv3.ibercivis.es**,** this is our PyBossa production server. We install only stable experiments and stable and tested versions of PyBossa at this server.
- **testsocientize.ibercivis.es:** PyBossa middleware is tested at this server. New features of the middleware are developed in this host, although our developers are moving to Vagrant[4]+ KVM[5], which allows server software to be developed easier using the developers personal computers. It is hosted under srv4.ibercivis.es.
- **mail.ibercivis.es:** Using Qmail[6], this host is used to serve the emails under the SOCIENTIZE domain name. It is hosted in srv3.ibercivis.es.
- **webs.ibercivis.es:** In this host, we have installed the Drupal CMS that serves the main page

2 http://backuppc.sourceforge.net/
3 https://github.com/PyBossa/pybossa
4 http://www.vagrantup.com
5 http://www.linux-kvm.org
6 http://www.qmail.org

of SOCIENTIZE project[7]. We have also a MySQL server to support some different applications. This is hosted also in srv3.ibercivis.es.

All hosts described above run under Debian 6.0 operating system.

3.3 Software components

We use several software components to support SOCIENTIZE project. Main components are described in the following:

- **Apache2:** Apache HTTP Server Project is an effort to develop and maintain an open-source HTTP server for modern operating systems including UNIX and Windows NT. The goal of this project is to provide a secure, efficient and extensible server that provides HTTP services in sync with the current HTTP standards. We use it in conjunction with PyBossa, Drupal, etc. to serve almost all the web pages provided by SOCIENTIZE project.
- **BackupPC** is a high-performance, enterprise-grade system for backing up Linux, WinXX PCs and laptops to a disk server. BackupPC is highly configurable and easy to install and maintain. Installed under backuppc.ibercivis.es (that is an apache2 VirtualHost directive under monitor.ibercivis.es) it is used to do daily incremental backups and weekly full backups of all of our virtual hosts.
- **Drupal** is a free and open-source Content Management System (CMS) written in PHP and distributed under the GNU General Public License. Under webs.ibercivis.es our Drupal CMS serves socientize.ibercivis.es.
- **EpiCollect** provides a web application for the generation of forms and freely hosted project websites (using Google's AppEngine) for many kinds of mobile data collection projects. We are starting to use it to create Android and iOS form-like apps.
- **Mailman** is free software for managing electronic mail discussion and e-newsletter lists. Mailman is integrated with the web, making it easy for users to manage their accounts and for list owners to administer their lists. Mailman supports built-in archiving, automatic bounce processing, content filtering, digest delivery, spam filters, and more. Two mailing lists of the SOCIENTIZE project are supported by this software.
- **MySQL** is the world's most widely used open source Relational Database Management System (RDBMS) that runs as a server providing multi-user access to a number of databases. Some of our projects (like HappyUp) are using this database. Also, Drupal CMS uses a MySQL database.
- **OpenVZ**, as aforementioned, is a container-based virtualization for Linux. OpenVZ creates multiple secure, isolated Linux containers on a single physical server enabling better server utilization and ensuring that applications do not conflict. Each container performance and execution is exactly like a stand-alone server; a container can be rebooted independently and has root access, users, IP addresses, memory, processes, files, applications, system libraries and configuration files.
- **PostgreSQL** is an Object-Relational Database Management System (ORDBMS) available for many platforms including Linux, FreeBSD, Solaris, Microsoft Windows and Mac OS X.

7 http://www.socientize.eu

It is released under the PostgreSQL License, which is a MIT-style license, and hence free and open source software. PyBossa middleware uses this database.

- **PyBossa** is an open source platform for crowdsourcing online (volunteer) assistance to perform tasks that require human cognition, knowledge or intelligence (e.g. image classification, transcription, information location, etc). It can be used for any distributed task application but was initially developed to help scientists and other researchers crowd-source human problem-solving skills. At this moment, PyBossa is one of the most important software of the SOCIENTIZE project.
- **Qmail** is a mail transfer agent (MTA) that runs on Unix. It is a more secure replacement for the popular Sendmail program. Qmail's source code is in the public domain.

4. MAINTENANCE AND APPLICATION PORTING POLICIES

We have defined a set of procedures that must be followed for the maintenance and upgrade of PyBossa servers as well as for the development and deployment of new applications.

PyBossa infrastructure is comprised of three distinct servers, each with its own purpose:

- PyBossa production server, hosted at societic.ibercivis.es, is used to deploy the validated SOCIENTIZE applications, and make them accessible to the general public through the project's web page;
- PyBossa alpha server, hosted at alfasocientize.ibercivis.es, is used to develop new applications and test their correct functioning before being deployed at the production server;
- PyBossa test server, hosted at testsocientize.ibercivis.es, is used to test new versions of the PyBossa middleware, and our own developments that are potential contributions to the middleware, before they can be deployed in the production and alpha servers.

The procedures that must be followed in order to avoid unnecessary downtime at the production server as well as erratic behavior of applications used by the general public are the following:

- Test server is used only to test upgrades in the PyBossa middleware. For a middleware update to be performed at the either the production or the alpha servers, this must be preceded by a successful deploy at the test server. Only after the update is tested and validated at the test server, a middleware upgrade in the production or alpha servers is allowed to occur. Test server has some simple applications deployed that are used to check if everything is working as expected.
- Alpha server must always have the same software version that is deployed at the production server. Upgrades in this server means upgrades in the production server, and vice-versa. This guarantees that an application that works in the alpha server will also work at the production server. Thus, before being deployed in the production server, applications must first be deployed and tested in the alpha server. Only after this validation is performed, can an application be deployed at the production server.

4.1 High Availability

Users tend to feel themselves frustrated when an online service is not available whatever is the reason: either because of a system failure or a denial of service given too many accesses at a time. Eventually, the result is the same, user is unable to access the service.

Availability is the grade in which an application or service is available when and how users expect to. Main features to be considered are:

- **Reliability:** Both hardware and software are critical elements in order for the system to work properly.
- **Recovery:** Is there a plan to make our application to keep working in case of a failure? How long would it take to restore the system in case of a disaster? These are some of the points to be studied and planned in order to minimize consequences of an unexpected event.
- **Error Detection:** It is necessary to know the status of an element (i.e. failed, saturated, etc) in order to fix it in case of failure. Monitoring is a key point to figure out that status.
- **Constant Improvement:** Maintenance tasks must be transparent for the end user.

In order to provide high availability of our services, we are using Keep Alive, HA Proxy and the Memcache tools. With the first one, the public IP is always up although even if one server is down; with the second one, we provide load balancing which allows us to scale the system as needed. The third one decreases the number of queries to our database servers, caching them. We have also an active-active database system, thus queries are balanced between them. The infrastructure (Figure 2) can continue to function properly even if one of this database servers fails.

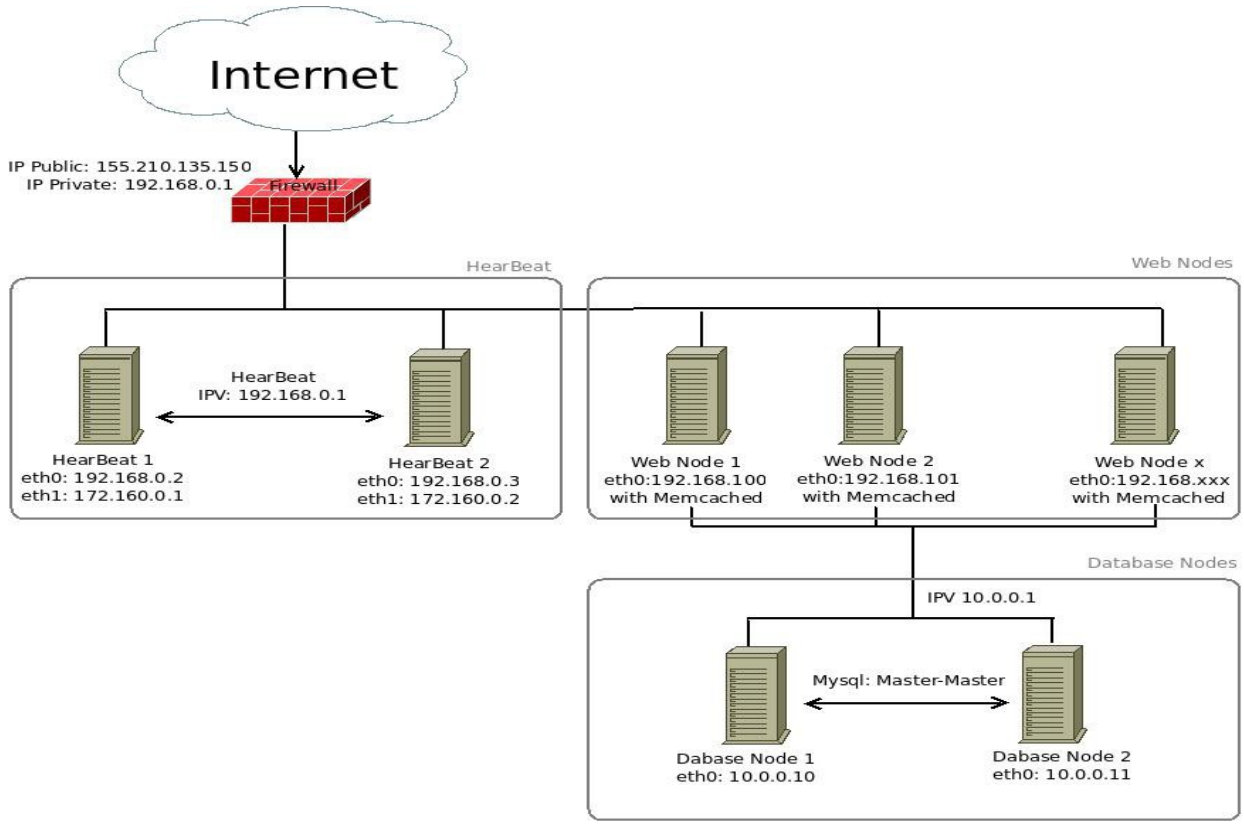


Figure 2. Topology of SOCIENTIZE infrastructure.

5. INTEROPERABILITY REQUIREMENTS

Much of the work of WP3 is to analyze and integrate existing technologies and resources with SOCIENTIZE components in order to set up the technological ecosystem of the project. This scenario raises the issue of interoperability, to be addressed at different levels:

- **Technical interoperability**, associated with hardware and software components, can be accomplished by adapting external infrastructure APIs to fit with SOCIENTIZE components, and, whenever possible, by addressing standardized protocols. Technological solutions explored within the scope of projects like EDGI or DEGISCO by some of the partners of this consortium are considered.
- **Syntactic interoperability** is related to data sharing and analysis. The fundamental goal in data interoperability is to facilitate and make transparent to end-users the extraction of information from multiple heterogeneous data sources residing in different locations. The design and management of schema mappings are the standard way to achieve data interoperability. A schema mapping is a specification of the relationship between two distinct file formats (XML, HTML,...) or database schema. Approaches for enhancing data interoperability are crucial when considering a collaborative environment of multiple citizen science projects.
- **Authentication** is required or recommended in most of citizen science projects. As in many other projects, applications running on SOCIENTIZE allow the participation of registered and unregistered volunteers.

  Currently, a volunteer can create an account by providing a valid e-mail address. Alternatively, OpenID intends to offer a unified "web identity" to each Internet user, allowing web sites and other people to connect different accounts, that the user has created online, into a more cohesive persona. Clearly oriented towards interoperability, both options could allow a volunteer to use the same identity across multiple (even if independent) citizen applications, thus allowing him/her the creation of a citizen scientist badge/profile.

6. INFRASTRUCTURE USAGE

Currently, SOCIENTIZE infrastructure supports two applications in production phase: semantic maps (*Mind Paths* – Deliverable D4.1) and cell images analysis (*Cell Spotting* – Deliverable D4.2). Additionally, two other applications are under implementation or test phases. The first being the Temperature Maps application which is included in the initial portfolio of SOCIENTIZE, and the second being new partner application named Sun4All. The main objective of the Sun4All application is the analysis of an asset of 15000 sun images (spectroheliograms) that are kept in the Astronomical Observatory of the University of Coimbra, as a result of a work of over 80 years of daily solar observations that started in 1926.

Different versions of these applications have been developed using HTML5 + JavaScript on the client side and Python on the server side running on top of PyBossa, the selected middleware for SOCIENTIZE Citizen Science infrastructure.

In the semantic maps application, a MySQL database – named *semantics* – contains the information on the list of words and the links. In the case of the cell image application, a database is used to store the information on the images under analysis.

7. AGGREGATED STATISTICS

One of the most important things that WP3 must ensure is that the whole infrastructure is running properly. In such a way, we collect some data of the usage of our platform and we have some triggers that warn us when something is not working as expected. Statistics for almost all of the tools that we are using are presented in the following. We focused on statistical aspects of these tools, but note we are using all these data to provide the best quality of service possible.

7.1 PyBossa

PyBossa, by default, presents some statistics of its usage. These statistics are available to the general public. In order to do that, PyBossa server collects some data from its database and plots them using Python libraries. Next, we present the PyBossa statistics for the usage of applications Mind Paths (Figures 3 and 4) and Cell Spotting (Figures 5 and 6).

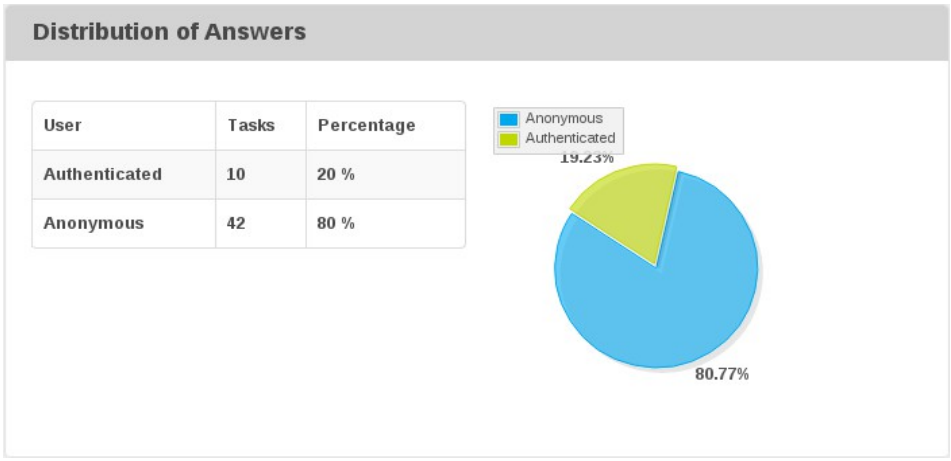


Figure 3. Global distribution of answers by authenticated and anonymous users for Mind Paths application running under PyBossa.

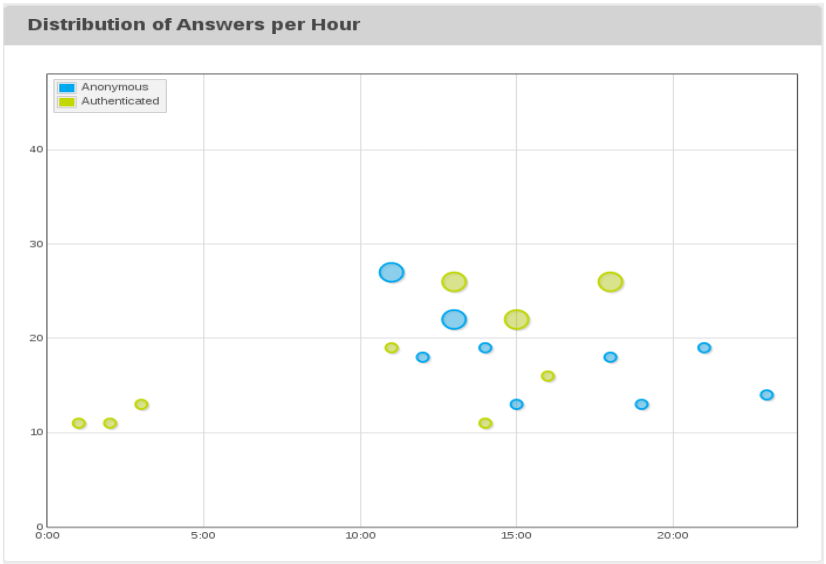


Figure 4. Distribution of answers per hour by authenticated and anonymous users for Mind Paths application running under PyBossa.

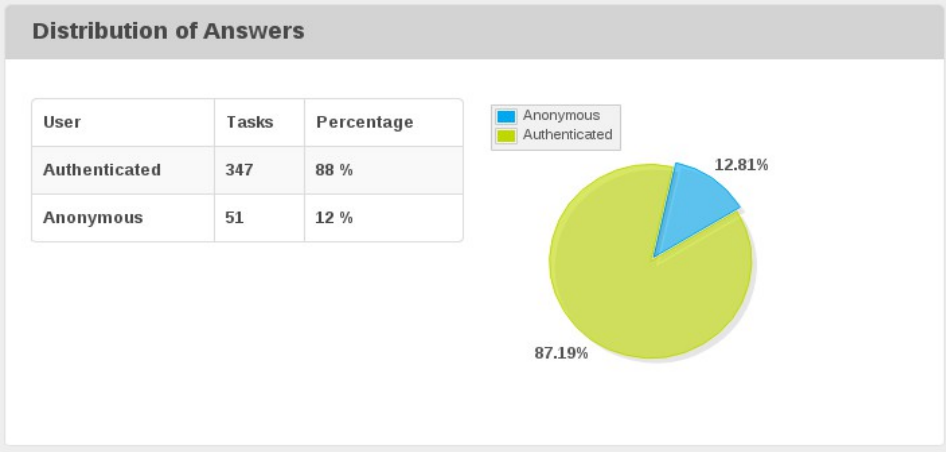


Figure 5. Distribution of answers by authenticated and anonymous users for Cell Spotting application running under PyBossa.

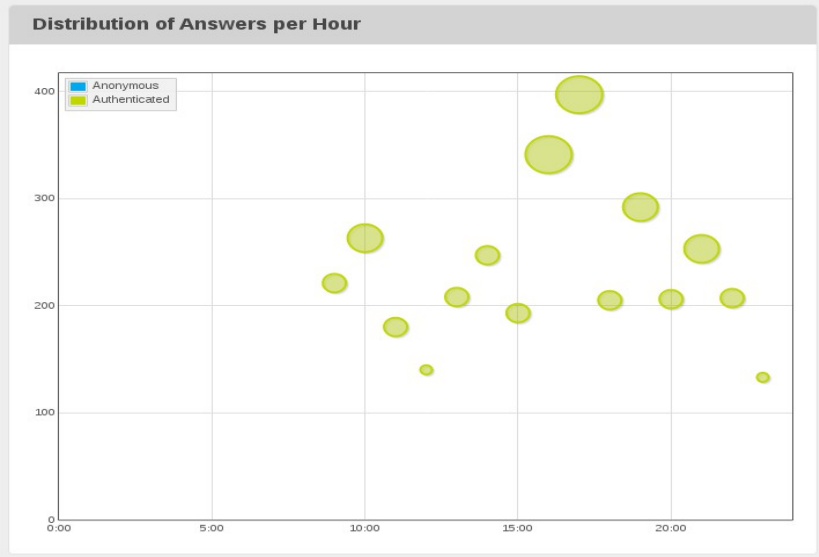


Figure 6. Distribution of answers per hour by authenticated and anonymous users for Cell Spotting application running under PyBossa.

7.2 Cacti

Cacti is an open-source and web-based network monitoring and graphing tool designed as a front-end application. It is generally used to graph time-series data of metrics such as CPU load and network bandwidth utilization. In Figure 7 some graphs obtained using this tool are depicted.

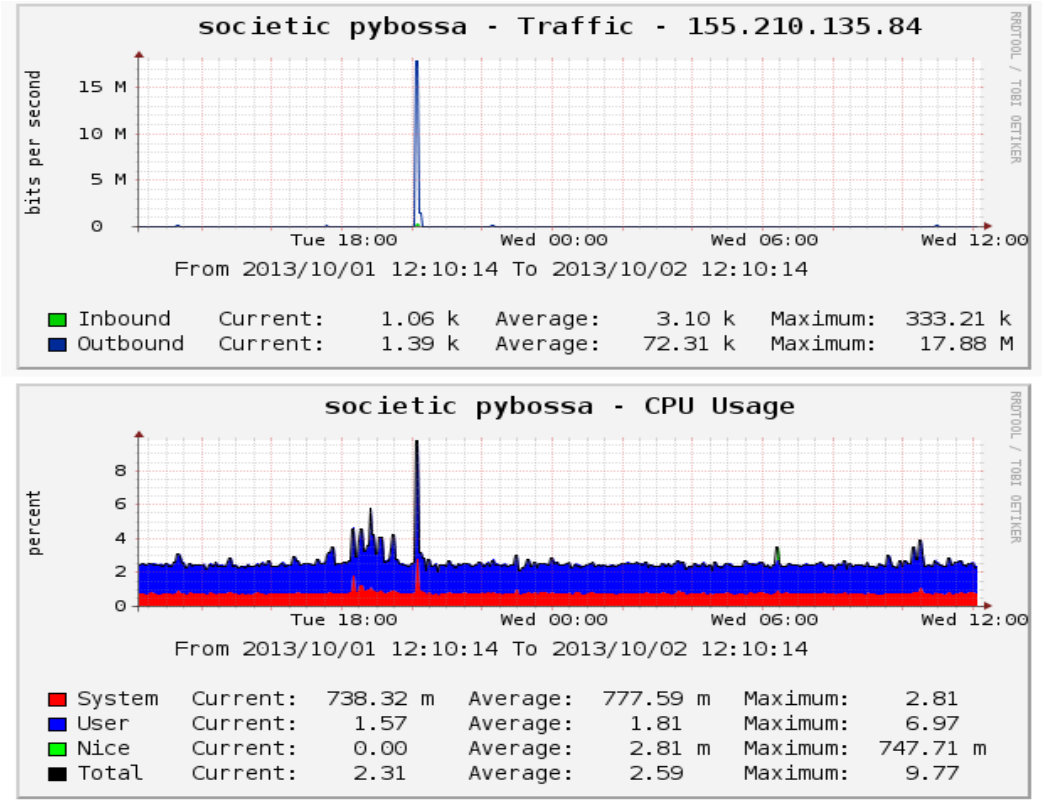


Figure 7. Cacti graphs for the network traffic and CPU usage of PyBossa server.

7.3 Nagios

Nagios is an open source system monitoring. Nagios offers monitoring and alerting services for servers, switches, applications and services. It alerts users when something goes wrong. We are using this tool mainly to check that everything is going fine, not for collecting data.

7.4 Google analytics

We started to capture some traffic analytics using the G*oogle analytics* tool by March 2013 in socientize.eu web page (Figures 8 and 9) and by April 2013 in pybossa.socientize.eu web page (Figures 10 and 11).

Google Analytics is a free service offered by Google. It generates some detailed statistics about a

website's traffic. It is implemented with *page tags*, i.e. a snipped of JavaScript, that the website owner adds to every page of the website. This tracking code runs in the client browser when the client browses the page and collects visitor data and sends it to Google data collection server as a part of a request for a web beacon.

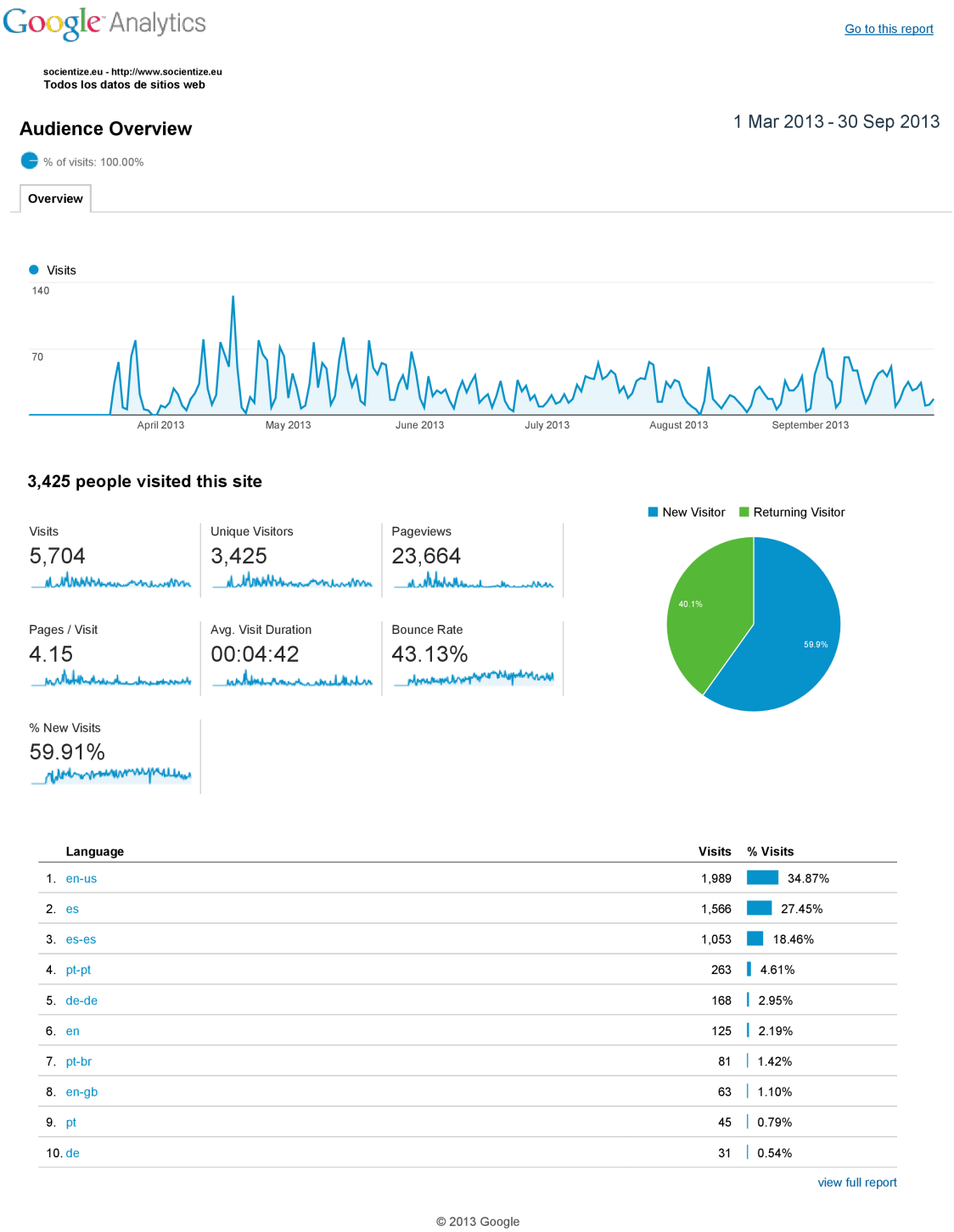


Figure 8. Number of visitors to website socientize.eu between March and September, 2013.

Google Analytics

Go to this report

socientize.eu - http://www.socientize.eu
Todos los datos de sitios web

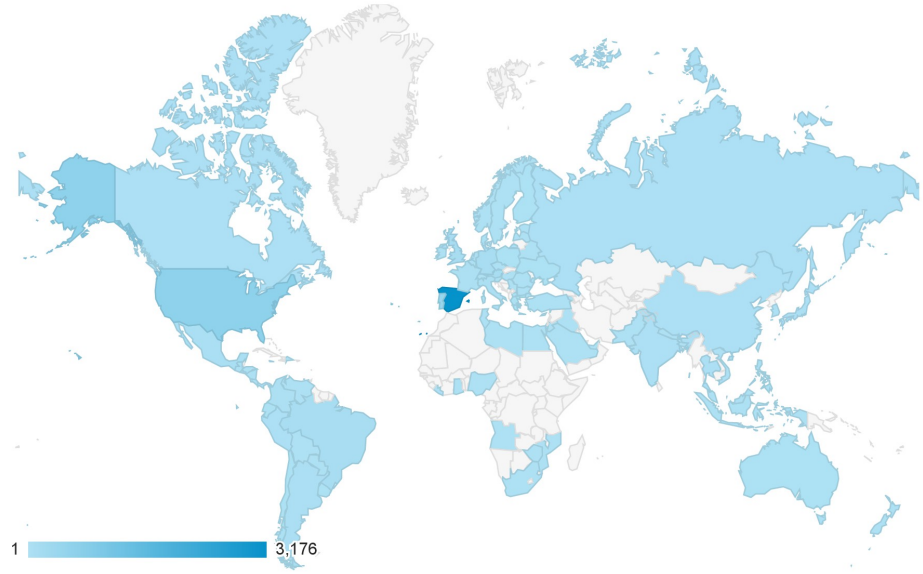
Location

1 Mar 2013 - 30 Sep 2013

% of visits: 100.00%

Map Overlay

Site Usage



Country/Territory	Visits	Pages / Visit	Avg. Visit Duration	% New Visits	Bounce Rate
	5,704 % of Total: 100.00% (5,704)	4.15 Site Avg: 4.15 (0.00%)	00:04:42 Site Avg: 00:04:42 (0.00%)	59.91% Site Avg: 59.91% (0.00%)	43.13% Site Avg: 43.13% (0.00%)
1. Spain	**3,176**	5.54	00:06:23	49.46%	26.10%
2. United States	**632**	1.35	00:00:35	93.35%	88.61%
3. Portugal	**430**	3.78	00:05:36	30.70%	33.72%
4. United Kingdom	**139**	3.17	00:02:35	69.78%	48.20%
5. Austria	**127**	4.11	00:06:06	34.65%	36.22%
6. Germany	**120**	2.74	00:02:36	65.83%	61.67%
7. Brazil	**113**	2.90	00:03:20	58.41%	35.40%
8. Mexico	**71**	1.11	00:00:13	94.37%	91.55%
9. Argentina	**65**	1.42	00:01:03	100.00%	87.69%
10. (not set)	**52**	1.56	00:00:10	92.31%	69.23%

Rows 1 - 10 of 90

Figure 9. Geographical distribution of the visitors of website socientize.eu between March and September, 2013

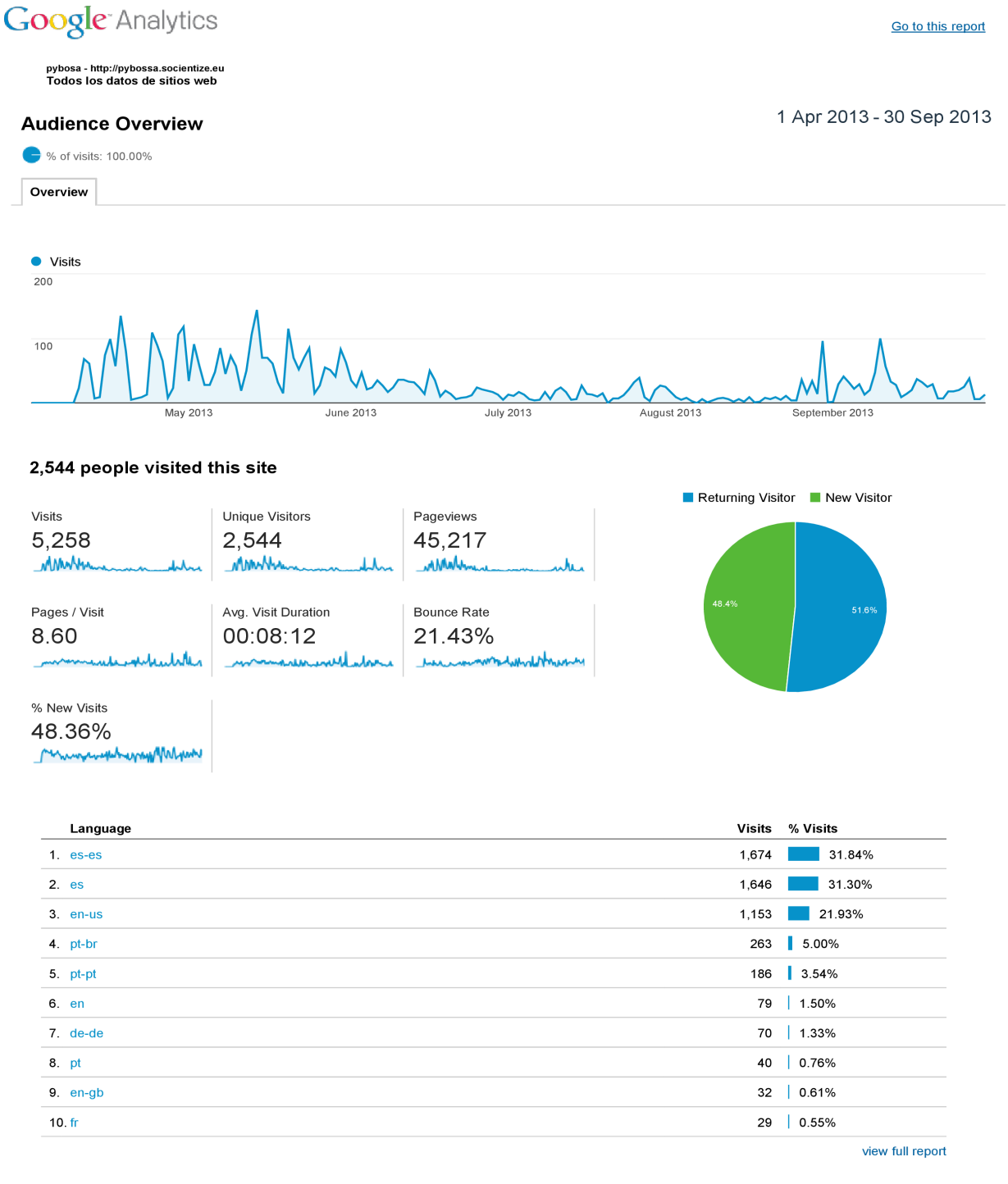


Figure 10. Number of visitors to website pybossa.socientize.eu between April and September, 2013.

Google Analytics

Go to this report

pybosa - http://pybossa.socientize.eu
Todos los datos de sitios web

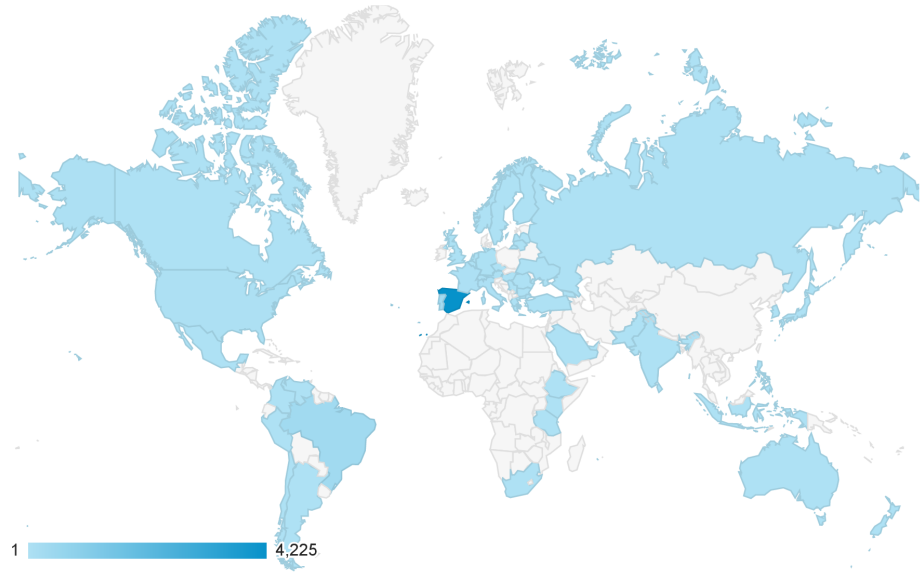
Location

1 Apr 2013 - 30 Sep 2013

% of visits: 100.00%

Map Overlay

Site Usage



Country/Territory	Visits	Pages / Visit	Avg. Visit Duration	% New Visits	Bounce Rate
	5,258 % of Total: 100.00% (5,258)	8.60 Site Avg: 8.60 (0.00%)	00:08:12 Site Avg: 00:08:12 (0.00%)	48.42% Site Avg: 48.36% (0.12%)	21.43% Site Avg: 21.43% (0.00%)
1. Spain	**4,225**	9.19	00:08:55	47.50%	18.79%
2. Brazil	**357**	8.76	00:07:57	40.34%	24.09%
3. Portugal	**313**	5.90	00:04:50	42.49%	25.56%
4. United States	**56**	2.86	00:02:10	83.93%	64.29%
5. Austria	**51**	4.24	00:04:35	54.90%	31.37%
6. Italy	**36**	7.64	00:05:40	25.00%	8.33%
7. France	**35**	3.20	00:03:00	65.71%	51.43%
8. United Kingdom	**24**	3.17	00:01:42	95.83%	33.33%
9. Germany	**21**	2.57	00:02:04	76.19%	42.86%
10. Belgium	**13**	2.31	00:03:18	61.54%	46.15%

Rows 1 - 10 of 58

Figure 11. Geographical distribution of the visitors of website pybossa.socientize.eu between April and September, 2013

From previous data and graphs, one can observe that the number of users is not high yet. During the time period here reported, none of the applications were officially launched, but monitoring system is ready to collect the necessary data in order to be analyzed by our partners. Moreover, the infrastructure is ready to alert us and adapt itself under possible failures.

8. OUTCOMES OF THE TECHNICAL EVENTS

Three technical events were organized by SOCIENTIZE consortium: the Citizen Science Open Technical Workshop held virtually using Google Hangout in January 2013 (month 4 of the project), and two Hackathons organized in Madrid and Zaragoza in May and June 2013 (months 8 and 9).

In the virtual meeting, we intended first to inform and discuss about existing tools related with citizen science. The major outcome of this hangout was the dissemination of open source tools related with Citizen Science which are ready to be used by developers, researchers and resource providers.

The hackathons were organized to work in hands-on developments with stakeholders such as policy makers, journalists, designers and developers communities (open source, DIYs, etc.). In the first day in Madrid, we followed the classical approach, with some experts presenting their project, but allowing the general public to make questions, which provided a very interesting environment for exchange of ideas and discussion. The second day was self-organized by attendants. This setup is commonly used by open-source communities. Our aim was to check if this format is valuable for citizen science. In the third one, the hackday of Zaragoza, we chose a lightweight event, very similar to the second day of the hackathon in Madrid.

From the two hackathons, very interesting ideas and new project proposals arose. Participants presented different technological solutions for the problems at hand. From SOCIENTIZE, we are providing support for the teams that are willing to continue working on their projects or apps beyond the events.

9. CONCLUSION

SOCIENTIZE infrastructure is fully operative and working properly, both from production and testing side, being capable to handle the demands of the applications developed or in development. Nevertheless, the infrastructure is flexible enough to incorporate new technologies and resources, if required throughout the execution of the project.

This is a realistic scenario for the SOCIENTIZE project as the process to select two new subcontracted applications is now undergoing, and that may require different solutions to support the underlying scientific experiments.