# EuropeanaLabs: An Infrastructure to Support the Development of Europeana

Nicola Aloia, Cesare Concordia, Carlo Meghini, and Luca Trupiano

Istituto di Scienza e Tecnologie dell'Informazione,
National Research Council, Pisa, Italy
{nicola.aloia,cesare.concordia,carlo.meghini,
luca.trupiano}@isti.cnr.it

**Abstract.** This document describes the Europeana Development and communication infrastructure, called EuropeanaLabs, built inside the EU projects Europeana and Europeana v. 2. The EuropeanaLabs consists of a number of servers, storages and communication devices; it is used to create Virtual Machines, called sandboxes, used by Europeana foundation communities. EuropeanaLabs provides a test environment, for applications and demos, to several cultural heritage and technology projects, most of them funded by EU, in addition it features a set of servers for cooperative work. In this paper we present the general architecture of the EuropeanaLabs infrastructure.

**Keywords:** Digital Library infrastructure, Europeana, sandbox.

## 1 Introduction

In the wide public, the Europeana Digital Library is primarily perceived as a portal exposing a great amount of cultural heritage information. Even though this perception is not entirely misleading, Europeana is much more. More precisely Europeana is an open services platform enabling users and cultural institutions to provide, manage and access a very large collection of information objects representing digital and digitized content [9]. Europeana is also a set of resources and tools open to the Community for developing, creating and disseminating new information resources. The set of these instruments provides the infrastructure, called EuropeanaLabs, on which is based the development of Europeana. The EuropeanaLabs provides development environments and various tools for building and managing digital libraries. These tools range from application servers for harvesting data and metadata (eg Repox), to Customers Relationship Manager (CRM), to collaborative work environments, products management, software validation, application showcasing, etc. From the organizational point of view the Europeana DL is the result of a number of activities run by different actors located across Europe, coordinated by the Europeana Digital Library Foundation (EDLF).

To define such a complex organization we can use the classical definition of Information System (IS): “a combination of Information Technology (IT) and people's activities that support operations, management and decision making [8]”.

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 53–58, 2014.
© Springer-Verlag Berlin Heidelberg 2014

According to this definition, the Europeana Portal is a component of the system, more specifically it is a web application using the Europeana API to provide services, in particular content discovery, to the Europeana Digital Library.

This paper focuses on, the computer system that act as a backend for EuropeanaLabs.

## 2 The EuropeanaLabs

As previously mentioned the Europeana DL is a collaborative work coordinated by EDLF and, up to today, the most part of the activities are carried on inside specific EU funded projects.

From the organization point of view this means that the community of users working on Europeana is composed by autonomous teams, usually geographically distributed, and number and needs of the community do vary over time.

We can individuated the following main features of the community of users working on Europeana:

- Distribution: there are geographically distributed teams, working on separate processes, often needing a close interaction each other to execute their activities.
- Heterogeneity: different teams can use different approaches for the same activity. For instance the various development teams could use different development methodologies.
- Scalability: at any moment new teams can join (or retire from) the organization
- Autonomy: every team must have a complete, autonomous working environment.

Up to November 2012 the community working on Europeana is composed by 1660 registered users, grouped in 42 main teams and each team works on one or more Europeana related activity.

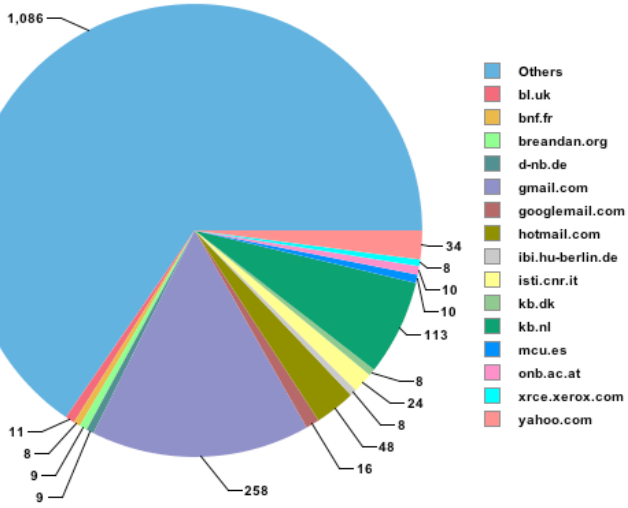


**Fig. 1.** The distribution of EuropeanaLabs community by email domains

In such a scenario, the role of EDLF is crucial. It is EDLF that coordinates activities of the Europeana community, collects and validates results, publishes them in Europeana (in form of new content or as new Europeana services).

The EuropeanaLabs has been designed to implement all needed tools and facilities to enable community users and EDLF to carry on their tasks.

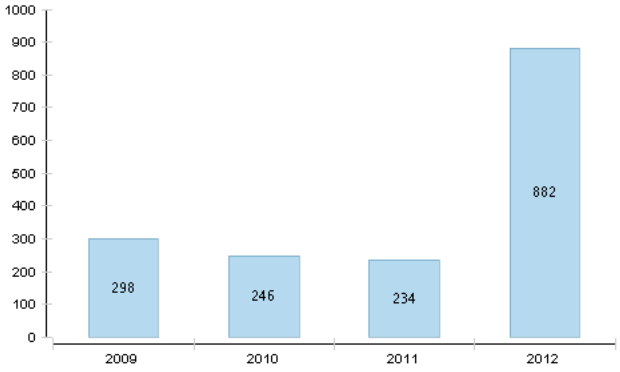


**Fig. 2.** New users registered in EuropeanaLabs every year

## 2.1 The Europeana Sandboxes

Basically the EuropeanaLabs provides computational and storage resources in form of autonomous systems called Europeana sandboxes.

Generally speaking, the term sandbox, in computing, may refer to:

- an isolated runtime environment used to run untrusted code (security sandbox)
- a testing environment that isolates untested code changes and outright experimentation from the production environment or repository, in the context of software development including Web development and revision control (development sandbox)[1].

We designed Europeana sandboxes by joining features of both categories; essentially a Europeana sandbox has the following features:

- it is a complete host computer on which a conventional operating system boots and runs
- it provides a controlled set of resources (main memory, storage, cpu, etc) and is accessible via the network
- it can be easily migrated between different hardware servers.

Every team working on one Europeana project, can ask for one or more sandboxes for tasks related to their activity; if the request is accepted, sandboxes are created and assigned to the team.

Europeana sandboxes are implemented using Hardware virtualization [6]: every sandbox is a Virtual Machine (VM). This is an obvious choice, virtualization paradigm

enable us to fully implement the specification defined by the project. The challenge is to define an architecture reliable, scalable and flexible enough to support the Europeana ambitious goals.

## 2.2 Architectural Description

The EuropeanaLabs infrastructure has been designed, developed and is maintained by CNR-ISTI. The work started in 2008: the initial configuration comprises a small server and a backup storage.

Over the years the complexity of the infrastructure has increased with the complexity of the project. It has finally conformed to a highly flexible, scalable and robust model. Among typical requirements for complex information systems (reliability, fault tolerance, etc) scalability is a major one for EuropeanaLabs: the infrastructure must be easily upgraded or remoulded to be adapted to the changing demands. The actual EuropeanaLabs model can really fulfil the current resources needs and easily meet the future ones, without requiring substantial changes. It comprises four main logical components:

1. the computational component,
2. the storage component,
3. the network component,
4. the administration component.

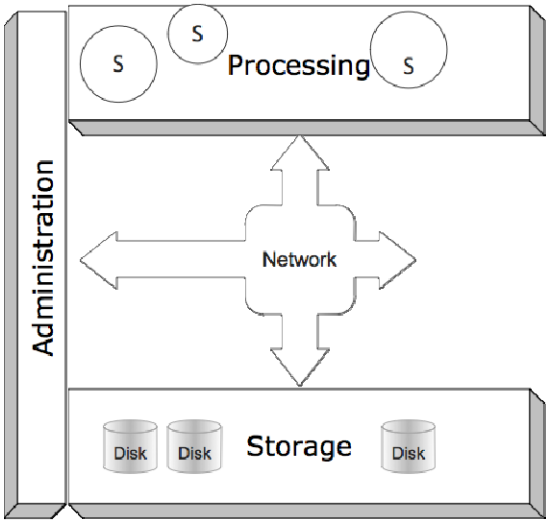


**Fig. 3.** EuropeanaLabs infrastructure components

The **computational component**, composed by 4 virtualization servers, is responsible of creating and running all the sandboxes and provides them with CPU and main memory. It also handles the storage space, remotely provided by the storage

component, making it available to the sandboxes. Finally, it provides Internet access to the sandboxes.

The **network component**, is composed by 2 high speed network switches. It provides LAN services and Internet connectivity to both the servers and the sandboxes. The data traffic between the virtualization servers and storage servers is done through standard protocols (AOE and iSCSI) and isolated from the Internet traffic, thus forming a storage area network (SAN). All network connections are made via a redundant structure both at the level of apparatuses that of physical links to provide fault tolerance and modularity.

The **storage resources**, composed by 2 servers, are redundant (all disks are in RAID configurations) to provide fault tolerance. They have grown with the project both in size and technology, and moved from general-purpose operating systems to dedicated storage systems with specialized hardware with the aim of improving performance and modularity.

The management and monitoring facilities, which make up the **administration component**, are provided by various software applications hosted on dedicated workstations. The management part makes use of standard hypervisor's tools, customized scripts and multiple servers management tools. The monitoring part uses Munin [3] to actively check all the relevant resources and services using a software agent installed on every virtualization and storage server. Another tool, Xymon [4], is used to check the status of the relevant network services of each sandbox without installing any software agent. The Internet traffic from and to the sandboxes is monitored in real time using tools like nProbe and nTop [5].

As all the sandboxes can be run on every virtualization server and have disk space provided by every storage server, this kind of architecture is reliable, flexible and easily expandable: virtualization servers can be added to increase the number of sandboxes, storage servers to increase the disks size, network switches to increase the number of connection links. If needed, all the sandboxes can be moved between virtualization servers, and the resources allocated to them can be changed.

### 2.3 EuropeanaLabs Software

The software used in the EuropeanaLabs infrastructure is open source. Besides from the servers operating system (Debian and OpenIndiana), the infrastructure needs hypervisoring, managing and monitoring software. “Hypervisor or virtual machine manager (VMM) is a piece of computer software, firmware or hardware that creates and runs virtual machines” [7]. As virtualization hypervisor Xen [2], being a widely tested and robust software, was chosen and installed on every virtualization server.

The grown of the infrastructure over time and the analysis of its usage by the community has resulted in the need for a dedicated middleware to manage and monitor all its resources, which can make the infrastructure still more flexible to adapt to the users requirements. We are currently working on this.

## 3 Conclusions and Future Works

The infrastructure, after five years of deployment, has been set up to fulfil all Europeana requirements. It currently provides enough computational, storage and communication resources to meet the needs of the project. It has the flexibility that allowed different reconfiguration to be done when needs have changed, has proved to be reliable with no data loss reported, has experienced minimal downtime without impacting the work of the Europeana community participants.

However some work still has to be done in order to make the infrastructure more flexible and manageable, in particular we're currently implementing the following features:

- live migration of sandboxes from any server to any other,
- monitoring and management operations done by a centralized middleware which would be able to define, create, migrate and destroy virtual machines and dynamically reallocate all the physical resources.

Another mayor activity being carried on by ISTI team is the design and implementation of a new and dedicated user interface for monitoring the infrastructure's resources status and usage, aggregating existing control data in a different way and paying particular attention to dependencies between sandboxes.

## References

[1] http://en.wikipedia.org/wiki/Sandbox_(software_development)
[2] Xen, http://www.xen.org/
[3] Munin, http://munin-monitoring.org/
[4] Xymon, http://xymon.sourceforge.net/
[5] nTop, nprobe, http://www.ntop.org/
[6] http://en.wikipedia.org/wiki/Hardware_virtualization
[7] http://en.wikipedia.org/wiki/Hypervisor
[8] Definition of Application Landscape. Software Engineering for Business Information Systems (sebis) (January 21, 2009)
[9] Concordia, C., Grad-mann, S., Siebinga, S.: Not just another portal, not just another digital library: A portrait of Europeana as an application program interface. IFLA Journal 36(1), 61–69 (2010)