Chapter 2
A Big Data Platform for Enhancing Life Imaging Activities

Leila Abidi
Université Paris 13, France

Hanene Azzag
Université Paris 13, France

Salima Benbernou
Université Sorbonne Paris Cité, France

Mehdi Bentounsi
Université Paris Descartes, France

Christophe Cérin
Université Paris 13, France

Tarn Duong
Université Paris 13, France

Philippe Garteiser
Université Sorbonne Paris Cité, France

Mustapha Lebbah
Université Paris 13, France

Mourad Ouziri
Université Sorbonne Paris Cité, France

Soror Sahri
Université Sorbonne Paris Cité, France

Michel Smadja
SISNCOM, France

ABSTRACT

The field of life imaging spans a large spectrum of scientific study from mathematics and computer science to medical, passing by physics, biology, etc. The challenge of IDV project is to enrich a multi-parametrized, quantitative, qualitative, integrative, and correlative life imaging in health. It deals with linking the current research developments and applications of life imaging in medicine and biology to develop computational models and methods for imaging and quantitative image analysis and validate the added diagnostic and therapeutic value of new imaging methods and biomarkers.

DOI: 10.4018/978-1-5225-4963-5.ch002

1. INTRODUCTION

The healthcare industry is a large generator of biomedical data. For instance, the U.S. healthcare system expected to reach the zettabyte (10^{21}) scale from electronic health records, scientific instruments, clinical decision support systems, or even research articles in medical journals (Raghupathi & Raghupathi, 2014).

In the last decade, we have witnessed the increasing resolution of imaging technologies which are considered as one of the most promising medical and health areas example and application of big data (e.g., NIH Brain initiative, n.d.) transforming case-based studies to large-scale, data-driven research (Luo, Wu, Gopukumar, & Zhao, 2016) and (Serrano, Blas, Carretero, & Desco,2017).

Interdisciplinary research in the field of imaging in the life sciences is essential. It requires the implication of different clinical and preclinical imaging departments yielding easy access to the state-of-the-art imaging equipment and patient data. Cooperative projects, including physicians, mathematicians, computer scientists, and physicists who are working closely together with bio scientists and clinicians are then launched in order to *(i)* develop computational models and methods for imaging and quantitative image analysis, and *(ii)* validate the added diagnostic and therapeutic value of new imaging methods and biomarkers.

Imaging is characterized by a large diversity in the types of data. Indeed, the data can originate from many different acquisition device, i.e., modalities, and the data format convention are quite loose with an important diversity in file formats and in completeness of annotation. The data themselves also strongly differ in their dimensionality, scale, size, and finality.

In such context, the life imaging project "IDV" (for Imageries Du Vivant) funded by University Sorbonne Paris Cité (USPC) launched the "Atlas IDV" initiative, which is a typical use case for data volume, variety and veracity in big data. The Atlas IDV initiative aims at *(i)* providing an integrated and agile environment supporting cooperation between scientists, and *(ii)* enabling to augment the research perimeter of imaging scientists and the extraction of new knowledge (data-driven research and images analytics) from the big multi-modal and multi-scale clinical and preclinical images available within the university.

A lot of studies in small animal imaging are hampered by small number of subjects, to the detriment of statistical quality of the findings. The junction of imaging data from a wide perimeter enables researchers to analyze a larger number of subjects, and hence to improve the statistical quality of their reports. Two use cases can be cited:

1. Many pathologies affect the normal physiology across many physical scales, and a complete understanding of these phenomena can only be obtained when examining images from a wide diversity of scales and contrasts. The "Atlas IDV" initiative, by virtue of its multidisciplinarity, will enable to put together datasets arising from a wide diversity of imaging techniques by the way of intelligent retrieval of heterogeneously stored data. As such, it will be a unique opportunity to make significant advances in the field of the life sciences.
2. The availability of such an infrastructure will also be helpful to facilitate standardization of the imaging methods. Indeed, an essential part of the validation process for imaging biomarkers is the possibility to share data obtained on a standardized object "phantom" across different vendors. The "Atlas IDV" initiative, by its distributed architecture, can allow to be carried out.

The "Atlas IDV" initiative brings together more than 200 scientists, affiliated with more than 20 research groups on 10 sites (Sts Pères Biomedical Center, HEGP, Necker Hosp., Bichat-Beaujon Hosp., Cochin Hosp., St Anne Hosp., Lariboisière Hosp., Cordeliers center, Chimie ParisTech, Villetaneuse Paris 13). The objective of the chapter is to give a global vision and feedbacks on an ambitious big data project for imaging research and what it brings to the IDV community.

Like organizations, one of the most important assets of any imaging research team is its image sets. Hence, the image sets are kept in two forms: using a set of dedicated cloud based operational systems of images records and processing supplying the "Atlas IDV" system (a.k.a data lake). The authors will show how the integrated imaging operational system promotes intra and inter-sites collaborative research work, while respecting data quality and privacy.

The Atlas IDV is based on the CIRRUS infrastructure available at USPC, more precisely the CUMULUS (n.d.) private cloud. Indeed, until recently USPC had no funding policy for platform development and sustainability, which has led to the need to disguise such development in call for proposals. With the CIRRUS platform, USPC has now made significant progress and the CIRRUS platform is used by projects in the 'deployment' phase. This implies requirements for stability through a deep control of the technical management to upgrade and to maintain the existing technology. In this chapter, a virtual datacenter inside CIRRUS is used to make concrete our ideas.

The business process as a service, i.e., BPaaS, implemented handles the imaging lifecycle of a variety of images' types (i.e., around 20 modalities acquired using heterogeneous imaging equipment's), with a large volume estimated at over 1 terabyte per day. For more information on BPaaS, see (Bentounsi, 2015).

To improve the images sets quality and indexing, native metadata (i.e., inserted by an imaging equipment) are enhanced using three main strategies:

1. A robust annotation scheme, which is the only way to ensure the scientific validity of the findings. Hence the availability of a web-based standardized annotation tools to enable researchers to associate datasets with accurate experimental information in a user-friendly manner, is of paramount importance (Kumar, Dyer, Kim, Li, Leong, Fulham, & Feng, 2016).
2. An enrichment of images sets via the crowdsourcing allows to harvest new annotations. Through an easy-to-use and intuitive interface, expert researchers (physicians, doctors, etc) submit their annotation tasks to non-expert contributors (students in medicine, biology, etc). Each task is a collection of one or more images. The contributors, categorized based on their experience, indicate their confidence for annotation. To collect high quality annotations, validation algorithm is used. It filters out annotations with low confidence and selects frequent and relevant annotations.
3. A semantic enrichment using linked open data and medical ontologies ("Radiology Lexicon", n.d.; "National Cancer Institute Thesaurus", n.d.; "Cell Ontology", n.d.).

Data enrichment strategies are helpful in big imaging data. This will improve images retrieval and analytics using innovative scalable algorithms based on MapReduce paradigm.

Privacy-enhancing technologies are used to preserve patient privacy by anonymizing images and metadata, and also by controlling the identification risk for the entire lifecycle of the "Atlas IDV" (Bentounsi & Benbernou, 2016) and (Bentounsi, Benbernou, & Atallah, 2016). In addition, a mechanism of images usage control in the Atlas is used leading to more transparency (Cao Huu, 2017).

E-Science has a better recognition of skills in software development as well as in large-scale infrastructure engineering in different disciplines. USPC should now recognize an interdisciplinary and complementary research on large scale systems and software in order to allow a sustainable development of the USPC platforms and to study a set of best (common) practices. The "Atlas IDV" initiative is a precursor project into that direction and their members are very excited by that idea.

The overall objective that we call for "interdisciplinary research" is preferably to apply well established methods inherited from e-Sciences for the building of systems for large scale computing and data management. The assumption is to check or to identify first what is common, in terms of Systems, between the disciplines versus the developments of new methods and technologies anchored in a unique discipline.

Finally, with such a System, USPC researchers can collectively be proactive in experimental science we obtain when using large scale platforms. We must not confuse the 'scientific problem' term and the 'System for scientific problem' term in order to serve the scientific community to solve scientific problems. We need scientists who develop software to support all the sciences we conduct at USPC. This cannot be the sole responsibility of computer scientists or engineers, but the informatics should rather percolate inside the disciplines, as we have shown in this chapter.

This Chapter is organized as follows. Section 2 introduces the Atlas IDV global architecture. Section 3 presents the project' private cloud and its architecture. Section 4 discusses how the crowdsourcing is used in the context of the project to annotate medical images. Semantic enrichment and Linked Open Data are presented in Section 5. Finally, Section 6 discusses images analytics in the context of the project and Section 7 concludes the chapter.

2. ATLAS IDV ARCHITECTURE

Like datasets for organizations, one of the most important assets of any imaging laboratory is its image sets, which represent the most visible part of a long (generally several years) and costly research process. First of all, like most areas of research, imaging researchers access to online research articles in medical journals and library reference collections to examine the state-of-art in a particular area. Based on their bibliographic studies, they develop research protocols specific to pathologies treated and results desired. The search protocol involves three key steps:

1. Develop new contrast agents or biomarkers.
2. Perform preclinical tests on cells and small animals to validate the efficiency and the non-harmfulness of the agents and/or biomarkers on living organisms.
3. If the first two steps have been conclusively established, clinical tests on humans are carried out before validating an agent or biomarker as a new medical diagnosis imaging.

During these stages, researchers need a set of data already available in several existing systems. In addition, the search process generates new information and knowledges to the community. Indeed, several operational systems and online databases are used to trace, index, and provide essential information to researchers and simplify the management of a research laboratory:

- Biobank databases which provide information on biological samples studied during the search protocol (Alfaro-Almagro, 2018).
- Animal facility management systems which provide information on small animals such as breed, weight, age, gender, model and lineage.
- Hospital Information Systems and Picture Archive and Communication System which provide patient data (Silva, Costa, & Oliveira, 2012).
- Materials for image acquisition which provide quantitative information on image acquisition process.

On the other hand, new knowledges and information are generated during the search protocol. They can be broken down into two categories. Experimental quantitative data and qualitative operational data on the functioning of the search protocol that assist the interpretation of quantitative finding (e.g., physical and biological properties of the contrast agent or the biomarker). This kind of data is at best saved in a Project Lifecycle Management system that keeps track of all the information generated during the search project (e.g., BenchSys, (2016) and Siemens Teamcenter (2018). Or at worst in a lab notebook or spreadsheet.

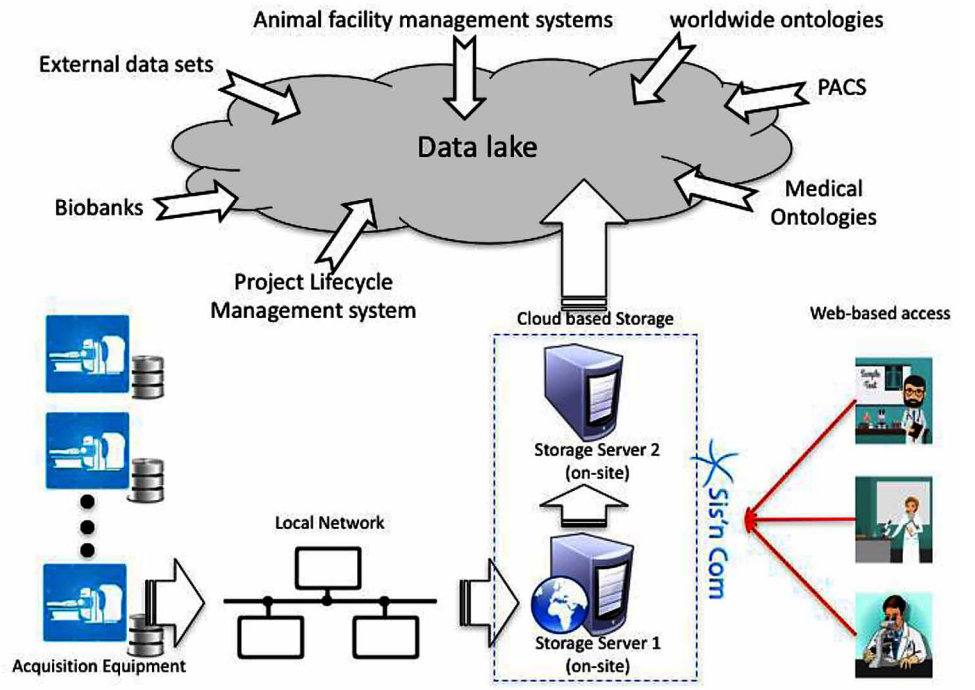
New knowledges take also the form of raw images acquired using a wide variety of medical imaging techniques (e.g, PET, CT, MRI, EPR), and also images processed using innovative image processing algorithms in order to detect the effect of studied contrast agents and biomarkers on organisms, and be able to diagnose pathologies as cancer, inflammation, etc.

Brutes and also processed images enrich the state-of-art in life imaging by providing online image bases allowing researchers to analyze a larger number of subjects, to test their image processing algorithms and hence to improve the statistical quality of their reports. This requires the implementation of image bases to index acquired images during research projects. We consider that a medical image may be indirectly identified by a combination of a set of metadata as for example a pathology, a contrast agent, and an organ. For that, we use all the metadata available to index medical images.

Implementing image bases can also reduce the costs and the duration of experimentations, and also proposing a new research paradigm in life imaging based on the statistical analysis of big image sets, i.e., data-driven research. However, the main obstacle to the implementation of image bases is the lack of storage resources at laboratories level. Often at the end of a project, the images are kept only for publications, then permanently deleted because of their volumes.

The “Atlas IDV” was implemented in the context of the project. Indeed, an integrated imaging operational system based on sis4web (n.d.) and connected to all systems and databases previously presented was developed. This leads to a base

Figure 1. Atlas IDV architecture



of indexed images using all operational and experimental information of the search protocol allowed the images' acquisition. The fact of having this mass of information related to the images makes feasible their re-use by researchers in future projects and feeds a large common base of images.

As depicted in Figure 1, at the end of the image acquisition process, a variety of raw images files are synchronized between acquisition servers and the storage server based on rsync (Mayhew, 2001).

Since health data are considered sensitive by the General Data Protection Reglement (Zerlang, 2017). Storing and processing health data requires a high level of security. Consequently, a dedicated private cloud based virtual servers have been used in the project (Krautheim, 2009).

The raw images have in their header a set of acquisition information (material, modality, date, time) and quantitative data (resolution, scale, etc.). These metadata are extracted from the DICOM and Bruker headers to be saved in the image base. In addition, the raw images are compressed using image compression algorithms in order to facilitate their transfer and visualization using web-based interfaces during collaborative work.

During the images storage process, researchers are asked to provide additional information such as: study objective, pathology studied, identifier of the object studied, the imaging agent characteristics, image processing algorithm, and physical measurements. These additional information enables the images base to query remote databases and operational systems via APIs in order to enrich images metadata. The proposed method based on internal and controlled enrichment allows researchers to index images through several attributes and be able to retrieve images via simple queries on a single attribute or via advanced filters on several attributes.

This architecture scheme is repeated on a multitude of laboratories and platforms within the project perimeter. However, slight modifications have been introduced in order to manage the volume and the velocity of acquired images and to face local network constraints. Indeed, in some cases, it was necessary to add an intermediate backup server on site which is synchronized in the night with the cloud so as not to congest the local network.

Periodically the compressed images and their metadata are extracted from distributed images bases, then integrated into a common images base. The images base stored in the cloud is accessible to all researchers on read only.

The export from private images bases is done in N3 format. The subject represents the image id, the predicate represents the attribute id, and the object represents the value. Records are stored in a csv file. Afterwards, the different csv files are imported into a single public NOSQL database accessible to researchers in order to make image analytics.

3. BUILDING THE DEDICATED CLOUD INFRASTRUCTURE

The current trend is to minimize the number of hosting datacenters while increasing the quality of data access. This principle is based essentially on the pooling of the resources of the various actors involved in the production and processing of health data. Another trend related to the previous one is to graft a scientific computing brick to the health data hosting part. This correlation will reduce the processing time of the data (reconciliation of the computing with the data) and minimize the security prerequisites for the transport of the health data between the hosting datacenters and those of processing. The adoption of the Cloud in the project is motivated, inter alia, by these two points.

Cloud computing technology potentially offers permanent access to data and services, from any device and anywhere at any time. Basically, it considers everything "as a service": computing, storage, network, and infrastructure. It pushes the user at the center of our concern by allowing him to deploy software on-demand,

development platforms or even the infrastructure he needs. Currently, this is exactly what scientific project partners ask for.

The goal is to build a software ecosystem and offer a research support service to make the best use of large infrastructures. This allows partners to access a wide range of software tools and also a large amount of image sets. The aim is also to ensure maximum comfort for researchers and engineers, i.e. do not disrupt their current working methods.

3.1 Technical Architecture

In order to be able to carry out the project's private cloud, we had to work in collaboration with the IT directors of Université Paris Descartes since the Cloud infrastructure is physically located there. We began by reinforcing existing hardware so that it could support the needs and expectations of researchers. In 2015, there was an investment of 1 million euros to reach a total of 4500 cores and 3 PB of storage to setup a private cloud and renovate two clusters.

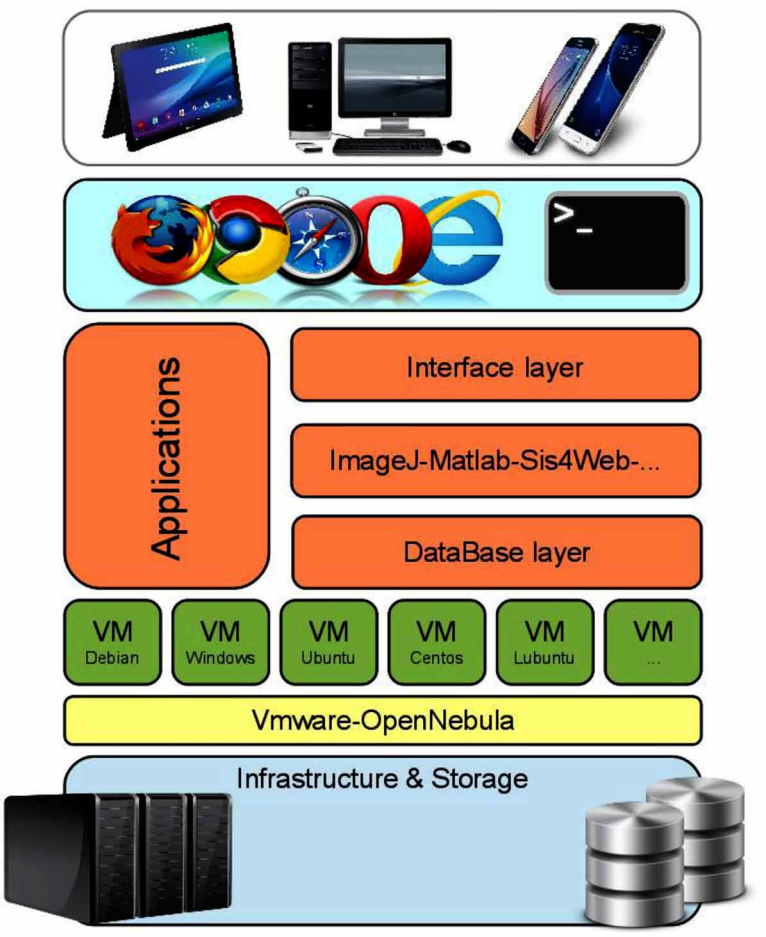
Conventionally, a virtualization-based solution has been chosen. This technique has the advantage of facilitating the cohabitation of several (operating) systems on the same physical support by providing complete isolation between systems and mutualized use of the resources of the system. The OpenNebula hypervisor, with a technical support, has been set up. OpenNebula operates as an orchestrator of the storage, network, supervision and security layers (Giovani, 2012).

Basically, a cloud-based solution consists in providing infrastructure, platform and software as services. Technically speaking, different templates of virtual machines (VM) are offered. These templates could be empty, i.e. containing only the OS, or predefined, i.e. containing a certain number of pre-installed tools. The selection of tools was not random; it was based on a survey carried out beforehand nearby the researchers involved in the project (Abidi, Cerin, Geldwerth-Feniger, & Lafaille, 2015).

Thus, we have set up VM templates and, for each of these templates, a virtual disk is defined, and an operating system is installed in this virtual disk. Different tools can be added to this system. Thereafter, one or many instances could be created from these templates. Each of these instances is associated with mandatory characteristics such as its name, memory size, the number of virtual processors…. One of the interests of the cloud is precisely to offer this type of mechanism which makes it possible to propose machines like a service i.e. "Infrastructure as a Service".

The Figure 2 summarizes the architecture adopted for the construction of the IDV cloud.

Figure 2. The different layers of the IDV cloud

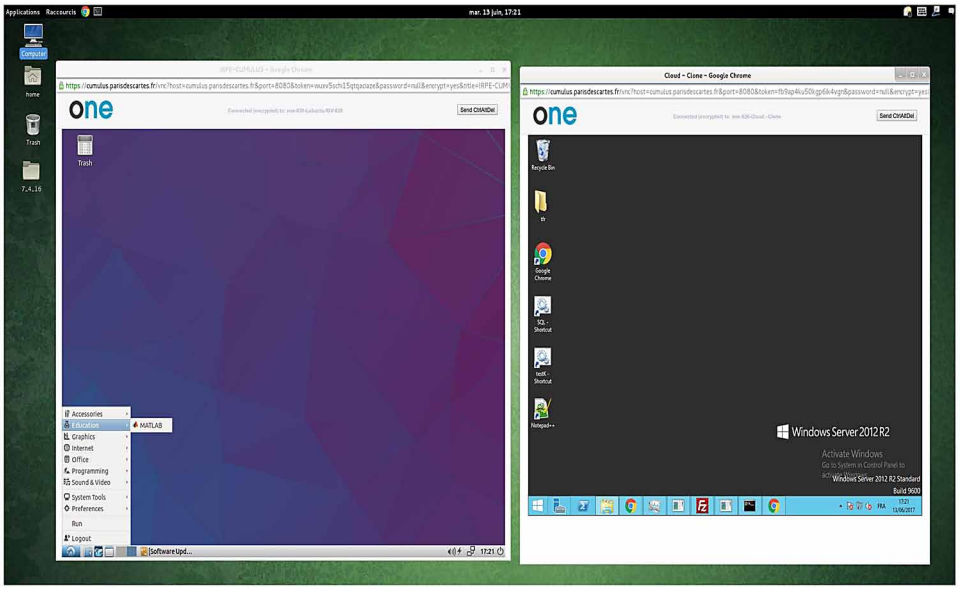


As shown in Figure 3, we have the possibility to use different virtual machines, simply through the browser. The researcher can deploy one or more virtual machines in one click and customize them with impressive characteristics (up to 64Go of memory, up to 20To of storage...).

Each researcher is the administrator of his own virtual machines. He have the possibility to customize his working environment, add/remove tools, add/remove users. There are also different ways to access the VM and work on it (VNC[1], SSH[2]…).

This first phase of the CUMULUS life allows to use the same tools as those used in laboratories but in a new operating context thanks to the deployment of virtual machines instead of machines isolated and disseminated in different physical places. The CUMULUS masks a lot of technical details. The centralization of data in the

Figure 3. Different operating systems are running in the browser's user



Cloud rationalizes the storage and avoid the multiplication of ad-hoc solutions to archive the data in the short term. This also facilitates access to data for the current atlas.

3.2 Discussion

3.2.1 High Performance Computing and Big-Data Convergence

Regarding the point of view of infrastructures for research in the academic sector, we still observe that large clusters, inherited from the HPC (High Performance Computing) community are still dominant. Large scale systems, such as clusters and clouds are systems with unprecedented amounts of hardware (>1M cores, >100 PB storage, > 100 Gbits/s interconnect). An ultra-large-scale system is an ecosystem with hardware, plus a large number of lines of source code, plus a large number of users, plus a large volume of data. In this context we have new issues, among them, how to build systems across multiple organizations; how to deal with conflicting purposes and needs; how to deal with heterogeneous parts with complex dependencies and emergent properties; how to deal with a continuous evolving; how to deal with software, hardware and human failures that are the norm, not the exception? Many answers can be found in the recent report from the Exascale (Asch & Moore, 2018) group of organizations, including academics and industrial partners.

Above all these considerations we still have issues regarding the problem of evaluating a large-scale system. If we evaluate clouds according to the metrics of the clusters, the comparison is not fair! We may evaluate both of them according to the productivity factor. Productivity traditionally refers to the ratio between the quantity of software produced and the cost spent for it. But many factors influence the decision, among them the programming language used, the program size, the experience of programmers and design personnel, the novelty of requirements, the complexity of the program and its data, the use of structured programming methods, the program class or the distribution method, the program type of the application area, the tools and environmental conditions, the maintaining of existing programs or systems, the reusing of existing modules and standard designs, etc

If we just make a focus on the programming language used, we may notice that the community of big-data do not use the same programming languages than the HPC community. However, the Harp project (Hadoop and Collective Communication for Iterative Computation) from the university of Indiana (n.d.) is a project that tries to "merge" two universes. The initial motivation is that communication patterns are not abstracted and defined in Hadoop, Pregel/Giraph, Spark. In contrary, MPI (Message Passing Interface) which is very used in HPC, has very fine grain based (collective) communication primitives (based on arrays and buffers).

Then Harp provides data abstractions and communication abstractions on top of them. It can be plugged into Hadoop runtime to enable efficient in-memory communication to avoid HDFS read/write. Harp works as a plugin in Hadoop. The goal is to make Hadoop cluster can schedule original MapReduce jobs and Map-Collective jobs at the same time. Collective communication is defined as movement of partitions within tables which are among the core data object. Collective communication requires synchronization. Hadoop scheduler is modified to schedule tasks in a bulk synchronous style (Wikipedia, n.d.) which allows predictions of the execution time (and in the theory). Harp also provides with fault tolerance with checkpointing. In other words, Harp attempts to realize a convergence between two layers in a system i.e. the programming layer (Hadoop like for the community of big-data) and efficient communication (MPI collective like communication for the community of HPC).

3.2.2 Cloud Computing

At this time the CUMULUS Cloud has been used to offer resources (CPU, RAM, disks...) to the project. But any project is faced to challenges such as communication and collaboration between product management, software development, integration, testing, deployment to cite a few. Our research group is currently focusing on the best methods and practices to put at the disposal of the IDV community, inside

CUMULUS, to deal with such issues that are not new issues. Indeed, we illustrate DevOps approaches that seems to be useful in this context because they are becoming increasingly widespread, especially in the community of practitioners.

DevOps (the contraction of Development and Operations) aims to establish a culture and environment where building, testing, and releasing software can happen rapidly, frequently, and more reliably (Samovskiy, 2010). There is no single "DevOps tool" but people consider "DevOps toolchains" consisting of multiple tools. Tools such as Docker (containerization), Jenkins (continuous integration), Puppet (Infrastructure as Code), Vagrant (virtualization platform) and Ansible (provisioning) -among many others- are often used and frequently referenced in DevOps discussions.

Let us comment two papers that exemplify the interest of DevOps approaches and DevOps toolchains to simplify the tasks of developers in large project. This part of the discussion is not addressed towards end-users but to developers in a broad sense. Final users do nothing, they just have to launch a script, in the worst case, if the job is not fully automated.

Stillwell & Coutinho (2015) deal with the problem to support the integration effort of HARNESS, an EU FP7 project. HARNESS is a multi-partner research project intended to bring the power of heterogeneous resources to the cloud. It consists of a number of different services and technologies that interact with the OpenStack cloud computing platform at various levels. Many of these components are being developed independently by different teams at different locations across Europe and keeping the work fully integrated is a challenge. Authors use a combination of Vagrant based virtual machines, Docker containers, and Ansible playbooks to provide a consistent and up-to-date environment to each developer. The same playbooks used to configure local virtual machines are also used to manage a static testbed with heterogeneous compute and storage devices, and to automate ephemeral larger-scale deployments to Grid'5000 (2017).

Indeed, authors present a development and operations (DevOps) workflow that allows: *(i)* teams of developers to work autonomously on specific parts of the software architecture; *(ii)* automated testing of individual projects as well as the integrated system deployments; and *(iii)* reproducible automated deployment on heterogeneous and large-scale testbeds. For that purpose, the authors introduce first a high-level view of the HARNESS DevOps workflow. Second, they describe how various deployment tools help to achieve reproducibility and why this is important for testing and quality assurance. The next topic covered in the article is to report the automated testing infrastructure and the novel methodology for testing full systems deployments. Finally, they describe two platforms where HARNESS is being deployed. To conclude, the paper demonstrates how to organize a research on DevOps approaches that serve multiple people, being working on the same project.

The second paper we would like to introduce is the very recent paper by Abidi, Saad, & Cérin (2017). In this paper, the authors are interested to deploying, in a multi-Cloud architecture, an infrastructure using the publish-subscribe paradigm for orchestrating the components of a framework that execute scientific workflows in highly heterogeneous and dynamic environments. More specifically they are in search of the adequate approaches to build deployment systems for heterogeneous and highly dynamic environments. The general objective is to offer « Workflow as a Service », as the concrete view, but we could imagine that the objective is to offer « X as a service », X being a utility function.

At least, we would like to mention the RosettaHub (n.d.) initiative that aims to facilitate the access to Amazon (AWS) computing and storage resources, from a basic user point of view. Services, now available free of charge to students and faculty members, include access to supercomputers, high-performance computing clusters, GPU and FPGA-accelerated machines, managed Hadoop and Spark clusters, storage services, databases and warehouses for big data, platforms for IoT and fog computing, machine learning services, etc.

The RosettaHUB platform creates, federates, manages and supervises the AWS accounts of some fifty higher education institutions around the world, totaling nearly 12,000 active AWS accounts of students and teachers. researchers. RosettaHUB also aims to establish a global, social and collaborative scientific meta-cloud. It simplifies and democratizes access to the infrastructures and tools needed for data science, big data and machine learning. Environments and tools such as Jupyter, RStudio, Zeppelin, TensorFlow, Spark, etc. are made instantly available as services to thousands of people. RosettaHUB enables real-time collaboration and sets up mechanisms for sharing all the scientific or educational artifacts produced on a cloud.

The platform exposes an API (Application Programming Interface) and federation model for clouds. Combined with the portability provided by the systematic use of container technologies (Docker), it makes possible and simple the transition from a 'classical' system (desktop like) to the cloud. The platform also exposes an API and federation model for compute engines and communicates R, Python, Scala, Mathematica, etc. in memory. through a common object model and deep integration of different virtual machines. RosettaHUB exposes above these federation models reactive programming frameworks for creating and publishing in multi-cloud and multi-language mode (R / Python / Scala) responsive microservices, collaborative interactive web interfaces, spreadsheets scientific and collaborative, data analysis workflows, etc. RosettaHUB's meta-formations capture all the dependencies of such data science / ML-oriented services and applications, allowing them to be reproduced and shared in one click. These meta-formations provide a foundation for reproducible research which is yet another issue that we need to address in the future.

3.2.3 Summary of the Work at the Infrastructure Level

The authors introduce a use case and they describe the different components of their DevOps architecture. They use the CUMULUS Cloud located at Université Sorbonne Paris Cité as the user's site. From CUMULUS the user starts a Vagrant script and this script automates all the deployment, installation, provisioning steps on multiple sites of the Grid'5000 tested for executing an application inside an infrastructure… that is also deploying. This is the key challenging point: how to automate the deployment of an infrastructure inside an infrastructure? DevOps approaches solve the problems in a convenient way.

4. CROWDSOURCING

Crowdsourcing is the process of outsourcing numerous tasks to an undefined, and generally large, network of people (the crowd). It is in widespread use in academic and industrial projects. Typical applications for the crowd include data collection, annotation or evaluation. For a wide array of tasks, the crowdsourcing paradigm has been shown to hold (Kittur, Chi & Suh, 2008).

In research domain, crowdsourcing platforms are a popular choice for researchers to replace expensive domain experts with crowd labour. It is particularly used to gather annotations quickly at scale. The most popular crowdsourcing scientific task is the categorization of galaxies (Raddick et al., 2013). In healthcare, great potential has been shown for various biomedical tasks, such as determination of protein folding (Eiben et al., 2012) and classification of malaria-infected red blood cells (Mavandadi et al., 2012). Crowdsourcing has also been used for clinical diagnosis (Nguyen et al., 2012) and for drug discovery (Lessl, Bryans, Richards & Asadullah, 2011).

Crowdsourcing was also recently used for image annotation in medical imaging. Many results showed that the annotation, via crowdsourcing, of a large amount of biomedical images can help at image classification. It has been shown in (Eickhoff, 2014), that the crowd can be much more effectively used to enhance the experts' performance and efficiency in detecting malignant breast cancer in medical images. In (Leifman, Swedish, Roesch, & Raskar, 2015), it has been demonstrated that crowdsourcing can be an effective, viable and inexpensive method for the preliminary analysis of retinal images. The work of (Herrera, Foncubierta-Rodríguez, Markonis, Schaer, & Müller, 2014) is based on the crowdsourcing for improving the quality of an automatic modality classification task based on the visual information of the images and the text of the figure captions. It particularly consists in verifying, by users familiar with medical images, the automatically detected modality of ImageCLEFmed images, and in reclassifying the images identified as wrongly classified.

In (Irshad, Montaser, Waltz, Bucur, Nowak, Dong, Knoblauch, & Beck, 2015), the crowdsourcing has been used for rapidly obtaining annotations by using the CrowdFlower platform for two core tasks in computational pathology: nucleus detection and nucleus segmentation. The results of this work show that the obtained annotations from crowdsourced non-expert-derived scores perform at a similar level to expert-derived scores and automated methods for nucleus detection and segmentation.

In our work and as part of the research project, we developed a crowdsourcing-based platform for the annotation of biomedical images. This platform referred as CrowdIDV, is an external module to the IDV atlas, that completes its semantic enrichment via the robust annotation scheme and the linked open data. A similar existing framework for the semantic enrichment of biomedical images, referred as SEBI, also includes a crowd-annotation module for biomedical images (Bukhari, Krauthammer, & Baker, 2014). Our CrowdIDV platform is a collaborative platform that is intended for students to annotate the biomedical images published by their expert researchers of the USPC community. It can also be considered as an educational platform.

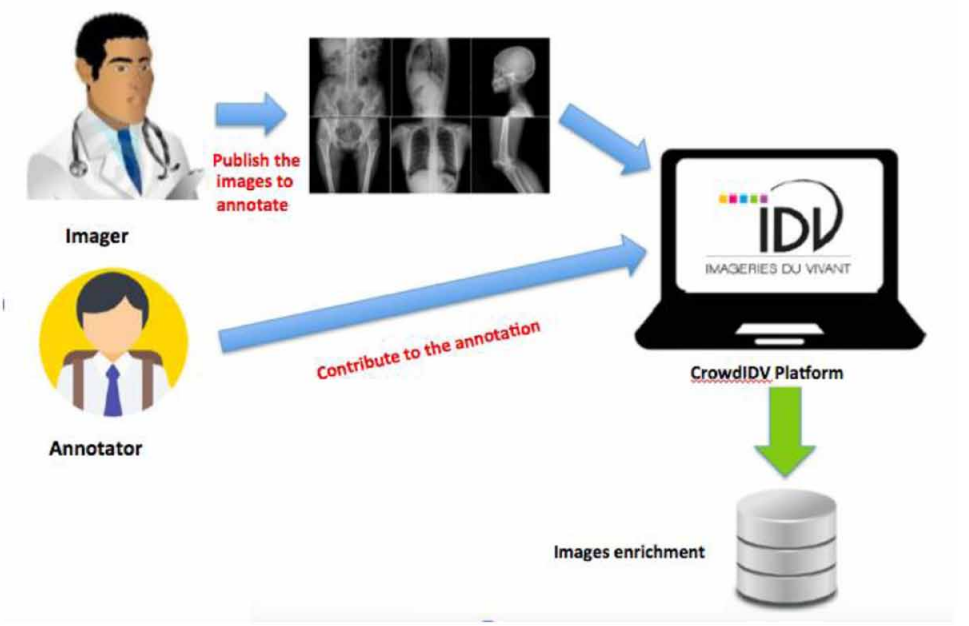
The researchers or experts, referred as imagers, are the requesters of the collaborative platform; and the students, referred as the annotators, are the crowders or participants.

As depicted in Figure 4, the imagers submit their annotation tasks and the annotators contribute to the annotation of the published images. The imagers and annotators must be registered via this platform.

In order to have relevant annotations, each imager submits its tasks for only its students. For instance, a biologist researcher publishes its images for only biologist students. To do so, students have access to a given task with a specific code given by its imager.

CrowdIDV incorporates algorithmic part to ensure the good quality of the collected annotations by students. We specifically developed an approach to validate the annotation input from various annotators. Our approach, inspired by Apriori (Agrawal & Srikant, 1994), the classic algorithm to mine frequent itemsets, consists in extracting the most frequent annotations for each published image while considering the confidence of the annotators. Indeed, at the annotation process, the student associates its confidence with the annotation he proposes (confident, high confident or hesitant). To do so, we follow these steps: First, we collect the annotations entered by students (or users). We assume that the annotation task is defined by a query image. To apply the Apriori algorithm, we model each image as a transaction database, where each transaction corresponds to student' annotations

Figure 4. CrowdIDV architecture



and then the annotations to the Apriori algorithm items. We recall that the Apriori algorithm have been proposed for generating association rules. The key idea of the algorithm is to begin by generating frequent itemsets with just one item (1-itemsets) and to recursively generate frequent itemsets with 2 items, then frequent 3-itemsets and so on until we have generated frequent itemsets of all sizes.

In our work, we use the Apriori algorithm to generate the frequent itemsets that represent the annotations. A wide variety of frequent itemsets mining algorithms exist in the literature such as Eclat, TreeProjection and FP-growth (Aggarwal, Bhuiyan, & Hasan, 2014). Apriori is the most commonly used algorithm and is a popular starting point for frequent itemset study (Heaton, 2016). Using Apriori, we consider the confidence of users at the generation of the frequent itemsets. The idea is to associate to each transaction (and then each user) a probability measure that corresponds to the user confidence. Consequently, only annotations resulting from this validation algorithm are considered and then proposed to the imager. This last can then lock the annotations if satisfied by the result or lets its images published for more annotations if not satisfied.

As future work, we will assess the accuracy of the annotations established by the crowd. Accordingly, we intend to consider the historical participation of each

annotator to further estimate the error rate of annotations and also for rewarding students. The reward for students could receive credits (EU project for example). CrowdIDV is currently deployed in IDV cloud. The next step is testing the platform by the community and then validate our approach and prepare the eventual enhancing steps. Once done, the collected annotations from CrowdIDV will be integrated in the IDV Atlas as triplestore for further activities of enrichment with the linked open data.

5. LINKED OPEN DATA BASED SEMANTIC ENRICHMENT

Today, the clinicians deeply rely on images for diagnosis, treatment planning and follow up. In fact, they deal with complex and heterogeneous data including image annotations allowing retrieving those images. In the previous section, we discussed a way to provide the annotations and enrichment of medical images using crowdsourcing paradigm in the IDV platform.

In other side, there exist other kind of data related to multi-modal and multi-scale life imaging including text, pre-clinical data and images of cells etc. Therefore, the production of life imaging is exponential. However, different banks of images are stored in their acquisition place, and such spatial fragmentation leads to under exploitation of those available huge amount of data where doctors and biologists do not collaborate and operate on their respective type of images. Then, the aim of the project is to allow medical scientists of pre-clinical and clinical researchers talking each through the fusion of big images sets (Dong & Srivastava, 2015) by using semantic interlinking between different types of images (Howe, Franklin, Haas, Kraska, & Ullman, 2017). So, the challenge is to gather richer medical information by connecting clinical images with pre-clinical images.

For clinicians, the aim is to enrich their clinical medical data and images with pre-clinical data and images, and vise versa for biologist. To do so, we use the Linked Data concept (Auer, Berners-Lee, Bizer, Capadisli, Heath, & Lehmann, 2017) (i.e., using the Web to create typed links between data from different sources) is a first step towards semantic-based data retrieval for semantic enrichment in life imaging domain using ontology mapping (Arnold & Rahm, 2014). Linked Data has the potential to provide easier access to significantly growing, publicly available related data sets, such as those in the healthcare domain (Sonntag, Wennerberg & Zillner, 2010).

Furthermore, the image interlinking does not deal only with proprietary data and images but may use other publicly available data called *open data* sources such as published data sets (Données publiques, n.d.) and worldwide ontologies including

Foaf (n.d.), dbpedia (n.d.), or medical ontologies (Radiology Lexicon, n.d.; National Cancer Institute Thesaurus, n.d.; Cell Ontology, n.d., ... etc.). Consequently, the project aims to use Linked Open Data (LOD) for life imaging.

Linked Data is essentially based on the RDF representation format for data representation, where the data can be linked using RDF/OWL links developed by W3C's Semantic Web Consortium, that become available during an advanced medical engineering process. In this project, we investigate how medical images and linked data sets can be used to identify interrelations that are relevant for annotating and searching medical images using RDF/OWL (W3C, 2004) for image representation though image annotations and SPARQL language for querying different data sources that are heterogeneous (Chekol & Pirrò, 2016) in order to search adequate data, and using ontology mapping process to discover the link between data (Mao, Peng, Spring, 2010) and (Verhoosel, Bekkum & Evert, 2015). Consequently, it is helping to enhance the patient diagnostics. As the produced images are massive, we are using new technologies such as Apache Spark (Amplab, 2018) (Zaharia Xin, Wendell, Das, Armbrust, Dave, Meng, Rosen, Venkataraman,Franklin, Ghodsi, Gonzalez, Shenker & Stoica, 2016)(Engle, Lupher, Xin, Zaharia, Franklin, Shenker & Stoica, 2012) and provide innovative algorithms related to Linked Open Big Images (LOBI), we wish to introduce.

Moreover, with the widespread use of PACS in the hospitals, the amount of medical image data is rapidly increasing (Silva, Costa, & Oliveira, 2012). Thus, the more efficient and effective retrieval methods are required for better management of medical image information. So far, a variety of medical image retrieval systems have been developed using either method (text-based or content-based) or combining two methods (Qi & Snyder, 1999), (Müller, Michoux, Bandon & Geissbuhler, 2004), and (Kyung-Hoon, Haejun, & Duckjoo, 2012). Each method has its own advantages and disadvantages. Text-based method is widely used and fast, but it requires precise annotation. Content-based approach provides semantic retrieval, but effective and precise techniques still remains elusive. The existing method are not handling the huge amount of data, in the project we are handling the volume and the heterogeneous variety of data sources using new technologies for big data.

Next, we provide a motivation example related to life imaging in order to enchich semantically data owner by solving the problem of semantic heterogeneity of different data sources.

5.1. Motivating Example: Linking Life Imaging

In our work, we assume that linked data contributes to solving the problem of structural heterogeneity as well as identifying entities to improve the quality of big data fusion

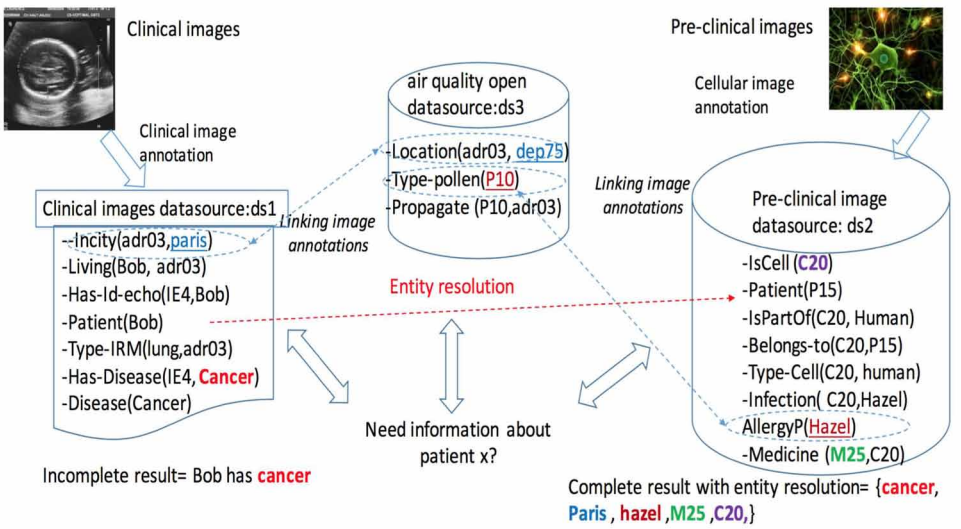
results (Benbernou & Ouziri, 2017). Nevertheless, RDF and its URI mechanism are not sufficient to solve the problem of semantic heterogeneity. For example, consider two data sources related to life imaging, clinical image representing the lung radiography of a patient namely Bob and a pre-clinical image representing an image of a cell extracted from human lung, as depicted in Figure 5 (All data sources are represented in RDF data). The challenge of our work is to connect clinical and pre-clinical images. For this aim, the images are annotated by doctors and biologists (or using a crowdsourcing system described previously). In the clinical source (at the top left side) is an IRM medical imaging on lung of a patient leaving in Paris area. A cancer is diagnosticated in the image. The second data source which is pre-clinical image represents a human cell infected with pollen of hazel. The annotation process generates databases containing descriptions of medical images.

Clinical and pre-clinical images are then linked through their respective databases. Linking can be made using direct connections between contents of databases or via intermediate other data sources namely open data sources dealing with air quality data.

In Figure 5, the concepts describing life imaging of data sources ds1 and ds2 are *Human* and *Cell*, respectively. The query over data sources is to retrieve all information related to the patient Bob.

Using an inference mechanism on clinical data source, we can infer only the knowledge *Bob has a cancer*. Such knowledge is *incomplete* because of the heterogeneity of the terminology used to describe the data from preclinical and

Figure 5. Example of linked data in life imaging



clinical sources and open data sources. Once a linking is processed between three data sources through *entity resolution* method between Paris and dept 75, Pollen P10 and Hazel related to the same entity, the query result becomes Complete result with entity resolution= {Bob has **cancer**, and is living in **Paris** having pollen **P10** and this type of pollen is **hazel** exists only in **Paris** and may can be applied on him a medicine **M25** to be cured because of its application on human cell **C20**}, Patient(P15) and Patient (Bob) are dealing with the same entity Bob. The entity resolution (ER) (Bhattacharya & Getoor, 2017), also known as record linkage or deduplication aims at *cleaning* a database by identifying tuples that represent the same external entity (Whang, Marmaros, & Garcia-Molina, 2013) and (Firmani, Saha, & Srivastava, 2016).

5.2. Open MEDICAL DATA and Semantics Representation (RDF and OWL)

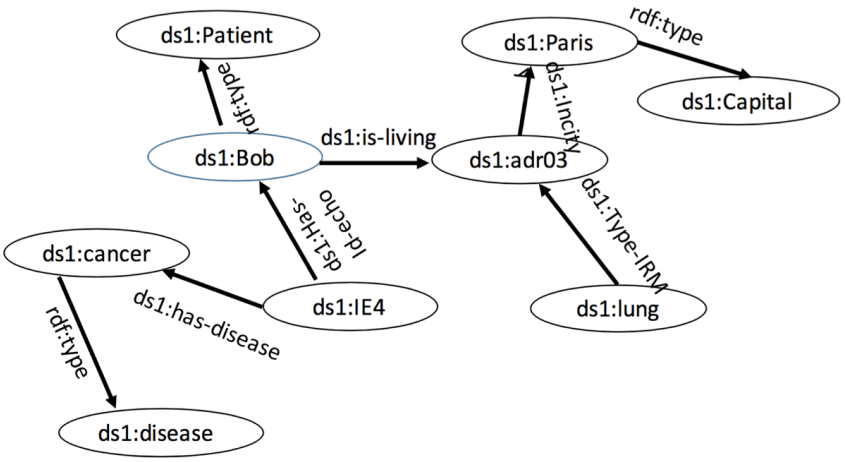
RDF is a standard data model, proposed by W3C for representing semantic Web data. RDF data is usually stored as statements in terms of triples subject, predicate, object, representing a relationship, denoted by the predicate, between the subject and the object. Subjects and predicates in triples are URIs when objects can be either URIs or literal values. An RDF data set forms a directed graph, where subjects and objects are vertices and predicates are labels on the directed edges from subjects to objects. OWL is a Semantic Web language designed to represent rich and complex knowledge about things, groups of things, and relations between things. It is a language allowing to represent knowledge for authoring ontology. An ontology a formal way to describe taxonomies and classification networks, it is a structuration of knowledge for various domains. The OWL languages are characterized by formal semantics based on Description logic [Baader, Calvanese, McGuiness, Nardi, Patel-Schneider, 2003]. The logical approach is used to verify the consistency of the knowledge or to make implicit knowledge explicit. Consider the data sources displayed in Figure 5, in Figure 6 is depicted its RDF knowledge representation of clinical data sources ds1 as facts i.e., Bob i*s-type* of Patient, Bob *has-id-echo* IE4, Bob *is-living* adr03, adr3 *in-city* Paris etc.

5.3. Semantic Linking of Life Imaging Data for Semantic Enrichment

In this section we will show the reasoning mechanism we can apply on different data sources to discover new knowledge and therefore enrich semantically the owner data.

Life imaging data sources are linked at two levels: *data level* and *semantic level:*

Figure 6. An example of life imaging data in RDF graph



- At the data level, when querying data sources, the data sources linking process aims to identify the same real-life entities (such as patients, a human organ) and connect them using the owl relationship *sameAs*. This process is named *Entity Resolution* (ER). The ER is the task of disambiguating manifestations of real world entities in various resources by linking through inference across networks and semantic relationships in application (ontology alignment) (Nentwig, Hartung, Ngomo, & Rahm, 2017) and (Shvaiko & Euzenat, 2013) and (Cheatham, Cruz, Euzenat, Pesquita, 2017).
- At the semantic level, in the era of big data and life imaging, the need for high quality entity resolution is growing as we are overwhelming with more and more data that needs to be i*ntegrated* (data fusion), *aligned* and *matched* before further utility can be extracted. Therefore, achieving inferences across the networks using semantic relationships between entities for a better high quality entity resolution become a great challenge. The aim of the semantic linking is to *make explicit the implicit data* across the network, thereby enriching life imaging semantically. The semantic linking is achieved using appropriate inference mechanisms to deal with big data and at the same time cleaning the big data produced when processing data fusion (Benbernou & Ouziri 2017).

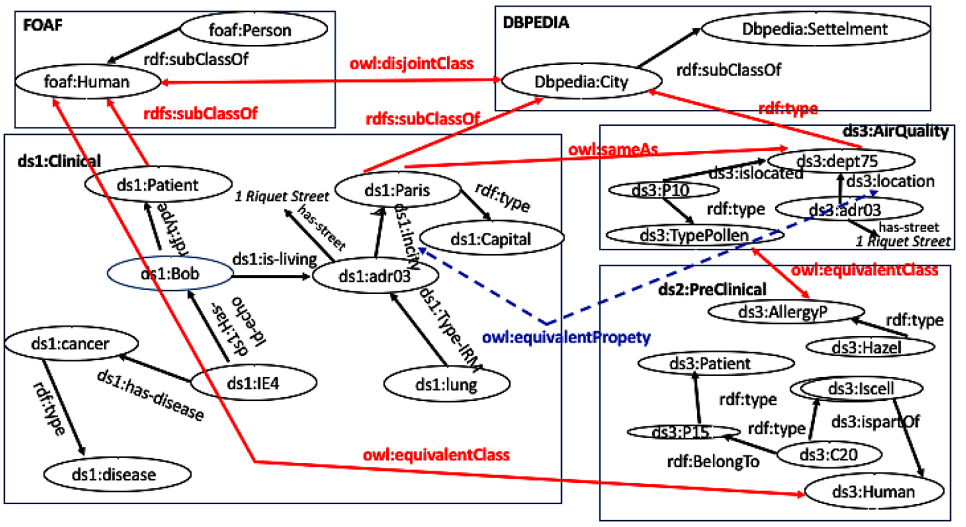
In order, to reconcile entities making entity resolution, we present in this section through an example, the inference mechanism that can be used to connect all heterogeneous RDF fragments of the same entity based on semantics and how

the enrichment of life imaging is processed. At the semantic level, concepts that are used to describe data sources are linked using semantic relationships (Shvaiko & Euzenat, 2013).

For illustration, Figure 7 presents three RDF fragments of the use case depicted in Figure 5, where concepts, describing life imaging, of datasources ds1 and ds2 are *Human* and *Cell*, respectively, data source ds3 is describing open data related to AirQuality, as well as two ontologies foaf and dbpedia as open data. The figure shows how the semantic linking between all those five sources using RDF/OWL languages is established. Some knowledge are extracted following the semantic linking. Furthermore, the entity resolution is operating at the conceptual (semantic) and entity (data) level. The W3C has defined axioms to make the semantic connections between RDFS and OWL standards. The most common are rdfs:subClassOf, owl:subClassOf, owl:equivalentClass and owl:disjointWith as pointed out by red and blue lines in Figure 7. In what follow we explain the inference reasoning that can be applied to discover new knowledge:

The RDF data are represented i.e., serialized by *facts*, some of them are listed: (1) ds1:*Bob* is a ds1:*patient* and is living at ds1:*adr03* (2) ds1:*adr03 inCity* ds1:Paris (3) ds3:adr03 *location* ds3:dept75 (4) ds1:Paris *sameAs* ds3:dept75 (5) ds2:*C20 belongTo* ds2:*P15* (6)ds2:*C20 infection* ds2:*Hazel* (7) ds2:*Hazel* sameAs ds3:*p10*

Figure 7. An example of how the life imaging data sources and open data are linked together using RDF/OWL



(8) ds3:p10 *propagate* ds3:adr03 (9) ds2:C20 *infection* ds2:Hazel. For instance the meaning of facts 7, the concept hazel is same as the concept p10, for fact 8, the concept p10 is propagated to adr03.

Besides, all those facts are along with *business rules* called also *domain rules* provided by the expert domain to make data compliant with them.

Therefore, the inference mechanism is processed as follow:The fact (4) when propagated (linked) to (2) and (3) infers that ds1:adr03 and ds3:adr03 are the same address and the given domain rule: *there can only one address street at a city*. The semantic linking in fact (7) when propagated to fact (8) infers that the adr03 is infected by hazel. By considering the fact (8) and given the domain rule: *there can only place infected by a Hazel*, the three resolutions can propagate to fact (1) and infer the place Bob is living is infected by Hazel. This resolution when propagate to fact (9) and (5) infers ds1:Bob and ds3:p15 is the same patient. And then the medicine can be applied to the patient to be cured.

The entity resolution is operating at the conceptual (semantic) and entity (data) level. The W3C has defined axioms to make the semantic connections between RDFS and OWL standards. The most common are rdfs:subClassOf, owl:subClassOf, owl:equivalentClass and owl:disjointWith as depicted in Figure 7. The inference mechanism is using such basic semantic connections to infer connections as well as operators of description logics to infer other knowledge for instance (1) dbpedia: City $\sqsubseteq$ ¬foaf: Human (2) ds1: Capital $\sqsubseteq$ dbpedia: City (3) ds2: Hazel $\sqsubseteq$ ds2: p10. For more details of different propagation that can be applied in the context of big data can be found in (Benbernou, Huang, Ouziri, 2017).

Once the owner data is enriched, it is ready to be translated to the image analytics process module discussed in the next section.

6. IMAGE ANALYTICS

In order to simplify the analysis for the experts, we present in this section through use case in image analytics our contributions in scalable machine learning. This task is important to simplify the inference mechanism that can be used to connect all heterogeneous RDF fragments presented in section 5.3. The global purpose is the enrichment of the expert knowledge.

In fluorescent medical imaging, a biological structure (e.g. gene, chromosome, cell, tissue or organ) is visualised by the expression levels of so-called marker proteins which attach specifically to this structure. A fundamental question in cellular biology is the identification of the regions of homogenous pixels delimited by these marker

proteins via image segmentation by unsupervised learning. A segmented image is the partition of these pixels into the clusters induced by a learning method. We focus on the class of modal clustering methods where clusters are defined in terms of the local modes of the probability density function which generates the data. The most well-known model clustering method is the k-means clustering (Lloyd, 1982). A segmentation based on solely on the fluorescence level is inadequate as the fluorescence level of marker proteins varies according to multiple factors, including most importantly the unavoidable natural, biological variation. A more adaptive segmentation would take into account other important data variables derived from the image other than the fluorescence levels, e.g. spatial localisation, topology etc. The *k*-means clustering currently used is capable of partitioning multi-dimensional variables but the resulting clusters are constrained to be ellipsoidal. These constraints imply that the *k*-means clustering is not well-suited to complex medical images. In our work, we decided to focus on density based algorithm, the most famous in that category is the DBScan (He, Tan, Luo, Feng & Fan, 2014) algorithm which is a density based algorithm which takes two parameters, ” which defines the radius of the hypersphere and minPts, which is the required number of points to consider the hypervolume as dense enough. Each times the density threshold is reached, dots are considered in the same clusters, the process is extended to catching points until the density is under the threshold. Rest of the points are considered as noise. One notable benefit of this algorithm is to detect automatically number of clusters with random shape but it remains hard to tuned efficiently.

On the other hand, Mean-Shift clusters are defined in a more flexible manner as the regions where these multi-dimensional variables are the densest. This leaves the clusters to evolve according to the local characteristics, which is well-suited to the task of delimiting pixels for complex biological structures. Mathematically these clusters are the basins of attraction to the local modes of the probability density function. For an image, each pixel is associated with a local mode by following the gradient ascent of the density function, and all the pixels associated with the same mode form a segmented region. The gradient ascent is estimated by the sequences of the nearest neighbour local means (Fukunaga & Hostetler, 1975).

On the other hand, Mean-Shift clusters are defined in a more flexible manner as the regions where these multi-dimensional variables are the densest. This leaves the clusters to evolve according to the local characteristics, which is well-suited to the task of delimiting pixels for complex biological structures. Mathematically these clusters are the basins of attraction to the local modes of the probability density function. For an image, each pixel is associated with a local mode by following the gradient ascent of the density function, and all the pixels associated with the same

mode form a segmented region. The gradient ascent is estimated by the sequences of the nearest neighbour local means (Fukunaga & Hostetler, 1975).

In our recent work, we demonstrated that the nearest neighbour Mean-Shift is effective for 2-dimensional non-medical images of a modest resolution. The factor which hinders a more widespread use of the Mean-Shift is that the sequential calculation of the nearest neighbours (via the dissimilarity matrix) quickly becomes computationally too onerous. To respond to the challenge, we have implemented a massively distributed version of it in the Spark-Scala Big Data ecosystem (Duong, Beck, Azzag, & Lebbah, 2016) and (Beck, Duong, Azzag & Lebbah, 2016).

We implement our algorithm in Scala because it is the native language in which Spark was implemented and so allows for optimal performance. Apache Spark is a fast-general purpose cluster computing system based on a master-slaves' architecture. The primary abstraction of Spark is a distributed collection of items called a Resilient Distributed Dataset (RDD) on which we apply the map and $reduce$ functions. Calculation times have been reduced by 10- to 100-fold, due notably to a rapid choice of the number of nearest neighbours and the calculation of approximate nearest neighbours via the random scalar projections of the "locality sensitive hashing" method (Indyk & Motwani, 1998) and (Slaney & Casey, 2008).

A part of experiments were realized on the Grid'5000 testbed which is the French national testbed for computer science research. It allows the deployment of a user's own operating system within the Grid'5000 hardware. We use a dedicated Spark Linux image1 optimized for Grid'5000 where Apache Spark is deployed on top of Hadoop YARN. Only the deployment of the image is automatized: we manually reserved the nodes as well as manually providing the Spark cluster with our executable code. For production purposes with CUMULUS we hide the configuration, deployment, provision of the services starting with a pool of dynamically allocated resources.

As mentioned previously, image segmentation can be made more precise by taking to account other available information. The Mean-Shift, as we have implemented it, uses the spatial localisation of dense data regions. Spatial coordinates are a continuous variable and could be utilised as the Mean-Shift clustering has been developed to treat continuous variables. On the other hand, incorporating categorical variables (e.g. yes/no, A/B/C) whenever intermediate values which have no intuitive interpretation is not yet feasible.

As future work, we propose to translate the successful technique for processing categorical data in other clustering methods via the recoding of categorical variables as a series of binary variables and replacing the mean with a median. Thus, we aim to develop a Median-Shift clustering method that is capable of handling composite data, consisting of a mixture of continuous and categorical variables, in the medical imaging and Big Data contexts.

7. CONCLUSION

We presented in this chapter a big data platform for enhancing life imaging activities in the setting of a multidisciplinary project IDV. We developed an ecosystem to the bio scientists and clinicians helping them to cooperate at the different levels for enabling to augment the research perimeter of imaging scientists and the extraction of new knowledge from the big multi-modal and multi-scale clinical and preclinical images available within the university. The platform is offering different services including sharing images in the same space, annotating their images using crowdscourcing system, linking their images through linked data technology, analysing their images, adding diagnostic and therapeutic value of new imaging methods and biomarkers. Several perspectives and enhancements raised regarding our experiences since the projects started three years ago, one can site enforcing the ethical and privacy aspects when sharing images and results, enhancing the multidisciplinary understanding through the platform still not mature and finally enlarge the platform to the international collaborations.

REFERENCES

Abidi, L., Cérin, C., Geldwerth-Feniger, D., & Lafaille, M. (2015). *Cloud Computing for e-Sciences at Université Sorbonne Paris Cité*. Taormina, Italy: Advances in Service-Oriented and Cloud Computing - Workshops of ESOCC.

Abidi, L., Saad, W., & Cérin, C. (2017). A Deployment System for highly Heterogeneous and Dynamic Environments. *International Conference on High Performance Computing & Simulation*, Genoa, Italy. 10.1109/HPCS.2017.98

Aggarwal, C., Bhuiyan, M., & Hasan, M. (2014). Frequent Pattern Mining Algorithms: A Survey. In C. Aggarwal & J. Han (Eds.), *Frequent Pattern Mining*. Cham: Springer. doi:10.1007/978-3-319-07821-2_2

Agrawal, R., & Srikant, R. (2014). *Fast algorithms for mining association rules*. VLDB.

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L. R., Griffanti, L., Douaud, G., ... Smith, S. M. (2018). Image processing and Quality Control for the first 10, 000 brain imaging datasets from UK Biobank. *NeuroImage*, *166*, 400–424. doi:10.1016/j.neuroimage.2017.10.034 PMID:29079522

Amplab. (2018). *Amplap UC Berkeley*. Retrieved from https://amplab.cs.berkeley.edu/tag/spark/

Arnold, P., & Rahm, E. (2014). Enriching ontology mappings with semantic relations. *Data & Knowledge Engineering*, *93*, 1–18. doi:10.1016/j.datak.2014.07.001

Asch, M., & Moore, T. (2018). *Big Data and Extreme-Scale Computing: Pathways to Convergence.* Retrieved from http://www.exascale.org/bdec/sites/www.exascale.org.bdec/files/whitepapers/bdec2017pathways.pdf

Auer, S., Berners-Lee, T., Bizer, C., Capadisli, S., Heath, K., & Lehmann, J. (2017). *Workshop on Linked Data on the Web co-located with 26th International World Wide Web Conference (WWW 2017). CEUR Workshop Proceedings*. CEUR-WS.org.

Baader, F., Calvanese, D., McGuinness, D., Nardi, D., & Patel-Schneider, P. (2003). *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press.

Beck, G., Duong, T., Azzag, H., & Lebbah, M. (2016). Distributed mean shift clustering with approximate nearest neighbours. *International Joint Conference on Neural Networks*. 10.1109/IJCNN.2016.7727595

Benbernou, S., Huang, X., Ouziri, M. (2017). Semantic-based and Entity-Resolution Fusion to Enhance Quality of Big RDF Data. *IEEE Transaction on Big Data*.

Benbernou, S., & Ouziri, M. (2017). Enhancing Data Quality by Cleaning Inconsistent Big RDF Data. IEEE Big Data conference, Boston, MA. doi:10.1109/BigData.2017.8257913

BenchSys. (2016). Retrieved from https://www.benchsys.com/

Bentounsi, M. (2015). *Business Process as a Service - BPaaS: Securing Data and Services* (PhD Thesis). Sorbonne Paris Cité - Université Paris Descartes, France.

Bentounsi, M., & Benbernou, S. (2016). Secure complex monitoring event processing. *NCA*, *2016*, 392–395.

Bentounsi, M., Benbernou, S., & Atallah, M. J. (2016). Security-aware Business Process as a Service by hiding provenance. *Computer Standards & Interfaces*, *44*, 220–233. doi:10.1016/j.csi.2015.08.011

Bhattacharya, I., & Getoor, L. (2017). Entity Resolution. Encyclopedia of Machine Learning and Data Mining, 402-408.

Brain Initiative. (n.d.). *What is the Brain Initiative?* Retrieved from https://www.braininitiative.nih.gov/

Bukhari, A. C., Krauthammer, M., & Baker, C. J. O. (2014). Sebi: An architecture for biomedical image discovery, interoperability and reusability based on semantic enrichment. *Proceedings of the 7th International Workshop on Semantic Web Applications and Tools for Life Sciences*.

Cao Huu, Q. (2017). *Policy-based usage control for trustworthy data sharing in smart cities* (PhD Thesis). Telecom & Management Sud, Paris, France.

Cell Ontology. (n.d.). Retrieved from https://bioportal.bioontology.org/ontologies/CL

Cheatham, M., Cruz, I. F., Euzenat, J., & Pesquita, C. (2017). Special issue on ontology and linked data matching. *Semantic Web*, *8*(2), 183–184. doi:10.3233/SW-160251

Chekol, M. W., & Pirrò, G. (2016). Containment of Expressive SPARQL Navigational Queries. *International Semantic Web Conference*.

CUMULUS. (n.d.). Retrieved from https://cumulus.parisdescartes.fr/

DBPedia. (n.d.). Retrieved from http://wiki.dbpedia.org/

Dong, X. L., & Srivastava, D. (2015). *Big Data Integration. Synthesis Lectures on Data Management*. Morgan & Claypool Publishers.

Données publiques. (n.d.). Retrieved from https://donneespubliques.meteofrance.fr/

Duong, T., Beck, G., Azzag, H., & Lebbah, M. (2016). Nearest neighbour estimators of density derivatives, with application to mean shift clustering. *Pattern Recognition Letters*, *80*, 224–230. doi:10.1016/j.patrec.2016.06.021

Eiben, C. B., Siegel, J. B., Bale, J. B., Cooper, S., Khatib, F., Shen, B. W., ... Baker, D. (2012). Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nature Biotechnology*, *30*(2), 190–192. doi:10.1038/nbt.2109 PMID:22267011

Eickhoff, C. (2014). Crowd-Powered Experts: Helping Surgeons Interpret Breast Cancer Images. *ECIR Workshop on Gamification for Information Retrieval*. 10.1145/2594776.2594788

Engle, C., Lupher, A., Xin, R., Zaharia, M., Franklin, M. J., Shenker, S., & Stoica, I. (2012): Shark: fast data analysis using coarse-grained distributed memory. *SIGMOD Conference*, 689-692. 10.1145/2213836.2213934

Firmani, D., Saha, B., & Srivastava, D. (2016). *Online Entity Resolution Using an Oracle*. PVLDB.

Foaf. (n.d.). Retrieved from http://www.foaf-project.org/

Fukunaga, K., & Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, *21*(1), 32–40. doi:10.1109/TIT.1975.1055330

Garcia Seco de Herrera, A., Foncubierta-Rodriguez, A., Markonis, D., Schaer, R., & Müller, H. (2014). Crowdsourcing for medical image classification. *Annual congress SGMI*.

Giovanni, T. (2012). *OpenNebula 3 Cloud Computing*. Packt Publishing Limited.

Grid5000: Home. (2017). Retrieved from https://www.grid5000.fr/mediawiki/index.php/Grid5000:Home

He, Y., Tan, H., Luo, W., Feng, S., & Fan, J. (2014). Mr-dbscan: A scalable mapreduce-based dbscan algorithm for heavily skewed data. *Frontiers of Computer Science*, *8*(1), 83–99. doi:10.100711704-013-3158-3

Heaton, J. (2016). Comparing Dataset Characteristics that Favor the Apriori, Eclat or FP-Growth Frequent Itemset Mining Algorithms. *Proceeding of the IEEE SoutheastCon*, 1-7. 10.1109/SECON.2016.7506659

Howe, B., Franklin, M. J., Haas, L. M., Kraska, T., & Ullman, J. D. (2017). *Data Science Education: We're Missing the Boat, Again*. ICDE.

Indyk, P., & Motwani, R. (1998). Approximate nearest neighbors: Towards removing the curse of dimensionality. *Annual ACM Symposium on Theory of Computing*. 10.1145/276698.276876

Irshad, H., Montaser-Kouhsari, L., Waltz, G., Bucur, O., Nowak, J. A., Dong, F., ... Beck, A. H. (2015). Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: Evaluating experts, automated methods, and the crowd. *Pacific Symposium on Biocomputing*, 294. PMID:25592590

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with mechanical turk. SIGCHI conference on human factors in computing systems, Florence, Italy. doi:10.1145/1357054.1357127

Krautheim, F. J. (2009). Private Virtual Infrastructure for Cloud Computing. *Proceedings of the 2009 conference on Hot topics in cloud computing HotCloud'09*, Article No. 5.

Kumar, A., Dyer, S., Kim, J., Li, C., Leong, P. H. W., Fulham, M. J., & Feng, D. (2016). Adapting content-based image retrieval techniques for the semantic annotation of medical images. *Computerized Medical Imaging and Graphics*, *49*, 37–45. doi:10.1016/j.compmedimag.2016.01.001 PMID:26890880

Kyung-Hoon, H., Haejun, L., & Duckjoo, C. (2012). *Medical Image Retrieval: Past and Present*. Healthc Inform Research.

Leifman, G., Swedish, T., Roesch, K., & Raskar, R. (2015). *Leveraging the Crowd for Annotation of Retinal Images*. EMBC. doi:10.1109/EMBC.2015.7320185

Lessl, M., Bryans, J. S., Richards, D., & Asadullah, K. (2011). Crowd sourcing in drug discovery. *Nature Reviews. Drug Discovery*, *10*(4), 241–242. doi:10.1038/nrd3412 PMID:21455221

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. doi:10.1109/TIT.1982.1056489

Luo, J., Wu, M., Gopukumar, D., & Zhao, Y. (2016). Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomedical Informatics Insights*, *8*, BII.S31559. doi:10.4137/BII.S31559 PMID:26843812

Mao, M., Peng, Y., & Spring, M. (2010). *An adaptive ontology mapping approach with neural network based constraint satisfaction*. J. Web Sem.

Mavandadi, S., Dimitrov, S., Feng, S., Yu, F., Sikora, U., Yaglidere, O., ... Ozcan, A. (2012). Distributed Medical Image Analysis and Diagnosis through Crowd-Sourced Games: A Malaria Case Study. *PLoS One*, *7*(5), e37245. doi:10.1371/journal.pone.0037245 PMID:22606353

Mayhew, A. (2001). File Distribution Efficiencies: cfengine Versus rsync. *Proceedings of the 15th Conference on Systems Administration LISA*, 273-276.

Müller, H., Michoux, N., Bandon, D., & Geissbuhler, A. (2004). A review of content-based image retrieval systems in medical applications. Clinical benefits and future directions. *International Journal of Medical Informatics*, *73*(1), 1–23. doi:10.1016/j.ijmedinf.2003.11.024 PMID:15036075

National Cancer Institute Thesaurus. (n.d.). Retrieved from https://bioportal.bioontology.org/ontologies/NCIT

Nentwig, M., Hartung, M., Ngomo, A. N., & Rahm, E. (2017). A survey of current Link Discovery frameworks. *Semantic Web*.

Nguyen, T. B., Wang, S., Anugu, V., Rose, N., McKenna, M., Petrick, N., ... Summers, R. M. (2012). Distributed human intelligence for colonic polyp classification in computer-aided detection for CT colonography. *Radiology*, *262*(3), 824–833. doi:10.1148/radiol.11110938 PMID:22274839

Qi, H., & Snyder, W. E. (1999). Content-based image retrieval in picture archiving and communications systems. *Journal of Digital Imaging*, *12*(S1), 81–83. doi:10.1007/BF03168763 PMID:10342174

Raddick, M. J., Bracey, G., Gay, L. G., Lintott, C. J., Cardamone, C., Murray, P., Schawinski, K., Szalay, A. S., & Vandenberq, J. (2013). *Galaxy Zoo: Motivations of Citizen Scientists*. Academic Press.

Radiology Lexicon. (n.d.). Retrieved from https://bioportal.bioontology.org/ontologies/RADLEX

Raghupathi, W., & Raghupathi, V. (2014). *Big data analytics in healthcare: promise and potential*. Health Information Science and Systems.

RosettaHub. (n.d.). Retrieved from http://www.rosettahub.com

Samovskiy, D. (2010). *The Rise of DevOps*. Retrieved from http://www.somic.org/2010/03/02/the-rise-of-devops/

Serrano, E., Blas, F.J.G., Carretero, J., & Desco, M. (2017). Medical Imaging Processing on a Big Data platform using Python: Experiences with Heterogeneous and Homogeneous Architectures. *IEEE/ACM CCGRID*, 830-837.

Shvaiko, P., & Euzenat, J. (2013). Ontology Matching: State of the Art and Future Challenges. *IEEE Trans. Knowl. Data Eng*.

Sis4web. (n.d.). Retrieved from http://www.sisncom.com/IMG/pdf/sis4web.pdf

Siemens. (2018). Retrieved from https://www.plm.automation.siemens.com/fr/products/teamcenter/

Silva, L. A., Costa, C., & Oliveira, J. L. (2012). A PACS archive architecture supported on cloud services. *International Journal of Computer Assisted Radiology and Surgery*, *7*(3), 349–358. doi:10.100711548-011-0625-x PMID:21678039

Slaney, M., & Casey, M. (2008). Locality-sensitive hashing for finding nearest neighbors. *IEEE Signal Processing Magazine*, *25*(2), 128–131. doi:10.1109/MSP.2007.914237

Sonntag, D., Wennerberg, P., & Zillner, S. (2010). Applications of an Ontology Engineering Methodology. *AAAI Spring Symposium: Linked Data Meets Artificial Intelligence*.

Stillwell, M., & Coutinho, J. G. F. (2015). A DevOps approach to integration of software components in an EU research project. *Proceedings of the 1st International Workshop on Quality-Aware DevOps*. 10.1145/2804371.2804372

University of Indiana. (n.d.). *Harp Project*. Retrieved from http://salsaproj.indiana.edu/harp/

Verhoosel, J. P. C., Bekkum, M. V., & Evert, F. V. (2015). Ontology matching for big data applications in the smart dairy farming domain. *International Semantic Web Conference*.

W3C. (2004). *World Wide Web Consortium Issues RDF and OWL Recommendations*. Retrieved from https://www.w3.org/2004/01/sws-pressrelease.html.en

Whang, S. E., Marmaros, D., & Garcia-Molina, H. (2013). Pay-as-you-go entity resolution. *IEEE Trans. Knowl. Data Eng*.

Wikipedia. (n.d.). *Bulk synchronous parallel*. Retrieved from https://en.wikipedia.org/wiki/Bulk_synchronous_parallel

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., ... Stoica, I. (2016). Apache Spark: A unified engine for big data processing. *Communications of the ACM*, *59*(11), 56–65. doi:10.1145/2934664

Zerlang, J. (2017). GDPR: A milestone in convergence for cyber-security and compliance. *Network Security*, *6*(6), 8–11. doi:10.1016/S1353-4858(17)30060-0

ENDNOTES

1 Virtual Network Computing.
2 Secure file transfer protocol.