From Data Islands to Sharing Data in the Cloud: the Evolution of Data Integration in Biological Data Repositories

Vinicius Vielmo Cogo[1], **Alysson Neves Bessani**[1]

[1]LaSIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal.

{vvcogo,anbessani}@ciencias.ulisboa.pt

Abstract. *Biological data repositories were often data islands with unharmonized formats, models, and protocols. Their integration evolved along the years and sharing data in multi-tenant infrastructures is a reality now. In this article, we illustrate this evolution by presenting real-world cases from the bioinformatics area and collect the best practices and current trends that future solutions should observe from these examples. Finally, we situate the platform being created by the BiobankCloud project in the scenario of integrating biological data.*

1. Introduction

Biological data repositories started by collecting and providing small public DNA sequences and related data to improve the scientific knowledge on genomics. However, they were isolated from each other and often overlapped. Their integration evolved along the years and several paradigm changes took place. For instance, the advent of the Next Generation Sequencing (NGS) reduced the costs of DNA sequencing exponentially in the recent years, which is increasing at the same pace the amount of data to be managed and stored [Marx 2013]. Sequencing the whole genome of a human being currently costs less than $1000, and the prices are expected to continue falling. Biological data repositories became responsible for storing also whole genomes instead of only small sequences.

Biobanks are repositories that store biological physical samples originally (e.g., in cryopreservation facilities), and are also becoming responsible for storing data about these samples. The availability of large sample collections accelerates medical breakthroughs, and researchers aim to analyze genomes from whole populations instead of from a few individuals [Muilu et al. 2007]. Fetching all genomes of interest to a private infrastructure (i.e., data shipping) before processing them is each time more impractical. Function shipping (i.e., sending the program to where data resides) is an efficient alternative.

In this article, we review the basic concepts of data integration (§2) and illustrate the evolution of integrating biological data repositories by presenting real-world cases from the bioinformatics area (§3). Additionally, we collect the best practices and current trends that future solutions should observe and learn from these examples (§4) and situate the platform being developed by the EU-funded BiobankCloud[1] project in the scenario of integrating biological data (§5).

2. Data Integration

Data integration roughly consists in providing a unified view and access of multiple data sets to users [Lenzerini 2002]. More specifically, it aims at (1) finding reliable

[1]http://biobankcloud.eu/

sources and pulling the needed information from the databases designed for this purpose and (2) understanding the power of data at large when many data sources are available [Brazhnik and Jones 2007]. Data from multiple sources can be heterogeneous, with different structure, formats, models, protocols, semantics, etc.

There are diverse integration approaches that consider the different ways to provide a global view of data. Earliest surveys [Hernandez and Kambhampati 2004, Stein 2003] defined three main models for data integration: data warehouse, mediator-based, and navigational integration. Another survey [Louie et al. 2007] separated mediator-based integration from federated databases and included a category for peer data management systems. Finally, a more recent survey [Mayer 2009] presented 9 categories of data integration, being those three (warehouse, mediator, and navigational) the main known data integration models, and considered other two (loosely coupled and semantic web) also noteworthy. Figure 1 contains a comparison on these five methods with an unintegrated scenario.

In the **unintegrated** case (Fig. 1(a)), clients must know the location of all data sources, query each one of them, and correlate the resulting data from each query to obtain the answer. Data sources do not exchange information in this scenario, which can lead to problems, such as, heterogeneous answers (unharmonized data), heterogeneous semantics, and duplicated answers.

The first data integration model considered in this article is the **data warehouse** (Fig. 1(b)) [Hernandez and Kambhampati 2004, Stein 2003, Louie et al. 2007, Mayer 2009]. This method retrieves, harmonizes, and stores data from multiple remote sources into a local central storage. A client needs to issue only one query to the data warehouse to receive an harmonized, deduplicated answer. This system does not rely on the network to access data during a query, since it previously fetched all data to the local storage—the dashed arrows marked with an *R* in Fig. 1(b). Furthermore, it avoids network bottlenecks, slow response times, and occasional unavailability of remote data sources. Queries can easily be optimized in execution time since there is only one local data repository. It also allows users to filter, validate, modify, and annotate data obtained from multiple sources. The main problems of this model are the cost of maintaining such system and the possible lack of freshness when accessing out-of-date information stored locally. Due to this second issue, warehouses need to regularly check the underlying sources for new or updated data to reflect them on the local copy.

The second data integration model we analyze is the **mediator-based** (Fig. 1(c)) [Hernandez and Kambhampati 2004, Stein 2003, Louie et al. 2007]. This method considers the existence of a mediator that maps each client query into a set of specific queries for the underlying sources at runtime and aggregate the replies in a single answer to the client. The main advantages of this model are that it does not need a large centralized storage system (it has a lower maintenance cost than warehouses) and the data is always up-to-date (it does not need the synchronization step as the previous model). The disadvantages are that it relies on the network to access data on-demand (the overall performance is equals to the slowest source) and optimizing queries for all external data sources is difficult. Federated databases [Louie et al. 2007, Mayer 2009] are a smooth case of mediator-based model since data sources collaborate among themselves.

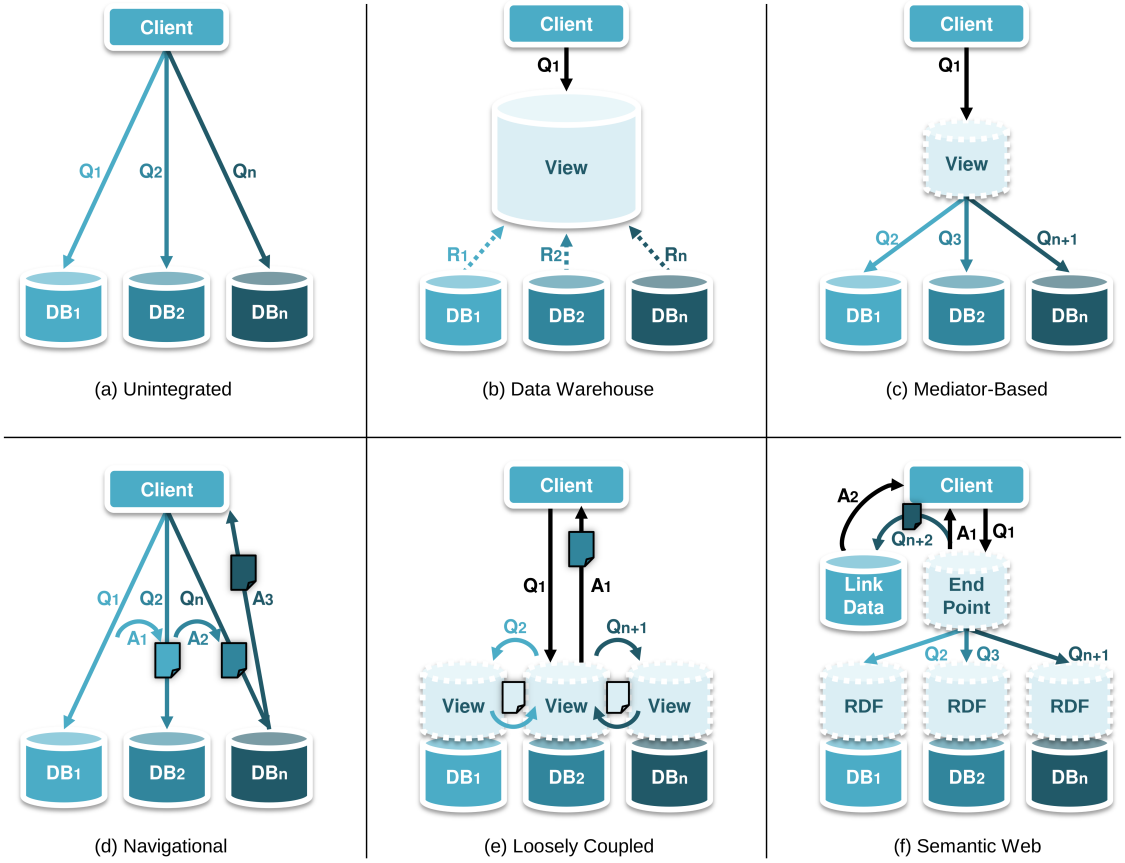

Figure 1. Comparison of an unintegrated scenario with five integration methods.

The **navigational integration** (Fig. 1(d)) [Hernandez and Kambhampati 2004, Stein 2003, Mayer 2009] provides interactions between users and pages similarly to a point-and-click web navigation. It is also known as link-based integration and relies in cross-references between services to allow users to navigate from one page to another in a different service. Workflows running over this method redirects the output of one service to the input of the service responsible for the subsequent workflow step. User queries are translated to path expressions that result in reaching pages containing the desired information and is only reachable through this particular path. Each data source is a set of pages with interconnections among them and specific entry-points. The main disadvantages of this method are the risks of ambiguous answers, broken links, and the fact that the onus of integration and interpretation is on the user side [Stein 2003].

The **loosely coupled** method (Fig. 1(e)) [Louie et al. 2007, Mayer 2009] integrates data from multiple sources using the minimal amount of knowledge as possible about each one of them. This method is also known as integration by peer management systems and normally uses flexible file formats such as XML or JSON. Each network contains a minimal mediated schema that represents its semantic knowledge. New sources entering in the network are mapped to provide information using this mediated schema. Internal modifications on each source are reflected in this mapping, making them still compliant with the mediated schema. Loosely coupled systems normally use a minimalistic approach to integrate diverse databases by mapping to only basic data types and

using only modest adaptations of existing web resources. A client issues requests to any network member, which are forwarded to other peers that will aggregate information to create the final answers.

The integration method based on **semantic web** (Fig. 1(f)) [Sheth 1999, Mayer 2009] allows the integration of semantically related information, regardless of distribution and heterogeneity. Clients issue queries to end points (e.g., a SPARQL engine) that launch queries to databases through languages for data representation (e.g., RDF—Resource Description Framework). The usage of languages for data representation allows the integration of very different data sources, for instance a web-server, a relational database, and a file server. RDF describes data in triplets containing: the subject (resource identifier), the predicate (an attribute name), and the object (the attribute value). The answer to each request from clients can contain diverse types of data representing the object, from simple numbers or ontology terms to unstructured data files. Linking data received from queries is one of the most important steps when integrating data through semantic web, because it is when complementary data are correlated to make sense for several applications.

3. Real-World Cases of Data Integration in Bioinformatics

There are several initiatives that aim to present an integrated view, at the metadata level, of samples available in several biobanks [Norlin et al. 2012, Müller et al. 2015]. In this section, we discuss the approaches in use to integrate real-world biological repositories at the data level.

We present one of the first integration initiatives for public DNA sequences (§3.1); an initiative to integrate biobanks from a country (§3.2); an initiative to integrate biobanks from a continent (§3.3); cases that integrate repositories with similar missions (§3.4); virtual biobanks (§3.5); and cloud-based initiatives for biological data integration (§3.6).

3.1. INSDC: A Pioneer in Integrating Public DNA Sequences

The INSDC—International Nucleotide Sequence Database Collaboration (`http://www.insdc.org/`) is a joint effort to collect and provide a globally comprehensive compilation of public domain nucleotide sequences and associated metadata [Cochrane et al. 2011]. Three organizations lead this initiative for more than three decades: the DNA Data Bank of Japan (DDBJ), the National Center for Biotechnology Information (NCBI), and the European Bioinformatics Institute (EBI).

When submitting a sequence to this system, a researcher must send the sequence to only one of the INSDC partners, which will be the authority for this sequence and will reserve a new *accession number* in the shared accessioning system. The three partners synchronize their databases daily by forwarding all new, modified, or removed entries to the other partners. As a consequence, INSDC can be seen as a geographically replicated repository deployed in three federated databases, since the data model and relationships among entities are exactly the same in all partners.

There is no need for a specific shared database or file system since INSDC relies on the universal accessioning name space and on a mutually understood presentation format for each type of data. The concurrency control is done through the shared acces-

sioning system at the allocation time. However, duplicated entries are not detected if the same sequence is submitted to more than one partner.

INSDC provides versioning of the entire database by allowing users to download it through the Sequence Version Archive service (`http://www.ebi.ac.uk/cgi-bin/sva/sva.pl`) and by scripts for updating local instances daily. Finally, the infrastructure allows an entry to be removed from their database, however it does not provide guarantees that the entry will disappear from everywhere, since other people may already have downloaded local copies from the database. The main limitation of INSDC is that it does not deal with private sequences: everything submitted and manipulated by the system is considered to be public.

3.2. UK Biobank: Integrating Biobanks from a Country

The UK Biobank (`http://www.ukbiobank.ac.uk/`) is a collaborative research project to recruit and follow longitudinally the health of 500 000 volunteers aged between 40–69 years from the UK [Ollier et al. 2005]. Collected physical samples are transported to a central site for processing and are stored on two geographically separate cryopreservation facilities. This infrastructure also stores the data sets associated with such physical samples. The data in UK Biobank is divided into three main categories: protected, managed, and open resources. The first contains participants' health data (including DNA) and medical records, the second one contains non-sensitive material that still needs to be protected for scientific and ethical reasons, and the latter can be freely available. DNA is sequenced from stored blood samples by a private company and returned to the UK Biobank.

Integration between entities occurs at the sample level since physical samples are sent to a central biobank right after their collection. There is also a central data repository that stores all data sets locally. Additionally, UK Biobank use a data warehouse model to integrate its local data with other national health systems. Researchers propose studies to obtain access to data, which must be approved by a council. There are costs involving this access, from proposal analysis fees to separated values for accessing only data or data and physical samples.

3.3. BBMRI and ELIXIR: Integrating Biobanks from a Continent

The BBMRI—Biobanking and Biomolecular Research Infrastructure (`http://bbmri.eu/`) is a pan-European research infrastructure aiming to improve the accessibility and interoperability of the existing collections of biological samples from different European populations. The preparatory phase of this project ended in 2011 with a joint knowledge about 311 European biobanks and more than 1.8 million DNA samples.

The next phase is the BBMRI-ERIC (European Research Infrastructure Consortium), which integrates all these resources into a hub-and-spoke [Muilu et al. 2007] network properly embedded into the European scientific, ethical, legal, and societal frameworks. The hub-and-spoke model consists in creating or choosing a central resource and connecting all others to it instead of creating one connection between each resource pair. BBMRI-ERIC intends to create a network of hub-and-spoke instances, by choosing major nodes as hubs in behalf of a region or a country, and local biobanks acting as end nodes (spokes). A researcher sends queries to one of the global portals that is connected to all

hubs, and consequently to all spokes. Each hub receives the query and forwards it to the end nodes of its responsibility. Each spoke provides the local biobank service that returns the result of those issued queries over the local database. Hubs aggregate these answers and present a final, integrated answer in the global portal. An integration prototype is already implemented in test-instances of existing biobanks.

ELIXIR (`http://www.elixir-europe.org/`) is an European project that aims to build and operate a sustainable infrastructure for biological information in Europe. It intends to support life-science research and its translation to medicine and environment, bio-industries and society [Crosswell and Thornton 2012]. Information gathered on data access and user requirements was central in designing ELIXIR as a distributed infrastructure with a central hub. The hub is connected to ELIXIR nodes (spokes) hosted at centers of excellence in universities and institutions across Europe. ELIXIR is expected to integrate access and data from 12 research infrastructures being funded by European Commission, including the BBMRI.

3.4. GenomEUtwin: Integrating Biobanks with Similar Missions

The GenomEUtwin [Litton et al. 2003, Muilu et al. 2007] is an international collaboration between eight repositories providing information about more than 600 000 human twins pairs. They propose the TwinNET, a federation of local data warehouses, combined with a global mediator that provides transparent access to them through database instances and the use of DiscoveryLink. The IBM DiscoveryLink is a database middleware that extracts data from multiple sources in response to a single query [Haas et al. 2001]. This system's architecture is also a hub-and-spoke, where the hub is the integration node and spokes are data-provisioning centers. Connections between hub and spokes are made using VPN tunnels, which are initiated from the spokes. Each spoke (data provider) contains a local data warehouse, which is fed with harmonized data from local production databases and LIMS—Laboratory Information Management System. Data is translated and transferred into this data warehouse (called TwinMart) located in a demilitarized zone within each spoke of TwinNET. Each subject receives a unique GenomEUtwin identifier and twins share portions of it. All databases and data sets maintained under the TwinNET are anonymous, where the only allowed identification is the GenomEUtwin identifier.

Some advantages of this model are the opportunity for query optimization and the transparent access to data. One major weakness of this approach is the enormous amount of money that must be invested before information can be queried and retrieved. Other problem is that partners must increase the maintenance cost of their local infrastructures since they need to control one more component, the TwinMart (a local data warehouse).

3.5. COMMIT: Integrating Classical Biobanks in Virtual Biobanks

A virtual biobank is a repository that provides data obtained by means of characterization and sequencing from samples stored in classical biobanks. The COMMIT project (`http://www.amolf.nl/research/bims/research-activities/e-biobanking/`) is a virtual tissue biobank that provides access to digital information about physical samples, which are harder to manipulate. In the specific case of COMMIT, the digital information are data and image sets obtained from mass spectrometry and tissue microarray experiments. These data sets are an important input to proteomics workloads, for instance, the discovery of amino acids composing a protein and the

validation of protein folding predictions. Additionally, the physical samples are breast cancer sections collected from medical institutions integrating COMMIT's partners. The fact that these samples are extracted from specific cases or diseases characterizes this initiative also as an example of integration of biobanks with similar missions. The infrastructure comprises a central repository for physical and digital samples in one project partner. Data is accessible from exterior through a web portal, after researchers being authorized through formal bureaucrat agreement protocols. The COMMIT project provide a workflow based management system and distributed processing resources. One may consider the goal of this project similar to the idea of storing and providing genome sequencing files (the raw input format for many bioinformatics workflows) instead of sequencing the entire genome each time one wants to execute an experiment [Verissimo and Bessani 2013].

3.6. Cloud-Based Initiatives: DNAnexus, BaseSpace, and Galaxy

This section analyzes three cloud-based solutions for storing and processing biological data: DNAnexus (`https://dnanexus.com/`), BaseSpace (`https://basespace.illumina.com/`), and Galaxy (`http://galaxyproject.org/`). We group these systems together because they are similar to each other: they are implemented in (mostly public) cloud-infrastructures and assume users can create a virtual infrastructure to store and process their data. Notice there is a change of paradigm here. Instead of downloading data sets, working on them locally, and producing new data sets to be inserted in shared infrastructures (e.g., UK Biobank or INSDC), these cloud infrastructures promote storing the data in the public cloud, and processing also there close to the data [Marx 2013]. This is done by using high-level tools (e.g., operated by web interfaces) or the usual tools deployed in cloud virtual machines.

DNAnexus is a startup that provides a cloud-based platform for genomic enterprises to expand their local infrastructures. It is an API-based infrastructure and a workflow-based tool. The entire solution is deployed over the Amazon Web Services (AWS). DNAnexus addresses some security concerns during data transfer and storage, namely: encrypted communication and storage, accountability, and two-factor authentication.

BaseSpace is a product from Illumina that allows their clients to directly connect their sequencing machines with the cloud. The idea is to avoid the need of a local IT infrastructure at the client side to store and analyze genomic data. A noteworthy aspect of BaseSpace is to provide an ecosystem in which third-party developers can create new tools for BaseSpace users, in a similar way to what is provided by Amazon in its market place. The BaseSpace also addresses some security concerns during data transfer and storage, namely: encrypted communication and storage and accountability.

Galaxy is an open-source software package that provides three possible usages: a free web-based service, a public cloud deployment, and a private cloud deployment [Goecks et al. 2010]. Such flexibility is a consequence from the fact that Galaxy is not a company or a cloud-based service, and is a software package to be deployed in physical or virtual machines. In the public cloud-based deployment, Galaxy provides a wizard for installing it on Amazon EC2 from AWS. A possible security threat is that users must provide their AWS credentials to deploy a cluster through the wizard. An

advantage of creating your own cluster is that you can increase or reduce the number of computing instances running your installation, as well as persistently terminate and relaunch the cluster. This solution is based on an Infrastructure-as-a-Service scenario since users determine how many resources are allocated to run their workloads.

Additionally, one may customize Galaxy to create its own instance. An example is the e-BioGrid project (`http://www.e-biogrid.nl/`), which supports life-science research through the preparation and maintenance of computing environments running over the BigGrid. The BigGrid is a national computing infrastructure created to support the execution of research experiments from Netherlands. Other countries also have similar infrastructures, for instance, the Grid'5000 (`http://www.grid5000.fr/`) from France. The main contribution of e-BioGrid project is that they provide specific computing environments that are functional by default for predefined experiments and studies. The custom Galaxy running over BigGrid is an example of these environments.

Comparison. Table 1 presents a comparison among the three cloud-based systems discussed in this section. The focus of these services is on providing a complete platform for managing and analyzing sequencing data rather than data integration. All three solutions are deployed in Amazon AWS, which means that they rely on a single point of failure and are not immune to vendor lock-in issues due to raises on prices or changes in policies. The table also includes a column about Amazon AWS for the sake of comparison with a public cloud provider.

Table 1. Comparison of the three cloud initiatives for storing and analyzing biological sequences when deployed in Amazon AWS.

	BaseSpace	DNAnexus	Galaxy	Amazon AWS
Goal	Cloud-based platform	Cloud-based platform	Cloud-based platform	Cloud provider
Type	Product by Illumina	Startup company	Academic project	Service
Filosophy	Proprietary	Proprietary	Open-source	Hybrid
Public cloud provider	Amazon AWS	Amazon AWS	Amazon AWS	—
Storage cost 1 TB	Free	$37.5/month	Same as AWS	$30/month
10 TB	$1500/month	$375/month	Same as AWS	$295.5/month
Processing cost (min-max)	Depend on the tool	$0.19 – $5.06/hour	Same as AWS	$0.013 – $5.52/hour
Predefined workloads	Yes	Yes	Yes	—
Custom scripts/workloads	Yes	Yes	Yes	Yes
Share data/workloads	Yes	Yes	Yes	Yes

By analyzing this table, we can infer that using BaseSpace and DNAnexus have a cost increase (e.g., 407% and 26.9% for storage) when compared with deploying bioinformatics tools directly on Amazon AWS or using Galaxy. However, most end-users of sequence analysis are biologists and biochemists who normally do not have enough expertise to build such infrastructure directly.

The table also shows that all infrastructures are quite flexible in supporting custom workloads, scripts, and data sharing. In this sense, these systems are leading the migration of computing to where data is [Marx 2013] and avoiding expensive local IT infrastructures. One of the main advantages of cloud-based solutions is that there is no need for a manual and time-consuming data-transfer step each time a workload is executed. After uploading the data to the cloud once, it is already up in there, accessible from anywhere.

4. Best Practices and Current Trends

Table 2 presents the main characteristics of the initiatives we discussed in the previous sections. More specifically, we compare the integration method employed in these systems (see Section 3), how the stored data can be accessed, if the managed data sets are public, private, or controlled, and if dependability measurements are employed in these systems.

Table 2. Comparison of the integration initiatives discussed in this article.

Initiative	Integration Method	Accessibility	Data Sets	Dependability
INSDC	Data warehouse	Web	Public*	3 replicas
UK Biobank	Data warehouse	Web	Private	In-site
BBMRI and ELIXIR	Mediator-based	Web	Public	In-site
GenoEUtwin	Mediator-based	Web	Private	In-site
COMMIT	Data warehouse	Web	Private	In-site
Cloud-based initiatives	Data warehouse	Web + API	Controlled**	AWS-based

* Data sets can be kept private temporarily until a paper publication. ** Even private data sets are accessible to the cloud provider.

This table shows that all initiatives employ the data warehouse approach ot the mediator approach. All in all, some interesting trends were observable through the analysis of these systems. An integrated system for attributing accessioning numbers is very important, but at this point it appears there is no evolved protection against duplicate entries, sequences, or individuals. All systems make their data sets available through web portals, that can be either freely accessible (e.g., if the data is public) or implement authentication and access control mechanisms to give access to certain data sets only to authorized users.

All mediator-based systems devise dependability mechanisms only in the end points (in-site). INSDC replicates all data in three globally distributed replicas, while UK Biobank uses two facilities to store only physical samples. Cloud-based solutions rely on the transparent replication and recovery mechanisms offered by the AWS. All cloud-based solutions focus on approaching computation to where data is located, which can bring dramatic performance improvements since data sets upload, download, and even normalization can be avoided.

5. A Hybrid Approach in the BionankCloud PaaS

Function shipping reduces the execution time of processing large quantities of data by approximating the computation to where data resides. On the other hand, accessing and fetching data from several sources accelerates medical breakthroughs by increasing the diversity of the processed sample collection. Balancing the levels of data and function shipping is a challenge in processing distributed big data, and is a key element in the Platform-as-a-Service (PaaS) being developed by the EU-funded BiobankCloud project.

The BiobankCloud PaaS is an open-source platform on top of Apache YARN that can be deployed in any private infrastructure (e.g., a biobank) for the secure storage, sharing, and parallel processing of genomic data [Bessani et al. 2015]. It integrates the data warehouse and mediator-based integration models, as well as provides data and function shipping to its users. The platform computes data available in a single cluster (i.e., a data warehouse), and allows function shipping through a scalable scientific workflow

engine [Bux et al. 2015] and a workflow description language [Brandt et al. 2015]. Additionally, the platform can lookup and fetch data from other biobanks automatically on execution time (i.e., a mediator).

Since the to-be-stored data can be huge, the platform allows biobanks to extrapolate their capacity by securely storing data in a cloud-of-clouds [Bessani et al. 2013]. The BiobankCloud PaaS is the first solution for biological data that uses multiple clouds to securely store private data in these multi-tenant infrastructures. It is secure and reliable because data is encrypted before its transfer and no single cloud stores the whole data set due to a secret sharing scheme used in our platform.

Most components of the BiobankCloud platform were already implemented and the current task force is focused on integrating them [Bessani et al. 2015]. Future work encompasses adding mechanisms for improved privacy-protection [Cogo et al. 2015] and compression of biological data [Alves et al. 2015] to strengthen the overall security and efficiency of the system.

6. Conclusion

In this article, we reviewed the basic concepts of data integration and presented the state-of-the-art in integrating biological data repositories through examples of initiatives from the bioinformatics area. We compiled the best practices and current trends that future solutions should observe and learn from the presented initiatives. Finally, we situated the BiobankCloud PaaS in the scenario of integrating biological data. More details on this platform can be found on the project's website and in a recently published joint paper [Bessani et al. 2015].

Acknowledgments. This work was partially supported by Fundação para a Ciência e a Tecnologia–PT, through the LaSIGE (PEst-OE/EEI/UI0408/2014), and by EU FP7, through the BiobankCloud project (ICT-317871).

References

Alves, F., Cogo, V. V., Wandelt, S., Leser, U., and Bessani, A. (2015). On-demand indexing for referential compression of DNA sequences. *PLoS ONE*, 10(7):e0132460.

Bessani, A. et al. (2013). DepSky: Dependable and secure storage in cloud-of-clouds. *ACM Transactions on Storage*, 9(4).

Bessani, A. et al. (2015). BiobankCloud: a platform for the secure storage, sharing, and processing of large biomedical data sets. In *the First International Workshop on Data Management and Analytics for Medicine and Healthcare (DMAH 2015)*.

Brandt, J., Bux, M., and Leser, U. (2015). Cuneiform: a functional language for large scale scientific data analysis. In *Proceedings of the Workshops of the EDBT/ICDT, vol. 1330*, pages 7–16.

Brazhnik, O. and Jones, J. F. (2007). Anatomy of data integration. *J. Biomed. Inform.*, 40(3):252–269.

Bux, M., Brandt, J., Lipka, C., Hakimzadeh, K., Dowling, J., and Leser, U. (2015). SAASFEE: Scalable scientific workflow execution engine. *Proceedings of the VLDB Endowment*, 8(12).

Cochrane, G., Karsch-Mizrachi, I., and Nakamura, Y. (2011). The international nucleotide sequence database collaboration. *Nucleic Acids Res.*, 39(suppl 1):D15–D18.

Cogo, V. V., Bessani, A., Couto, F. M., and Verissimo, P. (2015). A high-throughput method to detect privacy-sensitive human genomic data. In *Proc. of the Workshop on Privacy in the Electronic Society (WPES 2015)*.

Crosswell, L. C. and Thornton, J. M. (2012). ELIXIR: a distributed infrastructure for European biological data. *Trends Biotechnol.*, 30(5):241–242.

Goecks, J., Nekrutenko, A., Taylor, J., et al. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8):R86.

Haas, L. M. et al. (2001). DiscoveryLink: A system for integrated access to life sciences data sources. *IBM Syst. J.*, 40(2):489–511.

Hernandez, T. and Kambhampati, S. (2004). Integration of biological sources: current systems and challenges ahead. *SIGMOD Rec.*, 33(3):51–60.

Lenzerini, M. (2002). Data integration: A theoretical perspective. In *Proc. of the 21st PODS*, pages 233–246. ACM.

Litton, J.-E. et al. (2003). Data modeling and data communication in GenomEUtwin. *Twin Res.*, 6(5):383–390.

Louie, B. et al. (2007). Data integration and genomic medicine. *J. Biomed. Inform.*, 40(1):5–16.

Marx, V. (2013). Biology: The big challenges of big data. *Nature*, 498(7453):255–260.

Mayer, G. (2009). Data management in systems biology I - overview and bibliography. *CoRR*, abs/0908.0411.

Muilu, J., Peltonen, L., and Litton, J.-E. (2007). The federated database—a basis for biobank-based post-genome studies, integrating phenome and genome data from 600 000 twin pairs in Europe. *Eur. J. Hum. Genet.*, 15(7):718–723.

Müller, H. et al. (2015). State-of-the-art and future challenges in the integration of biobank catalogues. In *Smart Health*, pages 261–273. Springer.

Norlin, L. et al. (2012). A minimum data set for sharing biobank samples, information, and data: MIABIS. *Biopreserv. Biobank*, 10(4):343–348.

Ollier, W., Sprosen, T., and Peakman, T. (2005). UK Biobank: from concept to reality. *Pharmacogenomics J.*, 6(6):639–646.

Sheth, A. P. (1999). Changing focus on interoperability in information systems: from system, syntax, structure to semantics. In *Interoperating geographic information systems*, pages 5–29. Springer.

Stein, L. D. (2003). Integrating biological databases. *Nat. Rev. Genet.*, 4(5):337–345.

Verissimo, P. E. and Bessani, A. (2013). E-biobanking: What have you done to my cell samples? *IEEE Security&Privacy*, 11(6):62–65.