



D2.2

Initial infrastructure framework

Project number	826121
Project acronym	iPC
Project title	individualizedPaediatricCure: Cloud-based virtual-patient models for precision paediatric oncology
Start date of the project	1 st January, 2019
Duration	48 months
Programme	H2020-SC1-DTH-2018-1

Deliverable type	Demonstrator
Deliverable reference number	SC1-DTH-07-826121 / D2.2 / 1.0
Work package contributing to the deliverable	WP2
Due date	September 2020 – M21
Actual submission date	14 th October 2020 – M22

Responsible organisation	Barcelona Supercomputing Center (BSC)
Editor	Salvador Capella-Gutierrez
Dissemination level	PU
Revision	1.0

Abstract	An initial demonstrator of the iPC infrastructure is reviewed. The platform's architecture is based on modules, which allow parallel developments and integration of different open source-based software components. This allows us to leverage other efforts and contribute towards its sustainability and maintainability. The release of a minimum viable platform is allowing us to capture early feedback from researchers at iPC.
Keywords	data catalogue; cloud computing; single sign-on; virtual research environment



Editor

Salvador Capella-Gutierrez (BSC)

Contributors (ordered according to beneficiary numbers)

Alejandro Canosa (BSC)

Laia Codó (BSC)

Elena de la Calle (BSC)

José María Fernandez (BSC)

Laura Rodriguez (BSC)

Jolanda Modic (XLAB)

Martina Truskaller (TEC)

Disclaimer

The information in this document is provided “as is”, and no guarantee or warranty is given that the information is fit for any particular purpose. The content of this document reflects only the author’s view – the European Commission is not responsible for any use that may be made of the information it contains. The users use the information at their sole risk and liability.

Executive Summary

The iPC project aims to establish an integrated computational environment for data sharing and analysis to enable paediatric cancer research. Thus, the platform is an essential component to integrate diverse and heterogeneous data, generate ML-oriented models and facilitate their use for treating new patients.

An initial demonstrator of the infrastructure is reviewed. The platform's architecture is based on modules, which allow parallel developments and integration of different open source-based software components. This allows us to leverage other efforts and contribute towards the platform's sustainability and maintainability. The release of a minimum viable platform is allowing us to capture early feedback from researchers at iPC to ensure the expected functionality is implemented.

iPC platform components are listed below:

- An Access portal, which allows users to access different platform resources from one single place and one single login based on OpenIDConnect.
- A Data Catalogue to visualize, search, filter, and select all data managed by the project. Data sources include either newly produced data and data from existing repositories such as Kids-First, R2, dbGaP.
- Different computational platforms such as the openVRE (open Virtual Research Environment) and Cavatica, that will combine controlled access to data, analysis and visualization tools.

The iPC platform is based on cloud technologies provided by BSC and will connect in the future to the high performance computing (HPC) capacities of iPC partners. HPC will facilitate the training and adjustment of highly computational intense machine learning (ML) models for paediatric cancer research.

Table of Content

Chapter 1	Introduction	1
1.1.	Existing efforts on Cancer Research	2
Chapter 2	Initial computational framework	4
Chapter 3	Platform components.....	6
3.1.	Access portal.....	6
3.2.	Data Catalogue	7
3.1.1	Catalogue Portal.....	7
3.1.2	iPC Data Model for managing samples metadata	8
3.3.	Data analysis platforms	9
3.3.1	Cavatica	9
3.3.2	Open Virtual Research Environment (openVRE)	10
3.3.3	Exemplary analysis flow for iPC researchers	11
3.3.4	Sharing CWL-based workflows across iPC workbenches	13
3.4.	Data Storage	14
4	Summary and Conclusion	16
5	List of Abbreviations	17
6	Bibliography.....	18

List of Figures

Figure 1. iPC platform design. Components and interoperability channels in yellow will be developed in the next iteration. At least one version of the green components have been released while the gray boxes are external developments that have been identified as relevant to the platform.....	5
Figure 2. iPC Access portal which serves as a starting point to the different internal and external platforms used in the consortium.....	6
Figure 3. iPC login page. The platform supports different OIDC identity providers.	6
Figure 4. iPC Catalogue portal for searching, filtering, and selecting metadata through Arranger.	7
Figure 5. iPC Catalogue portal data management section: Expose data for analysis through the iPC Outbox REST API.....	8
Figure 6. iPC Catalogue outbox. iPC Data Catalogue Management section is shown on the left side, where a user chooses to expose three datasets to analyse them at the openVRE . Then, as shown on the right side of the figure, associated datasets metadata is displayed at the openVRE and the datasets are imported with a simple click.....	11
Figure 7. User's workspace on the VRE. Selecting DoRothea tool on the VRE workspace.	12
Figure 8: Running Dorothea on the VRE.....	12
Figure 9: Visualizing Dorothea's results on the VRE. Data is displayed using one the visualizers integrated at the VRE: an R-Shiny application.....	13
Figure 10. <i>Configuration screen for sample CWL tool.</i>	14
Figure 11. BSC's Nextcloud service for iPC: Upload/download data on a user-friendly interface.	15

List of Tables

Table 1: Initial iPC data model for samples management.....	9
---	---

Chapter 1 Introduction

Unlike the popular perspective and as defended by many (and pointed out by the North American [National Cancer Institute](#), NCI), **cancer** is not one but rather a collection of related diseases, usually known as types of cancer. They share the fact that, when sick, “[some of the body's cells begin to divide without stopping and spread into surrounding tissues](#).” All cancer types negatively affect the quality and the lifestyle of the patient and most are still very lethal. In 2018 alone, around [9.6 million people around the globe died from cancer](#) - 1 person every 3.3 seconds. In Europe, the cancer toll in 2018 was almost [2 million people](#). Cancer is also a great source of research problems in the biomedical domain, driving some of the biggest achievements in health during the XXI century.

Cancer affects everyone. People from different backgrounds, of different races, of both genders, and of all ages - even our youngest. Luckily, cancer is relatively rare among kids. Nevertheless, it is still a [major cause of death](#) (in some regions even the leading cause of death) in **children worldwide**. Each year, across the globe, [approximately 200.000 children from 0 to 14 years old are diagnosed with cancer and around 75.000 children with cancer die](#).

The majority of [paediatric cancers are different from cancer in adults](#). Children's bodies work in a unique way and so the type of the developed cancer, how far and how fast it spreads throughout the body, and how it responds to treatment is often **different than in adults**. Another differentiating factor is the cause. In adults, cancer is often a consequence of some environmental or lifestyle factor, like poor diet or smoking. But this is not the case with kids. Many times, the cause of a childhood cancer is not known. So studying the disease and preventing it is very challenging.

Paediatric cancer is rare. That's great! But it also means that no one hospital and no one research center in the world has **enough data** on a specific cancer type or subtype to perform meaningful research. To really understand the disease, draw the right conclusions, and win the cancer fight, researchers and clinicians need access to as much data from as many patients in as many institutions as possible.

The lack of data availability is tightly connected to a couple of other issues. Technology and regulation. Different organisations use different information systems and **different standards** for managing their data. So, whenever institutions want to share data sources, they usually encounter some [interoperability problem](#). The other problem they usually bump into are [legal barriers](#). Increasingly stricter **privacy** and data protection laws are raising barriers to collecting and sharing patients' data. Additionally, due to complex procedures and various policies that are in place in hospitals and organisations across different countries, sharing datasets usually comes with an incredible amount of bureaucracy.

The individualized paediatric cure (iPC) project is addressing the need to *gather, harmonize, and share high-quality, multi-disciplinary paediatric cancer data* - all this while simplifying data access procedures and at the same time ensuring compliance to various security policies and data protection laws. Making these resources easily findable and broadly available will support a wide community of cancer research experts in creating effective personalised therapies for kids with cancer. To this end, the consortium is building a cloud-based platform for storing, accessing, analysing, and sharing data and disseminating results relevant for paediatric cancer research. The platform will host newly produced datasets and will also link to other similar repositories and platforms hosting paediatric cancer data and offering data analytics and data visualisation tools. This connected data infrastructure will simplify sharing of resources from multiple locations and will make data work better for researchers and clinicians around the globe.

As paediatric cancers are rare, and despite the differences with the adulthood ones, it is important to examine the major past and current efforts in cancer research as a reference learning path. It is also important to review existing efforts to facilitate controlled data access to cancer-related

data sets. For instance, the KidsFirst initiative is an exemplary effort on how to push forward paediatric cancer research.

1.1. Existing efforts on Cancer Research

International Cancer Genome Consortium (ICGC)

The International Cancer Genome Consortium¹ (ICGC) is a global initiative to present an open access catalog of genomic data on the most common types of tumors, with the ultimate goal of developing new and improved screening and diagnosis tools, as well as personalized treatments. The ICGC Data Portal contains 22,330 donors with molecular data from 86 cancer projects, and allows an advanced search based on donor, specimen, genetic and / or mutations data and its associated metadata.

The ARGO project² (Accelerating Research in Genomic Oncology) within ICGC represents the next phase of the consortium to translate genomic knowledge into new methods for the benefit of cancer patients. This initiative is the product of summarizing many efforts to increase the number of donors compared to the initial ICGC project, with the aim of reaching one hundred thousand cancer patients participating in clinical trials. In this context, the consortium intends to develop a new portal, not available at the moment, that encompasses its predecessor in ICGC with the support of the Global Alliance for Genomics and Health (GA4GH).

Gabriella Miller Kids First pediatric research program

The Kids First Data Resource Center³ is a collaborative pediatric research effort using large genomic data sets to develop new precision medicine-based approaches for children diagnosed with cancer and / or structural birth defects. The DRC presents molecular data from over 19 thousand samples from more than 13 thousand patients within 19 pediatric cancer studies. Through this initiative, the platform allows to find, combine and compare data for cross-disease analyses to identify the genetic pathways that underlie childhood cancer. The analyses can be carried out locally or in Cavatica by easily sending the selected data into the environment's workspace.

The Cancer Genome Atlas (TCGA) and its Pan-Cancer project

The Pan-Cancer project, hosted by The Cancer Genome Atlas⁴ (TCGA) consortium is an initiative for observing common alterations between different lineages of tumour to pursue better treatments based on precision medicine, by designing effective therapies for a given type of cancer and translating them into other cancer types presenting similar genomic profiles. The project is based on the knowledge obtained from the analysis of over 11,000 tumors from 33 of the most prevalent forms of cancer.

Database of Genotypes and Phenotypes (dbGAP)

dbGAP⁵ is a public repository aimed to archive, curate and distribute controlled data produced by studies focused on the investigation of the interactions between genotypes and corresponding phenotypes.

¹ ICGC, <https://icgc.org/>

² ICGC ARGO, <https://www.icgc-argo.org/>

³ Kids First Data Resource Center, <https://kidsfirstdrc.org/>

⁴ TCGA, <https://gdc.cancer.gov/>

⁵ dbGAP, <https://www.ncbi.nlm.nih.gov/gap/>

European Genome-phenome Archive (EGA)

EGA⁶ is a long-term repository with controlled and secure access to store a wide range of data associated with all types of diseases. More than half of the data that EGA stores are related to cancer and constitute a very valuable information repository for the development of cancer research. In order to protect sensitive data, EGA follows strict protocols for data management and storage, and grants authorized accesses under certain terms and policies, in collaboration with data providers, to securely allow its distribution and sharing.

⁶ EGA, <https://ega-archive.org/>

Chapter 2 Initial computational framework

The iPC central computational and data platform is conceived as a reference place for the partners in the consortium. As such, the platform should **engage with different users**: wet-lab scientists, bioinformatics and technical engineers, and provide the mechanisms to guarantee a smooth interaction with its components.

The platform is developed using continuous interactions under the **minimum viable product** (MVP) paradigm, which has just enough core features to effectively deploy the platform, and no more. In this way, it is possible to capture frequent and valuable feedback before releasing a major version of the platform. Indeed, this strategy targets avoiding building a system that users do not want or find difficult to interact with.

Together with the MVP-based development approach, the iPC central platform strongly relies on existing **open-source components**, which contribute to accelerate its development and facilitate its sustainability and maintainability over time. To make this possible, the platform architecture depicted in Figure 1 is quite modular. This approximation allows to build in parallel different parts of the platform, integrate existing software components and test them without requiring having the full platform implemented. Such approach also contributes to make an optimal use of resources as they can be enabled to only those subparts that need them e.g. **computational cloud resources** for the analytical component. Modules are organized as **microservices**, which implies defining common interfaces among them and allowing them to have their own development processes without compromising the overall architecture.

To build the indicated functionalities, the following components have been designed and are currently at some implementation stage:

1. A main **access portal** providing a centralized and unified entry point to the platform. The web-based frontend is meant to link to the different components, but it also gives access to the project information, and general documentation.
2. A **data catalogue** built from a minimal harmonized metadata model derived from a list of relevant reference repositories like Kids First, EGA, ICGC and ICGC ARGO. The access to the catalogue data is public, although registered users are also enabled to use the portal functionalities for searching, filtering and selecting datasets for further analysis. This component relates to **task 2.2** on “Implementation of a central platform for data and metadata storage and models inferences.”
3. A **data portal** designed to store and manage iPC datasets according to a data access framework that will be implemented to facilitate the communication with Data Access Committees. Currently, the prototype integrates a cloud storage solution based on Nextcloud. This effort is done in the context of **task 2.3** on “Implement of a framework to facilitate communication with Data Access Committees.”
4. **Analysis portals** exploiting cloud computing capabilities. iPC platform meant to include two different data analysis frameworks, Cavatica and openVRE. Both are integrating analysis tools and workflows together with data services granting access to private workspaces and relevant research datasets. This part of the overall architecture could be considered part of **tasks 2.4** and **2.5** on “Implementation of interactive web-based portal for data and software access” and “Build a cloud-based technological solution for utilization of computational models”, respectively.
5. A **Common Identity** system based on the openIDConnect⁷ protocol that enables an integration of all platform components under the same authentication and authorization infrastructure.

⁷ OpenID Connect, <https://openid.net/connect/>

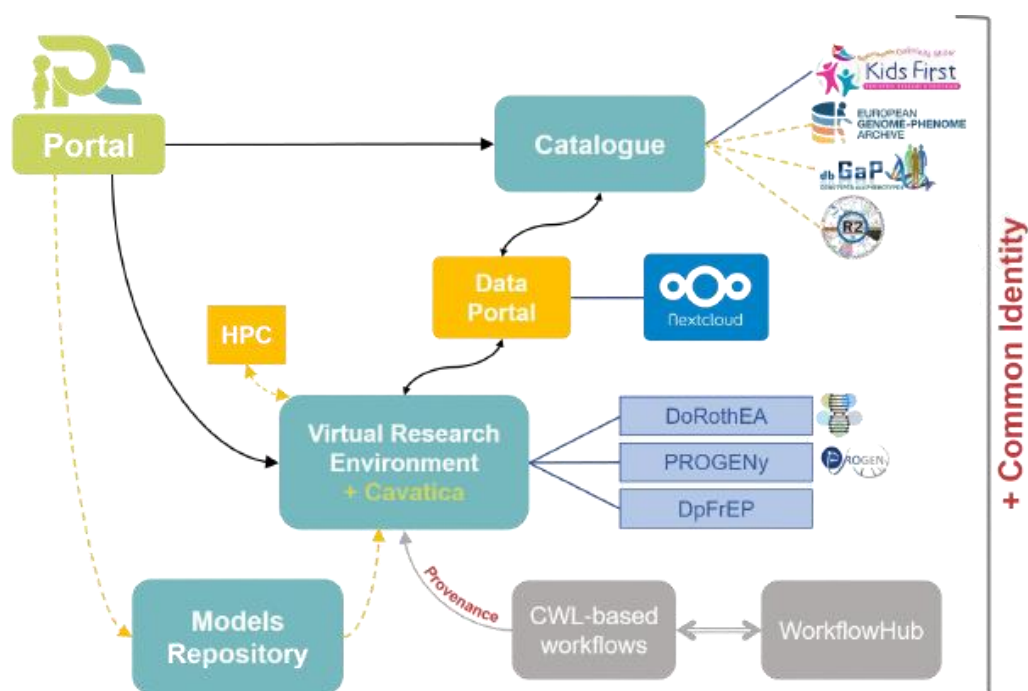


Figure 1. iPC platform design. Components and interoperability channels in yellow will be developed in the next iteration. At least one version of the green components have been released while the gray boxes are external developments that have been identified as relevant to the platform.

Chapter 3 Platform components

3.1. Access portal

The iPC Access Portal (<https://portal.ipc-project.bsc.es/>) is the **main entry point** for iPC users (depicted in Figure 2), which allows accessing the different platform components (Data Catalogue, Data Analysis platforms, Data Storage, Documentation) from one single place and one single login (**Single-Sign-On**). For that purpose, we have integrated an authentication and authorization system based on OpenIDConnect (shown in Figure 3) that goes from the access portal to the rest of platform components. We have also deployed a Keycloak instance, which supports multiple **OpenIDConnect** identity providers (idP), such as Elixir AAI (Authentication and Authorisation Infrastructure) [1], and Technikon Auth system among others, that will constitute the ground for a federated login schema.

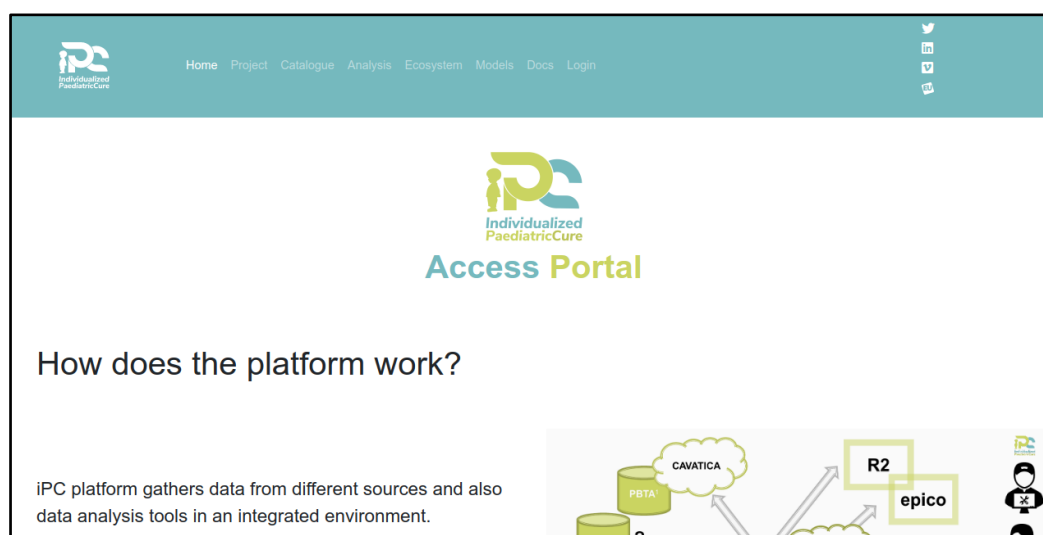


Figure 2. iPC Access portal which serves as a starting point to the different internal and external platforms used in the consortium.

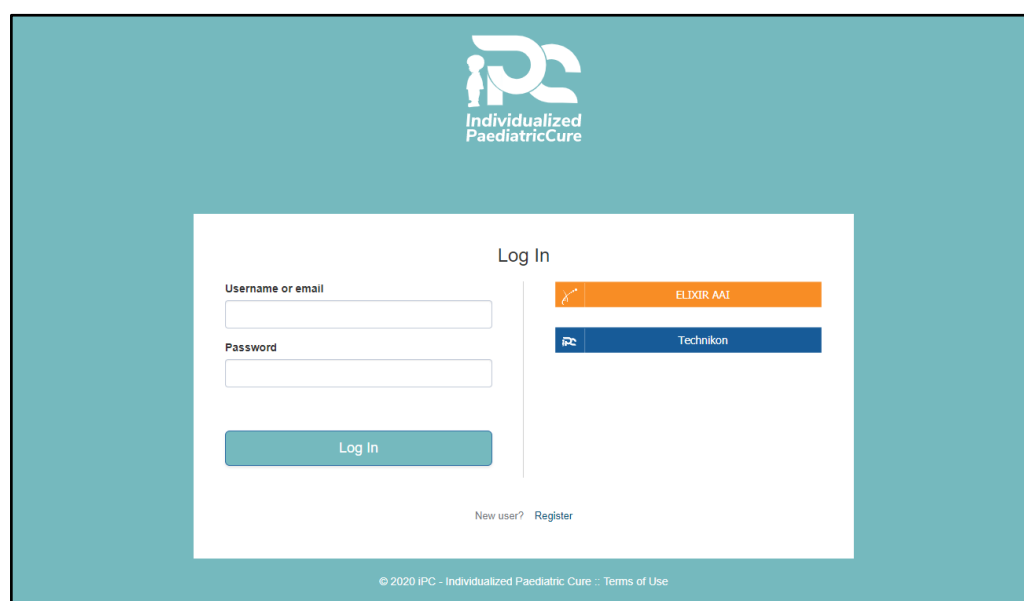


Figure 3. iPC login page. The platform supports different OIDC identity providers.

3.2. Data Catalogue

iPC Data Catalogue (<https://catalogue.ipc-project.bsc.es>) will **integrate paediatric cancer studies** coming from reference repositories like Kids-First, R2 or dbGap, as well as newly produced data from iPC partners. As detailed below, the catalogue currently includes metadata of the openPBTA initiative (Pediatric Brain Tumor Atlas), part of the Kids-First repository.

3.1.1 Catalogue Portal

Technically, the catalogue is deployed on Arranger⁸, a data search interface part of the Overture suite. It has facilitated the development of a data-based web portal for **searching, filtering, and selecting** paediatric cancer related datasets allocated on the catalogue database (as depicted in Figure 4).

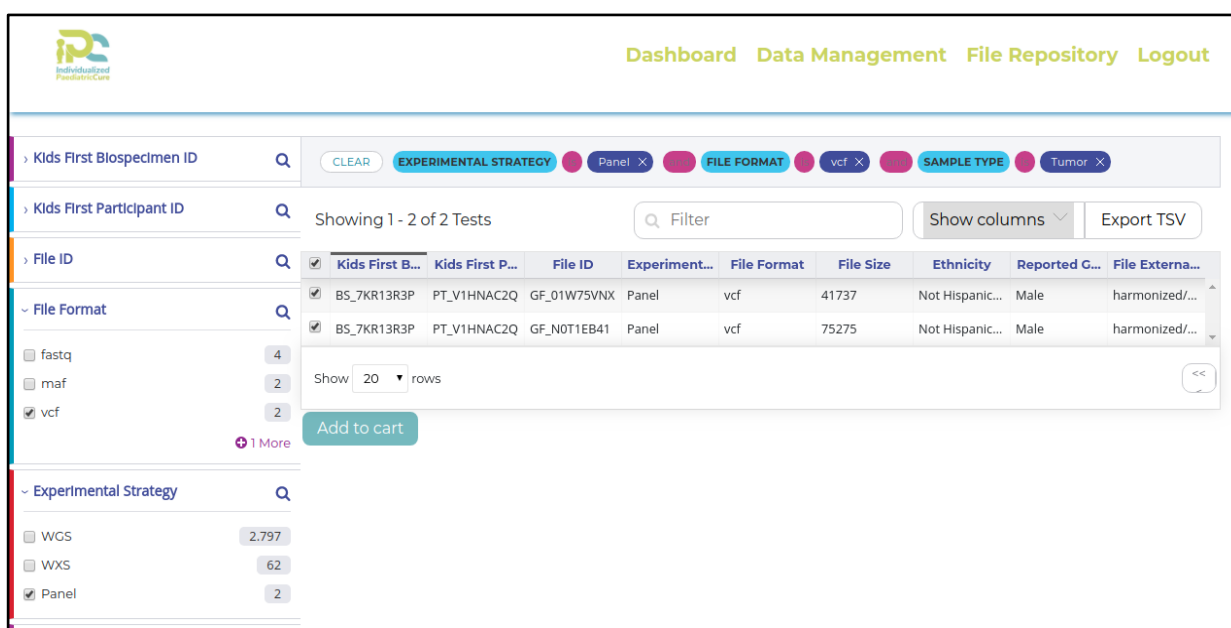


Figure 4. iPC Catalogue portal for searching, filtering, and selecting metadata through Arranger.

Users are also able to expose selected data to the different iPC Data Analysis platforms in an interactive and user-friendly way (as shown in Figure 5) for the latter data analysis stage. For that purpose, the iPC Catalogue Outbox (<https://github.com/inab/iPC-Catalogue-outbox-api>) has been implemented as a REST API that enables the **exportation** of metadata from the iPC Catalogue Portal, making it available to external applications/services. Interface accesses are controlled using the OpenIDConnect protocol in use for the iPC Common Identity system. As part of the demonstrator, the openVRE analysis platform is making use of the iPC Outbox to integrate catalogue's datasets as one of its reference repositories. In the next phase, a data access request mechanism will be integrated based on existing initiatives (i.e. GA4GH standards) in order to provide an additional authorization layer to the system.

⁸ Arranger (Overture stack), <https://www.overture.bio/products/arranger>

Data sets available to the iPC openVRE:

Inspect and/or remove already loaded data sets into VRE.
[Go to iPC openVRE](#)

fileID : GF_01W75VNX
file_locator : harmonized/simple-variants/70250648-82ee-4cfb-a997-1c4d25e497d5.mutect2_somatic.PASS.vep.vcf.gz
es_host : pbta_histologies_filecentric

[Unload data](#)
[Get Details](#)

Data sets available to Cavatica:

Inspect and/or remove already loaded data sets into Cavatica.
[Go to Cavatica](#)

fileID : GF_N0T1EB41
file_locator : harmonized/simple-variants/833d8c66-60a4-4612-ac67-9cbdba0cf6c4.strelka.PASS.vep.vcf.gz
es_host : pbta_histologies_filecentric

[Unload data](#)
[Get Details](#)

Figure 5. iPC Catalogue portal data management section: Expose data for analysis through the iPC Outbox REST API.

3.1.2 iPC Data Model for managing samples metadata

One of the objectives iPC is to maximize the reuse of data related to paediatric cancers, in order to enhance new techniques for the development of treatments within personalized medicine. Thus, it is necessary a formal and flexible **data model that integrates existing data** from other repositories and provides a guide on what metadata is needed for new generated samples. The interoperability of the models used in reference repositories such as ICGC, ICGC ARGO and EGA has been analysed, studying their content and structure as well as the current and potential use of ontologies that are included.

In order to establish the central axis of the unified model that is capable of capturing data from ICGC, ICGC ARGO and EGA in a harmonized way, the metadata included in each of these models has been compared, to select those that are common and cross-compatible. As a result of this, the **proposed metadata** to be included in the data model of the catalogue, as considered elemental for the future functionalities of the platform, is shown in Table 1.

Name	Description
donor_id	iPC identificator for the donor
sex	Biological sex of the donor
diagnosis	General term for detecting and classifying cancer in patients
age_at_diagnosis	The age of an individual at the time of initial pathologic diagnosis
disease_stage	An adjectival term that can specify or describe a disease stage
vital_status	The state or condition of being living or deceased; also includes the case where the vital status is unknown
sample_id	iPC identificator for the sample
sample_type	Whether the sample corresponds to a tumour or not
sample_age	Age of enrollment: the age of a subject when entering a group, catalog, list, or study

Name	Description
source_tissue	A clinical site that collects and provides patient samples and clinical metadata for research use
cell_type	An established cell culture that has the potential to propagate indefinitely
file_id	iPC identifier for the file
sequencing_platform	The name of the technology platform used to perform nucleic acid sequencing
sequencing_strategy	Sequencing library strategy
file_format	File extension format
file_size	File size
file_source	Source repository where the file is located
file_external_id	External identifier of the file
file_md5	MD5 checksum of the file
file_route	Route of the file

Table 1: Initial iPC data model for samples management.

Furthermore, after the study of the mentioned repositories, it has been found that most the vocabulary used in the three data models can be standardized into the National Cancer Institute Thesaurus (NCIT) ontology terms, which is the reason why the proposed metadata matches some definitions shown in the NCIT ontology.

The basic data model has been evaluated by calculating the coverage of data corresponding to the **Pediatric Brain Tumor Atlas (PBTA)** initiative provided by the partner Children's Hospital of Philadelphia (CHOP), included also in Kids First (<https://github.com/AlexsLemonade/OpenPBTA-analysis>): 18 out of the 37 metadata fields contained in PBTA could be covered by the proposed set of metadata for iPC, which represents 49% of the original variables. Further efforts need to be performed on this subject, considering the addition of other data from partners that should be included in the catalogue in the future. To this point, the Data Catalogue is populated with only the PBTA metadata and a few open access data samples used for the demonstrator of the VRE, and, due to the lack of more data from other sources to be harmonised with, the original data model of PBTA has been maintained in the iPC Catalogue.

3.3. Data analysis platforms

3.3.1 Cavatica

Cavatica (<https://cavatica.squarespace.com/>) is one the analysis portals of the iPC platform. It is designed as a **data analysis and sharing platform** meant to accelerate discovery in a scalable, cloud-based compute environment where data, results, and workflows are shared among the world's research community. Hosted at CHOP, the framework is one of the Data Resource Center (DRC) part of the NIH Common Fund's Gabriella Miller Kids First Pediatric Research Program (Kids-First).

Cavatica platform provides seamless access to some reference datasets, including, apart from Kids-First studies, the dbGaP repository. In their cloud workspaces, researchers launch analysis workflows that will run on **Amazon** Elastic Compute Cloud (Amazon EC2), a powerful and public cloud provider that offers an scalable and secure computational environment. Compute units correspond to virtual machines with dockerized applications described using **CWL** (Common Workflow Language).

3.3.2 Open Virtual Research Environment (openVRE)

iPC openVRE [2] (<http://vre.ipc-project.bsc.es/>) is one the analysis portals of the platform. It exposes the researcher to a fully intuitive and project-tailored **Virtual Research Environment (VRE)**, while transparently the platform manages job-based executions on the underlying cloud infrastructure in an elastic manner. openVRE cloud layout was first developed under the context of MuG H2020 project, and is being further developed and adapted by other computational research infrastructures.

openVRE interacts with any OCCl-compliant [3] cloud middlewares, like OpenNebula⁹ or OpenStack¹⁰. Currently, the system runs on top of an **‘on premise’ cloud** hosted at the Barcelona Supercomputing Center (BSC) based on OpenNebula. Compute units correspond to virtual machine instances either deployed on-demand by an enacting remote system (PMES [4]), or directly connected as in cluster environment using a local queueing batch system (OGS¹¹).

VRE offers a project-customized repository of **analysis and visualization tools**, which are brought into the platform by *third party* collaborators (*i.e.* iPC partners). Currently, iPC openVRE integrates:

- KFDRC Whole-Genome alignment workflow (CHOP), available at <https://github.com/kids-first/kf-alignment-workflow>
- DoRothEA (UKL-HD): transcription Factor activity inference from scRNA-seq data, available at <https://saezlab.github.io/dorothea/>
- PROGENy (UKL-HD): pathway activity inference from scRNA-seq data, available at <https://saezlab.github.io/progeny/>
- Shiny Visualizer (BSC, UKL-HD): creation of interactive web applications from R, available at https://github.com/inab/ipc_shiny_server/ and <https://github.com/saezlab/ShinyFUNKI>

Specific **iPC interoperability channels** with external repositories are also implemented to offer a seamless interaction with reference datasets. The following interconnections are currently in place:

- Querying datasets’ metadata from iPC Data Catalogue:
Relevant dataset’s metadata is extracted from the Data Catalogue using the iPC Catalogue Outbox REST API. The interface exposes only those datasets the user has previously selected on the Catalogue for outboxing - provided the appropriate data access permissions.
- Importing datasets from the iPC Data Portal:
Authenticated and robust data transferances are enabled with the WEBDAV server exposed by the current implementation of the iPC Data Portal. The prototype is using a privileged account to read exposed datasets.

⁹ opennebula, <https://opennebula.io/>

¹⁰ openstack, <https://www.openstack.org/>

¹¹ OGS, Open Grid Scheduler (former SGE, Sun Grid Engine) <http://gridscheduler.sourceforge.net/>

3.3.3 Exemplary analysis flow for iPC researchers

A usual researcher's flow on the analysis portal starts by loading into the cloud's workspace the datasets willing to be analysed. At openVRE, if the dataset is part of the iPC catalogue, the researcher can directly load it, as VRE is able to query the iPC Catalogue Outbox (see Figure 6).

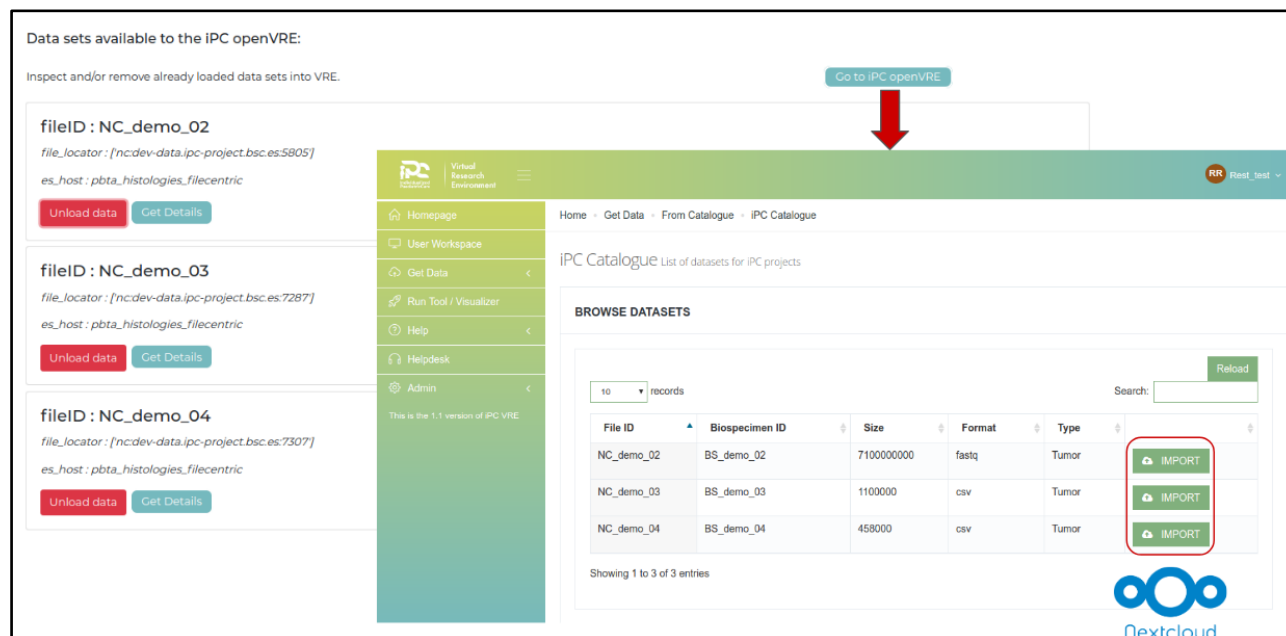


Figure 6. iPC Catalogue outbox. iPC Data Catalogue Management section is shown on the left side, where a user chooses to expose three datasets to analyse them at the openVRE. Then, as shown on the right side of the figure, associated datasets metadata is displayed at the openVRE and the datasets are imported with a simple click.

Once the data is transferred to the VRE workspace, it becomes eligible to carry out an analysis on the cloud. In particular, when selecting the dataset, the VRE will show only those analyses compatible with that data selection thanks to the dataset' metadata (see Figure 7).

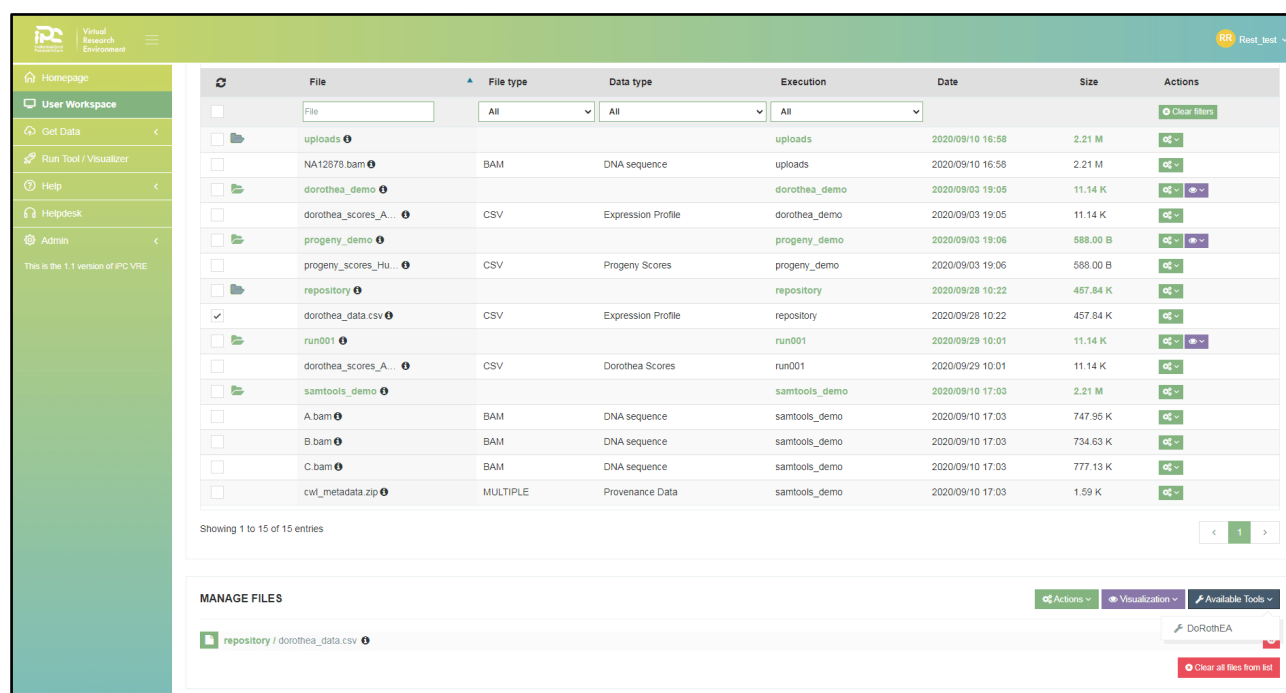


Figure 7. User's workspace on the VRE. Selecting DoRothEA tool on the VRE workspace.

The next snapshot (Figure 8) shows the tool configuration for the execution job of DoRothEA using the dataset imported from the iPC Data Catalogue.

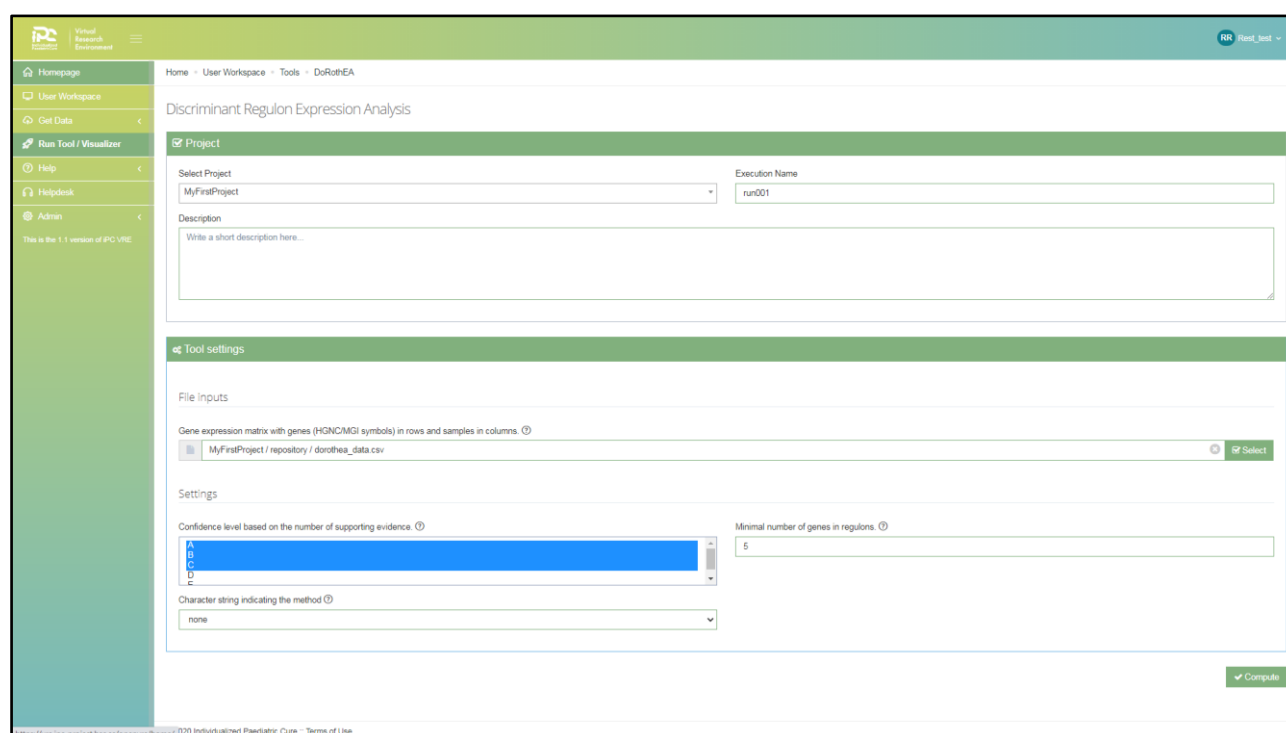


Figure 8: Running Dorothea on the VRE

After the job is completed, the results are displayed on the VRE workspace, and they can be downloaded, visualized online (see Figure 9) or further analysed with other pipelines.

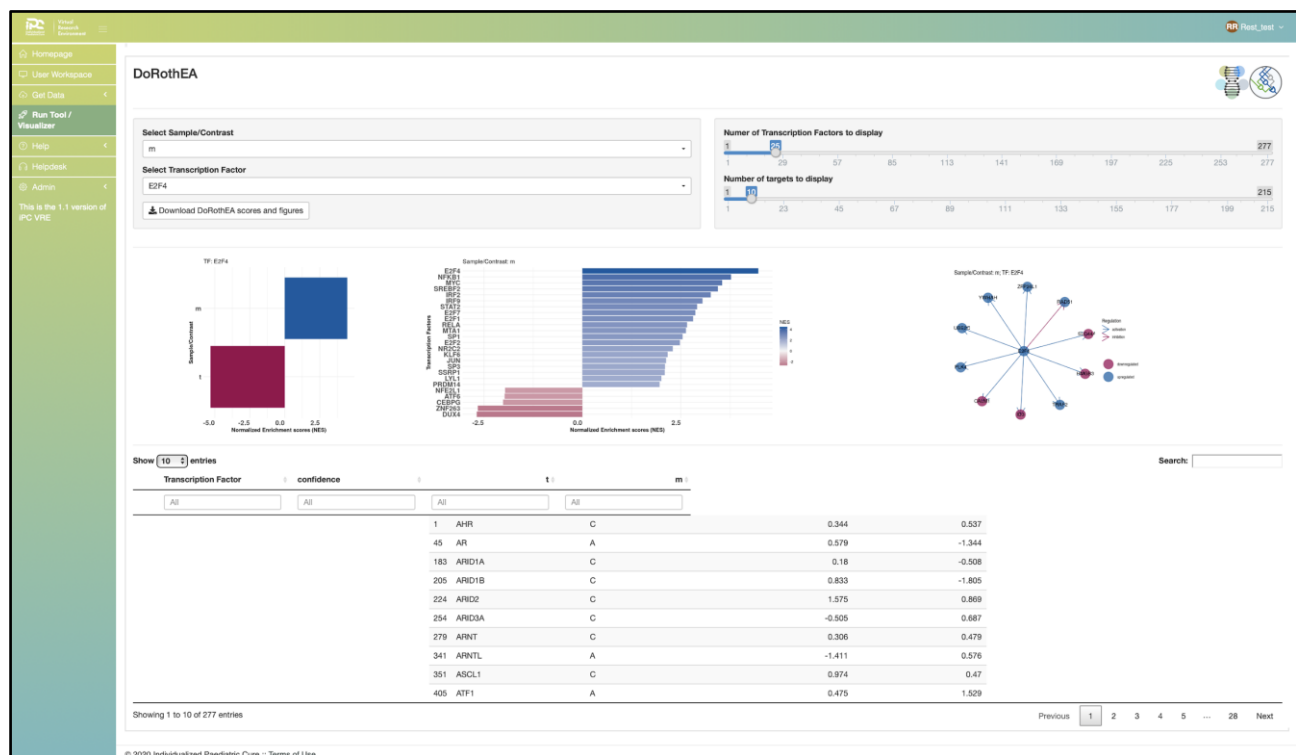


Figure 9: Visualizing Dorothea's results on the VRE. Data is displayed using one the visualizers integrated at the VRE: an R-Shiny application.

3.3.4 Sharing CWL-based workflows across iPC workbenches

The Common Workflow Language (CWL) is an open specification for data analyses and workflows that fosters pipeline sharing in a reproducible manner. Cavatica applications are natively described using CWL, while openVRE tools are specified using a domain specific language (DSL). Efforts are being dedicated to implement a CWL-conformant execution engine on the VRE, so that a **common iPC repository of analyses** becomes feasible. In collaboration with CHOP and as a proof of concept, an specific Cavatica application, the 'KFDRRC Whole-Genome alignment workflow', is loaded into the openVRE application by importing only its CWL description file. Results are proved to be comparable between the two analysis platforms.

The **openVRE CWL adapter** developed provides openVRE with a flexible CWL runner able to consume different CWL workflows without requiring modifications. The pipeline is materialized when launching the application and is executed in a controlled environment, as any other tool. The adapter receives VRE infiles and parameters, dynamically builds a CWL parametrization file and fires a CWL-compliant engine. Furthermore, the results are not only registered to the VRE workspace, but also are accompanied with a set of provenance metadata derived from the CWL execution. Indeed, the mechanism to capture the provenance is being developed in the context of Research Object Crate¹² (**RO-Crate**) in the EOSC-Life project. RO-Crate is being used as the minimal unit for providing rich metadata for workflows deposited in **WorkflowHub**, which is also being developed in the context of the H2020 EOSC-Life project.

In the sample CWL tool (Figure 10), users can assign selected data files to the appropriate input parameters and arguments after tool configuration, to send execution jobs to the VRE backend who executes the workflows. Progress and results of each execution job can be followed in the VRE workspace.

¹² <http://www.researchobject.org/ro-crate/>

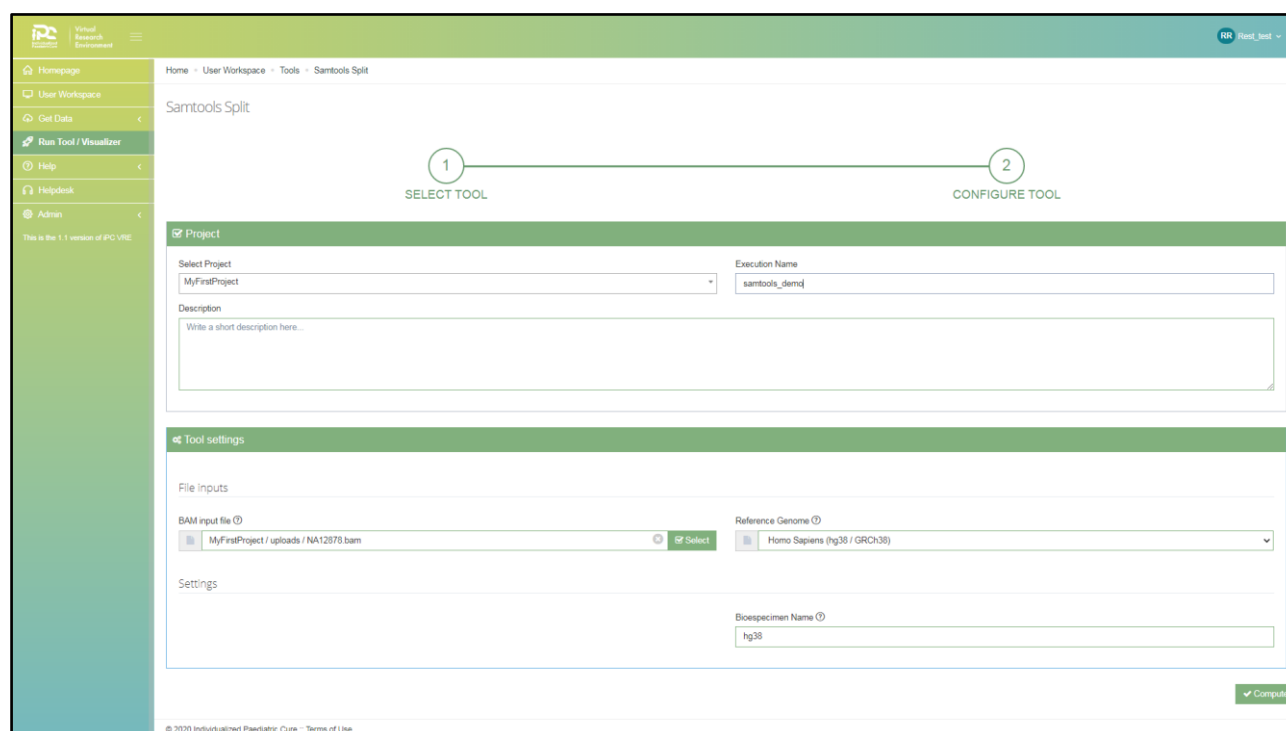


Figure 10. Configuration screen for sample CWL tool.

3.4. Data Storage

iPC Data Storage system comprises different technologies either for storing metadata and its associated primary data (i.e. genomics data). An **ElasticSearch**¹³ engine has been put in place to search and store the metadata that fuels iPC Data Catalogue. This service represents an open-source, distributed, and document oriented NoSQL database that allows real-time searching, which is tightly coupled to Overture's Arranger, able to natively interpret Elasticsearch mappings.

On the other hand, a **Nextcloud**¹⁴ service (<https://data.ipc-project.bsc.es>) has been deployed to the iPC platform and proposed as the main primary data storage system. Nextcloud is an open-source project which provides an user interface for uploading and downloading data at high speed rates, while retaining control over users data (Figure 11). Therefore, a very interesting solution not just for sharing data among iPC partners, but also, a powerful way for making primary data available on the different data analysis platforms.

¹³ ElasticSearch, <https://www.elastic.co/>

¹⁴ Nextcloud, <https://nextcloud.com/>

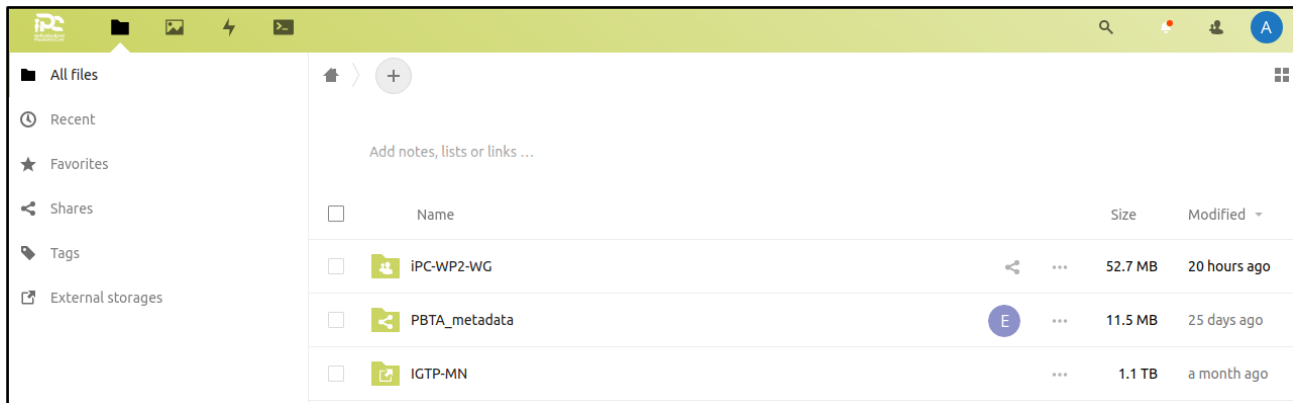


Figure 11. BSC's Nextcloud service for iPC: Upload/download data on a user-friendly interface.

4 Summary and Conclusion

Here we have provided a review of the initial infrastructure for iPC. The iPC central computational and data platform aims to enable paediatric cancer research by bringing distributed and heterogeneous data together with generic and specific software solutions in the frame of HPC and cloud computing infrastructures. Such a massive and complex platform has a number of different components that have been reviewed in this deliverable (data catalogue, data portal and virtual research environments). The use of a modular architecture allows the independent development of the different major components. Moreover, the adoption of a MVP strategy allows to quickly gather feedback about any of those components from the iPC partners. In this way, it is possible to develop a fully functional platform that can be used from early on in the project.

There are a number of challenges ahead of us including the full development of the data access manager to facilitate the access to the data (and metadata) available at different data repositories such as Kids First, dbGAP, and EGA; in a transparent, unify and friendly manner. To ensure the use of community-led protocols and standards, we are already engaged with the GA4GH. We are specifically looking at the GA4GH passports and visas standards and reference implementations to facilitate the access to data sets using a single sign-on in our platform. We are also considering the use of the GA4GH Data Usage Ontology (DUO) as a mechanism to gain quick access to data sets for paediatric cancer research. GA4GH DUO will facilitate automated access to those data sets for which the use level and the iPC researchers level match while for other data sets will be necessary to follow the established procedures via the Data Access Committees (DACs).

Current deployed tools are oriented to secondary data analyses. Those tools are so far computationally less intensive than the tools used for the primary data analyses, which make the initial processing of the raw data. During the next period, our efforts will be focused into two workstreams. First, we will continue the advance of primary data analysis workflows that can be easily deployed across relevant partners, e.g. CHOP, BSC. Such efforts will contribute to the initial analysis of massive amounts of data that can be subject to tighter regulations regarding data privacy and confidentiality. Resulting data, often one or more order of magnitude smaller in size, is easier to share and transfer across different jurisdictions, e.g. USA and Europe. Second, we will work towards enabling HPC capabilities in the platform to allow the training and refinement of machine learning models, which tend to integrate different data types.

5 List of Abbreviations

Abbreviation	Translation
AAI	Authentication and Authorisation Infrastructure
API	Application Programming Interface
ARGO	Accelerating Research in Genomic Oncology
BSC	Barcelona Supercomputing Center
CHOP	Children's Hospital of Philadelphia
CWL	Common Workflow Language
DAC	Data Access Committees
dbGAP	Database of Genotypes and Phenotypes
DRC	Data Resource Center
DSL	Domain Specific Language
DUO	Data Use Ontology
EGA	European Genome-phenome Archive
GA4GH	Global Alliance for Genomics and Health
ICGC	International Cancer Genome Consortium
idP	Identity Providers
iPC	individualized Paediatric Cure
KFDRC	Kids First Data Resource Center
MD5	Message-Digest Algorithm 5
MuG	Multi-Scale Complex Genomics project
MVP	Minimum Viable Product
NCI	North American National Cancer Institute
NCIT	National Cancer Institute Thesaurus
NIH	North American National Institutes of Health
OCCI	Open Cloud Computing Interface
OGS	Open Grid System
PBTA	Pediatric Brain Tumor Atlas
PMES	Programming Model Enactment Service
R2	Genomics Analysis and Visualization Platform by the Academic Medical Center
REST	Representational State Transfer
RO-Crate	Research Object Crate
TGCA	The Cancer Genome Atlas
UKL-HD	University Hospital of Heidelberg
VRE	Virtual Research Environment
WebDAV	Web Distributed Authoring and Versioning

6 Bibliography

- [1] Linden M et al., “Common ELIXIR Service for Researcher Authentication and Authorisation” F1000Research. 7. 2018. doi:<https://doi.org/10.12688/f1000research.15161.1>.
- [2] Open Cloud Computing Interface (OCCI). Open Source Cloud Computing Platforms (2010). <https://doi.org/10.1109/GCC.2010.77> . <https://www.ogf.org/documents/GFD.227.pdf>
- [3] L. Codó et al., “MuGVRE. A virtual research environment for 3D/4D genomics,” bioRxiv, p. 602474, Apr. 2019 doi: <https://doi.org/10.1101/602474>.
- [4] F. Lordan et al., “ServiceSs: An Interoperable Programming Framework for the Cloud,” J. Grid Comput., vol. 12, no. 1, pp. 67–91, Mar. 2014. doi: <https://doi.org/10.1007/s10723-013-9272-5>