D2.3

Recommended metadata standards and portal prototype

Project number	826121
Project acronym	iPC
Project title	individualizedPaediatricCure: Cloud-based virtual-patient models for precision paediatric oncology
Start date of the project	1st January, 2019
Duration	53 months
Programme	H2020-SC1-DTH-2018-1

Deliverable type	Demonstrator
Deliverable reference number	SC1-DTH-07-826121 / D2.3 / 1.0
Work package contributing to the deliverable	WP2
Due date	May 2021 - M29
Actual submission date	28th May, 2021

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

Abstract	We report on the selection of the appropriate data models to handle the available data and metadata to the iPC Central Computational and Data platform. We also report on the current status of the development for the iPC Data portal.
Keywords	Metadata; Data models; Standards; Data Catalogue

The project iPC has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826121.

Editor

Salvador Capella-Gutierrez (BSC)

Contributors (ordered according to beneficiary numbers)

Alejandro Canosa (BSC)

Elena De La Calle (BSC)

Laia Codo (BSC)

Dmitry Repchevsky (BSC)

José María Fernández (BSC)

Alfonso Valencia (BSC)

Jolanda Modic (XLAB)

Aleš Černivec (XLAB)

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

This document reviews the different approaches on metadata representations within the iPC platform, as well as the efforts dedicated to integrate and exploit them at the iPC Catalogue and the overall iPC Central Computational and Data Platform. The proposed data models (**chapter 2**) are designed for enabling meaningful research data management. Research data is essential for answering scientific questions advancing the existing knowledge on the molecular basis of cancer. Ideally, data models should facilitate the management of data files while enabling rich metadata descriptions. Metadata that enhances the use of phenotypic descriptors represents an excellent source of knowledge for researchers, as they provide information mostly focused on clinical aspects (**chapter 2.1**). Other descriptors are focused on providing experimental information and raw data processing approaches (**chapter 2.2**). As experimental data can be obtained from a variety of sources, data models should be able to handle such complexity to enable researchers to find the necessary data for conducting their investigations. Further, different cancer-type data models have been proposed as an alternative approach for representing the iPC use-cases (**chapter 2.3**). Additionally, the iPC project aims to ensure data interoperability among different resources, and therefore, principles and well-defined standards on data accessibility, usability, and registry must be enforced by the platform (**chapter 3**). Finally, this document reviews the latest development efforts made on the iPC Catalogue portal (**chapter 4**), and also, the planned strategy for further developments on the iPC Central Computational and Data Platform (**chapter 5**). The present deliverable reviews, builds on and extends previous ones, especially D2.2 - Initial infrastructure framework, where some of the aspects discussed here had previously been elaborated.

In summary, specific guidelines to support the development of the metadata registry are provided, which consist of the recommendation of standards to fulfill both technical and scientific requirements of the platform. Nevertheless, it is important to remark that the data model features that are proposed here, will be reviewed and agreed upon by all partners, and further developed based on their specific needs. Future efforts will be directed on obtaining feedback and consensus from the consortium.

Table of Content

List of Figures

List of Tables

Chapter 1 Introduction

The initial infrastructure of the iPC Central Computational and Data Platform is discussed in detail at deliverable D2.2 “IPC Initial infrastructure framework”, where the main functionalities and components are presented to serve as a reference framework for the partners in the consortium. Designed as a one-stop shop, the main portal (https://ipc-project.bsc.es) integrates under a central authentication service the main platform’s components: the data catalogue, the data portal and several analysis frameworks. The development of the platform is an on-going process where new features are being implemented and components reinforced under the minimum viable product (MVP) paradigm. Therefore, the infrastructure is currently most suited for hosting genetics/genomics data although it can eventually handle other data types, which will require following a similar approach to identify the best way to represent and handle those datasets. The establishment and implementation of a unique data model that enhances a unified and integrated use of the platform components is one of the key aspects for accomplishing the objectives formulated in WP2.

This document focuses on the iPC guidelines proposed for developing an extensible metadata model that annotates genomic data available to the consortium for fostering its reusability and exploitation not only within the platform, but also for the overall research community. Along with the development of new data structures, new communication interfaces have been implemented to securely expose it across the different platform components. Additionally, part of this deliverable is dedicated to the new implementations and progress on the portal prototype, which is under continuous development.

Chapter 2 Data models

For reinforcing the sharing, re-use, and aggregation of pediatric cancer data across iPC platform's researchers and beyond, it is essential facilitating a formal and flexible data model that enables data distribution in a structured and standardized manner. On this account, a comprehensive comparison of reference data models adopted by several well-established initiatives and data repositories like the International Cancer Genome Consortium[1] (ICGC), the Accelerating Research for Genomic Oncology[2] project of ICGC (ICGC-ARGO) and the European Genome-phenome Archive[3] (EGA) was presented as part of D2.2. Not only data structures were considered, but also the use of controlled vocabulary with ontological references. A mapping of the terminology used in enumerated fields against standard cancer ontologies - *i.e* National Cancer Institute Thesaurus[4] (NCIt) - was performed using Zooma[5]. An overall summary of the analysis is shown in Table 1. As a result, and based on a minimal consensus approach, an elementary common data model was proposed by iPC.

Table 1: Key features overview of the reference repositories.

	ICGC Data Portal	ICGC ARGO	EGA
Purpose	genomic oncology research		general genomics
Website	www.icgc-argo.org	dcc.icgc.org	ega-archive.org
Access	public	public	controlled
Unique fields	193	103	38
Fields with enumerated content	64	30	14
Clinical / phenotypic fields	55	61	3
Molecular data fields	112	19	16
Ontologies used/recommended	ICD-10, ICD-O-3	ICD-10, ICD-O-3	EFO
Enumerated content mapped into	NCIT	NCIT	NCIT, EFO

However, the initial model has been subject to sequential redesign iterations that lead to a natural growth and extension of the metadata model in order to cover the integration of new data types and categories at the iPC catalogue. Figure 1 depicts the interrelation of the basic entities modeled. This design is articulated in four metadata objects corresponding to *Samples*, *Files*, *Datasets* and *Donors*, which are the basic schemas to describe the entities that these represent.

However, in order to evolve the model introduced there, it seemed mandatory to contrast with other data expected to be included in the iPC Catalogue. For that purpose, we used a collection of datasets

[1] ICGC Data Portal, https://dcc.icgc.org

[2] ICGC ARGO, https://www.icgc-argo.org

[3] EGA, https://ega-archive.org

[4] NCIt, https://ncithesaurus.nci.nih.gov

[5] Zooma. Ontology Annotation Service, https://www.ebi.ac.uk/spot/zooma

(see Table 1) corresponding to different cancer types, that were provided by partners from AMC, who curated and hosted in the R2 platform[6] the data that was originally in the Gene Expression Omnibus[7] (GEO). These datasets correspond to a series of normalized and anonymized matrix files from microarray and RNA-Seq experimental data. Metadata accompanying these files in GEO was reviewed in order to find additional information to be considered as part of the metadata hosted by the platform. As a result of this effort, the set of metadata fields has been extended and is presented in this chapter (see Tables 3-7).

Table 2: Datasets from R2 added to the iPC Data Catalogue.

Cancer type	GEO Accession	Number of samples	Data Type	Sample Type	Author
Neuroblastoma	gse62564	498	RNA-Seq	Tumor	Wang
Neuroblastoma	gse73517	105	RNA Expression by microarray	Tumor	Henrich
Neuroblastoma	gse49710	498	RNA Expression by microarray	Tumor	Wang
Neuroblastoma	gse3960	101	mRNA Expression by microarray	Tumor	Maris
Neuroblastoma	gse19274	100	RNA Expression by array	Tumor +cell lines	Jagannathan
Ewing Sarcoma	gse68776	74	RNA Expression by array	Tumor +control	Lawlor
Ewing Sarcoma	gse17679	117	RNA Expression by array	Tumor +control	Savola
Ewing Sarcoma	gse7007	39	Total RNA Expression by array	Tumor	Tirode
Ewing Sarcoma	gse63157	85	RNA Expression by array	Tumor	Volchenboum
Ewing Sarcoma	gse34620	117	RNA Expression by array	Tumor	Delattre
Leukemia	gse87070	654	RNA Expression by array	B-ALL	Polak
Leukemia	gse68790	283	RNA Expression by array	B-ALL	Loh
Leukemia	gse11877	207	RNA Expression by array	B-ALL	Harvey
Leukemia	gse7440	98	RNA Expression by array	B-ALL	Carroll
Leukemia	gse26713	117	RNA Expression by array	T-ALL	Meijerink
Leukemia	gse8879	55	RNA Expression by array	T-ALL	Mullighan
Leukemia	gse10255	157	RNA Expression by array	ALL	Sorich
Leukemia	gse68720	97	RNA Expression by array	ALL	Chen
Medulloblastoma	gse10327	62	Total RNA Expression by array	Tumor	Kool

[6] R2, https://hgserver1.amc.nl/cgi-bin/r2/main.cgi

[7] GEO, https://www.ncbi.nlm.nih.gov/geo/

Cancer type	GEO Accession	Number of samples	Data Type	Sample Type	Author
Medulloblastoma	gse85217	763	Total RNA Expression by array	Tumor	Cavalli
Medulloblastoma	gse37382	285	Total RNA Expression by array	Tumor	Northcott
Medulloblastoma	gse21140	103	Total RNA Expression by array	Tumor	Northcott

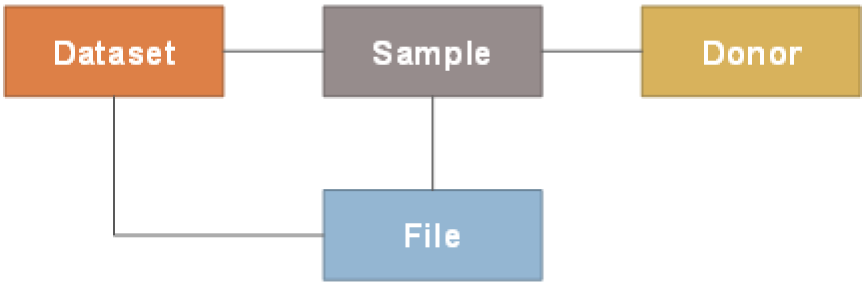

Figure 1: Informative schema showing the interrelationships between the high-level elements of the data model for the iPC platform

The *Sample* entity in Figure 1 refers to the biological material taken from a donor with the purpose of performing a genomic analysis. It is intended to collect any relevant information related with the biospecimen from which the genomic data is generated. To this point, we identified some basic metadata based on the reviewing of reference repositories and the related metadata of the datasets collection in GEO. It is then proposed a set of fields and its suggested ontological annotation, which is presented in Table 3. However, this metadata may be extended to contain other information such as additional descriptors, collection, handling and preparation of the samples.

Table 3: Proposed fields for the Sample entity with suggested ontology annotation.

Field	Type	Description	Ontology term	Values	Ontology terms
Sample Id	text	Identifier for the sample			
Sample Type	enum	Whether the sample corresponds to a tumor or to a normal tissue	NCIT:C70713	normal tumor (neoplasm)	NCIT:C14165 NCIT:C3262
Age	number	Age of enrollment or sample acquisition, meaning the age of a subject when entering a study.	NCIT:C164338	years	
Tissue	enum	An anatomical site from which the sample was obtained.	NCIT:C12801	SC of NCIT:C12801	
Histology	enum	Histological classification of the sample.	NCIT:C61478	SC of NCIT:C4741 for tumors and SC of NCIT:C12578 for normal tissues (*)	
Files	list	File identifiers associated with the sample.			

Note from the interrelations shown in Figure 1 that *Sample* is a core entity that connects the *Dataset* and *Donor* metadata objects. A *Dataset* contains a set of *Samples*, which in turn, each *Sample* is associated to a specific *Donor*. Additionally, a *Sample* may have one or multiple *Files* associated.

The files themselves refer to the genomic data that is hosted within the platform. Metadata related to data files is described within the *File* entity (see Table 4), which contains the necessary information to describe the files located in the repository. This metadata is also intended for internal use, as the platform components will process the files depending on some of these features and will require others to enable its functionality.

Table 4: Proposed fields for the File entity with suggested ontology annotation.
Terms marked with (*) may differ depending on the cancer type; SC=Subclass.

Field	Type	Description	Ontology term	Values	Ontology terms
File Id	text	Identifier for the file			
Data Type	enum	Structural format of the data carried	NCIT:C42645		
Sequencing Platform	enum	The name of the technology platform used to perform nucleic acid sequencing.	NCIT:172274		
Sequencing Strategy	enum	Sequencing strategy.			SC of NCIT:C153598
File Format	enum	The format of the file.	NCIT:C171252		SC of NCIT:C171252
Size	number	The size of the file.	NCIT:C171192		
Source	text	Source repository where the file is originally located.			
File External Id	text	External identifier of the file.			
MD5	text	MD5 checksum of the file.			
File Path	text	The specification of a node in a hierarchical file system, usually specified by listing the nodes top-down.	NCIT:C47922		

2.1. Dataset-centric data model

The most direct way to represent currently available data is in the form of datasets, since this is the most common way to share anonymised data. The metadata contained within the *Dataset* object (see Table 5) will describe features and conditions that are relevant for understanding how the experiment was performed in terms of experiment design and analysis to obtain the sample data files that are contained within the datasets. The *Dataset* metadata object also represents the series matrix files obtained after the normalisation of a set of samples.

Table 5: Proposed fields for the Dataset entity.

Field	Type	Description	Values
Dataset Id	text	Identifier for the dataset	
Series Title	text	Official descriptive name of the series matrix.	
Cancer Type	enum	The type of cancer object of the study from which the dataset was obtained.	
Category	enum	Whether the dataset contains normal, control samples or both.	Tumor Control Tumor + Control
Number Of Samples	number	Number of samples contained within the dataset.	
Normalization	text	Normalization algorithm applied to obtain the series matrix.	
Author	text	Individual, group, or organization primarily responsible for the content of the dataset	
Release Date	date	Date on which the dataset is due to become available for the public.	
Samples	list	Sample identifiers associated with the dataset	
Files	list	File identifiers associated with the dataset.	

2.2. Patient-centric data model

The metadata of the patient-centric data model is intended to describe phenotypic and clinical information related to the subject to whom each sample corresponds, and is represented within the *Donor* object. Being able to select sample files according to patient characteristics is crucial to address most of the questions and applications stated within the project, such as response prediction to therapies to provide personalized treatments. In this sense, defining a standard based on the consensus of the expected and existing information, having in count the needs of the platform users, will be key to its usefulness.

Table 6: Proposed fields for the Donor entity with suggested ontology annotation.
Terms marked with (*) may differ depending on the cancer type; SC=Subclass.

Field	Type	Description	Ontology term	Values	Ontology terms
Donor Id	text	Identifier for the donor			
Sex	enum	Biological sex of the donor	NCIT:C28421	male female unknown	NCIT:C20197 NCIT:C16576 NCIT:C17998
Diagnosis	enum	General term for detecting and classifying cancer in patients.	NCIT:C16213	SC of NCIT:C48233	
Age At Diagnosis	number	The age of an individual at the time of initial pathologic diagnosis.	NCIT:C156420	years	
Disease Stage	text	An adjectival term that can specify or describe a disease stage.	NCIT:C28108	SC of NCIT:C48698 (*)	
Vital Status	enum	The state or condition of being living or deceased; also includes the case where the vital status is unknown.	NCIT:C25717	dead alive unknown	NCIT:C28554 NCIT:C37987 NCIT:C17998
Overall Survival Time	number	Measure of time until the donor is deceased.	NCIT:C125201		
Event Free Survival	number	The length of time after treatment during which a patient survives with no sign of a particular complication of disease.	NCIT:C125201		
Samples	list	Sample identifiers associated with the donor.			

2.3. Metadata related to specific cancer types

Although there is general phenotypic and clinical information that is significant for conducting biomedical research of any kind, the study of specific cancers equally relies on distinctive characteristics and biomarkers related to each tumor type. For this matter, we note the importance of identifying specific metadata that is relevant for each type of paediatric cancer considered in the iPC project. It is important to remark that some of the metadata content may also differ in terminology depending on the cancer type to which the *Sample* or the *Donor* is related. However, these aspects will require to be discussed with the rest of the partners.

Due to the lack of practical data to work with, we aimed to determine a set of tumor-specific metadata for each type of cancer by using public information of the series matrices available in GEO corresponding to the datasets in Table 1. The resulting initial proposal is shown in Table 7.

Table 7: Cancer-specific proposed fields with suggested ontology annotations.

Field	Type	Description	Ontology term	Values	Ontology terms
Neuroblastoma					
MYCN Status	enum	Result of the laboratory test to determine if the diagnosed tumor is found to present MYCN gene amplification, in which case the tumor is more likely to spread in the body and less likely to respond to treatment.		Amplified	
				Non-Amplified	NCIT:C116945
Risk Classification	enum	A classification system developed to establish a consensus approach for pretreatment risk stratification of neuroblastomas.	NCIT:C102563	Low	
				Intermediate	
				High	NCIT:C150281
Leukemia					
Subtype	enum	Subtype characterization of Acute Leukemia		ALL	NCIT:C3167
				AML	NCIT:C3171
				Other	
White Blood Cells	number	White Blood Cells count		mcL	
Medulloblastoma					
Subtype	enum	Subtype characterization of Medulloblastoma		WNT	NCIT:C129440
				SHH	NCIT:C129441
				Group 3	NCIT:C129445
				Group 4	NCIT:C129446
Hepatoblastoma					
Vena Cava Invasion	yes/no	Whether there is an Hepatic Vein/Inferior Vena Cava Thrombus discovered by Imaging			
Blood Transfusion	yes/no	Whether the patient has received an injection of whole blood or a blood component directly into the bloodstream.	NCIT:C15192		
Focality	enum	The characterization of the location of the tumor.	NCIT:C157425	Solitary	
				Multifocal	

Chapter 3 Metadata standards

There are several aspects from both a technical and scientific perspective that need to be considered for the development of the iPC Central Computational and Data platform. The first requirements, which are independent of the specific purpose and content of the matadata, consist in methodologies to manage authentication, roles and access rights from users, as well as restrictions of data usage, file format standards and conceptual structure of the scientific metadata registry. The second are focused on methodologies to standardize experimental, phenotypic and clinical metadata associated to the data stored within the file repository of the iPC platform, which involve the usage of controlled vocabularies. Both types of specifications establish the foundations of data interoperability that, in this case, will allow exchanging and reusing research data in the context of paediatric cancer through the integration of metadata standards.

3.1. Technical standards

Data Use Ontology (DUO) is a GA4GH technical standard that provides **ontological terms** such as data use restrictions, geographic restrictions or intended research use. These terms can be used as a metadata associated with datasets in order to implement data access policies via matching data access restrictions with researchers' consents. For instance, a dataset may be annotated to be for "genetic studies only".

The ISO/IEC-11179 is an international metadata registries standard supported by the International Standards Organization (ISO) and the International Electrotechnical Commission (IEC) that establishes guidelines on the standardization, representation and registration of metadata within registries that gather metadata from different sources, in order to make the data understandable and shareable. It prescribes a **conceptual model** for structuring descriptive metadata and defining how the metadata is shared, but is not intended to offer specific guidelines for the physical implementation of a metadata registry [1].

Schema.org[8] is a **structured data vocabulary** developed and maintained within a collaborative community for enriching Internet content with metadata. It defines entities, actions and relationships that can be used with different encodings, such as JSON-LD. Bioschemas[9] uses the Schema framework by extending its features and reusing existing standards in the Life Science community in order to address the specific needs in the field, making data Findable through search engines based on the structured information contained as metadata, and therefore improving the Reusability of the data (see FAIR Principles [2]).

3.2. Specific standards for cancer and genomics research

There are also important considerations in order to implement and extend the data models proposed in the previous chapter, which involve optimizing the degree of standardization of both genomic data and clinical and phenotypical metadata associated.

Concerning genomic data, there are a number of conventions for describing roles and locations of higher order sequences of genomic domains and elements, such as those proposed by the International Nucleotide Sequence Database Collaboration (INSDC), a collaborative effort between the DNA Data Bank of Japan (DDBJ), the European Bioinformatics Institute (EMBL-EBI) and the European Nucleotide Archive (ENA) and the GenBank as a part of the National Center for Biotechnology Information (NCBI) of United States[10]. The Sequence Read Archive (SRA) XML Schemas that describe metadata and handles submissions and downloads in the SRA repository

[8] Schema.org, https://schema.org/

[9] Bioschemas.org, https://bioschemas.org/

[10] The INSDC Feature Table Definition, http://www.insdc.org/documents/feature-table

was developed in collaboration with the INSDC and also serves as a base for the International Human Epigenome Consortium[11] (IHEC) Metadata Specification that defines metadata standards oriented to the purposes of the NIH (National Institutes of Health of the United States) Roadmap Epigenomics Project.

The Phenopacket Schema[12] is being developed under one of the initiatives of the Global Alliance for Genomics and Health (GA4GH), whose objectives rely on the standardization of genomic data for enhancing its exchangeability and interoperability in biomedical research. Phenopackets serves as an open standard for capturing and sharing clinical and phenotypic information between information systems, based on data concept definitions and a standardized information model. In detail, Phenopackets define requirement levels for metadata concepts and contents are supported by the use of ontologies, presenting a complex schema with the ability to be adapted to various applications, including cancer research. Phenopacket building blocks cover several fields such as those proposed in the previous chapter and more, allowing the flexibility needed for the aims projected in the iPC platform. The schema is interoperable with other standards such as the ISO TC215 committee and the HL7 Fast Healthcare Interoperability Resources Specification (FHIR). Moreover, Phenopackets is in process to be implemented for the clinical and phenotypical data submission in EGA.

With the purpose of standardizing contents of the metadata, there are available multiple biomedical controlled vocabularies such as the Medical Subject Headings (MeSH), Logical Observations, Identifiers, Names and Codes (LOINC), the Systematic Nomenclature of Medicine (SNOMED), the Gene Ontology (GO) and the NCIt ontology. The most relevant standardized ontologies oriented to cancer research are brought together in the Unified Medical Language System (UMLS) of the National Library of Medicine in United States (NLM) and can also be found in biomedical ontology repositories such as BioPortal and the Ontology Lookup Service (OLS). As noted in the previous chapter, most of the terms included in the data model were mapped to NCIt ontology. However, the use of other ontologies with broader terminology as those mentioned should also be considered, since the purpose of this effort is to establish the most standardized vocabulary possible to facilitate the interoperability of the metadata. In this regard, the use of cross-reference tools for ontologies, such as the EMBL-EBI Ontology Xref Service[13] (OxO) in OLS, may result in a great advantage for exchanging terms with equivalent meanings [3].

[11] IHEC, https://github.com/IHEC/ihec-ecosystems

[12] Phenopackets, http://phenopackets.org/

[13] Ontology Xref Service, https://www.ebi.ac.uk/spot/oxo/

Chapter 4 Portal prototype

4.1. iPC Catalogue portal overview

The initial infrastructure framework was already discussed on D2.2. Since then, several improvements have been made to the platform at a technical level. Such enhancements fundamentally aimed at achieving better reproducibility, scalability, and also, improving the user interfaces. On the other hand, new datasets have been added to the iPC platform data storage system based on Nextcloud[14] , which are available on the iPC Catalogue (see 4.2. Section).

Some of the improvements are listed below:

- **Catalogue deployment and development**

  The whole set of the data catalogue components run with the Docker[15] technology, which greatly improves the project's portability and reproducibility. The catalogue portal now runs as a submodule of the entire project, easing the testing of the different Arranger releases, while ensuring the catalogue's repository integrity.

- **User interfaces**

  The data catalogue portal technology stack has been extended by the implementation of modern web technologies, such as Redux[16] and SASS[17]. These will significantly improve the code readability and scalability.

- **Hardware**

  Resources have also been increased for a better user experience on the iPC Catalogue portal.

- **New features**
  - ❖ Authorization layer based on roles.
  - ❖ Admin section: Dedicated view for users with an admin role.
  - ❖ Data Management section: More intuitive user interfaces, download CSV functionality, …
  - ❖ Several bugs have been fixed.

4.2. Meta(data) included in the iPC catalogue

Since the first release of the iPC Computational Platform (D2.2), new datasets have been added to the catalogue (https://catalogue.ipc-project.bsc.es/). Here below is shown a list of relevant datasets currently indexed to the iPC Catalogue:

A. Data from the R2 platform (AMC)

BSC in collaboration with different institutions (AMC, CURIE) have chosen a small group of datasets to be incorporated to the iPC Data Catalogue, that comprises several cancer-types (see Section 2. Table 2). In total 22 expression matrices have been added to the iPC catalogue, which are already available for their analysis on the Virtual Research Environment.

B. Data from the OpenPBTA project (CHOP)

Clinical metadata from 950 participants were added as part of the D2.2. demonstrator. Additionally, two expression matrices have been added to the platform that encloses all the OpenPBTA participants. The aforementioned matrices are stored at Nextcloud, and therefore, accessible for their analysis on the Virtual Research Environment.

[14] Nextcloud, https://data.ipc-project.bsc.es

[15] Docker, https://www.docker.com/

[16] Redux, https://redux.js.org/

[17] SASS, https://sass-lang.com/guide

Chapter 5 Future developments

We envision a platform where end-users (researchers) will be able to visualize, filter, and select datasets from the iPC Catalogue portal, and perform their analysis from the Virtual Research Environment and/or Cavatica platforms. Moreover, users will be able to request access for protected datasets indexed to the iPC Catalogue, in a very intuitive and simple way. For achieving this objective, the current platform architecture has to be further developed for handling private datasets permissions appropriately. In this regard, there are some improvements planned for the iPC Computational Platform (Figure 2). This task will require the deployment of new services on the iPC Computational framework, such as the Permissions API, which will register user permissions over private datasets. This component will be highly dependent on the Data Access Portal, which will handle data access requests from iPC users, and dispatch them to the proper Data Access Committee, that will validate users requests (Figure 3). Eventually, dataset permissions will flow from the Data Access portal to the Permissions API transparently.

This task will require collaboration between different partners - Task 2.3 (XLAB as leader and BSC) - to detail the use cases, define the workflow specifications, and implement the services.

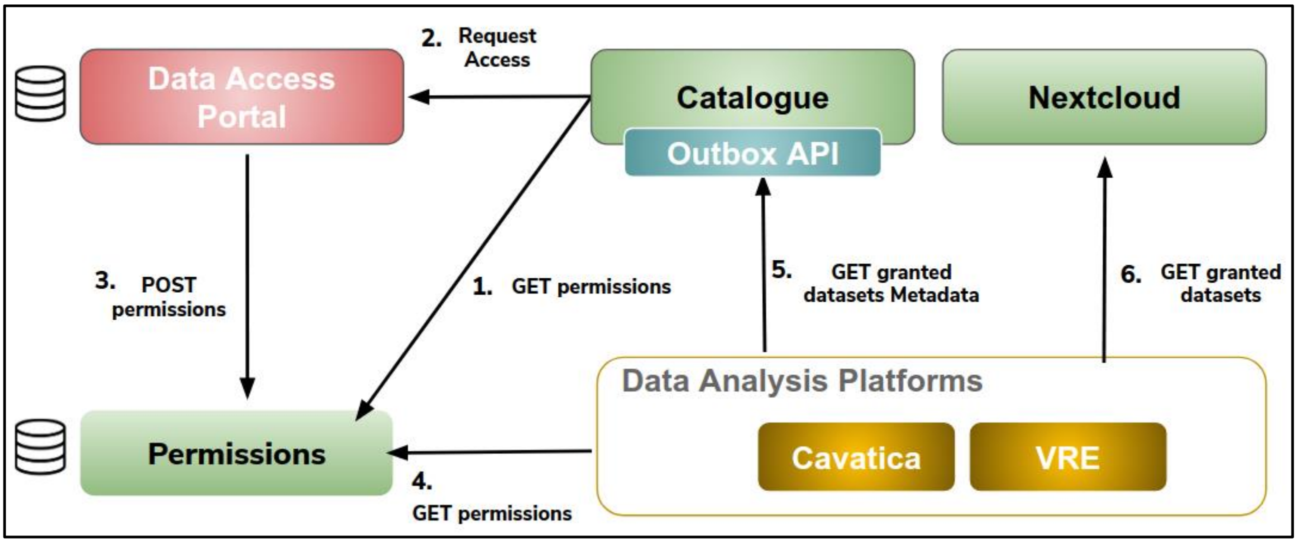

Figure 2: Tentative proposal for the data access management on the iPC Computational Platform.

After dataset/s selection, a request is made from the iPC Catalogue portal to the Permissions API for checking user's data permissions (1.). In case the user does not have access to the selected dataset/s, then, data access requests can be triggered from the Catalogue Portal to the Data Access Portal (2.), that will deliver such request/s to the proper Data Access Committee (DAC). After DAC approval (3.), the user will be able to access these datasets from the analysis platforms, that will check user's data permissions (4.), retrieve metadata from the Outbox API (5.), and import primary data from Nextcloud data storage system to the user workspace (6.).

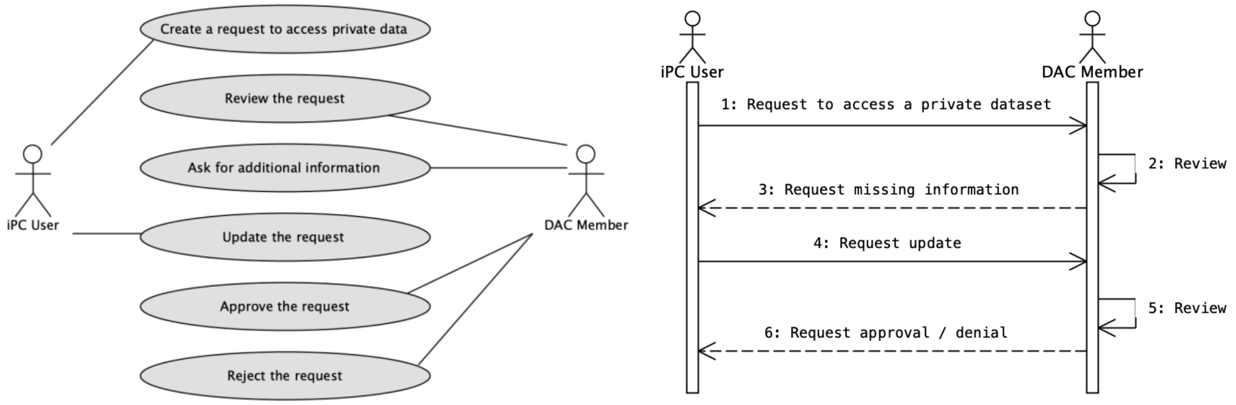
Create a request to access private data
Review the request
Ask for additional information
Update the request
Approve the request
Reject the request
iPC User
DAC Member
iPC User
DAC Member
1: Request to access a private dataset
2: Review
3: Request missing information
4: Request update
5: Review
6: Request approval / denial

Figure 3: Overview of the data access management use cases (left) and workflow (right) managed through the DAC Portal.

The DAC Portal will offer a simple interface for (1) the iPC users to issue requests to access private datasets and (2) the DAC members (and data owners) to review the requests, manage further communication with the iPC users about the requests, if needed, and, finally, approve or reject the requests.

List of abbreviations

Abbreviation	Translation
AMC	Academic Medical Center (Amsterdam)
ARGO	Accelerating Research in Genomic Oncology
BSC	Barcelona Supercomputing Center
CURIE	Curie Institute (Paris)
CHOP	Children's Hospital of Philadelphia
DA	Data Access
DUO	Data Use Ontology
EBI	The European Bioinformatics Institute
EFO	Experimental Factor Ontology
EGA	European Genome-phenome Archive
GEO	The Gene Expression Omnibus database
ICGC	International Cancer Genome Consortium
iPC	individualized Paediatric Cure
MVP	Minimum Viable Product
NCI	National Cancer Institute (United States)
NCIT	National Cancer Institute Thesaurus
R2	Genomics Analysis and Visualization Platform by the Academic Medical Center
XLAB	XLAB - Innovative IT solutions

Bibliography

[1] Ngouongo, S. M., Löbe, M., & Stausberg, J. (2013). The ISO/IEC 11179 norm for metadata registries: Does it cover healthcare standards in empirical research?. *Journal of biomedical informatics*, *46*(2), 318-327.

[2] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1-9.

[3] Côté, R. G., Jones, P., Apweiler, R., & Hermjakob, H. (2006). The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC bioinformatics*, *7*(1), 1-7.