Ref. Ares(2021)2915974 - 30/04/2021

COVID eXPONENTIAL PROGRAMME

GRANT AGREEMENT ID: 101016065

D2.1 – Sandbox Design & Datalake Creation and Ingestion

Revision: v.1.0

Work Package	WP2
Due date	30/04/2021
Submission date	30/04/2021
Deliverable lead	INTRASOFT
Version	1.0
Authors	Themistoklis Anagnostopoulos, Serafeim Tsironis, Sofia Tsekeridou (INTRASOFT) Alexandros Karatzos, Despina Gkatzioura (8Bells) Borja Aroyo, Silvia Uribe, Gustavo Hernandez (UPM) Henar Gonzalez, Elena Arredondo Lillo (SERMAS) Giulio Pagliari, Carlotta Cattaneo (ICH) Sokratis Nifakos (KI)

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

© 2020-2022 COVID-X

Reviewers	Antonio Damasceno (F6S)

Abstract	This deliverable presents the COVID-X Sandbox design, architecture, and internal services. Moreover, it describes the available datasets that are used for generating the initial COVID-X data model and will feed the Sandbox data lake.
Keywords	Sandbox, Data lake, Data Management, Data model, Data sources

DOCUMENT REVISION HISTORY

Version	Date	Description of change	List of contributors
V0.1	08/02/2021	1st version of the ToC	8BELLS
V0.2	09/02/2021	Revised ToC	All WP2 Partners
V03	20/02/2021	Added Section 2.2, 2.3 and 5.1	8BELLS, SERMAS, ICH, KI
V0.4	07/03/2021	Completed Chapter 2	INTRA, UPM, 8BELLS
V0.5	21/03/2021	Initial definition of System Requirements	8BELLS
V0.6	25/03/2021	Refined version of System Requirements	8BELLS, INTRA, UPM
V0.7	20/04/2021	Finalized System Requirements, finalized section 5.1	INTRA, 8BELLS, UPM
V0.8	27/04/2021	Added section 5.2, Updates in all chapters, introduction and conclusions added, submission for internal review	INTRA
V0.9	29/04/2021	Internal review	F6S

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

V1.0	29/04/2021	Addressing internal review comments, updates, finalization for formal submission	INTRA

DISCLAIMER

The information, documentation and figures available in this deliverable are written by COVID-X project's consortium under EC grant agreement 101016065 and do not necessarily reflect the views of the European Commission. The European Commission is not liable for any use that may be made of the information contained herein.

COPYRIGHT NOTICE

© 2020 - 2022 COVID-X Consortium Reproduction is authorised provided the source is acknowledged

Project co-funded by the European Commission in the H2020 Programme		
Nature of the deliverable:	**R***	
Dissemination Level		
PU	Public, fully open, e.g., web	X
CL	Classified, information as referred to in Commission Decision 2001/844/EC	
CO	Confidential project and Commission Services	

** R: Document, report (excluding the periodic and final reports)*
DEM: Demonstrator, pilot, prototype, plan designs
DEC: Websites, patents filing, press & media actions, videos, etc.
OTHER: Software, technical diagram, etc.

EXECUTIVE SUMMARY

The main challenge of the COVID-X Sandbox is to integrate healthcare datasets from various sources. These datasets are characterized by multiple types of heterogeneity. Fully anonymized data will be available to be collected in various formats through multiple different methods. These data must be curated, harmonized, annotated, and structurally and semantically enriched to feed third-party applications and services. The current deliverable presents the design, architecture, and internal services of the COVID-X Sandbox. In addition to that, it provides an overview of the data collection process that leads to the creation of the Sandbox Datalake. Initially, it identifies the user scenarios and the collected and analysed user requirements that the COVID-X Sandbox must fulfil and defines the identified legal and operational constraints for the implementation of the Sandbox services. Subsequently, it analyses the main concepts and existing solutions of data integration, management, harmonization and visualization, cybersecurity and federated learning fields. Based on the above analyzis and the user requirements, the overall system requirements are defined and described in a structured and detailed manner, with assigned priorities. This leads to the definition of the initial sandbox architecture for both release A and release B, including an extensive design specification of the Sandbox internal services of data ingestion and annotation, data lake, security, visualization, federating learning, that consitute the core building blocks on which third party solutions will be integrated and take advantage of, and which will be used to ingest heterogeneous data from a multitude of data providers. Finally, the document presents the available data sources, at this phase, from COVID-X clinical partners, as well as open ones, that will feed the COVID-X Sandbox and are used for generating the 1st version of the COVID-X common data model.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

TABLE OF CONTENTS

Contents

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

LIST OF FIGURES

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

LIST OF TABLES

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

ABBREVIATIONS

Accelerate	Transfer technical, business, and ethical knowledge
AI	Artificial Intelligence
API	Application Programmable Interface
CD	Continuous Deployment
CI	Continuous Integration
COVID-X	Innovation action supported by the European Commission in the framework of the EC call SC1-PHE-CORONAVIRUS-2020-2B
CSV	Comma-Separated Values
DB	Database
ELK	Elasticsearch Logstash Kibana
ES	Elasticsearch
ETL	Extract Transform Load
FL	Federated Learning
GDPR	General Data Protection Regulation
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
ICH	Instituto Clinico Humanitas
JSON	JavaScript Object Notation
KI	Karolinska Institute
MoSCoW	MUST, SHOULD, COULD, WOULD
OC	Open Calls
OWID	Our Word In Data
RBAC	Role-Based Access

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

REST	Representational State Transfer
SERMAS	Servisio Madrileno de Salud
Sigle Solution (SS)	Open call line for technology providers
Single player	*See Single Solution*
Team Solution (TS)	Open call line for clinical partners working in a team with technology providers
TLS	Transport Layer Security
WHO	World Health Organization
XML	Extensible Markup Language

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

1 Introduction

1.1 Purpose of the Document

The current deliverable presents the COVID-X Sandbox service-oriented architecture and design specification, detailing all its internal services for data ingestion, annotation, data storage, security, visualization and federated learning. In addition to that, it provides an overview of the data collection process that leads to the creation of the Sandbox Datalake.

The main challenge of the COVID-X Sandbox is to integrate healthcare datasets from various sources. These datasets are characterized by multiple types of heterogeneity. Fully anonymized data will be available to be collected in various formats through multiple different methods. These data must be curated, harmonized, annotated, and structurally and semantically enriched to feed third-party applications and services. Identifying the existing user scenarios and user requirements, as well as the overall system requirements, is a fundamental first step towards designing and developing the services that will be provided for addressing the challenges described above. Moreover, state-of-the-art analysis of the main concepts and existing solutions in the data integration and management area, as well as the cybersecurity and federared learning ones, provides a clearer insight into the tools that can be utilized for the implementation of the COVID-X services. This process leads to the initial design of the COVID-X Sandbox, which will be published in two major releases, following a Service-Oriented Architecture approach.

Another critical challenge for developing the COVID-X Sandbox is to identify the available datasets that will be used for feeding healthcare data into the system. The characteristics of each dataset are specified, including the size of the data, their type, and format. At the initial stage of the data collection process, these datasets are either provided by the clinical partners within the project consortium or collected by open data sources. At a later phase of the project, healthcare stakeholders that will join the project through the Open Calls will feed the system with additional datasets. Finally, the COVID-X common semantic data model is defined based on the available datasets variables and capitalizing on existing relevant metadata standards and ontologies. The semantic data model provides a high-level structural and semantics-based representation of the data catalog used to collect and store healthcare and other required data. The Sandbox primarily uses the semantic data model to facilitate data annotation, harmonization, and cataloging in a unified and common way for all types of aggregated heterogeneous datasets, targeting data interoperability .

1.2 Structure of the Document

The Deliverable is structured as follows:

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

- **Chapter 1** introduces the document and explains the overall purpose and structure.
- **Chapter 2** identifies the user needs and usage scenarios of the COVID-X clinical consortium partners for which the COVID-X Sandbox must provide the enabling core services to be used by third parties joining through Open Calls to implement these needs. Moreover, it defines the legal and operational constraints that need to further be addressed in the design and implementation approach of the COVID-X Sandbox. Finally, it briefly analyses the main concepts and existing approaches for data integration, management, cybersecurity and visualization areas.
- **Chapter 3** defines the general system requirements, in a prioritized manner, that lead the design and implementation of the COVID-X Sandbox.
- **Chapter 4** presents the envisioned architecture of the COVID-X Sandbox for both releases, analyzing the system actors and the Sandbox components/services, and detailing a delivery plan for both releases
- **Chapter 5** describes the identified datasets that are used for ingesting data into the system. Moreover, it presents the initial COVID-X semantic data model used by the Sandbox for data integration, harmonization, and annotation processes.
- **Chapter 6** concludes the document and specifies the follow-up activities.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

2 State-of-the-Art Analysis and User Needs

2.1 Relevant Methods and Approaches

2.1.1 Data integration, management & visualization

Data in business decisions enables organizations and companies to retrieve the best actionable and useful insights to create new business opportunities, generate more revenue, predict future trends and optimize operational efforts. Due to this importance and based also on the data explosion we have nowadays, data-driven services have become very crucial both from a business perspective and for the potential impact that they could bring for societies.

A data-driven approach is when decisions are based on analysis and interpretation of hard data rather than on observation. A data-driven approach ensures that solutions and plans are supported by sets of factual information. The meaning of data-driven is the practice of collecting and analysing data to derive insights and solutions. Data management plays a dominant role in the continuous lifecycle of data-driven solutions, while data integration drives the quality preservation of the data, and data visualization provides all these tools that show the importance and the value of this data.

2.1.1.1 Data Integration

Data integration is the problem of combining information coming from different sources and providing the user with a unified view of these data. Data integration as a process is highly correlated with the business needs of the data-driven solution since the last will drive the decision making on how source data should be combined and represented efficiently to support informed decisions.

Data source evaluation is the first step that needs to be taken into consideration while designing the data integration part of a system. It consists of a careful selection of the data sources that are necessary to match the needs and the purposes of the whole solution. These sources can be either limited access data coming from a specific provider (e.g., a clinical partner), data coming from open access sources (e.g., WHO, ECDC, etc.) and data coming from other parts/components of the internal system (e.g., statistical models or predictive data necessary for future needs).

After the definition of all these sources that will feed the entire system with data, the next step is to define and build those tools and mechanisms responsible for ingesting this information and fetching it into the system. Data ingestion relies on the establishment of all the connection endpoints necessary for harvesting data (structured or unstructured) from data sources and the development of the tools responsible for reading the information from these endpoints.

During the ingestion step, it is possible that several instances of data may be corrupted or irrelevant to the purposes of the entire system and therefore can be omitted. Data cleansing is the process responsible for filtering out any information coming from data sources that do not match the business requirements for the underlying system.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

All the “clean” and relevant information that need to be kept and processed further, is passed to the data transformation step. The data transformation process includes all the actions that can map (or combine and map) the source data to the internal components and schemas that are used to support the business logic of the entire system.

The last component of the data integration part is the data loading step which makes all the transformed data available to the data storage layer and its components. Data ingestion, data cleansing, data transformation and data loading are all building blocks of a procedure called ETL (Extract, Transform, Load).

To summarize, the data integration part is important to data-driven solutions because it improves collaboration and unification of systems, it saves time and boosts efficiency in further data processing or analysis components, it reduces the errors or any further rework and delivers more valuable data.

2.1.1.2 Data Management

Data management is the core component in data-driven solutions, and since it describes any possible ways to manipulate data for the purposes of the system, it uses the data integration part as the starting point of the whole approach.

After data ingestion is completed, data management stores and processes data using databases, big data environments, data warehouses and data lakes in an efficient, secure and reliable way in respect to the purpose of the system that uses it. The selection of what tools or methodologies should be used is related to the volume of the data as well as the business cases that need to be supported.

The most common scenario in database management is the usage of either relational or NoSQL databases based on the existing needs. The case of relational databases is a bit straightforward, as the different entities in the data model are organized into a set of formally described tables from which data can be accessed or reassembled in many ways without having to reorganize those tables.

On the other hand, NoSQL databases form a more aggregated-oriented approach as they allow us to build more complex data records suitable for the internal functionality of the system or the business requirements and any insights derived from it. The main features of NoSQL databases are : (i) multiple data model support, which allow the usage of the same data in different data model types without having to manage a completely different database, (ii) easy scalability that adapts to the data volume and complexity of cloud applications, improves performance and allows the continuous data availability, (iii) flexibility on how data will be stored and represented in order to improve efficiency, (iv) ability to distribute data in multiple locations and keep multiple copies of the same information to support continuous availability, v) zero downtime, which derives from the fact that multiple copies of data in several locations can afford the cost of any possible failure without affecting the continuous availability of data at all.

Data is ingested in its raw state regardless of format, structure or lack of structure and can be used and reused for differing purposes to match business needs. In addition to databases, data lakes consolidate data into a governed and well-managed environment that supports both analytics

COVID-X has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

development and production workloads. Like NoSQL databases, data lakes can be distributed across several locations and furthermore access can be restricted to applications or people that need them.

2.1.1.3 Data Visualization

Data visualization describes any effort that helps people to understand the significance of data by placing it in a visual context. Patterns, trends, and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization tools and techniques.

Data visualization tools have evolved during the last years, going beyond the standard charts and graphs and moving to displaying data in more sophisticated ways. Typical examples of this evolution are the infographics, dials and gauges, geographic maps, sparklines, heat maps and detailed bar, pie and fever charts. Images may include interactive capabilities, enabling users to manually manipulate them or drill into the data for querying and analysis. Indicators designed to alert users when data has been updated or predefined conditions occur can also be included.

Data visualization is a vital component in a data-driven solution, as it can effectively support decision making by highlighting the value and the importance of the data while at the same time extracting meaningful semantics from them.

2.1.2 Cybersecurity

Electronic healthcare technology is prevalent around the world and creates potential to improve clinical outcomes and transform care delivery. However, there are increasing concerns related to the security of healthcare data and devices. Increased connectivity to existing computer networks has expanded the attack opportunities, thus exposing medical devices to newly introduced cybersecurity threats. Healthcare remains an attractive target for malicious parties for two fundamental reasons: it is a rich source of valuable data and there are a lot of outdated and legacy devices vulnerable to a wide range of cyber-attacks.

Despite continuous innovations of cybersecurity in the scope of healthcare technology, the number of health data breaches climbs higher each year. Malicious cybersecurity incidents can reduce patient trust, cripple health systems and threaten human life. More alarming than the sheer number of reported attacks is the virtue of stolen property. Health data is personal and irreplaceable, that is why it is extremely important to address a system's cybersecurity with the aim to protect the sensitive data it contains [1] . In order to facilitate a secure infrastructure, changes to human behaviour, technology and processes are required.

The term cybersecurity fundamentally indicates the set of procedures and methodologies used to defend computers, servers, mobile devices, electronic systems, networks, and data from malicious attacks. Cybersecurity is therefore applied to various contexts:

- Network security: the procedures for using the network safely;

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

- Application Security: the procedures and solutions for using applications safely;
- Information security: the management of information in a secure way and in a privacy sensitive manner in accordance with pre-established regulations;
- Operational security: the security in IT operations;
- Disaster recovery and operational continuity: the procedures for restarting after problems that have affected the regular/routine operation of a system and to ensure operational continuity;
- Users' training: specific training for the actors involved in the use of the systems, which where necessary, must also include the citizen. In this framework, institutions should ensure all staff are aware of common cyber-attacks including; (i) luring victims into downloading malicious apps, (ii) phishing emails disguised as official outbreak updates which distribute malware via attachments or links, and (iii) embedded spyware or malware in publicly available interactive websites. Moreover, good "cyber-hygiene" should be incorporated into everyday working patterns, including (i) use of strong passwords, (ii) avoiding opening unknown emails and links, (iii) enablement of firewall protection at work and home, and (iv) delivery of effective staff training [2].

Cybersecurity with regards to health data is even more complex. In addition to the previously mentioned subdomains, it aims to address the following four main aspects:

- Data preservation (strong need for availability for a prolonged period).
- Data access and modification (requires authentication and authorization).
- Data exchange (requires secure data transfer).
- Legal compliance (regulatory standards, such as the GDPR law) [3].

The COVID-X Sandbox data-driven platform constitutes a turn-key technology for the COVID-X project. A tool that allows third-party software providers to connect with a variety of healthcare datasets to validate their AI models or data-driven solutions. As a technological innovation in the healthcare sector, the COVID-X Sandbox is still part of the interconnectivity between network, medical systems and devices, thus inheriting all their potential cybersecurity vulnerabilities. It encapsulates privacy by design and is equipped with security necessities in order to face the ever-changing threat landscape ahead.

According to the Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [http://data.europa.eu/eli/reg/2016/679/2016-05-04]:

- '<u>personal data</u>' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

- 'data concerning health' means personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status;

Personal health data fall under the GDPR legislation and bind Data Controllers and Data Processors with specific obligations with regards to the patients to whom the personal data belong. Nonetheless, in the framework of the COVID-X project, all data ingested in the COVID-X Sandbox will be anonymized, based on the Anonymization Guideline provided in the framework of Deliverable 1.1.

However, since the risk of de-anonymization always remains, it is crucial to secure the COVID-X Sandbox in a way to prevent any malicious incident that could lead to a data leak and could later on compromise the security of personal data.

The three main security pillars are: confidentiality, integrity and availability. Loss of data confidentiality occurs when data is exposed to those who are not authorized to use it or released prematurely ahead of its time of use or disclosure. Health data is said to have lost its integrity when it is modified by unauthorized means or persons, or under illegal circumstances, placing the original data owner at risk of misdiagnosis. Health information is said to have lost its availability when it proves difficult to access in a timely manner or when it is completely inaccessible due to system failure, virus, cyberattack, power supply, network fault, ransomware, sabotage, etc.

2.1.3 Federated learning

The most usual schemas for Machine Learning models formulation are based on the idea of moving data to computation, since they are focused on providing centralized data training models that are executed on a simple machine or a data centre by applying consecutive data manipulation techniques. This is a common solution for many different scenarios in the Artificial Intelligence (AI) research area, but it may bring some security gaps, especially for those applications which need a reliable data exchange.

In this context, privacy-preserving concerns about typical artificial intelligence solutions have contributed to the definition and development of the Federated Learning paradigm (FL) [4], which is based on turning "the data to computation movement" idea around. Now the computation is moving while the data stay, appearing a new distributed approach that enables different entities to collaboratively define a shared model without exchanging any training data, that is, while storing them into their local premises. This helps train on large corpus in a decentralized way [5], addressing the main problems about privacy, ownership and locality of data, which is vital for COVID-X.

Following this, FL aims at solving the problems associated with centralized learning such as data limitation, storing and privacy that may appear in common cloud-based systems. For doing so, this new approach is based on two main properties, that is, the separability and the linearity of standard multivariate modelling functional, that allow to divide a functional across different data partitions.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

In this regard, the capability of FL for firstly distributing the models' execution and then merging the partial results to obtain a more complex one can be considered as a special advantage for environments such as COVID-X, since it enables health organizations located at different geographical locations to develop AI models in a collaborative way without providing direct access to potentially sensitive or classified users' data. With this in mind, several researches have explored this privacy preservation while maintaining a high model performance during the last years, as can be seen in [6], where the authors used real electronic data of 1 million patients, and in [7], where a new strategy for the use of large amount of data for knowledge creation in radiotherapy field was proposed.

This new paradigm involves a corresponding network infrastructure with an arbitrary topology and the definition of different APIs for authorized access to local data, security and encryption mechanisms, model transfer protocols, and model update procedures based on weight aggregation and model validation to be applied after the training. *FIGURE 1* shows a simple overview of a FL infrastructure that can be applied to the COVID-X scenario, where we will have a primary node and then the clinical partners' sites acting as secondary nodes, each of them having access to local datasets for privacy concerns.

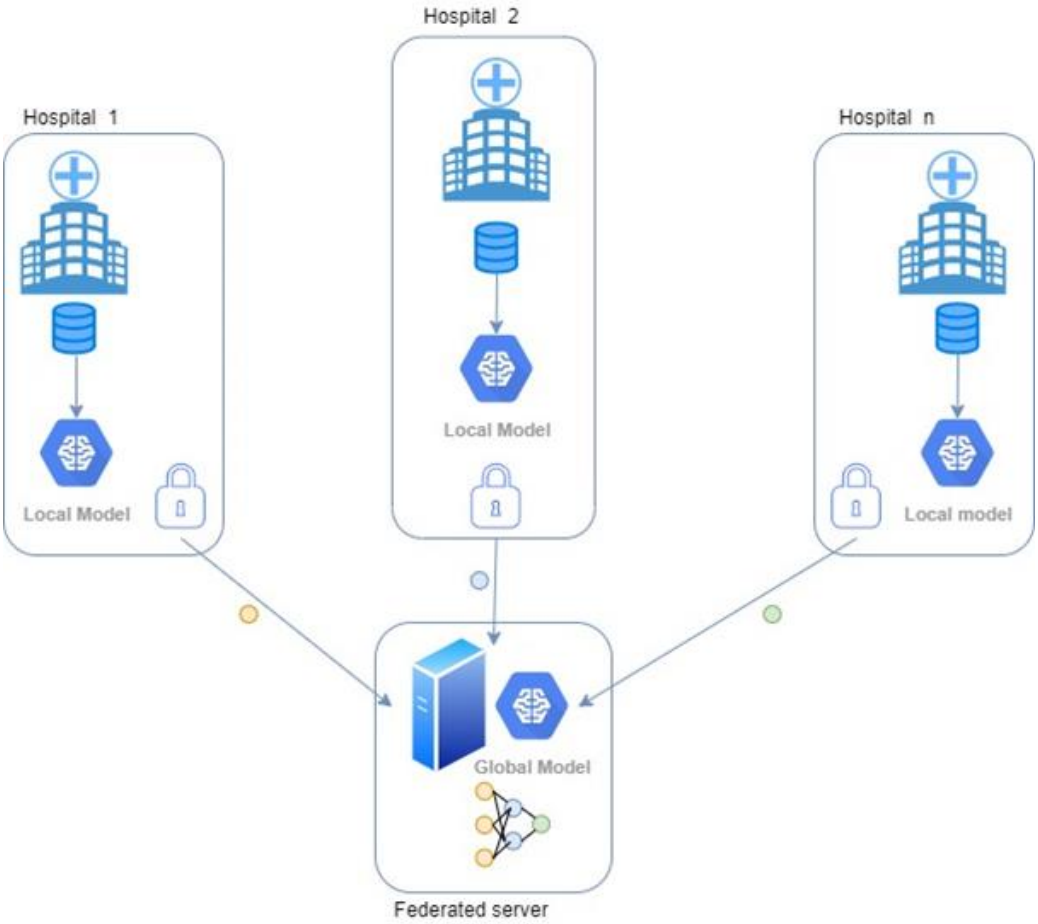

FIGURE 1: FEDERATED LEARNING OVERVIEW IN HEALTH SCENARIO

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

In general terms, current federated learning approaches can follow one of the three next different categories, as can be seen in *FIGURE 2*:

- Horizontal federated learning, when the datasets from the different decentralized nodes have the same feature space.
- Vertical federated learning, when these spaces are different, but they are used for jointly training a global model.
- Federated transfer learning, which is similar to the previous one but with a pre-trained model trained on a similar dataset but in a different problem.

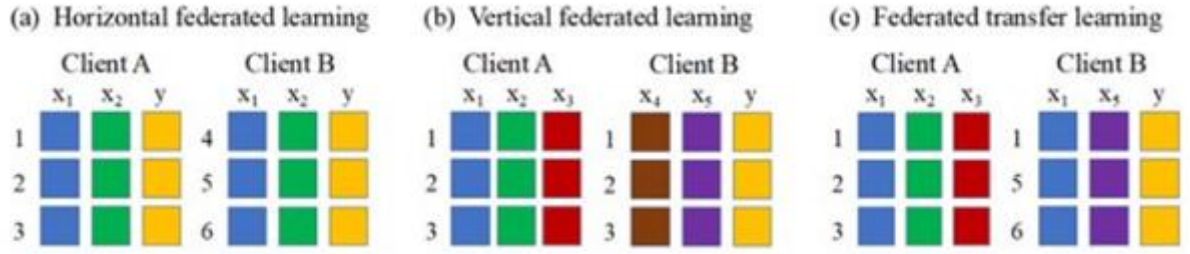

FIGURE 2: CATEGORIZATION OF FEDERATED LEARNING APPROACHES (BY [8]*)*

As can be seen, this new AI approach provides interesting capabilities, but it also includes several challenges that need to be addressed mainly related to the system and statistical heterogeneity, privacy concerns, communication efficiency, etc. [9]. In the case of COVID-X, the analysis of the user requirements together with the definition of the second version of the platform will help us define which of those challenges need be considered.

2.2 User Needs and Usage Scenarios

This section presents the user needs and targeted usage scenarios for each clinical challenge, as defined by the COVID-X clinical consortium partners (ICH, SERMAS, KI). More specifically, it analyses which are the user needs and interactions in each case and what is required to meet those needs.

2.2.1 Istituto Clinico Humanitas (ICH)

Istituto Clinico Humanitas (ICH) is located in Milan, at the North Italy area, and provides healthcare services for about 1 million patients/year. During COVID-19 pandemic, ICH treated 3,000 patients (about 150 ICU, 1,000 hospitalized and 2,000+ managed by ER).

The main objective of ICH is to provide fast, accurate and combined analysis of clinical data available in ER, diagnostic imaging (CT for ICH) and historical profile of patients. Moreover, ICH aims at detecting,

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

classifying, and treating infected patients as early as possible. Finally, it envisions efficient patient-centred diagnostic and AI-based solutions for cross analysis between clinical and radiological data to assess personalized clinical paths and treatments.

For this purpose, ICH utilizes a combination of CTs and aggregated datasets (90+ features per patient), viral load, clinical records, and real historical data from the first COVID-19 wave from ICU, hospitalized patients, and the ER department.

2.2.2 Servicio Madrileño de Salud (SERMAS)

The Hospital Clínico San Carlos is a tertiary care center from Madrid (Spain) belonging to SERMAS. It covers a population of more than 300k citizens. From the beginning of the SARS-CoV2 pandemic in Madrid, this center has attended more than 3000 hospitalized patients diagnosed with COVID-19.

The main interest of the health professionals caring for hospitalized patients with COVID-19 is to be able to swiftly and precisely diagnose this condition and assess its prognosis. In particular, we are interested in being able to identify “exceptions to the rule”, i.e. subjects belonging to risk groups but who will fully recover with no or minor complications, and vice versa.

In order to be able to achieve these aims, SERMAS offers real world data generated during the hospitalization on COVID-19 patients, at the emergency department, hospitalization and intensive care unit (ICU). Data includes codified diagnoses, comorbidities, and procedures at discharge (ICD10), codified treatment prescribed during the stay (Spanish Agency of Medicines and Medical Products code), administrative data regarding date of admission and discharge (and motive), admission in the ICU, lab work, radiology images, and microbiology tests (e.g. Polymerase Chain Reaction).

2.2.3 Karolinska Institutet (KI)

Karolinska Institutet (KI) is one of the world's leading medical universities, it accounts for over 40% of the medical academic research conducted in Sweden and offers the country's broadest range of education in medicine and health sciences.

The management of patients in the Emergency Departments (ED) when COVID-19 should be considered or even patient can be infected, for instance patients with chest pain with COVID-19, is chaotic with negative consequences not only for patients with chest pain not needing emergency care and for ED patients generally, who need emergency care, that became more complicated. Data generated with COVID-19 patients in ED and other clinical manifestations is a useful asset. KI is interested in establishing collaboration with solutions that can provide evidence for recommending and supporting both health professionals and patients in an ED to potential COVID-19 and providing appropriate guidelines for the attention.

COVID-X has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

2.3 Legal and Operational Context and Constraints (for 1st and final releases)

This section presents the legal and operational constraints that need to be considered in the design and implementation of both the 1st and final releases of the COVID-X Sandbox.

2.3.1 Sandbox - Legal context and constraints

The main legal constraint to solve is to ensure compliance with the guidelines contained in the GDPR regarding the use of personal data. Usually patient data are obtained in a healthcare context from which the collaboration of the patient to participate in a clinical trial may or may not be obtained. In the latter case, the patient signs a consent (which he/she can revoke at any time) by which she/he authorizes his/her data to be used for a specific purpose.

Unfortunately, the subset of data for which such consent is available is small compared to the large volume of data that could be used. In addition, the treatment of this subset of data for which there is consent is subject to the limits of the clinical trial and at most it could be extrapolated to a similar pathology (for example, prostate cancer, extrapolated to the specialty of oncology), being prohibited that they are used for a different purpose for which they were collected. However, in a research process it is frequent that the results of the same can derive benefits towards other pathologies that could not be related to the initial purpose and could be of great interest.

These restrictions greatly limit the possibilities and benefits that could be obtained from large-scale data processing. To resolve these limitations, the most appropriate strategy is to subject the data to an anonymization process that avoids the identification of patients by any mechanism, after which they are no longer considered personal data by the GDPR and can be used more freely, establishing crosses and associations on them, in short what is the essence of the research.

Other considerations that must be taken into account such as, the analysis of the risks to which the data sets are subjected, the different roles and responsibilities that intervene in their management and are part of the organization's security policy, the measures of security (technical and organizational) and the additional guarantees of control and audit are developed in detail in the deliverable *D1.1 - Ethical and Legal Framework* [10].

Deliverable D1.1 deals with the following issues:

- Chapter 2 provides an overview about legal issues, regulatory aspects, and problems about biomedical research.
- Chapter 3 provides a description of applicable EU legislation and local normatives.
- Chapter 4 provides a description of general regulations about scientific investigation.
- Chapter 5 provides the privacy requirements for information systems and principles from GDPR and other constraints that must be taken into account.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

2.3.2 Sandbox - Operational conditions and constraints

The operation of the COVID-X systems must be managed diligently, taking the appropriate measures to protect them against accidental or deliberate damage that may affect the availability, integrity, confidentiality, intended use and value of the information processed or the services provided.

TABLE 1 summarizes the major system conditions and constraints that need to be taken under consideration when designing the COVID-X Sandbox.

TABLE 1: SANDBOX OPERATIONAL CONDITIONS AND CONSTRAINTS

Req.ID	Description
[COV_CON_001]	One COVID-X Sandbox instance may be created for each clinical partner on own controlled access infrastructure, for which the required legal certifications apply and which are managed by the respective organization/data controller.
[COV_CON_002]	One COVID-X Sandbox instance may be created for each third-party team solution on the infrastructure of each team's healthcare provider if relevant legal restrictions apply for the third-party healthcare provider to move its data to a cloud-based deployment of the COVID-X Sandbox.
[COV_CON_003]	Only authenticated and authorized users can access the Sandbox.
[COV_CON_004]	Users' activities are monitored, and malicious behaviour is detected.
[COV_CON_005]	Users are assigned to the minimum privileges needed.
[COV_CON_006]	Source code requires unit and integration testing before it's deployed.
[COV_CON_007]	Only trusted container images from the project's private docker registry are allowed to enter the cluster.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

3 General System Requirements

3.1 Requirements Gathering Methodology

The elicitation or communication of end-user requirements comprises an early and critical but highly error-prone stage in system development. Socially oriented methodologies provide more support for user involvement in design than the rigidity of more traditional methods, facilitating the degree of user–designer communication and the 'capture' of requirements. A more emergent and collaborative view of requirements elicitation and communication is required to encompass the user, contextual and organisational factors.

During the requirement's gathering methodology phase, the main focus is on understanding users, their needs and how they operate within the context of the proposed system, which benefits greatly the gathering of technical and functional requirements of the system. Part of this process also involves specifying the criteria that can be used to evaluate the operation of the system. That is to say, how the system will attain a greater degree of useability, resource management, security and privacy, scalability, maintainability will be described by a set of non-functional requirements. To that end, it is necessary to understand the challenges, the stakeholders, and their actions within the scope of each challenge, as well as the application of the available tools, that are needed to tackle these challenges.

The requirements methodology phase of a development project is also characterised by intense communication activities and involves a diverse range of people who differ on levels of background, skill, knowledge, and status. The goal of such activities, as already asserted, is to achieve an understanding of the problem and one that must be shared between disparate people, a task made all the more difficult by the complexity, vastness and volatility of the requirements [11].

Modern data platforms aim to establish a centralized system which combines scalable flexibility with distributed data storage and computational power for acquiring and analysing large data sets to provide users with reliable and accurate data, visualizations, and analytics [12]. The COVID-X Sandbox aims to cover the aforementioned demands, by incorporating a wide spectrum of innovative technologies, security practices and regulations regarding personal data protection and scientific investigation. The role of the COVID-X Sandbox is to bridge the gap between technology and healthcare and it acts as a mediator between technical and healthcare providers, providing the first with datasets from the latter in order to validate their analysis and accelerate their go-to-market. These datasets set the requirements for the database schemas and defined the necessity for a user group management system within the COVID-X Sandbox.

Continuous contact and communication with partners helped to capture its unique features. As mentioned earlier, to gather and identify the needed requirements, it is deemed necessary to examine the stakeholders involved in each clinical challenge of the project, understand their needs and their actions, since they will be mainly interacting with the COVID-X Sandbox. A first approach of the COVID-X Sandbox was developed as a conceptual design. As new challenges made their appearance,

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

continuous improvements over its architecture took place, leading to the creation of the First Release plan.

The above methodology was implemented as it is necessary to gather the requirements for the hardware infrastructure of the COVID-X Sandbox, its functionalities and the data structure of private and public datasets which will be used in the implementation of the clinical challenges.

All gathered and analyzed requirements are presented in subsequent sections in a tabular format, containing the requirement ID, description and priority. Priorities will be attributed to each requirement following the MoSCoW prioritization method in order to deliver the greatest and most immediate benefits early.

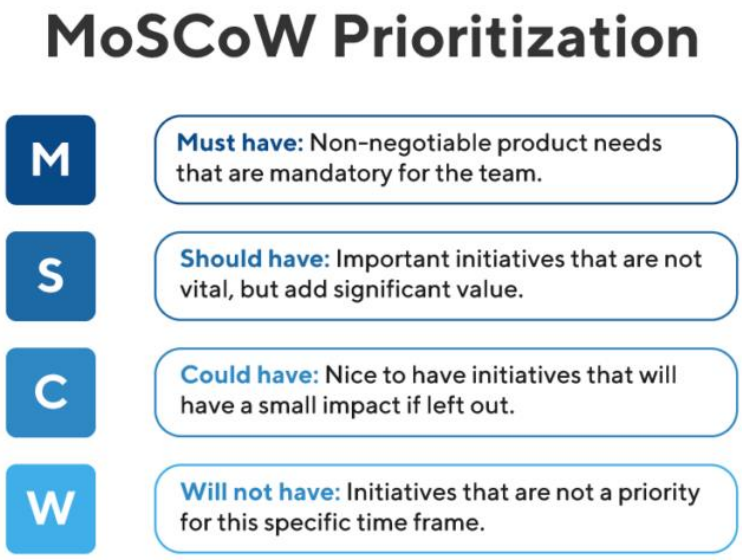

FIGURE 3: THE MOSCOW PRIORITIZATION METHOD

3.1.1 Hardware infrastructure

Clinical partners may be providing the hardware infrastructure that will host the COVID-X platform. Therefore, after reviewing the needs and challenges of the COVID-X project, it was concluded that a well-tempered data platform requires as an initial local deployment a multi-core machine with multiple drives to store volumes of the data that will be stored through the lifetime of the project.

3.1.2 Clinical challenges

The clinical challenges that delineate the functionalities of the COVID-X Sandbox were defined by the clinical partners, based on the goals they aim to achieve throughout the lifetime of the COVID-X Project. These challenges are defined in detail in Deliverable *D1.1 - Ethical and Legal Framework* [10], but are also briefly presented in Section 2.2 of the current Deliverable.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

3.1.3 Open data sets

The COVID-X Sandbox envisions to access easily, uniformly and securely various health data sources (e.g. Electronic Health Records, medical images) – including both public data available through respective open access repositories such as the COVID-19 Section of the European Data Portal, or the COVID-19 Data Portal, and private databases supplied by the clinical partners.

After the initial selection of the available open datasets related to Covid-19, instructions/feedback was provided by the clinical partners (KI, HUM, SERMAS) regarding which of the datasets were more suitable in addressing each challenge. Moreover, only datasets from trusted sources, containing structured or image data, were chosen to be ingested in the COVID-X Sandbox.

3.1.4 Private datasets

Concerning the private datasets supplied by the clinical partners, discussions took place with respect to their characteristics and the challenges that they aim to tackle. In order to give us further information on these characteristics and facilitate the gathering of the requirements with respect to the initial design and implementation of data ingestion tools, the clinical partners provided us with synthetic information having the exact same format/structure with the real data.

In order to ensure COVID-X Sandbox compliance with GDPR and other national regulations regarding the protection of personal data, a number of internal discussions/interviews took place and a common solution accommodating all clinical partners needs was established. More specifically, the ingestion of *only* anonymized data within the COVID-X Sandbox was decided. In this regard, a comprehensive and detailed **Anonymization Guideline** was released in the framework of *D1.1 - Ethical and Legal Framework* [10]. The said Guideline proposes an anonymization technique to follow before ingesting data in the COVID-X Sandbox, while it also defines a Privacy Impact Assessment (PIA) methodology in order to carry out a Risk Analysis of the anonymization process to subsequently manage the resulting risks with technical, contractual, organizational or any other measures. This is required to address the compliance of the COVID-X Sandbox with relevant regulations. To that end, the above document took into consideration all applicable European (EU) Legislation and local normatives on data protection, along with EU and national regulations on Scientific Investigation.

3.2 Major System Capabilities

3.2.1 Performance

Performance is a pervasive quality of software systems and it is affected by every aspect of the design, code, and execution environment. Before measuring performance, it is mandatory to define and analyse system requirements in order to identify concerns or factors that will affect its performance. Part of this process entails the definition of the system's operational profile, its workload intensities as

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

well as a description of its behaviour in various scenarios. After their interactions between system resources are defined with respect to the architecture, it is necessary to proceed to performance tests. These involve stressing the system's individual units in order to measure performance and validate that each part of the system performs as expected. This is an ongoing process, since more additions and changes will be added through the duration of the project. The above procedures provide an overall perspective about the performance capabilities of the system which are mainly tied to its requirements.

The overall performance capabilities of the COVID-X Sandbox are tightly related to the capabilities of the Elastic Stack and are presented in the table below:

TABLE 2: SANDBOX PERFORMANCE CAPABILITIES

Req.ID	Description	Priority
[COV_CAP_001]	Sandbox Services support high availability.	M
[COV_CAP_002]	Sandbox services support fault tolerance.	M
[COV_CAP_003]	Near-real time data ingestion.	M
[COV_CAP_004]	High throughput for data queries.	M
[COV_CAP_005]	System latency should be minimal.	M
[COV_CAP_006]	System can handle multiple logins.	M
[COV_CAP_007]	System can host large datasets	M
[COV_CAP_008]	System can scale on multiple nodes.	M

3.2.2 Manageability and Maintainability

The manageability and maintainability of COVID-X Sandbox are main characteristics of a functional system and describe the robustness of its tools and services. COVID-X Sandbox software can be configured for optimal performance, availability. In terms of security, access control lists and authorization are set up, ensuring protection against hostile activity.

The CI/CD pipeline benefits the process of managing and maintaining the software as the integration, testing and deployment procedures become automated. In addition, setting up each pipeline in the development environment becomes easily repeatable. CI/CD stack includes monitoring tools and logs, which assist in the process of troubleshooting during the development. To ensure that a section of an application meets its defined requirements and behaves as intended we will also employ a unit testing process.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

TABLE 3: SANDBOX MANAGEABILITY & MAINTAINABILITY CAPABILITIES

Req.ID	Description	Priority
[COV_MNT_001]	Highly configurable and autonomous way to build/test and deploy through a CI/CD pipeline	M
[COV_MNT_002]	CD tools support version control repositories for managing code changes.	M
[COV_MNT_003]	Managing the environment / system variables and configuring them for the target environment.	M
[COV_MNT_004]	Executing continuous tests and rollback environments if tests fail.	M

3.2.3 Portability

The COVID-X Sandbox will require multiple deployments, since many participants will join with their solutions over the course of the COVID-X program. The robustness of a system is a key concern when it comes to multiple deployments in different environments, introducing a requirement for the system to always produce the desired results. On a larger scale, this is rather costly and requires lots of resources for each deployment. Thus, the COVID-X Sandbox was designed to be portable in order to be deployed at the local environment of each healthcare provider in an automated way. It was also designed to be highly configurable in order to adapt to different operational environments, thus reducing the development cost. Tools and services provided by third party SMEs will be part of the integration, testing and deployment process of the COVID-X Sandbox as well. For this purpose, a CI/CD approach was adopted, utilizing virtualization technologies, which offer logical isolation and increase system interoperability and robustness, as it is easier to implement them remotely with a small overhead.

TABLE 4: SANDBOX PORTABILITY CAPABILITIES

Req.ID	Description	Priority
[COV_POR_001]	The CI/CD pipeline allows copying selected configuration settings from one system to another system.	M
[COV_POR_001]	Containerized deployment allows Sandbox to be independent of the underlying platform	M

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

3.3 Operational Requirements

The operational requirements refer largely to the mission and purpose of the COVID-X Sandbox and aim to describe and communicate the end state of the world after the system is deployed and operated [13]. The COVID-X Sandbox is to be used by the following user groups: clinical partners, third-party technical partners, third-party healthcare providers (in the case of team solutions) and administrative users (technical partners of the COVID-X consortium). Each user group is accompanied by a set of operational requirements that delineate its activities. The operational requirements are listed in the table below:

TABLE 5: OPERATIONAL REQUIREMENTS

Req.ID	Description	Priority
[COV_OP_001]	Clinical partners and third-party healthcare providers should connect to the COVID-X Sandbox through secure REST APIs in order to ingest their data.	M
[COV_OP_002]	Clinical partners and third-party healthcare providers should provide technical infrastructure with a minimum set of requirements in order to deploy the COVID-X Sandbox.	S
[COV_OP_003]	Third-party healthcare providers should provide only the necessary and minimum anonymized data required for the third-party technical providers to perform their analysis, as GDPR implies.	M
[COV_OP_004]	Third-party technical providers should be able to query the data that they have access to in order to visualize COVID-X Sandbox data and have control over its services via interactive and customizable dashboards.	M
[COV_OP_005]	Administrative users should connect to the COVID-X Sandbox in order to process the control-access lists, logs and backups.	M

3.4 Functional Requirements

The COVID-X Sandbox fulfils the following functional requirements:

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

TABLE 6: SANDBOX FUNCTIONAL REQUIREMENTS

Req.ID	Description	Priority
[COV_FUN_001]	Role-based access control system (RBAC) for user management.	M
[COV_FUN_002]	Data transfer based on encryption protocols.	M
[COV_FUN_003]	Data ingestion in batch mode for static data and in streaming mode for data coming from IoT devices, sensors or wearables.	M
[COV_FUN_004]	Data harmonization / transformation using filtering tools.	M
[COV_FUN_005]	Data cataloguing along with metadata.	M
[COV_FUN_006]	Data indexing for fast querying and retrieving.	M
[COV_FUN_007]	Users are entitled to access only the part of the data necessary for their analysis (principle of least privileges).	M
[COV_FUN_008]	All inbound and outbound operations are logged.	M
[COV_FUN_009]	Activity monitoring for abnormal behavior detection.	S
[COV_FUN_010]	Provide the tools to perform federated validation.	M
[COV_FUN_011]	Unified summary of data catalogues should be supported.	M
[COV_FUN_012]	Visualization tools to display aggregated data, infographics.	M
[COV_FUN_013]	Full text search is supported.	M
[COV_FUN_014]	Supported input data connectors should be at least through file ingestion, database dumps and REST APIs.	M
[COV_FUN_015]	Structured and unstructured data are supported.	M
[COV_FUN_016]	Ingestion of various data formats is supported.	M

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

3.5 Policy and Regulation Requirements

Since the COVID-X Sandbox entails the collection, process, storage, and cataloguing of a large variety of healthcare data sets from health stakeholders, it is required to be compliant with a number of policies and regulations at European, national and, even, enterprise level. These policies and regulations are mainly associated with data protection regulations [3] and the medical device regulations [14], [15]; however, general regulations about scientific investigation also apply.

All relative regulations can be found in the Deliverable *D1.1 - Ethical and Legal Framework* [10] - (Section 2 - APPLICABLE EUROPEAN UNION (EU) LEGISLATION AND LOCAL NORMATIVES & Section 3 - GENERAL REGULATIONS ABOUT SCIENTIFIC INVESTIGATION) [10].

3.6 Non-Functional Requirements

3.6.1 Safety Requirements

The table below we describe the safety requirements that entail the use of the COVID-X Sandbox and aim to cover the Policy and Regulation Requirements, presented in the previous section:

TABLE 7: SANDBOX NON-FUNCTIONAL SAFETY REQUIREMENTS

Req.ID	Description	Priority
[COV_SAF_001]	Proper training on system tools & services should be provided	M
[COV_SAF_002]	According to GDPR regulations sensitive health data should be anonymized and managed in a secured and-safe environment.	M
[COV_SAF_003]	According to GDPR regulations, only the necessary and minimum data required for the analysis performed by the third-party technical providers should be provided to them.	M
[COV_SAF_004]	Third-party healthcare providers should follow the Anonymization Guideline that was published in the framework of Deliverable 1.1.	S

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

[COV_SAF_005]	Only anonymized data, that comply with medical protocols (HL7 FHIR, ICD 9-10-11, DICOM), are hosted by the COVID-X Sandbox.	M
[COV_SAF_006]	Partners joining the program should sign a contract that will hold them responsible for their data anonymization procedures.	M
[COV_SAF_007]	Information considered confidential cannot be disclosed.	M
[COV_SAF_008]	In case of a data breach, third parties participating in the program should be notified immediately.	S
[COV_SAF_009]	COVID-X Sandbox should store the anonymized data provided by the clinical partners and third-party healthcare providers until the end of the contract signed by these parties.	M
[COV_SAF_010]	In case of improper use of the allocated resources, certain actions will be taken to notify both the users and the administrators, thus preserving the smooth operation of the infrastructure.	C

3.6.2 Performance Requirements

Performance requirements constitute one of the main drivers of architectural decisions. As the majority of performance problems have their roots in architectural decisions, and since poor performance is a principal cause of software project risk, it is essential that performance requirements are set early in the software lifecycle, and that they be clearly formulated [16].

In order to assess the performance of a system the following four types of requirements must be clearly specified.

3.6.2.1 Response Time

In order to delineate the response time requirements of the COVID-X Sandbox, the following *use-case - requirement* pairs need to be considered:

- Data coming from medical and IoT devices and wearables, need to be served in a near real-time manner.
- Response to an end-user's request to access data should be real-time.
- Database queries need to be promptly served, in order to enhance user experience.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

The COVID-X Sandbox aims to meet the real-time requirements by capitalizing on Elastic Stack, which provides a scalable and fast way of ingesting data from various sources and storing them in one place, where they are indexed in order to be easily, securely and quickly accessed.

3.6.2.2 Workload & load balancing

The performance of the COVID-X Sandbox depends on how the load is distributed within the system. Therefore, it is important that the workload profile is defined. The workload could refer to simultaneous users, storage capacity, the maximum number of simultaneous transactions handled, or anything else that pushes the system past its original capacity. The workload peaks around lunchtime and late evening but activity is very quiet during the night; hence, administrative tasks, such as backups, should be made during low-workload times.

To define the workload profile, we must take into account all scenarios executed by users, such as data ingestion, data access, execution of database query, but also all administrative tasks that should take place, like data transformation/harmonization, security logging, alert handling and backups. We should also consider error scenarios and handling.

3.6.2.3 Scalability

System's scalability refers to the extent to which an increase to the system's workload can be handled by the system. The response time requirements should still be met as the workload scales. A system is considered scalable when it doesn't need to be redesigned to maintain effective performance during or after a steep increase in workload. In our case, the COVID-X Sandbox has a limited and well-defined number of users. Also, since it will be employed in a federated manner, the amount of data accommodated in each COVID-X Sandbox instance will not escalate uncontrollably. Therefore, scalability should not present any issue with regards to the federated deployment of the COVID-X Sandbox. The requirements of the centralized COVID-X Sandbox, although different, should not raise any significant scalability concern as well, since the data that it will host, along with the security services (logs, alerts, backups) that will accompany it, will be limited in comparison to the local Sandbox deployments.

3.6.2.4 Platform

In order to achieve the desired performance for the COVID-X Sandbox, we also need to consider the platform on top of which our system will be built. A platform is defined as the underlying hardware and software (operating system and software utilities) which will host the system. First of all, the COVID-X Sandbox is deployed using the Elastic Stack, a reliable and secure way to take data from any source, in any format, then search, analyze, and visualize it in real time. Secondly, the COVID-X Sandbox has a highly modular architecture as different services are deployed in separate Docker containers that will communicate with each other through an overlay network. Moreover, the platform consists of external resources, such as connections to external databases (e.g MongoDB), therefore the response times of these external resources must also be specified. In our case, possible

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

connection to third-party IoT devices needs to be considered as well when trying to achieve the desired performance.

The overall performance requirements of the COVID-X Sandbox are presented below:

TABLE 8: SANDBOX PERFORMANCE REQUIREMENTS

Req.ID	Description	Priority
[COV_PER_001]	High availability.	M
[COV_PER_002]	Responsiveness and reliability.	M
[COV_PER_003]	COVID-X Sandbox services should be easily accessible to the users.	M
[COV_PER_004]	Data coming from medical and IoT devices and wearables should be served in a near real-time manner.	M
[COV_PER_005]	Database queries should be served quickly, so as to enhance user experience.	M
[COV_PER_006]	Administrative tasks, such as backups, can be made during low-workload times.	S

3.7 Security Requirements

This section summarizes both functional and non-functional security requirements. These requirements aim to protect the COVID-X Sandbox from unauthorized data access, data leaks, excessive usage of resources and other vulnerabilities a data platform may exhibit. Along with the Safety Requirements presented in the previous section, they aim to cover the Policy and Regulation Requirements of the COVID-X Sandbox.

TABLE 9: SANDBOX SECURITY REQUIREMENTS

Req.ID	Description	Priority
[COV_SEC_001]	The COVID-X Sandbox has a well-defined number of users.	M
[COV_SEC_002]	Users need to insert their credentials (username & password) in order to access the COVID-X Sandbox.	M

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

[COV_SEC_003]	The system verifies users' credentials in order to provide them access to its services.	M
[COV_SEC_004]	Backups of the configuration files, access control lists and other important files should be kept.	S
[COV_SEC_005]	All data transfers should be encrypted.	M
[COV_SEC_006]	Every user is entitled to specific privileges that are the minimum privileges needed to perform their analysis.	M
[COV_SEC_007]	User roles describe all privileges a user with a certain role is entitled to.	M
[COV_SEC_008]	Users should only be able to perform actions dictated by their assigned roles.	M
[COV_SEC_009]	The production of faulty software that could compromise security should be avoided.	S
[COV_SEC_010]	All activities that take place within the COVID-X Sandbox are monitored.	S
[COV_SEC_011]	In case malicious activities are detected inside the COVID-X Sandbox, alerts are generated.	S
[COV_SEC_012]	Only authorized modules/containers should be able to connect to the containerized environment of the COVID-X Sandbox.	M
[COV_SEC_013]	The COVID-X Sandbox should use exclusively genuine software and ban all illegitimate software and applications [14].	M

3.8 Interface Requirements

The Interface Requirements Specification (IRS) specifies the requirements imposed on one or more systems, subsystems, Hardware Configuration Items (HWCIs), Computer Software Configuration Items (CSCIs), manual operations, or other system components to achieve one or more interfaces among these entities. The IRS can be used to supplement the Requirements of a System or Software as the basis for design and qualification testing of systems and CSCIs [16].

The table below summarizes all interface requirements related to the COVID-X Sandbox, its internal components' interactions, and its interaction with external systems.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

TABLE 10: SANDBOX INTERFACE REQUIREMENTS

Req.ID	Description	Priority
[COV_INT_001]	APIs follow a REST API architecture.	M
[COV_INT_002]	API Gateway should enable authorized external users to securely connect to Sandbox Services.	M
[COV_INT_003]	API to retrieve data from Sandbox datalakes.	M
[COV_INT_004]	Sandbox APIs should enable internal communications between Sandbox services and the deployed third-party containers implementing application-oriented solutions.	M

3.9 Other Considerations

The considerations that have emerged during the designing process of the COVID-X Sandbox mainly describe the need for proper user training in order to cultivate a "security culture" and build a more robust system.

In order to ensure proper usage of the COVID-X Sandbox, it is extremely important to ensure that all its user groups are properly and adequately trained with respect to their view of the system.

All users are encouraged to employ cyber smart behaviour, such as paying attention to privacy, being aware of suspicious messaging, and browsing responsibly. Instruction for use should include the necessary information so that users can be up-to-date with the latest version of software, use sufficiently complex passwords, that are frequently being changed, secure the computer where the COVID-X Sandbox will be implemented and use backups and protection of their healthcare data, in case of on-premise implementation for Team Solutions. They should also be aware of general EU regulations and laws associated with data protection.

Moreover, best practices regarding cybersecurity of healthcare applications dictate to install only software programs necessary for the intended use of the operating environment [17], thus minimizing dependence on third-party software vulnerabilities.

A summary of these considerations is presented below:

- Privacy awareness concerning suspicious messaging and browsing.
- Refer to the COVID-X Sandbox User Guide located in Deliverable2.2 which contains information for the use of the COVID-X Sandbox software.
- Awareness of 3rd party software dependencies and vulnerabilities.
- Use of sufficiently complex passwords for user accounts in each user group

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

- Passwords should be frequently changed.
- Clinical partners and healthcare providers should keep backups of their data.
- Awareness of General EU regulations and laws associated with data protection.
- Installation of only the necessary software for the intended use of the operating system.
- The COVID-X Sandbox should have technical support throughout the project's lifetime.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

4 Covid-X Sandbox Architecture and Design Specification

4.1 Design of the Sandbox Architecture

The purpose of the COVID-X Sandbox is to create a technological platform that enables the collection, process, storage, and cataloguing of a large variety of healthcare data sets from health (emergency) stakeholders as well as open data sets. Sandbox utilizes various services that work collectively to fulfil the functional and non-functional requirements presented in Chapter 3. This section describes the main principles and design of the Sandbox architecture. Firstly, it presents the Service-Oriented Architecture style that is adopted for the Sandbox system and software design. The following section presents the timetable for the Sandbox implementation, which will produce the two planned major releases. Furthermore, it provides a high-level overview of the envisioned Sandbox architecture for both releases. Finally, it describes the design specification of each internal component.

4.1.1 Service-Oriented Architecture

The COVID-X Sandbox introduces a combination of different functions and services that collectively aim to enable seamless access to a set of healthcare data sources. The Sandbox development's main axes allow the seamless digestion of information between various services and data sources and allow the system's maximum modularity in its integration and deployment. The Sandbox as a toolset must offer flexibility and scalability in order to be able to fulfill the needs and the requirements of the use case.

To accomplish the objectives mentioned above and provide a fully integrated and robust system, we decided to follow the ISO/IEC 18384 Service Oriented Architecture (SOA) standard [18]. SOA defines a way to make software components reusable via service interfaces. These interfaces utilize common communication standards in such a way that they can be rapidly incorporated into new applications without having to perform deep integration each time.

Each service in an SOA embodies the code and data integrations required to execute a complete, discrete functionality. The service interfaces provide loose coupling, meaning they can be referenced with little or no knowledge of how the integration is implemented underneath. Each service is exposed using standard network protocols, such as SOAP (simple object access protocol)/HTTP or JSON/HTTP, to send requests to read or change data. The services are published to enable developers to find them and reuse them to assemble new applications quickly.

Implementing SOA in the COVID-X Sandbox enables the following features:

- Promotes the use of open standards and interfaces to achieve interoperability and location opacity

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

- Provides clear descriptions of the service offered
- Offers microservices and processes designed to mirror real-world business activities
- Requires appropriate governance of service representation and implementation
- Provides criteria to allow service consumers to determine whether the service has been thoroughly and adequately executed following the service description.

SOA utilizes "service" as its primary element to construct information systems to suit various solution requirements. "Service" from a business perspective is the delivery of business outcomes; "service" from an IT perspective is the IT implementation of those business processes. The development process of an SOA solution may be private to a specific organization (e.g., deploying a service), collaborative between a set of business entities (e.g., service invocations and choreographies), or joint activities for maintaining the viability of the service ecosystem (e.g., publishing new services).

Some of the intended system benefits of using SOA are:

- Efficient development of information systems
- Efficient system integration
- Efficient reuse of resources

A set of SOA technical principles, specific norms, and standards have been established to deliver these efficiencies. The unified set of terms, principles, and concepts ensure standardization and, potentially, the quality of solutions, promoting effective large-scale SOA adoption. It should be highlighted that these principles apply to both software engineering and systems engineering to formalize service-based systems (i.e., complex systems, a federation of systems, systems of systems, enterprise architecture).

In general, a role is defined by a set of tasks and activities that serve a common goal. Roles are assigned to specific entities, who then are responsible for performing the activities defined by the role. The role definition specifies whether all the activities are required and the schedule they must follow. SOA defines two fundamental roles, namely service providers and service consumers. An SOA ecosystem comprises services that deliver functionality, service consumers who interact with services, and service providers who develop and host the consumer's services. However, service providers and service consumers' roles span various activities and may be accomplished by different parties responsible for different activities. For example, the entity that has operational responsibility may not have contractual responsibility and vice versa. Moreover, a party assuming the service provider role in one context may assume the service consumer role in another. In the diagram depicted in the following picture, the term service provider indicates an entity performing an activity associated with the service

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

provider role. In contrast, the term service consumer is an entity performing an activity associated with the service consumer role.

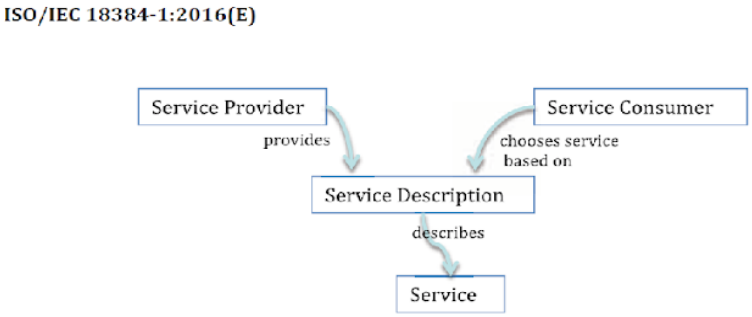

FIGURE 4:SOA SERVICE PROVIDER & CONSUMER ROLES

Figure 5 presents the primary SOA elements, including services, human actors, tasks, and systems. Any of these elements may be responsible for assuming a role and performing services.

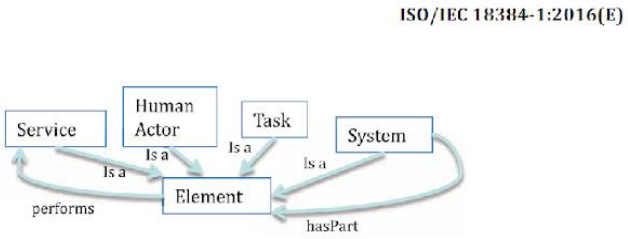

FIGURE 5: SERVICES & ELEMENTS OF SOA

The overall goal is to structure the COVID-X Sandbox as a collection of services that are: highly maintainable and testable, loosely coupled, independently deployable, organized around business capabilities, and autonomously developed by a small team. It is expected that each service is running in its own process and communicates with other services and clients using lightweight mechanisms such as well-defined HTTP resource APIs. These APIs frequently and, depending on the case, employ authenticated and encrypted communications. The services are implemented in ways that enable the independent deployment from each other while ensuring fully automated deployment approaches following CI/CD best practices. This particular approach further allows for minimum centralized management of the available services that can also be implemented by different teams, using different programming languages and potentially employ different data storage approaches if necessary.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

By adopting the SOA approach, development teams of the COVID-X project are able to develop and deploy the necessary services independently and, therefore, introduce functionality to the system simultaneously with minimal dependency between them once the interfaces and integration points among the components have been identified. The technologies and programming languages employed for implementation can be the most suited for the service being developed in each case. The implemented functionality can be deployed in a containerized fashion, which can minimize overhead and increase the portability between environments while allowing the realization of CI/CD pipelines.

4.1.2 Delivery Plan

This section presents the integration methodology and plan, that will drive the development, integration, and testing activities. This is an iterative process which is expected to lead to two releases of the COVID-X platform on M06 and M12 of the project. The integration process of the defined services of the Sandbox follows the CI/CD best practices [19], aiming at reducing the errors during development, integration and deployment, while increasing project's velocity. The development team of each component utilizes the COVID-X CI/CD Stack in order to proceed with the development and integration tasks.

The delivery plan includes two development cycles with several stages for each cycle that produces two releases of the Sandbox. presents an overview of the delivery plan:

TABLE 11: SANDBOX DELIVERY PLAN

Iteration	Integration point	Components	Partners	Date
Initial Phase	Installation of COVID-X CICD Stack	-	INTRA	M01-M02
1st Dev Cycle	Main development phase of software services, including unit tests and bilateral integration tests prepared for the 1st release	(i) Data collection, management, and harmonization service (ii) Security services (iii) APIs	INTRA, UPM, 8BELLS	M03-M05
1st Platform Integration	Functional Tests, End-to-End integration tests	Sandbox platform	INTRA, UPM, 8BELLS	M06
1st Release	First Sandbox Integrated release	Sandbox platform	INTRA, UPM, 8BELLS	M06
2nd Dev Cycle	Additional development phase of the core Sandbox services – Bug fixes and new features – additional integration tests	(i) Data collection, management, and harmonization service (ii) Security services (iii) APIs	INTRA, UPM, 8BELLS	M07-M11
Third Party Integration	Integrate components and services provided by third parties into the 1st Sandbox release	Third party components	INTRA, UPM, 8BELLS, Third Parties	M07-M11

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

3rd Dev Cycle	Additional development phase of the core Sandbox services of the 2nd release – Bug fixes and new features	(i) Data collection, management, and harmonization service – extended for third party clinical partners data sets (ii) Extended security services (iii) Visualization services (iv) Federated learning services (v) Updated APIs	INTRA, UPM, 8BELLS	M07-M11
2nd Platform Integration	Functional tests – End-to-end integration tests ready for the 2nd release	Sandbox platform	INTRA, UPM, 8BELLS	M11-M12
2nd Release	Final Sandbox integrated release	Sandbox platform	INTRA, UPM, 8BELLS	M12
Third Party Integration	Integrate components and services provided by third parties into the 2nd Sandbox release	Third party components	INTRA, UPM, 8BELLS, Third Parties	M12-M24

The first step of the integration process is the deployment and configuration of the COVID-X CI/CD Stack. This task has been successfully implemented by Month 2 of the project, in order to be able to support the two development cycles, described in the following two sections.

4.1.2.1 First Development Cycle

The first development cycle, which will run from month 3 to month 5, includes the following steps for each service of the Sandbox:

- Planning and Requirement Analysis
- Defining the Requirements
- Designing the service internal architecture and external interfaces
- Developing/Building the service
- Testing the tool

The first development cycle can be further divided into the following phases:

- **First development phase** runs from month 3 to month 5 and it is responsible for producing the initial release of all the software modules that comprise the 1st release of the COVID-X Sandbox.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

- **First platform integration** takes place on month 6 and it is responsible for executing the necessary end-to-end integration and functional tests on the Sandbox.

The outcome of the first development cycle will be the first release of the Sandbox, which is presented in D2.2 [20]. It will be used for integrating the components provided by the third parties that will be onboarded during the first open call. This process will be described in D2.4, due month 12.

4.1.2.2 Second Development Cycle

The second development cycle will run from month 7 to month 11 and includes the following phases:

- The **final development phase** follows the 1st release of the Sandbox and implements additional features and bug fixes based on insights and user feedback. Following the same approach with the previous development phase, unit tests, as well as bilateral integration tests, are executed during this phase.
- The **third-party integration phase** during which applications and components provided by third parties are integrated into the 1st release of the Sandbox.
- The **final platform integration** runs on months 11 and 12 and it is responsible for executing the necessary end-to-end integration and functional tests on the final version of the Sandbox services.

The second development cycle will produce the final release of the Sandbox, which will be described in D2.3, due month 12. It will also be used for integrating components and applications provided by third parties.

4.1.3 Release A Architecture

The Sandbox design is based on a highly modular architecture, built as a collection of services, following the SOA approach. Each service might be further divided into building blocks comprising microservices. The microservices approach adopts a strategy of putting together a large and complex application from small individual building blocks. Based on this design perspective, the architecture of the Sandbox release A is depicted in Figure 6. The core services of closed and open data integration, real-time data ingestion, data annotation, indexing, cataloguing and storing, security, data access/filtering/search/querying and data visualization, along with an API Gateway compose this release.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

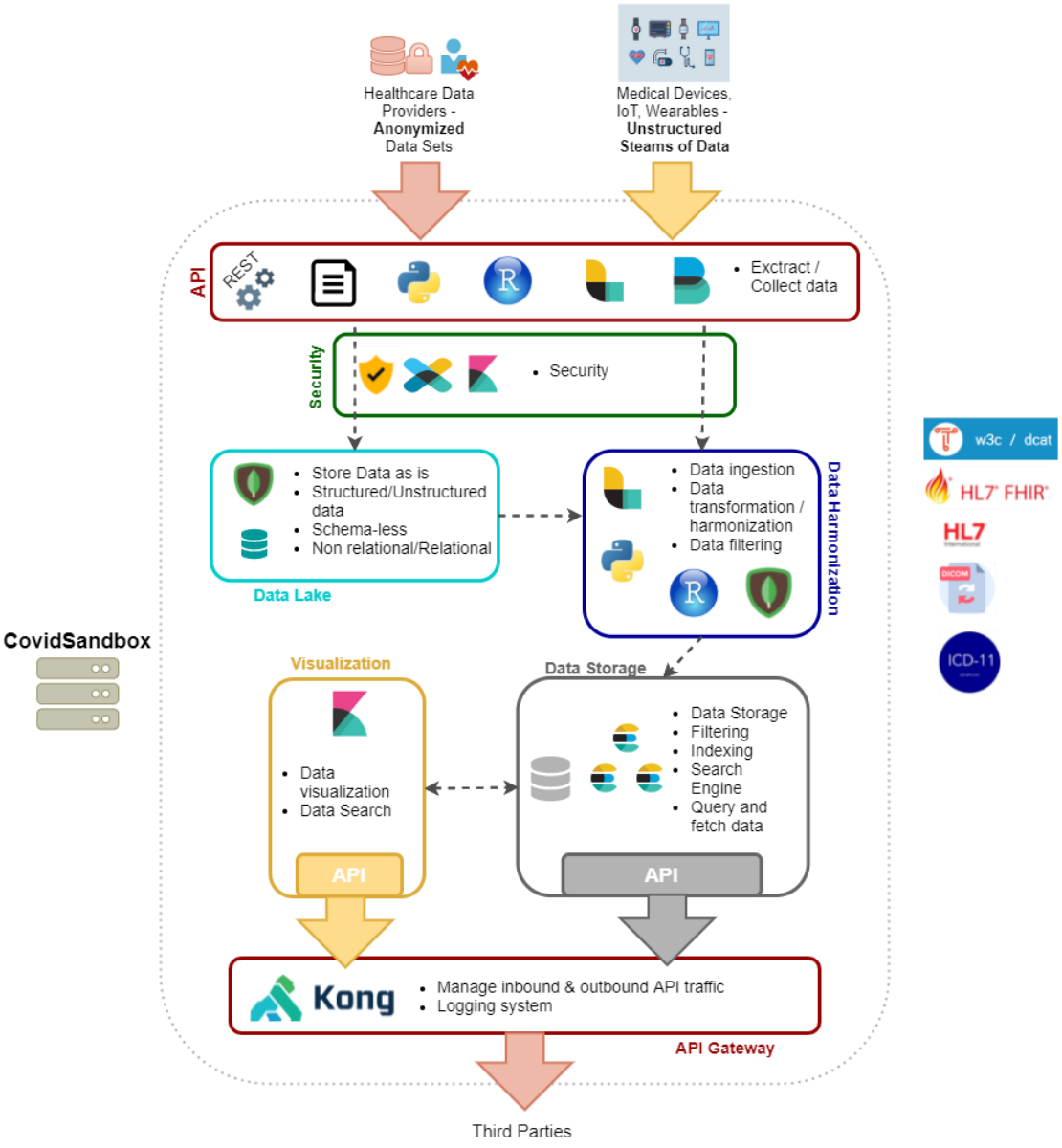

FIGURE 6: SANDBOX RELEASE A ARCHITECTURE

Service-to-service communication is a challenging task that lies at the heart of the SOA implementation. Sandbox adopts the Representational State Transfer (REST) architectural approach for realizing internal communication among the multiple services, a design pattern for implementing HTTP resource APIs. Each service is configured internally to expose RESTful web endpoints based on the technology and the programming language that has been employed for its implementation. These endpoints expose information about the service's resources. The typical interaction between client service and a RESTful API begins with the former providing the URL related to the resource of interest and the server's operation on that resource. The operation is in the form of an HTTP method such as GET, POST, PUT and DELETE, while the JSON format is used for exchanging data.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

4.1.4 Release B Architecture

Release B of the Sandbox will maintain the same design approach, following the highly modular and containerized implementation. Figure 7 shows the envisioned Sandbox architecture in Release B. All the individual components and services included in Release A will be further developed to introduced enhanced functionality and new features, based on insights and user feedback. Moreover, identified bugs and issues from Release A will be fixed.

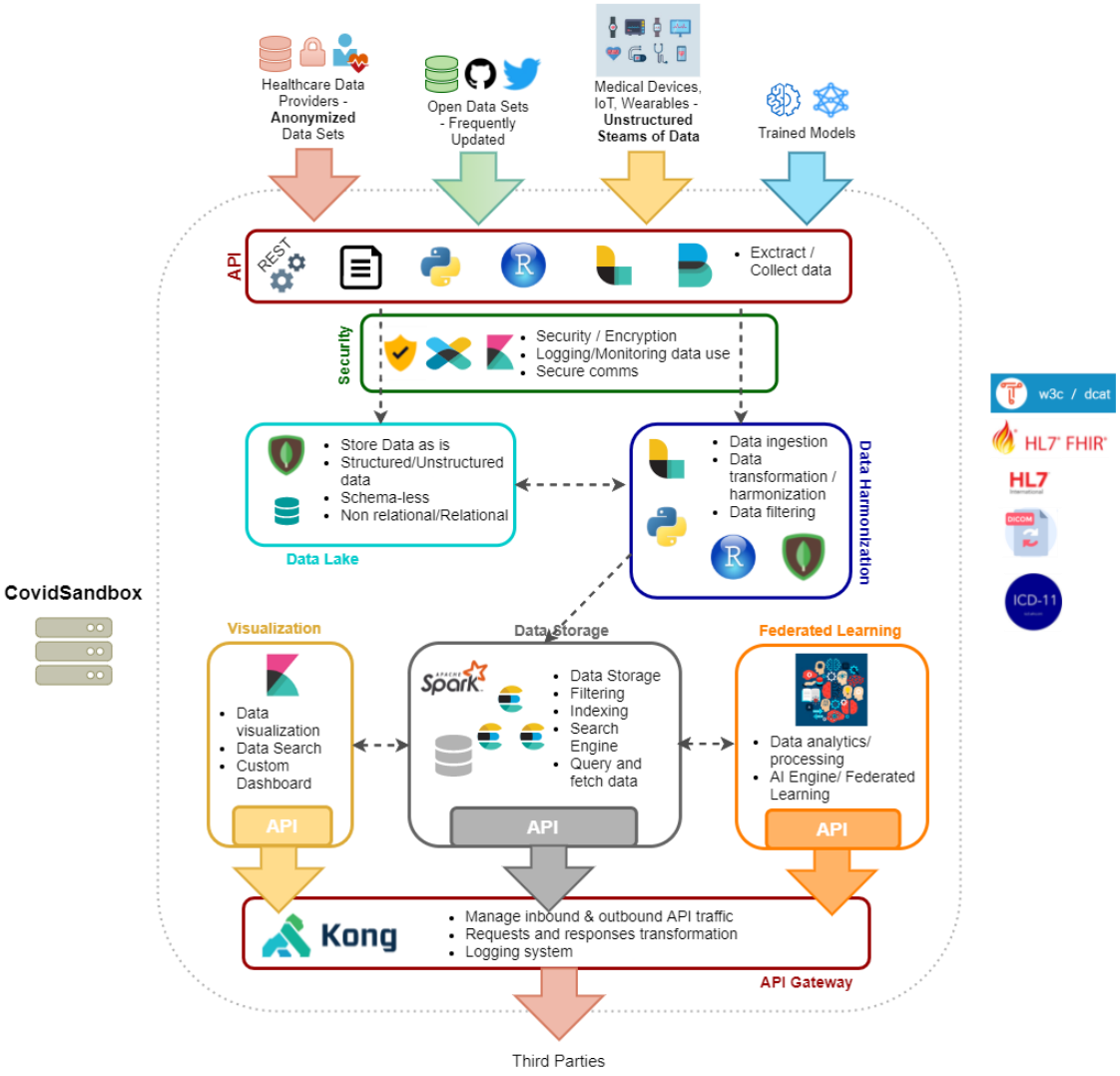

FIGURE 7: SANDBOX RELEASE B ARCHITECTURE

In this release, additional open and publicly available datasets will be ingested in the Sandbox. Finally, the federated learning module will be introduced, to permit the deployment of federated learning edge nodes in the targeted hospitals enabling remote and secure training of relevant AI algorithms over diverse types of data (numerical, textual, time-series, images such as scans, etc,) while

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

maintaining the integrity and security of data and meeting the hospitals needs to properly handle personal health data. The federated learning module will encompass another counterpart at the centralized cloud-based Sandbox system to train and retrain AI models with data coming from open data sources.

4.1.5 Sandbox Components

COVID-X sandbox introduces a combination of different functions that target the continuous integration and access to a set of healthcare data sources. All the architecture of COVID-X sandbox is mainly based on the ELK stack components to facilitate data ingestion, data harmonization and cataloguing, data storage, data management and access, data visualization and security. It is based on a highly modular architecture, as shown above, and each component is responsible for carrying out and delivering one or more services. In more detail, the core components that support the functionality of COVID-X sandbox are described in the following subsections.

4.1.5.1 Data Ingestion & Harmonization

This part is responsible for all the necessary operations for information coming from data sources and data lakes and then becoming available in the data storage layer for further usage by either internal or external services of the COVID-X sandbox. This component implements ingestion of data either from data lake or through API layer (Extract), converts them into a consistent format in compliance to the COVID-X Semantic Reference Model, 1st version (Transform) and then loads the transformed information into the data storage layer (Load). To do this, data harmonization provides the mapping between the data sources and the common schema/data model used by other internal components of COVID-X sandbox. This whole process is known as ETL process. The data ingestion process supports various methods that can be used for extracting and collecting data: REST HTTP endpoints, websockets, SFTP servers, HTTP polling mechanisms, database connectors and message broker connectors.

This component realizes the following Sandbox capabilities and requirements:

- [COV_CAP_003]: Near-real time data ingestion.
- [COV_PER_004]: Data coming from medical and IoT devices and wearables should be served in a near real-time manner.
- [COV_FUN_003]: Data ingestion in batch mode for static data and in streaming mode for data coming from IoT devices, sensors or wearables.
- [COV_FUN_004]:Data harmonization / transformation using filtering tools.
- [COV_FUN_014]: Supported input data connectors should be at least through file ingestion, database dumps and REST APIs.
- [COV_FUN_016]: Ingestion of various data formats is supported.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

4.1.5.2 Data Storage & Management

This component can be described as a centralized data repository that stores all the information coming from data ingestion & harmonization layer. If possible, this component also offers the capability of accessing data through a search engine that enables the execution of complex and fast queries. Several APIs can be served at this layer both of indexing/cataloguing and querying/retrieving data. In addition, a data lake is implemented as part of the Sandbox Architecture. The datalake is a repository that is used for archiving any useful data that COVID-X sandbox needs to keep for future processing. Data are stored in a schema-less manner, allowing the combination of both structured and unstructured format. The purpose of data lake is mostly on keeping data in its primitive format rather than representing relations between them.

This component realizes the following Sandbox capabilities and requirements:

- [COV_CAP_001]: Sandbox Services support high availability.
- [COV_CAP_002]: Sandbox services support fault tolerance.
- [COV_CAP_004]: High throughput for data queries.
- [COV_CAP_005]: System latency should be minimal.
- [COV_CAP_006]: System can handle multiple logins.
- [COV_CAP_007]: System can host large datasets.
- [COV_CAP_008]: System can scale on multiple nodes.
- [COV_FUN_005]: Data cataloguing along with metadata.
- [COV_FUN_006]: Data indexing for fast querying and retrieving.
- [COV_FUN_011]: Unified summary of data catalogues should be supported.
- [COV_FUN_013]: Full text search is supported.
- [COV_FUN_015]: Structured and unstructured data are supported.
- [COV_PER_001]: High availability.
- [COV_PER_002]: Responsiveness and reliability.
- [COV_PER_005]: Database queries should be served quickly, so as to enhance user experience.
- [COV_PER_006]: Administrative tasks, such as backups, can be made during low-workload times.

4.1.5.3 Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. This component includes all the different tools that can be applied on top of data that exist on the data storage layer and allow users to create custom

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

visualizations, such as pie charts, line charts, heat maps, and gauges. Various customizable and interactive dashboards are established based on COVID-X sandbox needs to support business or any other decisions that COVID-X third parties may have.

This component realizes the following Sandbox capabilities and requirements:

- [COV_FUN_012]: Visualization tools to display aggregated data, infographics.

4.1.5.4 Security

To ensure data confidentiality, integrity and availability within the COVID-X Sandbox, multiple layers of security are applied. First of all, a risk assessment process based on the Anonymization Guideline will take place. Secondly, personal healthcare data will undergo an anonymization process that aims to remove personal identifiers from the data, hindering the retrieval of a person's identity from his/her data. Then, a set of human centric and technical measures are defined and applied in order to minimize the impact that a threat factor can potentially cause to the health sector. These measures entail any cybersecurity action, process, device, or system that can prevent or reduce the effects of threats, vulnerabilities, and attacks on health data that could be significantly compromised due to its sensitive nature [21].

The selected measures should address all relevant aspects of cybersecurity presented in the beginning of this section. One major practice to enhance the robustness of the infrastructure is to focus on the weakest link which is the human factor. Specifically, it is encouraged to establish a "security culture" amongst staff that includes continuous cyber-security training for all employees from all organizational levels and roles. Additionally, organizations should ensure meticulous tracking of who is accessing health record systems. Furthermore, they should also provide reliable and frequent backups of the systems. All mobile devices containing personal medical information should be protected with encryption and third-party software should not be installed by staff without prior consent. Staff working on remote devices should be enabled to connect to a virtual private network (VPN) to maintain a secure connection over unsecured internet infrastructure. Apart from the countermeasures taken, one should be prepared to handle the short- and long- term impacts of any attack; to that end, continuous monitoring of networks and systems managed by internal IT staff is crucial.

The COVID-X Sandbox will be locally deployed in each Clinical partners' premises. Each host will be utilizing a containerized environment. The main aim of the containers' utilization is to gather all the required software dependencies, so that the COVID-X Sandbox can run reliably from one computing environment to another. In order to manage the containerized environment, an orchestrator will be used which automates the scheduling, deployment, networking, scaling, health monitoring, and management of the various containers as well as it provides a fault tolerant and controlled environment where there is visibility across containers' activity. Additionally, containers will only use trusted images avoiding vulnerabilities from public image sources. From the orchestrator's side, as a runtime monitoring application, it provides interoperability and robustness, maintaining system

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

integrity and availability. Furthermore, it adds to the manageability of available resources without burdening the system. Last but not least, regular backups are also supported by these technologies.

For the integration, testing and deployment process a Continuous Integration/Continuous Delivery (CI/CD) and DevSecOps approach is followed by COVID-X Sandbox. It is an automated pipeline that supports built-in security features. Services follow the least privilege possible in order to minimize unauthorized connections and access, containers are utilizing only certified images and logs are being produced during deployment. Last but not least communication between CI/CD infrastructure makes use of secure API calls that use HTTP over Transport Layer Security (TLS), a standard that keeps an internet connection private and checks that the data sent between two systems (a server and a server, or a server and a client) is encrypted and unmodified.

Diving deeper into the COVID-X Sandbox functionalities the main focus is on the layer of security. As part of the professional medical system, it interacts with its network or any medical devices connected to it. Furthermore, it provides access to users that want to interact with its API. Aside from the external connections, the COVID-X Sandbox is also a distributed technology which entails an interconnectivity between modules in a parallel way. It is however a double-edged sword since it also provides an opportunity for attacks to a malicious actor. The need to protect the system and its users is addressed by the encryption mechanism that the system utilizes. In more detail, the COVID-X Sandbox supports encryption protocols and standards, such as TLS. When establishing a connection with an external client, the communication in between is always encrypted, protecting the COVID-X Sandbox from eavesdropping, and preventing exposure of sensitive data during transfer. Internally, there is always the possibility of someone trying to attack a part of the system (e.g., hijacking a node), thus the same standards are also enforced when it comes to the internal COVID-X Sandbox communications, i.e., communications with every module.

Once Authenticated, users gain access to the surface layer of APIs where they can interact with the COVID-X Sandbox functionalities. Furthermore, they have available views of the part of data that they will use to perform their validations. Even though data is completely anonymized, there is no added value to having users access the whole data lake; on the contrary, it adds more to the attack surface and burdens the users with unnecessary data volumes. Therefore, the principle of least privilege is followed. A user must be able to access only the information and resources that are necessary for his/her legitimate purpose. In more detail, all users should have discrete access rights to the data lake and should be able to view only the data indices that correspond to their data model. Particularly, a role with the appropriate rights and privileges will help regulate a group of users which act in the same data ecosystem. A role-based access control (RBAC) system is implemented, which is responsible for user roles regulating user interaction with data indices. It not only adds to the layer of security, but also enhances system stability and data integrity since it enables the introduction of a limited and well-defined number of users accessing the same source.

This component realizes the following Sandbox capabilities and requirements:

- [COV_FUN_001]: Role-based access control system (RBAC) for user management.
- [COV_FUN_002]: Data transfer based on encryption protocols.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

- [COV_FUN_007]: Users are entitled to access only the part of the data necessary for their analysis (principle of least privileges).
- [COV_FUN_008]: All inbound and outbound operations are logged.
- [COV_FUN_009]: Activity monitoring for abnormal behavior detection.
- [COV_SEC_001]: The COVID-X Sandbox has a well-defined number of users.
- [COV_SEC_002]: Users need to insert their credentials (username & password) in order to access the COVID-X Sandbox.
- [COV_SEC_003]: The system verifies users' credentials in order to provide them access to its services.
- [COV_SEC_004]: Backups of the configuration files, access control lists and other important files should be kept.
- [COV_SEC_005]:All data transfers should be encrypted.
- [COV_SEC_006]: Every user is entitled to specific privileges that are the minimum privileges needed to perform their analysis.
- [COV_SEC_007]: User roles describe all privileges a user with a certain role is entitled to.
- [COV_SEC_008]: Users should only be able to perform actions dictated by their assigned roles.
- [COV_SEC_009]: The production of faulty software that could compromise security should be avoided.
- [COV_SEC_010]: All activities that take place within the COVID-X Sandbox are monitored.
- [COV_SEC_011]: In case malicious activities are detected inside the COVID-X Sandbox, alerts are generated.
- [COV_SEC_012]: Only authorized modules/containers should be able to connect to the containerized environment of the COVID-X Sandbox.
- [COV_SEC_013]: The COVID-X Sandbox should use exclusively genuine software and ban all illegitimate software and applications [14].

4.1.5.5 APIs Gateway

This layer defines and implements all the necessary endpoints that will be exposed to data source providers and third parties to use COVID-X sandbox services. Various REST API endpoints will be created for data ingestion (data providers), data integration, and data access (third parties), including access through querying the data storage layer directly and using the tools provided by the data visualization layer.

This component realizes the following Sandbox capabilities and requirements:

- [COV_INT_001]: APIs follow a REST API architecture.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

- [COV_INT_002]: API Gateway should enable authorized external users to securely connect to Sandbox Services.
- [COV_INT_003]: API to retrieve data from Sandbox datalakes.
- [COV_INT_004]: Sandbox APIs should enable internal communications between Sandbox services and the deployed third-party containers implementing application-oriented solutions.
- [COV_PER_002]: Responsiveness and reliability.
- [COV_PER_003]: COVID-X Sandbox services should be easily accessible to the users.

4.1.5.6 CI/CD Stack

The CI/CD Stack is implemented as a collection of open-source software components, which enable an automated build system capable of integrating changes performed by developers working on individual tools of the COVID-X Sandbox. In addition to that, the CI/CD Stack is utilized for integrating software applications provided by third-party SMEs. The CI/CD Stack supports CI pipelines that automate the parts of software development related to building, testing, and deployment of applications. Each integration cycle introduces automated builds and unit tests on the latest code changes to immediately surface any errors . Moreover, it enables automated deployment that follows the CD best practices. The deployment comprises a fully containerized environment to allow the software to run consistently in the same way, independently of the underlying infrastructure. Containerization provides Operating System-level virtualization, allowing multiple applications to run in containers on a single host.

This component realizes the following Sandbox capabilities and requirements:

- [COV_MNT_001]: Highly configurable and autonomous way to build/test and deploy through a CI/CD pipeline
- [COV_MNT_002]: CD tools support version control repositories for managing code changes.
- [COV_MNT_003]: Managing the environment / system variables and configuring them for the target environment.
- [COV_MNT_004]: Executing continuous tests and rollback environments if tests fail.
- [COV_POR_001]: The CI/CD pipeline allows copying selected configuration settings from one system to another system.
- [COV_POR_001]: Containerized deployment allows Sandbox to be independent of the underlying platform

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

4.2 System Actors

The COVID-X sandbox is a solution that provides various functionalities and use cases both for internal and external purposes. COVID-X sandbox requires a set of system actors in order both to support its functionality and at same time to define and controls the accessibility to its internal components. In general, system actors may represent roles played by human users, database systems, clients and servers, cloud platforms, devices, etc. COVID-X consists of four fundamental system actor categories: Data Provider, Component Provider, Sandbox User and Sandbox Supervisor.

4.2.1 Data Provider

Data providers are usually organizations that provide data to for the other actors of the system. The provided data can be raw data, refined data/information or analyzed information. Data providers can be divided into two groups according to their motive: organizations that provide "for free" without any conditions or with some licenses that restrict the use of data (open data sources for COVID-X project purposes) and organization that offer their data under agreed and very specific conditions (clinical partners for COVID-X project purposes).

4.2.2 Component Provider

The main purpose of COVID-X sandbox solution is to provide all the necessary tools and services that will highlight the value and importance of data coming from clinical partners and will help third party users to take advantage from this functionality. Component providers (technical partners for COVID-X project purposes) are responsible for the implementation of services that must: a) identify sandbox users' needs, b) produce relevant data from input data to a particular context and c) represent the produced data in a usable way. However, component providers do not necessarily provide a complete service for the user but can simply provide a part from a service chain. These providers may provide ready-made service chains, or these service chains may be composed at runtime.

4.2.3 Sandbox User

Sandbox users access data with the help of data-based applications and services created by component providers. A user can be part of an organization that is authorized to access the exposed services of the entire sandbox system (third party organizations for COVID-X project purposes). Depending on the setup and implementation of the system, sandbox users may have limited access to the data, or the services provided.

4.2.4 Sandbox Supervisor

Sandbox supervisor offers all the necessary hardware and software tools that ensure the continuous support and functionality of the entire system. The relevant roles, related to this category, are : a) system and operation administrators who are responsible for system maintenance and receiving usage

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

information about the deployed services, b) tool providers that provide tools to develop, configure services for different user needs and execute/control these, and if applicable c) cloud service providers who provide the facilities that host all the different component and services of the system.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

5 Data Collection and COVID-X Data Model

5.1 Clinical and Other Data Sets and Data Sources

This Section provides a high-level description of the data sets that are utilized by the COVID-X Sandbox, along with the associated metadata. These data sets are either provided by the clinical data sources available in the COVID-X project or collected by open and publicly available data sources.

5.1.1 Clinical Data Sources

5.1.1.1 ICH

ICH feeds the COVID-X Sandbox with data and metadata from two main datasets, namely the ***Covid Clinical DB*** and the ***COVID CT DB***. The following table provide an overview of the available datasets:

TABLE 12: ICH DATASETS

Name	**Covid Clinical DB**
Description	All clinical data collected from the Istituto Clinico Humanitas (ICH) electronic health records for Covid-19 patients, from the admission to the emergency room (ER) to hospital discharge. The dataset contains about 1.200 patients by December 2020 and includes more than 90 features.
Purpose	The dataset aims to provide structured information about patients' clinical condition, hospitalization journey and laboratory results. Longitudinal highly granular data has proven of critical importance in devising prescriptive analytic systems. This database allows to evaluate specific patients' characteristics, particularly suited for precision medicine.
Format	CSV
Anonymization	The dataset is fully pseudonymized and de-identified. A further anonymization process is going to be performed when specific data requirements from SMEs will be available, according to the selected subset's content and dimension.
Storage	The dataset is stored in ICH Cloud Environment.
Access	The access is granted to authorized users through the COVID-X sandbox.
Name	**Covid CT DB**

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

Description	A collection of chest CTs acquired from Covid patients at different stages of the illness. All the included CTs are provided together with a collection of relevant DICOM metadata. Image annotations and radiomic features are available for a subgroup of the included CTs.
Purpose	Purpose: the dataset aims to provide a collection of relevant Covid CT scans, including multiple stages and severity of the infection. The Covid CT DB is suitable for development of computer vision tools to characterize individual patients' patterns of the Covid phenotype. Since CTs were collected from different machines, this dataset can be additionally used to improve generalization abilities of algorithms.
Format	CSV
Anonymization	The dataset is fully pseudonymized and de-identified. A further anonymization process is going to be performed when specific data requirements from SMEs will be available, according to the selected subset's content and dimension.
Storage	The dataset is stored in ICH Cloud Environment.
Access	The access is granted to authorized users through the COVID-X sandbox.

5.1.1.2 SERMAS

SERMAS provides the Sandbox with data coming from twelve datasets, all hosted at their local clinical premisses. The access to the datasets is granted to authorized users through the COVID-X Sandbox. The following table provide an overview of the available datasets:

TABLE 13: SERMAS DATASETS

Name	Paciente
Description	Administrative database. Record of all patients ever attending the hospital. This database contained all personal and demographic data from the patients, including CIPA (identifier to match hospital data with primary care data: universal patient code for the Madrid Regional Health System).
Period	Historic data.
Purpose	All data from the different data sources has been generated during the health care process of subjects attending our tertiary care centre at

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

	different services, including the Emergency Department. Relevance: Demographic information
Format	CSV
Anonymization	Anonymized
Storage	SERMAS local premisses.
Access	The access is granted to authorized users through the COVID-X sandbox.
Name	**CMBD_H (Minimum basic data set_inpatients)**
Description	Administrative database. Record of hospital inpatients.
Period	2016 - 2021
Purpose	Hospitalization dates and diagnosis and procedures at discharge
Format	CSV
Anonymization	Anonymized
Storage	SERMAS local premisses.
Access	The access is granted to authorized users through the COVID-X sandbox.
Name	**CMBD_A (Minimum basic data set_ambulatory patients)**
Description	Administrative database. Record of ambulatory patients.
Period	2016 – 2021
Format	CSV
Anonymization	Anonymized
Storage	SERMAS local premisses.
Access	The access is granted to authorized users through the COVID-X sandbox.
Name	**CMBD_U (Minimum basic data set_visits to emergency room)**
Description	Administrative database. Record of emergency room activity.
Period	2016 – 2021
Purpose	Relevance: Emergency Department attending dates and diagnosis and procedures at discharge.
Format	CSV

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

Anonymization	Anonymized
Storage	SERMAS local premisses.
Access	The access is granted to authorized users through the COVID-X sandbox.
Name	**Lisencam**
Description	Administrative database. Record of dairy occupied beds.
Period	July 2019 – 2021
Purpose	Relevance: Bed and department where the patient was hospitalized.
Format	CSV
Anonymization	Anonymized
Storage	SERMAS local premisses.
Access	The access is granted to authorized users through the COVID-X sandbox.
Name	**CEX (Outpatient visits)**
Description	Administrative database. Record of outpatient visits.
Period	2016 – 2021
Purpose	Relevance: administrative data covering outpatient activity within the hospital
Format	CSV
Anonymization	Anonymized
Storage	SERMAS local premisses.
Access	The access is granted to authorized users through the COVID-X sandbox.
Name	**Farmacia**
Description	Dispensated drugs record. Since 2020 we have a more extensive database, with not only dispensation but prescriptions.
Period	2016 – 2021
Purpose	Relevance: Medication prescribed during admissions.
Format	CSV
Anonymization	Anonymized

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

Storage	SERMAS local premisses.
Access	The access is granted to authorized users through the COVID-X sandbox.
Name	**Laboratorio**
Description	Relevant Covid19 Clinical labwork.
Period	2020
Purpose	Relevance: Labwork carried out during admission or during outpatient activities
Format	CSV
Anonymization	Anonymized
Storage	SERMAS local premisses.
Access	The access is granted to authorized users through the COVID-X sandbox.
Name	**PCR**
Description	SARS-CoV-2 PCR results
Period	March 2020 – 2021
Purpose	Relevance: Classification as SARS-CoV-2 positive or not
Format	CSV
Anonymization	Anonymized
Storage	SERMAS local premisses.
Access	The access is granted to authorized users through the COVID-X sandbox.
Name	**UCI**
Description	Data from ICU admitted patient
Period	March 2020 -July 2020
Purpose	Registry of COVID-19 patients seen at ICU. Relevance: Detailed clinical data from COVID19 patients.
Format	CSV
Anonymization	Anonymized
Storage	SERMAS local premisses.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

Access	The access is granted to authorized users through the COVID-X sandbox.
Name	**Biobanco**
Description	Bio bank
Period	Under Request
Format	CSV
Anonymization	Anonymized
Storage	SERMAS local premisses.
Access	The access is granted to authorized users through the COVID-X sandbox.
Name	**Imagen**
Description	Chest CTs
Period	March 2020
Purpose	Relevance: Imagine data from patients, requested based on clinical manifestations and medical decisions
Format	DICOM (.dcm)
Anonymization	Anonymuzed
Storage	SERMAS local premisses.
Access	The access is granted to authorized users through the COVID-X sandbox.

5.1.1.3 KI

KI feeds Sandbox with data through the Clinical History Taking Program (CLEOS). CLEOS is an interactive knowledge base that performs all cognitive tasks within the clinical method except for physical examination, as depicted in Figure 8. The system gathers clinical data directly from primary sources, e.g., patient and laboratories. The program starts by collecting the patient's complete medical history because a detailed medical history is the sine qua non for clinical outcomes that match what is possible to achieve by applying state-of-the-art knowledge. The findings so far showing that diagnoses can be made more than 80% of the time from history alone, for example, and history defines significant co-morbid states, significant past medical events, environmental risks, and family history.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

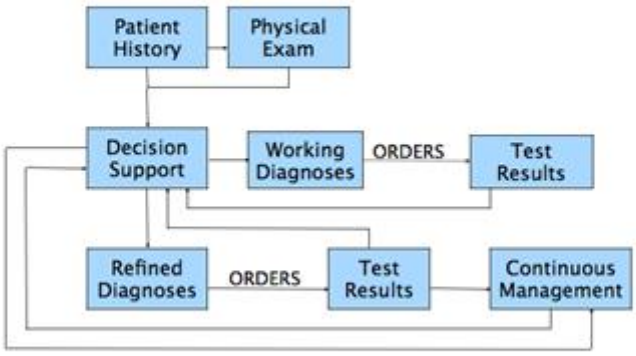

FIGURE 8: FLOW OF CLINICAL INFORMATION

CLEOS, uses the patient's time to collect data for a range of clinical issues physicians ignore, e.g., detailed histories of life style, functional capacity in daily living, diet, energy expenditure, mental health, emotional well-being and occurrence of self-limited, intercurrent illness, which is thought to be but may not be irrelevant to the incidence and progression of chronic disease. Moreover, the system acquires histories from all patients, whether for an initial history-taking session or interim history, according to the single protocol embedded in its semantic network. This means that questions not posed to a patient are not posed for valid medical reasons. Data elements "missing" from a CLEOS history can be used confidently to define phenotype for all patients. The system can define, through the data it is directed to collect, a large number of different clinical phenotypes in groups of patients with identical diagnoses or healthy people deemed to have identical levels of risk(s) for one or another disease.

The history-taking component is built in modules but functions as a single, seamless decision graph at the patient interface, which is initiated when the patient selects an answer to "Why do you need medical care now?" Subsequent questions are generated in the context of the answer to this question and continuous evaluation of the clinical significance of the sum of answers to any point in the interview. Interim histories are triggered by the patient's known medical problems.

Coverage in the current version is only for adults. The current version covers in depth all chronic illness and their complications and common acute illnesses that present to emergency rooms. The knowledge base of the current version is represented by > 18,200 decision nodes in ~ 475 graphs that direct collection of data to populate > 30,000 data fields, an example of which is shown in Figure 9.

Decision nodes direct history-taking according to established, empirical rules that relate symptoms [what a patient experiences] to pathologic states. The system directs the line of questioning through

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

continuous data analysis as data is collected. CLEOS also analyzes the clinical significance of all data at the conclusion of an interview to report diagnostic and therapeutic recommendations, organize data elements as clinical narratives for the physician, and generate orders for diagnostic tests and therapeutic interventions. Every data field has a unique code; and all data are saved. The program also saves the pathway by which the data are collected in each interview. The current program has >7500 explicit rules for interpreting and organizing the data it collects. Rules-analysis operates iteratively to refine diagnoses and adjust treatment as new clinical data feed-back from testing regimes, therapeutic results, interim histories and the natural course of the patient's illness.

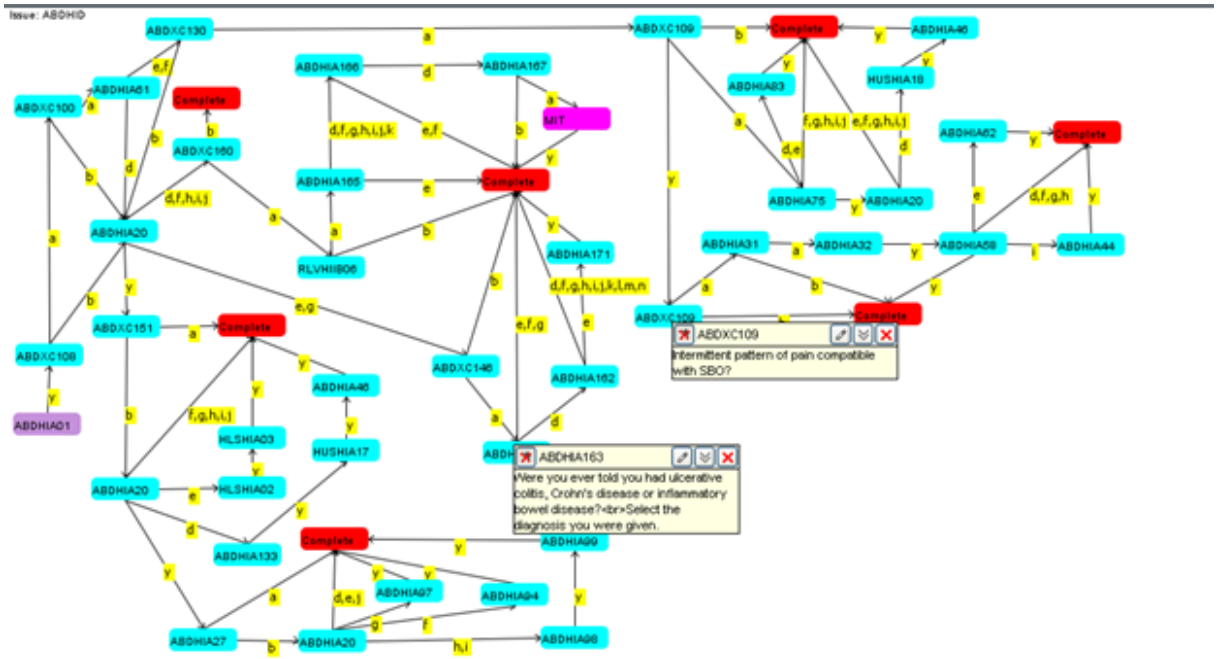

FIGURE 9: DECISION GRAPH FOR COLLECTING HISTORICAL DATA ABOUT ABDOMINAL PAIN

5.1.2 Public Data Sources

Since the appearance of Sars-Cov-2 at the very beginning of 2020, worldwide efforts to obtain reliable and valuable data to be used in different research areas have been enormous. In this regard, different entities, from world-renowned organizations such as WHO (World Health Organization) to more local institutions such as many national universities have tried to contribute in the search for solutions as well as in broadening the general knowledge of the disease and its evolution by defining different datasets.

The need to quickly find effective solutions to alleviate the current worldwide situation has made it vital to share, in an open way, this type of database. From those which are in charge of reflecting the monitoring of the general situation (which are mainly statistical and demographic-based data) to those which provide more specific data such as the medical one, all these open databases have a common

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

objective: to facilitate the access to existing information to all of those stakeholders who try to find a key to reverse current situation.

With this in mind, one of the first phases of the project was the identification of the most relevant worldwide data sources for the fight against the pandemic. From this analysis emerged a wide list of datasets that we decided to classify according to the type of information they provide:

- **Demographics:** These data sources include general information about the situation of the pandemic such as number of people affected, number of deaths, etc.
- **Research documentation:** These datasets aim at including the different scientific papers and computational resources that have been published about Covid in order to have them identified and classified.
- **Clinical Image:** These datasets include specific health images (such as radiography) that can be used in signal processing-based solutions.
- **Patient data:** These databases include different clinical information of the patients in an anonymized way.
- **Social networks:** These data sources include information about the progression of the disease appearance in the social networks.

Following this classification, Figure 10 presents the most relevant data sources that have been analysed for the project:

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

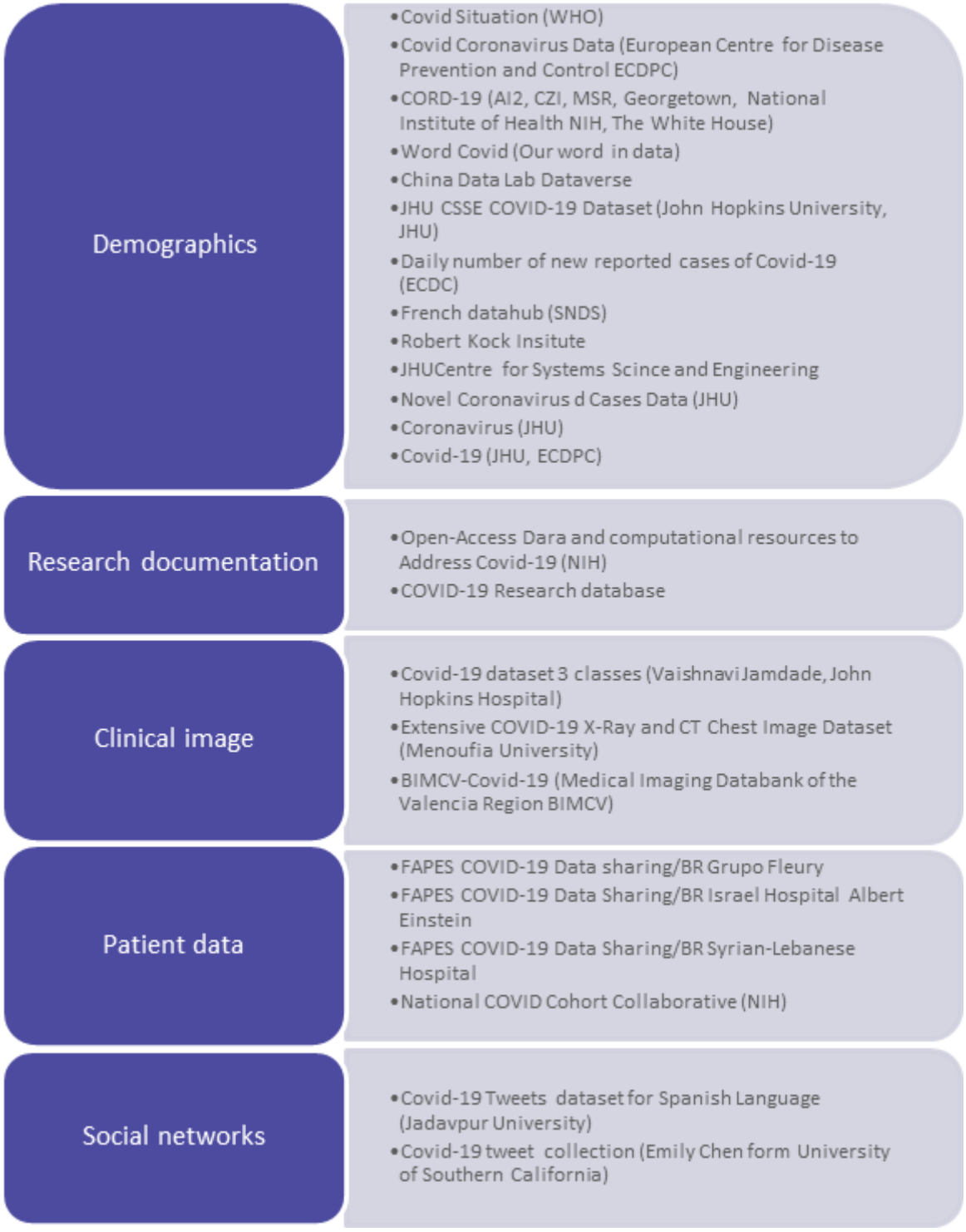

FIGURE 10: IDENTIFIED OPEN DATASETS

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

After further analysing the identified open datasets, the datasets that will be utilized for feeding the COVID-X Sandbox in Release A are described in the table below:

TABLE 14: OPEN DATASETS

Name	**Extensive COVID-19 X-Ray and CT Chest image Dataset**
Publisher	Menoufia University
Content	Non-COVID and COVID cases of both X-ray and CT images
Purpose	Use the images to develop AI based approaches to predict and understand infection
Categories	X-Ray, CT Images
Link	https://github.com/ieee8023/covid-chestxray-dataset
Name	**FAPESP COVID-19 Data Sharing/BR - Dados COVID Grupo Fleur**
Publisher	Fleury Group
Content	(1) Data about patients (serology or PCR), (2) Respective results of laboratory tests, (3) Data dictionaries: spreadsheet in which each tab describes, respectively, all fields in the Patients and Exams spreadsheets.
Purpose	Publish open COVID-19 data to contribute to and foster research related to this topic.
Categories	Serology, laboratory tests
Link	https://repositoriodatasharingfapesp.uspdigital.usp.br/handle/item/99
Name	**BIMCV-COVID19+: a large annotated dataset of RX and CT images of COVID19 patients**
Publisher	BIMCV
Content	Large annotated dataset with chest X-ray images CXR (CR, DX) and (CT) imaging of COVID-19 patients along with their radiographic findings, pathologies, (PCR), G (IgG) and M (IgM) diagnostic antibody tests and radiographic reports from Medical Imaging Databank in Valencian Region Medical Image Bank (BIMCV)
Purpose	Collect and publish chest X-ray images, coming from hospitals affiliated to the BIMCV, to which data that allows their identification will be erased for the purpose of training Deep Learning (DL) models
Categories	CR, DX, CT

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

Link	https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/#1590858128006-9e640421-6711
Name	**Covid-19-data**
Publisher	Our World in Data (OWID)
Content	A collection of the COVID/19 data that includes confirmed cases, deaths, hospitalizations and testing, as well as other variables of potential interest
Categories	Demographic data
Link	https://github.com/owid/covid-19-data/tree/master/public/data
Name	**Novel Coronavirus (COVID-19) Cases Data**
Publisher	John Hopkins University
Content	New COVID-19 cases
Categories	Epidemiological data
Link	https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases

5.2 COVID-X Data Model

and semantically annotating raw and aggregated data of diverse types, as well as for harmonizing and enriching data. The common semantic data model is an ontology defined as the COVID-X ontology. This ontology provides a high-level design of the main concepts and entities interacting in the Covid-19 pandemic management knowledge domain and defines the information that needs to be captured in the schema of the database. The current version of the ontology has been driven by the data sources identified and collected at the current timing of the project. As more data sets will become available by third parties, the COVID-X ontology will be refined to follow the advancements in the system design.

5.2.1 Methodology

A methodology is the definition and organization of a number of fundamental phases that ensure the correct completion of deliverables, in this case the COVID-X ontology, and how we get there. The steps that were followed for the development of the COVID-X ontology are listed below:

- **Information-gathering**
- **Reuse existing ontologies**
- **Initial structuring (lightweight model)**

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

- **Formalization (heavyweight model):** Encode the lightweight knowledge models in order to be employed by information systems. Formalization requires using a formal knowledge representation language used to express the elements and properties of the model.

Each phase is further analyzed in the following subsections.

5.2.2 Information Gathering

The information-gathering phase is responsible for identifying the main elements, concepts, and data fields that are part of the domain of interest. For the development of the COVID-X common semantic data model, the first step was to identify the available data sources that will feed the COVID-X Sandbox. Each clinical partner has created a list that contains the available datasets and the associated metadata for each one, as presented in section 5.1. Furthermore, clinical partners extensively described the data fields and variables contained in each dataset. These variables were further analyzed and characterized for defining the attributes of the COVID-X common data schema. The COVID-X data schema is implemented as a superset of the identified datasets, where the attributes of this superset are used to construct the datatype properties of the COVID-X ontology. The data ingested in the Sandbox will be transformed and curated to comply with this superset, and they will be semantically enriched based on the COVID-X ontology.

5.2.3 Reuse Existing Ontologies

Ontologies are the Semantic Web basis [22], and a diverse range of ontologies of different domains have been constructed and shared by the ontology community. These ontologies have been described and published online, with the sole purpose of being reused as building blocks. In this regard, it is beneficial to reuse existing ontologies that are available publicly instead of constructing an ontology from scratch. COVID-X makes extensive use of terms and relations from other ontologies and vocabularies. Standardized ontologies are identified by namespaces that are based on the domain name system of the Web. The namespace is added as a prefix to the defined elements of the ontology to make them unambiguously identified over the web. For the COVID-X ontology, the namespace http://www.covid-x.eu/ontology has been selected. Table 15 summarizes the ontologies and vocabularies used in the COVID-X ontology.

TABLE 15: NAMESPACES USED IN COVID-X ONTOLOGY

Ontology	Prefix	Namespace
COVID-X	covidx	http://www.covid-x.eu/ontology
FOAF	foaf	http://xmlns.com/foaf/0.1/
OWL	owl	http://www.w3.org/2002/07/owl#
RDF	rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

RDFS	rdfs	http://www.w3.org/2000/01/rdf-schema#
PROV-O	prov	http://www.w3.org/ns/prov#
SKOS	skos	http://www.w3.org/2004/02/skos/core#
XSD	xsd	http://www.w3.org/2001/XMLSchema#
ORG-O	org	http://www.w3.org/ns/org#
HL7v3	hl7	http://purl.bioontology.org/ontology/HL7/class_code_classified_by
FHIR	fhir	https://www.hl7.org/fhir/
ICD-11	icd	https://icd.who.int/browse11/l-m/en
DCAT3	dcat	http://www.w3.org/ns/dcat#
VSO	vso	http://purl.obolibrary.org/obo/RO_0002001
Dublin Core	dc	http://purl.org/dc/elements/1.1/

5.2.4 Initial Structuring

The initial structure defines very general concepts of the target domains. It provides semantic interoperability across these concepts and builds a common starting point for the formulation of definitions. Graphical languages are commonly used to model lightweight ontologies visually. A graphical language is a schematic specification that can be used to model the building blocks of an ontology. The most critical elements that can be represented in knowledge modeling are: (i) classes, which are the meaningful groupings or sets that can contain individuals, (ii) relations/properties, which are associations involving two classes/subclasses, (iii) individuals, which are the instances of classes, and (iv) datatype properties, which relate classes or individuals to literal data, such as strings, numbers, dates, etc. Unified Modelling Language (UML) [19] was the preferred graphical language for creating the initial lightweight model of the COVID-X ontology. UML is a widely used graphical modeling language that allows users to plan and model systems of almost any kind.

Figure 11: Covid-x ontology lightweight model depicts the high-level lightweight model of the COVID-X ontology. This diagram defines the main classes and relations comprising the COVID-X ontology. The diagram employs and reuses many entities from standardized ontologies described above. These elements are identified with the respective ontology prefix. Table 16 summarizes the main classes and their relations that are defined in the COVID-X ontology.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

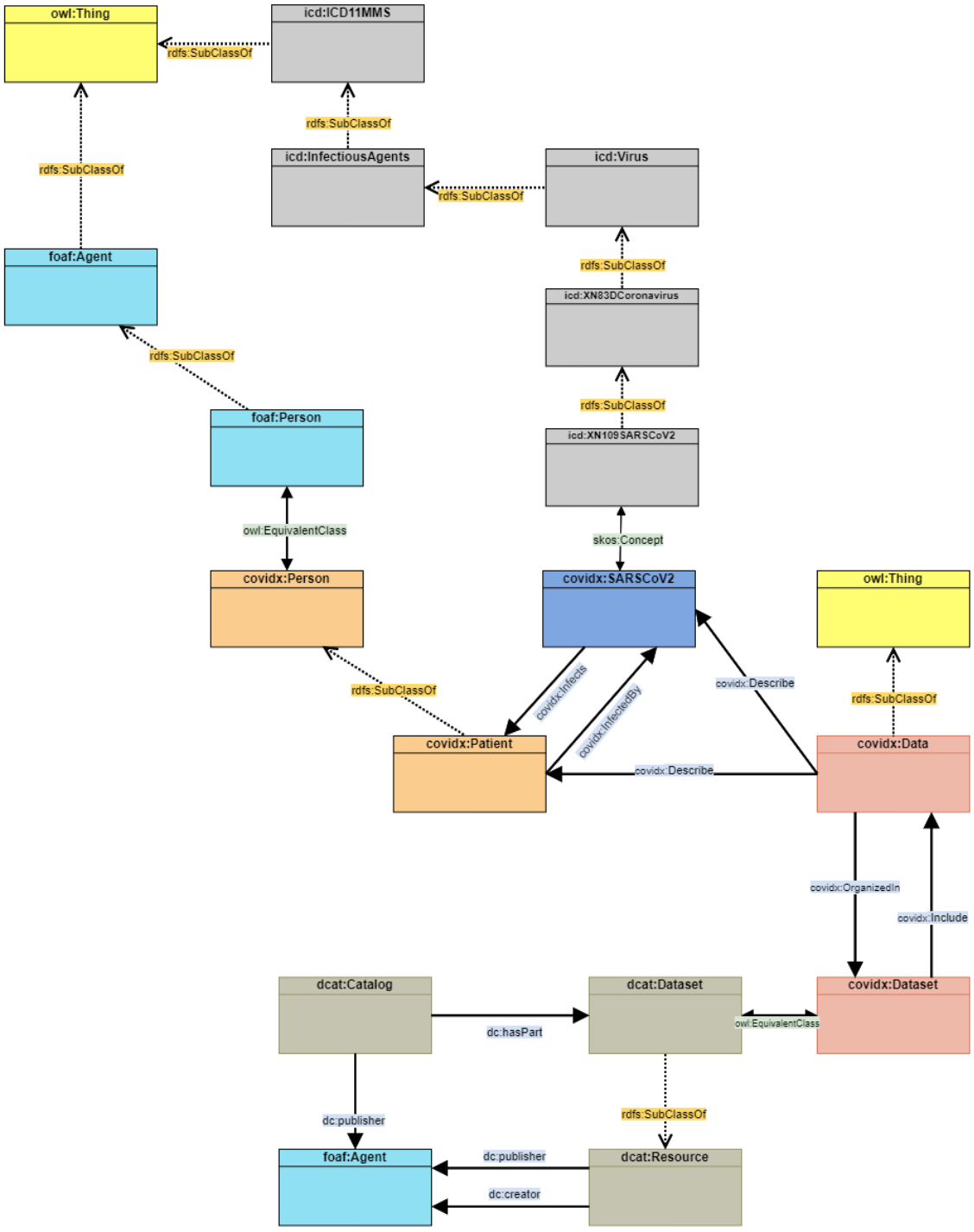

FIGURE 11: COVID-X ONTOLOGY LIGHTWEIGHT MODEL

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

TABLE 16: COVID-X ONTOLOGY MAIN CLASSES

Class	Subclass of	Relations	Description
covidx:Person	foaf:Agent		Equivalent class to **foaf:Person**. It represents a single person that can be alive, dead, real, or imaginary.
covidx:Patient	covidx:Person	• covidx:InfectedBy covidx:SARSCoV2	A person that has been infected by SARS-Cov-2 virus. This class inherits the datatype properties from covidx:Person class.
covidx:SARSCoV2	icd:XN83DCoronavirus	• covidx:Infects covidx:Patient	Equivalent class to **icd:XN109SARSCoV2**. The virus that causes Covid-19 disease
covidx:Data	owl:Thing	• covidx:Describe covidx:SARSCoV2 • covidx:Describe covidx:Patient • covidx:OrganizedIn covidx:Dataset	Data available in the COVID-X project and characterize patients and the SARS-CoV-2 virus
covidx:Dataset	owl:Thing	• covidx:Include covidx:data	Equivalent class to **dcat:Dataset**. It is defined as a collection of data published by a single agent, and available for access or download in one or more representations.

5.2.5 Formalization

The Formalization phase is about encoding the lightweight knowledge models in order to be employed by information systems. Formalization requires using a formal knowledge representation language for

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

expressing the elements and properties of the model. For the development of the COVID-X ontology, Web Ontology Language[1] (OWL) has been utilized. OWL is intended to be used when the information contained in documents needs to be processed by applications. It can be used to explicitly represent the meaning of terms in vocabularies and the relationships between those terms and define ontologies in a formalized format. A knowledge model expressed in OWL is constructed by combining elements and relations in a descriptive manner. Classes are specified and organized into flexible hierarchies, while relations that are called properties in OWL describe how classes and their individuals behave.

For the development of the formal ontology, Protégé has been employed. Protégé is an open-source OWL ontology editing tool. The created ontology is exported in XML/RDF syntax it is stored on a centralized Gitlab repository. Figure 12 depicts a snapshot of the COVID-X Ontology in Protégé, while Figure 13 shows the visualization of COVID-X Ontology built with Protégé VOWL plugin. The full list of elements comprising the COVID-X ontology is presented in APPENDIX 2.

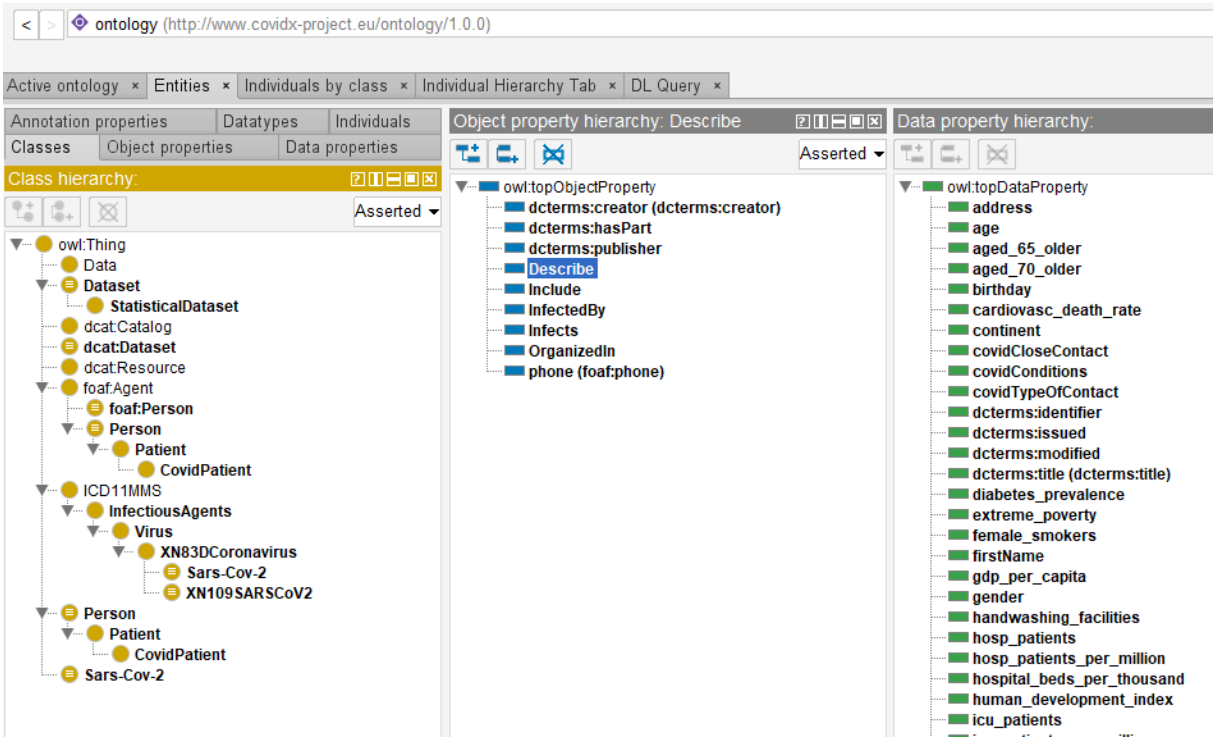

FIGURE 12: COVID-X ONTOLOGY IN PROTÉGÉ

[1]https://www.w3.org/TR/owl-features/

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

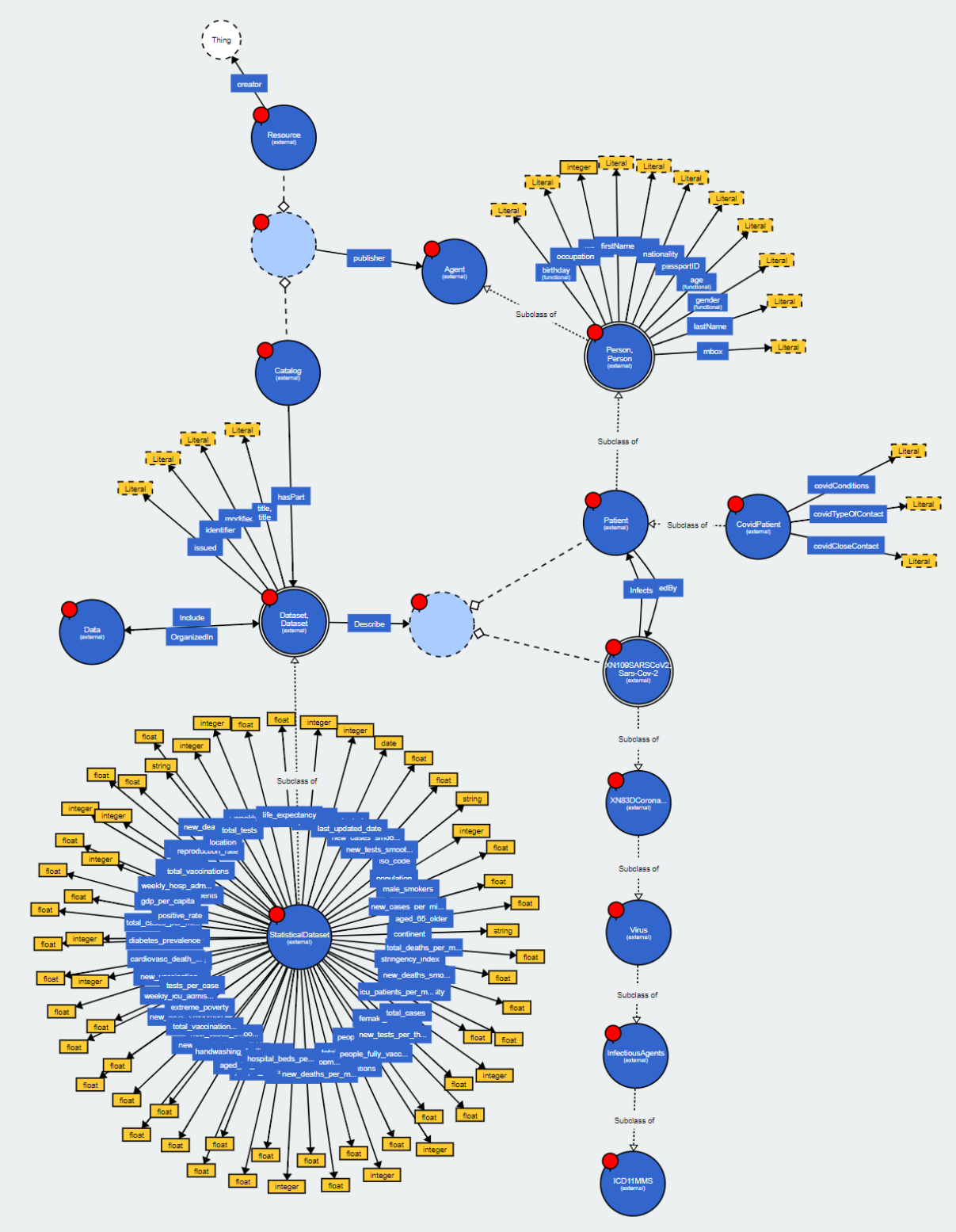

FIGURE 13: COVID-X ONTOLOGY VISUALIZATION

.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.

© 2020-2022 COVID-X

6 Conclusions

This document presented the user need and analyzed system requimrents that further drove the definition of the initial design of the envisioned COVID-X Sandbox architecture for both release A and release B. Sandbox follows the Service-Oriented Architecture approach, implemented as a combination of different components that collectively aim to enable seamless access to a set of healthcare data sources. Each component within the architecture realizes and delivers one or more of the aggregated services. The key advantages of this architectural approach are that it offers simplicity in building and maintaining applications, flexibility, and scalability, while the containerized approach makes the applications independent of the underlying system. The critical user requirements and general system requirements are also presented in this document, driving the design of the Sandbox Architecture.

The second part of this document presented a summary of the identified datasets that will be used for feeding the Sandbox. These datasets are either provided by the clinical data sources available in the COVID-X project or collected by open and publicly available data sources. Based on the data schema and variables of the available datasets, the COVID-X common semantic data model is created. The semantic data model provides a high-level structural and semantics-based representation of the data sources used to collect and store healthcare and other required data. It is implemented as an ontology, defining the main entities that exist and interact within the Covid-19 pandemic management knowledge domain. It is essential to mention that the current version of the ontology has been driven by the data sources identified and collected at the current timing of the project. As more data sets will become available by third parties, the COVID-X ontology will be refined to follow the advancements in the system design.

The outcomes of the current deliverables will be used to guide the development and implementation of both Sandbox releases. The releases will be presented in deliverables *D2.2 - First Sandbox implementation and services provision* and *D2.3 - Final Sandbox implementation and services provision*, respectively.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

7 References

[1] Frank Luh, Yun Yen, "Cybersecurity in Science and Medicine: Threats and Challenges," *Trends in Biotechnology,* vol. 38, no. 8, pp. 825-828, 2020.

[2] Muthuppalaniappan, Menaka, LLB & Stevenson, Kerrie, "Healthcare cyber-attacks and the COVID-19 pandemic: an urgent threat to global health," *International Journal for Quality in Health Care,* vol. 33, 2020.

[3] "REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL (GDPR)," 2016. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

[4] Konečný, Jakub, et al, "Federated learning: Strategies for improving communication efficiency.," in *NIPS Workshop on Private Multi-Party Machine Learning*, 2017.

[5] Choudhury, O., Gkoulalas-Divanis, A., Salonidis, T., Sylla, I., Park, Y., Hsu, G., and Das, A., "Differential privacy-enabled federated learning for sensitive health data," 2019.

[6] Li Huang, Andrew L. Shea, Huining Qian, Aditya Masurkar, Hao Deng, Dianbo Liu, "Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records," *Journal of Biomedical Informatics,* vol. 99, 2019.

[7] Skripcak, Tomas, et al, "Creating a data exchange strategy for radiotherapy research: towards federated databases and anonymised public datasets," *Radiotherapy and Oncology,* pp. 303-309, 2014.

[8] Shaoqi Chen, Dongyu Xue, Guohui Chuai, Qiang Yang, Qi Liu, "FL-QSAR: a federated learning-based QSAR prototype for collaborative drug discovery," *Bioinformatics,* 2020.

[9] Viraaji Mothukuri, Reza M. Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha, Gautam Srivastava, ""A survey on security and privacy of federated learning," *Future Generation Computer Systems,* vol. 115, pp. 619-640, 2021.

[10] COVID-X Consortioum, "D1.1 - Ethical and Legal Framework," COVID-X, 2021.

[11] Coughlan, Jane & Macredie, Robert, "Effective Communication in Requirements Elicitation: A Comparison of Methodologies. Requirements Engineering," 2002.

[12] "The next generation data platform," [Online]. Available: https://www2.deloitte.com/lu/en/pages/technology/articles/next-generation-data-platform.html.

[13] Medical Device Coordination Group (MDCG), "MDCG2019-16 Guidance on Cybersecurity for medical devices," 2019. [Online]. Available: https://ec.europa.eu/docsroom/documents/41863.

[14] THE EUROPEAN PARLIAMENT AND OF THE COUNCIL, "REGULATION (EU) 2017/745 on Medical Devices," April 2017. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02017R0745-20170505&from=EN.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

[15] THE EUROPEAN PARLIAMENT AND OF THE COUNCIL, "REGULATION (EU) 2017/746 on diagnostic medical devices," April 2017. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02017R0746-20170505&from=EN.

[16] "The Importance of Scalability In Software Design," [Online]. Available: https://www.conceptatech.com/blog/importance-of-scalability-in-software-design .

[17] A. Bondi, "Best practices for writing and managing performance requirements: a tutorial," 2012.

[18] IETF, "Information technology — Reference Architecture for Service Oriented Architecture (SOA RA) — Part 1: Terminology and concepts for SOA," ISO/IEC JTC 1/SC 38 Cloud computing and distributed platforms, 2016.

[19] Mojtaba Shahina, Muhammad Ali Babara, Liming Zhu, "Continuous Integration, Delivery and Deployment: A Systematic Review on Approaches, Tools, Challenges and Practices," *IEEE Access,* 2017.

[20] COVID-X Consortium, "D2.2 - First Sandbox implementation and services provision," COVID-X, 2021.

[21] Okereafor, Kenneth & Marcelo, Alvin, "Addressing Cybersecurity Challenges of Health Data in the COVID-19 Pandemic," 2020.

[22] Tim Berners-Lee, James Hendler and Ora Lassila, "The Semantic Web: A New Form of Web Content That is Meaningful to Computers Will Unleash a Revolution of New Possibilities," *Scientific American,* 2001.

[23] Giansanti, Daniele, "Cybersecurity and the Digital-Health: The Challenge of This Millennium," 2021.

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

8 Appendix A: Data Schema of the Available Datasets

SERMAS

Data Source 1 – Paciente

Variable	Description	Type
NHC_HCSC	Patient's medical history number	number
sexo	Sex	date/time
fechanac	Date of birth	character
paisnac	Country of birth	character
paisresi	Country of residence	character

Data Source 2 – CMBD_H & Data Source 4 - CMBD_A

Variable	Description	Type
NHC_HCSC	Patient's medical history number	number
PROCED	Origin of the admission.	character
HOSPROC	Hospital where the patient comes from	character
FECURG	Date and time of arrival at the emergency department	date/time
FECING	Date and time of hospital admission	date/time
TIPING	Circumstances of admission	character
SERVING	Service where the patient is admitted	character

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

SECCING	Code that identifies the sections or units of the hospital's services where the patient is admitted	character
SERVALT	Service that discharges the patient. In the event of voluntary discharge, death or leakage, the last service where the patient was admitted will be listed.	character
SECCALT	Code that identifies the sections or units of the hospital's services.	character
FECALT	Date and time of discharge.	date/time
TIPALT	Circumstances of discharge and continuity of care	character
FECINT	Only to be completed if the patient has undergone a surgical intervention or any procedure carried out in the operating theatre, including Caesarean sections. If there is more than one operation in this field, the date and time of the first one will be recorded.	date/time
FECINT2	Only to be completed if the patient has undergone MORE than one surgical intervention or any procedure carried out in the operating theatre. In this field, the date and time of the last one will be recorded.	date/time
M1-M7	Morphology codes of neoplasias. These are obligatory.	character
C1	Main diagnosis	character
POAC1	Existing or present diagnosis at entry or start of contact	character
CE1	This variable (EC1) will be used exclusively as complementary information to the main diagnosis in order to identify the external cause that has motivated it. Its completion is obligatory in these cases.	character

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

POACE1	Existing or present diagnosis at admission or at the beginning of the contact	character
C2-C20	Other diagnostics	character
POAC2-POAC20	Existing or present diagnosis at entry or start of contact	character
P1-P22	Surgical and/or obstetric procedures and other procedures	character
FECTRAS1-FECTRAS8	Date and time of transfer	Fecha/Hora
SERVTRAS1-SERVTRAS8	Identification of the transfer service. This is the service responsible for assisting the patient once an inter-service transfer has taken place within the centre.	Carácter
SECCTRAS1-SECCTRAS8	Identification of the section of the transfer. This is the section responsible for the care of the patient once a transfer has taken place within the centre.	character
HOSPDEST	Hospital of destination, code of the hospital to which the patient is referred. In the case of hospitals dependent on the Ministry of Defence, they will be identified with the corresponding code according to the National Hospital Catalogue.	character
CIE	CIE version	character
NUMICU	Episode Identifier	number

Data Source 4 – CMBD_U

Variable	Description	Type

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

NHC_HCSC	Patient's medical history number	number
MOT_ASISTENCIA	Reason for attendance	number
PROCEDENCIA	Identify the origin of the episode.	character
AREA_ING	Code of the Area that attends the patient on admission to the Emergency Room.	number
AREA_ALTA	Code of the Area that discharges the patient from the Emergency Room.	number
SERVICIO_ING	Code of the Regional Service that attends the patient on admission to the Emergency Department	character
SERVICIO_ALTA	Code of the regional service that discharges the patient	character
PRIORIDAD	In case of pre-evaluation, indicate status code	number
FECHA_URG	Date and time of entry into the emergency room	date/time
FECHA_TRIAJE	Date and time of triage	date/time
FECHA_INT_QUIR	Date and time of entry to the operating room	date/time
FECHA_ALTA	Date and time of medical discharge indicated in the Emergency Department Discharge Report. If there is an Admission Order, it will coincide with the FECHA_ORDENING	date/time
TIPO_ALTA	Type of discharge	character
SERV_ING	Service in charge of the admission	character
FECHA_SALIDA	Date and time of emergency exit (or hospital admission)	date/time

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

C1	Main diagnosis	character
CE1	This variable (EC1) will be used exclusively as complementary information to the main diagnosis in order to identify the external cause that has motivated it. Its completion is obligatory in these cases.	character
C2-C10	Other diagnoses	character
P1-P10	Surgical and/or obstetric procedures and other procedures	character
NUMICU	Episode Identifier	number

Data Source 5 – Lisencam

Variable	Description	Type
NHC_HCSC	Patient's medical history number	number
SERVPAC	Service where the patient is admitted	character
CAMA	Bed he was admitted to	character
T_FINAN	Type of financing	categories
PROC	Procedure	categories
AREA	Healthcare area of the Community of Madrid	categories
SERV_CAMA	Service where the patient is admitted	character
COD_ING	Code of admission	Numérica

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

DIAG_ING	Admission diagnosis	character
DIRECCION	Health Centre/Consultancy	categories
UNID_ENF	Nursing unit	character
TIP_ING	type of admission	character
HOSP_REF	Reference hospital	character
DESC_PROV	Province	character
TIP_CIUDADANO	Citizen type	character
FEC_CORTE	Cut-off date	date

Data Source 6 – CEX: Administrative database. Record of outpatient visits

Data Source 7 – Farmacia

Variable	Description
NHC_HCSC	Patient's medical history number
grupo1	Service of origin
grupo2	Diagnosis
grupo4	active ingredient
cantidad	number of pharmaceutical forms dispensed

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

centro	Centre
codigo1	Service of origin (abbreviation)
codigo4	pharmacist internal code
ddd	defined daily dose
fecha	date of prescription
dosis	total dose per dispensing (quantity*unit dose active ingredient)
dospresen	Units/doses per commercial package
dias_dispensados	Duration of treatment (some are blank as they are 1-dose treatments)
unidad_med	unit of measure of the active ingredient
pacientes	Patients reached (some blank as it is not an adapted field for data extraction)
pacientes_tot	patient linked to that movement (it is always 1)
numicu	Episode Identifier
fecha_ingreso_u	date of admission to hospital
procesos_tot	administrative identifier?
n_dispensaciones	number of dispensations
dosis_pres	unit dose of the active ingredient

Data Source 8 – Laboratorio

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

Variable	Description	Type
NumeroHistoria	Patient's medical history number	number
Nombre	Lab test name	Carácter
LOINC_NUMBER	Lab test LOINC code	Numérico
Petición	Lab test unique identifier	Numérico
Fecha	Lab test date and time	Fecha/hora
Medición	Value	Numérico
Unidad	Unit	Carácter
Centro Pide	Sample request service	Carácter

Data Source 9 – PCR

Variable	Description	Type
NHC_HCSC	Patient's medical history number	number
TIPO_MUESTRA	Sample type	character
FEC_MUESTRA__	Date of sample	date
PETICIONARIO	Sample request service	character

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

RESULT_COVID_19	Results of the COVID-19 sample	character

Data Source 10 – UCI: Clinical variables

Data Source 11 – Biobanco: Different variables under request

Data Source 12 – Imagen: Under request

ICH

Dataset ICH 1 - Covid Clinical DB:

Data format: CSV

Data structure: data is arranged in a relational database, split in different thematic tables. Following the tables of the DB:

- ER Admissions;
- Hospital Admissions;
- Patients;
- Transfers;
- Procedures;
- Diagnosis;
- Chartevents;
- Labevents;
- Microbiologyevents;
- Inputevents;

Part of the tables includes the following dictionaries: D_chartevents, D_labevents, D_inputevents, D_procedures, D_diagnosis, D_microbiologyevents

Please note the Covid Clinical Dataset is modeled as the MIMICIII Publicly available Dataset.

Data variables:

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

- Administrative data: length-of-stay, Service, internal transfers, cause of discharge;
- Diagnosis and procedures at discharge: ICD-9;
- Treatment prescribed and administered during admission;
- Labwork: complete blood cell count, chemistry, coagulation testing, D-Dimer, IL-6;
- Imaging test (chest X-ray, CT scan): non protocolized, performed by medical decision based on the clinical manifestations and CT at admission in the emergency department.
- Clinical notes: unannotated and reported in Italian;
- COVID-19 related clinical manifestations: comorbidities, symptoms;
- Vital Signs: including heartbeat rate, blood pressure and body temperature;
- Treatment: prescribed and administered during admission.

Dataset ICH 2 - Covid CT DB:

Data format: DICOM, CSV, JSON

Data Structure: data is arranged in a master CSV file. A reference to DICOM UIDs is provided and allows links with each CT slice, exam and study.

Variables: a master CSV file with patients information indexes the CT DICOMS. Each CT scan DICOM including a subcollection of relevant metadata and the image data. The dataset also includes morphological annotations of covid-19 related anomalies for a subset of the available CT scans (included as JSON file).

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

9 Appendix B: COVID-X Full Ontology

Entity	SubClass of	Equivalent to
Data		
Dataset		dcat:Dataset
Person	foaf:Agent	foaf:Person
Patient	Person foaf:Agent foaf:Person	
CovidPatient	Patient Person :foaf:Agent foaf:Person	
StatisticalDataset	Dataset dcat:Dataset	
Sars-Cov-2		XN109SARSCoV2

Object Property	Domains	Ranges
Describe	Dataset	Sars-Cov-2 Patient
Include	dcat:Dataset	Data
InfectedBy	Patient	Sars-Cov-2
Infects	Sars-Cov-2	Patient
OrganizedIn	Data	Dataset

Datatype Property	Domain	Range
identifier	Dataset	literal
issued	Dataset	literal
modified	Dataset	literal

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

title	Dataset	literal
iso_code	StatisticalDataset	string
continent	StatisticalDataset	string
location	StatisticalDataset	string
last_updated_date	StatisticalDataset	date
total_cases	StatisticalDataset	integer
new_cases	StatisticalDataset	integer
new_cases_smoothed	StatisticalDataset	float
total_deaths	StatisticalDataset	integer
new_deaths	StatisticalDataset	integer
new_deaths_smoothed	StatisticalDataset	float
total_cases_per_million	StatisticalDataset	float
new_cases_per_million	StatisticalDataset	float
new_cases_smoothed_per_million	StatisticalDataset	float
total_deaths_per_million	StatisticalDataset	float
new_deaths_per_million	StatisticalDataset	float
new_deaths_smoothed_per_million	StatisticalDataset	float
reproduction_rate	StatisticalDataset	float
icu_patients	StatisticalDataset	integer
icu_patients_per_million	StatisticalDataset	float
hosp_patients	StatisticalDataset	integer
hosp_patients_per_million	StatisticalDataset	float
weekly_icu_admissions	StatisticalDataset	float
weekly_icu_admissions_per_million	StatisticalDataset	float
weekly_hosp_admissions	StatisticalDataset	float
weekly_hosp_admissions_per_million	StatisticalDataset	float
new_tests	StatisticalDataset	integer
total_tests	StatisticalDataset	integer
total_tests_per_thousand	StatisticalDataset	float

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

new_tests_per_thousand	StatisticalDataset	float
new_tests_smoothed	StatisticalDataset	float
new_tests_smoothed_per_thousand	StatisticalDataset	float
positive_rate	StatisticalDataset	float
tests_per_case	StatisticalDataset	float
tests_units	StatisticalDataset	integer
total_vaccina	StatisticalDataset	integer
people_vaccinated	StatisticalDataset	integer
people_fully_vaccinated	StatisticalDataset	integer
new_vaccinations	StatisticalDataset	integer
new_vaccinations_smoothed	StatisticalDataset	float
total_vaccinations_per_hundred	StatisticalDataset	float
people_vaccinated_per_hundred	StatisticalDataset	float
people_fully_vaccinated_per_hundred	StatisticalDataset	float
new_vaccinations_smoothed_per_million	StatisticalDataset	float
stringency_index	StatisticalDataset	float
population	StatisticalDataset	integer
population_density	StatisticalDataset	float
median_age	StatisticalDataset	float
aged_65_older	StatisticalDataset	float
aged_70_older	StatisticalDataset	float
gdp_per_capita	StatisticalDataset	float
extreme_poverty	StatisticalDataset	float
cardiovasc_death_rate	StatisticalDataset	float
diabetes_prevalence	StatisticalDataset	float
female_smokers	StatisticalDataset	float
male_smokers	StatisticalDataset	float
handwashing_facilities	StatisticalDataset	float
hospital_beds_per_thousand	StatisticalDataset	float

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X

life_expectancy	StatisticalDataset	float
human_development_index	StatisticalDataset	float
address	Person	literal
covidCloseContact	Patient	literal
covidConditions	Patient	literal
covidTypeOfContact	Patient	literal
nationality	Person	literal
occupation	Person	literal
passportID	Person	literal
weight	Person	integer
title	Dataset	literal
age	Person	literal
birthday	Person	literal
firstName	Person	literal
gender	Person	literal
lastName	Person	literal
mbox	Person	literal
phone	Person	literal

COVID-X has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101016065.
© 2020-2022 COVID-X