Technical Specification

ISO/TS 8376

Genomics informatics — Requirements for interoperable systems for genomic surveillance

First edition
2023-12

Reference number
ISO/TS 8376:2023(en)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

ISO draws attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO takes no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents. ISO shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC215, *Health informatics*, Subcommittee SC 1, *Genomics Informatics*.

Any feedback or questions on this document should be directed to the user’s national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

0.1 Rationale

In a world where international travel and trade is essential, a pathogen affecting one region or country can rapidly spread around the globe. One of the most important tools in responding to infectious disease is genomic surveillance, the process of constantly monitoring pathogens, and analysing their genetic similarities and differences.

Genomic surveillance is transforming public health action by providing a deeper understanding of pathogens, their evolution and proliferation. Alongside clinical, epidemiological and other multi-source data, genomic data for potentially dangerous pathogens informs risk assessments, enables governments or non-governmental organizations (NGOs) to track emerging or spreading infectious diseases, and supports tailored recommendations for prevention. For example, governments can use genomic surveillance to guide policy or public health measures, academic and research organizations can interrogate the pathogen and understand its impact, and individuals can be better informed about potential risks.

To make use of pathogen genomics data, it must be interpreted using contextual data, such as sample metadata, laboratory methods, patient demographics, clinical outcomes, and epidemiological information, underscoring the importance of incorporating a variety of data in surveillance tools. Due to the importance of contextual data in the interpretation of pathogen data collected by a network of independent data acquisition nodes, when this document refers to “data”, it means both genomic and contextual data. Similarly, all derivatives such as “data access” and “data sharing” include contextual data as well.

The focus of this document is genomic surveillance of pathogens; however, it is important to consider the role of multi-source data when building federated surveillance systems, including health records and administrative data, in understanding the clinical significance of a given pathogenic variant or prevention strategies. Scientists are also increasingly looking at the environment, including human host genetics, to identify biomarkers that can explain susceptibility to as well as severity of disease.

The emergence of SARS-CoV-2 and impact of the COVID-19 pandemic crystalized the need for coherent regional, national and global genomic surveillance systems. The world needs timely, high quality and geographically representative data in as close to real-time as possible. To realize the benefits of genomic surveillance data needs to be shared across jurisdictions, both within and between countries, through networks, systems and platforms.

Sharing pathogen genome data is critical for preventing, detecting and responding to epidemics at national and international level, as well as monitoring and responding to endemic diseases and tracking antimicrobial resistance. However, genomic surveillance presents challenges, in terms of the infrastructure, capacities and capabilities needed, and the harmonization across systems and countries to be able to compare and use the data effectively. Digital systems for genomic surveillance are becoming increasingly available, however, they are not being built on common design principles and rarely use standards that enable them to interact as nodes in a national interoperable digital network and even less so in a highly dynamic international ecosystem.

The generation of high-quality pathogen genomic data that can be shared quickly and effectively in a global system requires capacity and infrastructure. Building upon current and new advances in genomic data sharing digital systems and platforms and to enable future effective, quality, safe, understood, timely and accurate genomics data sharing and surveillance, a common set of design principles, services requirements and standards must be rapidly determined, quickly adopted through consensus and widely published to realize such large-scale interoperability. As countries and organizations begin to build a stronger global architecture for health emergency preparedness and response, global standards are critical to support interoperability and collaboration, particularly in a federated model.

0.2 Importance of sharing data and benefits to regional, national, and international response

The COVID-19 pandemic underscores the importance of interoperable solutions to facilitate rapid data sharing and data governance to support critical pandemic response activities. In a globalized world, successful pandemic response depends on nation states and regions rapidly communicating accurate information, including pathogen identification, incidence, transmission patterns and mortality to the international community. This information allows regional and national jurisdictions, as well as international organizations, to implement targeted and comprehensive control measures as quickly as possible, protecting at the utmost the health and safety of the citizens or populations they serve.

The importance of sharing data to address global health priorities, including informing responses to outbreaks and epidemics, is now widely recognized. In the context of COVID-19, widespread mandates for data sharing were established by several international stakeholders including national governments, global health NGOs, scientific journals, research funders, and research institutions. In epidemics and pandemics, the case for such practices is especially urgent and required to develop much-needed vaccines, therapeutics and diagnostics.

For example, researchers in China sequenced the SARS-CoV-2 genome — the virus that causes COVID-19 in humans — and made the data publicly available through an open access platform in January 2020, which sped up the development of critical diagnostic assays. As the virus spread and mutated, becoming more virulent and more transmissible, data from around the world enabled countries to rapidly change public policy and enabled vaccine development at an unprecedented pace or scale.

Further, widespread public and private data sharing across domains and borders during the COVID-19 pandemic generated insights never seen before at this scale. Through linkages between viral genomic and other types of data (such as policy or mobility data), public health bodies and decision makers could model the potential impact of border or workplace closures.

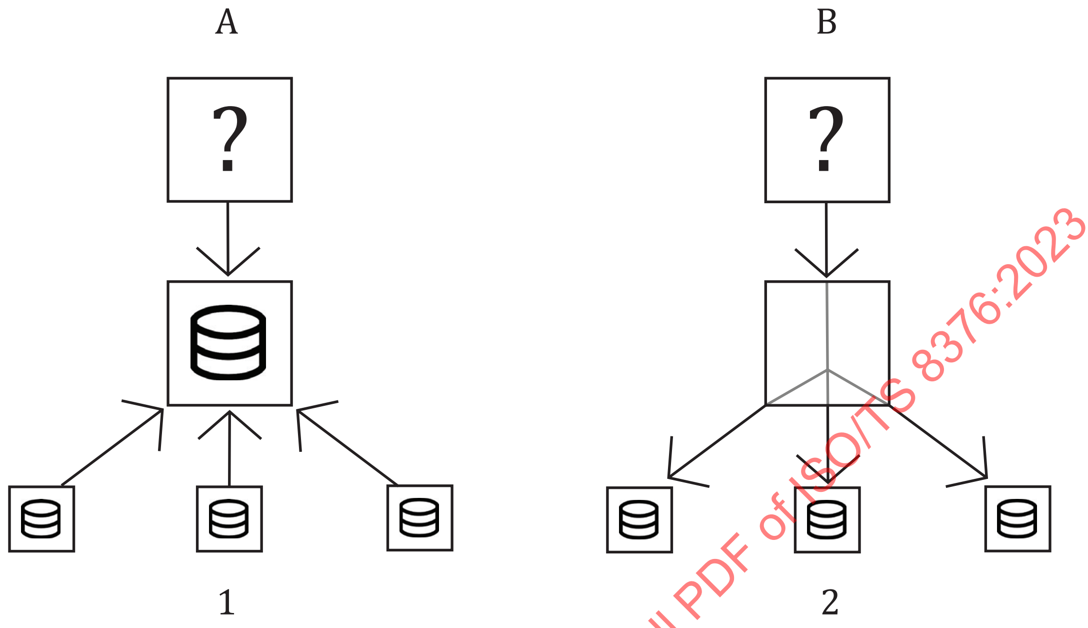
0.3 Level of interoperability

This document is intended to highlight the requirements for interoperability between networks, systems and platforms required to share data within and between organizations and countries. It focuses on the technical, structural and syntactic levels of interoperability, while acknowledging the importance of semantic interoperability. It does not however address business, organization, or other types of interoperability that aim to address differences in organizational perspectives within the genomic surveillance domain. It does not describe methods or best practices for data management, which might need to be harmonized to make use of data shared within systems. The problem spaces, use cases and data access patterns in which genomic data plays a key role also continue to grow, resulting in a need to build an accompanying framework of extensible biomedical workflow profiles, which would then provide the specific contexts of system interoperability. Such frameworks and profiles are not the topic of this document.

Additionally, this document does not model, represent or manage genomic surveillance as a system, framework, or domain; nor does it provide a data model or information model for genomic surveillance; nor is it intended as a generic system architecture for genomic surveillance knowledge level interoperability. It is understood such generic and formal genomic surveillance modelling work is seen as a potential valuable and useful standards activity and would be greatly aided and founded in the formal system-oriented, architecture-centric, ontology-based and policy-driven approach as standardized in ISO 23903. Modelling genomic surveillance for multi-domain interoperability requires the advancement from data models, information models and Information Communication Technology (ICT) domain knowledge perspectives to the knowledge perspective of genomic surveillance with an abstract, domain-independent representation for genomic surveillance systems. Such work can generate an ISO 23903 interoperability and integration reference architecture instance. That said, the model and framework for formally modelling and managing genomic surveillance and its behaviour and creating an interoperability and integration reference architecture instance is beyond the scope of this document. It is recognized and acknowledged that should such genomic surveillance standard modelling processes and the associated, explicit, formalized ISO 23903-based integration and reference architecture instance for genomic surveillance knowledge level interoperability and harmonization be an agreed, completed and published ISO specification, it may require adaptation or revision of this document.

0.4 Data federation

Data federation is a technique that allows search and data analysis to be performed across multiple distributed datasets, with each individual dataset remaining in its protected local environment, instead of copying or moving data into a single centralized location. The centralized and federated models of data sharing are described in Figure 1.



Key

A centralization

B federation

1 Data from multiple sources are moved into a central location to be queried. Data custodians relinquish control over the data and cannot directly enforce access policies.

2 Data from multiple sources are queried through a system, network, or platform that facilitates access, enabling each data custodian to maintain control of the data.

Figure 1 — Overview of the centralized and federated models of data sharing

Since the location of the data does not change, the data custodian responsible for the dataset maintains administrative control of the data, including privacy, security and access based on consent. Further, data generators have transparency into how the data are used and can enforce attribution policies. In a federated system, researchers send their questions to the data and do not have direct access to the data. Instead of creating and distributing multiple copies of large files to researchers looking to analyse the data, a single copy of sensitive data is created and stored in the same region as the data was generated. Through a federated system, network, or platform, distributed data can be linked to other relevant data, for example connecting genomic data with clinical or administrative data.

Data federation is particularly valuable in genomic surveillance, where a large volume or diversity of data is required to generate insights, and where real time data and regional representation benefit the global community and regional response. However, data federation requires navigating different data policy laws, security and privacy protocols, and data interoperability challenges.

0.5 Technological foundation for secure data sharing

The need to share data between and within organizations in the healthcare and life sciences sector has long been recognized, however broad sharing of data has been limited due to privacy and security considerations, and interoperability across systems.

The necessary technologies to address interoperability are being developed and implemented in the healthcare and life sciences sector of cloud computing, application programming interface (API) management, cybersecurity, as well as access to lightweight health resources, such those defined by HL7®[1] FHIR®[2]. A central component of federated data systems is the use of APIs and foundational architecture, which enables a scalable, secure and reliable means of accessing data from data custodians, particularly as data sources likely use different underlying technologies and data formats.

Various solutions have been developed and most, if not all of them, have used the paradigm of data "exchange" and even led to creating an entire segment in digital health called electronic Health Information Exchange (HIE). In such an exchange, data practically changes hands and is transferred from a "data provider" or "data custodian" to a "data consumer". With all the benefits and simplicity of exchanging data, it also poses challenges — duplicating large amounts of data on both sides of the exchange, keeping the data in sync, keeping track of all transformations that data goes through, enforcing rules on transitive downstream data exchange, as well as passing the control of contextual data that might contain identifiable personal information with all the privacy, security and data governance concerns associated with it.

The technological developments in the recent years have enabled us to begin to talk more about "data sharing" and "data access", which is substantially different from exchanging copies of data. One sharing data approach uses a centralized system acting as a broker or intermediary in the form of a "data union" or a "data cooperative" where data governance rules and "data use contracts" can be defined and strictly imposed. Another one is via using a truly distributed, federated approach, like the InterPlanetary File System (IPFS) and blockchain.

Blockchain has enabled a promise of building a new Internet (often referred to as Web3) where digital assets can not only be read, written and accessed but also be owned, thus giving their "owners" the exclusive decision of how to share their data and extract value from it. This is extremely important for the contextual data discussed before, as in some cases it might contain very sensitive information. Blockchain has also enabled the tracking of the many transformations data goes through and allows the reproduction of analytical results with confidence guaranteed by data's cryptographic immutability.

Another foundational technology that enables true federation in an open network is decentralized identifiers (DID), self-sovereign identity (SSI), verified credentials (VCs) and Trust over IP (ToIP). These technologies enable dynamically adding trusted nodes to a network as well as uniquely identify datasets and their derivatives.

0.6 Use cases

Pathogens such as viruses and bacteria are constantly evolving in response to selective pressures, and these changes result in different characteristics, such as a pathogen being more or less transmissible, detectable and deleterious. Once a pathogen is identified in the human population, ongoing sequencing and genomic surveillance facilitates tracking geographic distribution and spread, as well as monitoring genomic alterations that change characteristics of the pathogen and its impact on the host.

Genomic surveillance tools can be developed or used by countries or governments, non-governmental organizations (NGOs) or global health initiatives, or industry providing solutions or whose business is impacted by infectious disease, as well as individuals conducting research in an academic setting. Further, data or insights facilitated by such tools can be consumed by individuals through mainstream media or research to understand the personal impact.

[1] HL7 is the registered trademark of Health Level Seven International. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of the product named.

[2] FHIR is a trademark of HL7®. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of the product named.

A generic use case for a genomics surveillance tool could include a public health agency and their data scientists and epidemiologists, with genome sequence data sources from multiple local or provincial (state) jurisdictions and multiple additional county ministries of health. Such a use case would demonstrate a federated surveillance system, built to compare data from local or provincial (state) jurisdictions to open data from all other jurisdictions in the country, and other jurisdictions worldwide. While a generic use case is not included in this document, Annex A provides actual in use project examples of federated networks to support genomic surveillance as well as human genomic research.

It is important to note that given the nature of pathogen surveillance, data collection is ongoing, and both data and insights are constantly changing. While one of the benefits of data federation is that data can be shared in near real-time, this also means that the results of a query are reflective of that point in time — the same query might return different results at different times. Further, in order to generate accurate insights, data must be transformed and harmonized in such a way that it can be analysed alongside other data, however this document focuses on interoperability for data sharing at the systems level and does not address the requirements for data.

Genomics informatics — Requirements for interoperable systems for genomic surveillance

1 Scope

This document outlines the design principles and the service and standards requirements to enable an interoperable system for genomic surveillance (herein referred to as “federated surveillance system”), including data representation, discovery and analysis, and data linkage.

Using select profiles this document applies to genomics digital systems, networks and platforms that enable a federated approach for researchers, clinicians, and patients in both the private and public sector at the local, regional, and international levels.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
genomic surveillance
pathogen surveillance
sequencing of genetic material of pathogens to identify and monitor genetic changes linked to the origins or characteristics of a disease afflicting different people

3.2
interoperability
ability of a system or a product to work with other systems or products without special effort on the part of the customer

Note 1 to entry: Under traditional ICT focus, interoperability is the ability of two or more systems or components to exchange information and to use the information that has been exchanged.

[SOURCE: ISO/IEC 2382: 2015, 2120585]

3.3
federated system
federated network
federated database
collection of independent but co-operating database systems that are distributed, autonomous and heterogeneous

Note 1 to entry: in a federated network, “shared” data are not moved into a centralized location for analysis, but rather queries are distributed across data sources

[SOURCE: ISO 19297-1:2019, 3.2, modified — “Independent” and “distributed” were added to the definition.]

3.4
data
representation of information in a formal manner suitable for communication, interpretation, or processing by human beings or computers

[SOURCE: ISO 10303-1:2021, 3.1.29]

3.5
data custodian
custodian
person or entity that has custody, control or possession of electronically stored information and is responsible for the safe implementation of data access policies

Note 1 to entry: in a federated model, there is only one copy of the data and thus only one data custodian who is responsible for adjudicating and granting access.

[SOURCE: ISO/IEC 27050-1:2019, 3.2, modified — Access policies were added to the definition and to Note 1.]

3.6
data provider
organization which produces data or reference metadata

[SOURCE: ISO 17369:2013, 2.1.18]

3.7
data consumer
individual or organization that uses data as a starting point

Note 1 to entry: In the research domain, a data consumer is a scientist or research group for commercial or non-commercial purposes.

Note 2 to entry: In the medical domain, a data consumer can be a physician or patient. In some cases, consumer can also be payer for commercial or non-commercial purposes.

Note 3 to entry: A data consumer may be any entity that uses data as an input in any form at any time.

[SOURCE: ISO/TR 3985:2021, 3.4]

3.8
data sharing
access to or processing of the same data by more than one authorized entity. In a federated system, sharing denotes access to the data to be viewed, queried, or analysed without making copies

[SOURCE: ISO/IEC TS 38505-3:2021, 3.7, modified —Additional context included in the federated model.]

3.9
data access
process by which a user or a system can retrieve or read published data on another system

Note 1 to entry: This data access happens over a network connection and the data typically does not persist after the connection is terminated.

[SOURCE: ISO 5127:2017, 3.1.11.17, modified — Note to entry has been added (taken from ISO/IEC 22624:2020, 3.5).]

3.10
data use
handling or dealing with information for a specific purpose

Note 1 to entry: This includes reproducing the information but does not include disclosing the information.

[SOURCE: ISO/TS 14265:2011, 2.11]

3.11
ontology
logical structure of the terms used to describe a domain of knowledge, including both the definitions of the applicable terms and their relationships

[SOURCE: ISO/IEC/IEEE 24765:2017, 3.2691]

3.12
variant
alteration in the most common DNA nucleotide sequence

Note 1 to entry: can describe an alternation that may be benign, pathogenic, or of unknown significance.

Note 2 to entry: variant implies deletion, insertion, indel or SNP.

[SOURCE: ISO 4454:2022, 3.23]

3.13
data linkage
aggregating data on one topic or data subject from several sources to create a new, richer dataset.

[SOURCE: ISO 5127:2017, 3.1.11.12, modified — Additional context included.]

4 Abbreviated terms

AAI	Authentication and Authorization Infrastructure
API	application programming interface
COVID-19	Coronavirus disease
CWL	Common Workflow Language
DID	decentralized identifiers
DRS	Data Repository Service
DUO	Data Use Ontology
FHIR®	Fast Healthcare Interoperability Resources
GA4GH	Global Alliance for Genomics and Health
HIE	electronic health information exchange
HPC	high performance computing
IPFS	InterPlanetary File System
NGOs	non-governmental organizations
OMOP	Observational Medical Outcomes Partnership
SaRS-CoV-2	Severe acute respiratory syndrome coronavirus
SSI	self-sovereign identity
ToIP	Trust over IP
TRS	Tool Registry Service

URL	Uniform Resource Locator
VCs	verifiable credentials
VRS	Variation Representation Specification
WDL	Workflow Description Language
WES	Workflow Execution Service

5 Design principles

5.1 Overview

In the development of federated surveillance systems, which requires cooperation with other such tools, there are dependencies that necessitate the designer to consider certain principles. Any large-scale system with participation from numerous stakeholders and client devices can face constraints from insufficient network communication or service reliability, especially in a highly distributed model.

While data harmonization is not the focus of this document, a number of these principles speak to requirements of the system to enable distributed data collected by different data custodians for different purposes to be used together, which is particularly important given the significance of contextual data to inform genomic surveillance. Missing data should be accounted for, beyond simple default assignments, using available tools (such as separating data fields like Year/Month/Date), so as not to impact interpretation and to enable better completeness, accuracy, and explicitness.

The following principles have been identified to make any genomic surveillance tool usable in a federated model, however this list is not exhaustive. These principles are distinct but interrelated and overlapping in nature and have inherent variability in their application. While the application of each of these principles is necessary in the development of a federated surveillance system, the manner in which they are incorporated can and shall vary in response to a system's purpose, circumstance, and the data type(s) and sensitivity (i.e. whether it contains personal health information).

5.2 Explicitness

Federated surveillance systems shall be explicit about the roles they play in data transactions as these roles can change over time and from one use case to another. Each participant in a federated network making a data request shall explicitly state the intent of the request just like each data provider needs to explicitly state the purposes for which the data has been collected. In developing genomic surveillance tools, it should be clear what is happening and how it is happening – for example, what data is being gathered and how it is being gathered – to ensure that the surveillance analysis is receiving the right data for the right purpose.

The utility of data shall be maximized, including accounting for the data's analytical completeness and validity, while minimizing disclosure risk. Data shall refer to explicit coding schemes and terminologies, so it is described in such a way that the data is defined before it is being processed. Federated systems shall also be explicit about what data processing, transformations, and manipulations are performed to prepare the data for analysis and to ensure that all data derivatives have explicit references to the original data sources. Unique dataset identification across nodes is another example of explicitness. Further, as data is distributed in a federated model, data access, management and retention policies shall be clearly accessible, interpretable and enforceable to ensure that each new network node that is onboarded can be checked and monitored for compliance thus maintaining the network's integrity.

5.3 Scalability

Federated surveillance systems shall consider the design of the model's framework to ensure that its structure has sufficient flexibility to adjust to changes and upgrades to a variety of components. Systems shall be able to increase capacity and functionality based on user demand, remaining stable while adapting to changes, upgrades, overhauls and resource reduction. Federated systems should balance performance and resource utilization, accounting for different volumes from a variety of clients, and be able to respond to changing needs, including the introduction of new servers or data.

Federated systems rely on a variety of data, including those in object stores and relational databases, requiring flexibility and scalability to be thoroughly considered. Distributed systems shall be built using reliable fault tolerance mechanisms to ensure continuity and reduce the risk of failure. For example, loose coupling can be used to minimize dependencies so modifications or faults in one system will have fewer consequences on other systems. Further, given the fast-paced nature of genomic surveillance, a federated system should be able to respond to rapid changes in resource availability, data type and structure, and provide the user with clear feedback.

To enable scalability across a network, designers should prioritize code that can be readily adapted to these changes while still being subject to amenable ongoing testing and maintenance.

5.4 Transparency

In a federated surveillance system with distributed data, the sources of information shall be transparent, facilitating reliability and trust across information systems. There shall be transparency in the origin of data elements (in a database, document, or repository), as well as clarity around cooperation and communication between systems, enabling data provenance and accession to be verified and facilitating reproducibility.

Transparency in connections between systems and data sources should enable visibility into compliance with applicable data protection laws and governance structure, including data sovereignty and residency. There are privacy considerations surrounding genomic and contextual data. As data is subject to the laws and governance structures within the jurisdiction in which it is collected and/or physically located, the legal rights of data subjects and data protection requirements are dependent upon the location in which data is stored and used, underscoring the need for transparency within the system.

5.5 Extensibility

Extensibility in a federated model is key to ensuring the long-term sustainability of federated surveillance system. Because software systems can be long lived, and both features and data sources will change over time, extensibility enables new or expanded software's capabilities and facilitates systematic reuse, without impairing existing system functions.

Genomic surveillance tools shall be built as a system of independent and connected modules with interfaces that are compliant with open standards, such as those outlined in Clause 6, to make it easy to add new modules.

5.6 Trust and cooperation

Organizations participating in a federated system shall establish trust, particularly as federated networks can be large-scale, loosely-coupled, heterogeneous distributed systems. In a federated surveillance system, collaboration and cooperation between data generators/data custodians and data consumers in different institutions is critical to sustained and predictable operations. Effective genomic surveillance requires collaboration regionally and internationally. As such, there is significant benefit from such cooperation, including building capacity and supporting equitable access in the regions and countries around the world, and including infrastructure, informatics, and data management, which further facilitates trust.

Many of the challenges related to trust in a federated system involve ongoing coordination and management of identities, attributes, roles, and privileges. Services in one node within a federated system necessarily depend on the terms of their trust relationships with other nodes when making security decisions. Expressed as policies, trust relationships instruct services on how to respond to a variety of security issues, including whether to fulfil a request or to permit the exchange of information from another node. Trust should be established through relationships between organizations and/or through the adoption of open services and standards that facilitate shared understanding and interoperability.

The level of trust can vary, but typically includes authentication and almost always authorization. A typical federation might include several organizations that have established trust for shared access to a set of resources and may include formal consortia or collaborations between governments. Ideally, each data sharing transaction should be governed by an explicit data use agreement, which is expressed in machine-interpretable format and is enforced by the participants in a verifiable and trustworthy manner. In certain cases, a federated network functions with public data that does not require all trust elements, however some level of cooperation between the nodes is still essential.

6 Service and standards requirements and recommendations

6.1 Overview

The application of each of these requirements is necessary in the development of a federated surveillance system, the manner in which they are incorporated can and must vary in response to a system's purpose, circumstance and the data type(s) and sensitivity (i.e. whether it contains personal health information).

6.2 Data representation

6.2.1 Data identifiers

Data identifiers shall include unique, discrete identification of data objects, resources, and provenance.

6.2.2 Platform/vendor agnostic data access and retrieval

Federated surveillance systems should use a platform or vendor agnostic data access and retrieval service.

One such example is Data Repository Service (DRS), a generic interface to data repositories so data consumers, including workflow systems, can access data in a single, standardized way regardless of where it's stored or how it's managed. DRS standardizes information about a file (e.g. name, size, URL), making it easier for data to be used.

6.2.3 Standard data models and formats

Data shall be represented, where available, using standard data models and formats suitable for the type of data being represented. Examples of such models and formats include:

— Phenopackets: a human and machine-readable way to structure phenotypic data about an individual, acting as a common model that can capture data from many sources. Phenopackets captures patient information in a structured format, which makes patient data more computable and shareable between clinical information systems and with related information systems. This information can then be shared across clinical and research environments or used for computational analyses.

— Fast Healthcare Interoperability Resources (FHIR): a representation for electronic health records that captures elements like patients, admissions, diagnostic reports and medications in a digital format so that they are more easily exchanged.

— Observational Medical Outcomes Partnership (OMOP) Common Data Model: a common data model for the purpose of systematic, standardized and large-scale analytics applied to clinical patient data.

— Variation Representation Specification (VRS): a specification for the exchange of genetic variation data.

6.3 Data discovery

6.3.1 Web interfaces for data search and discovery

Federated surveillance systems shall, where possible, use technology stack-agnostic web interfaces for data search and discovery. Examples of such web interfaces include:

— Data Connect: a simple, uniform mechanism to publish, discover, query and analyse any format of biomedical data. Data Connect empowers researchers to ask sophisticated discovery or analytical questions across distributed networks of different biomedical data types (e.g. genomics, clinical) and different storage systems (e.g. electronic health record, cloud storage).

- Beacon: a standard API that improves genomic data discoverability by enabling researchers to query genomic data collections, such as population based or disease-specific genome repositories. Beacon gives researchers the ability to discover whether a data collection contains information about a specific genetic mutation, and additional metadata, facilitating aggregate information about genomic datasets to be shared while maintaining privacy and access control.

6.3.2 Discoverability and networking web service

Federated surveillance systems should use a discoverability and networking web service. Two such examples are Service Info and Service Registry.

- Service Info: an endpoint for describing Global Alliance for Genomics and Health (GA4GH) service metadata, designed for extension and inclusion in other APIs. Service Info is used to describe a single service, while Service Registry is used to describe multiple services. It allows systems to discover basic information about a software service, like what kind of service it is, what version it implements, and who owns it. Service Info is a valuable feature for building networks of software systems.

- Service Registry: provides information about other services, primarily for the purpose of organizing services into networks or groups and for service discovery across organizational boundaries. It's a minimalistic, simple, lightweight, read-only API for listing services and their metadata, as described by service-info, making it easy to create and manage data sharing networks.

6.4 Data access

6.4.1 Researcher authorization

Federated surveillance systems shall use a widely supported digital identity to support trust within a federated network by enabling researcher verification and authorization to access data. Examples of such tools include:

- Passport: a digital identity that defines a standard way of communicating a researcher's data access authorization based on their role, affiliation, or access status. The specification aims to streamline the data access process and support researchers' ability to access and aggregate controlled data efficiently, and enables the automation of time- and resource-intensive data access reviews.

- Authentication and Authorization Infrastructure (AAI): profiles the OpenID Connect protocol to provide a federated (multilateral) authentication and authorisation infrastructure for greater interoperability between institutional systems in a manner specifically applicable to (but not limited to) the sharing of restricted datasets. The AAI makes it possible to enact sophisticated access policies that can be enforced by computers.

6.4.2 Data access decision making

Federated surveillance systems shall use tools to support data access decision making for datasets with usage restrictions. Examples include:

- Data Use Ontology (DUO): a dictionary of terms that describe how data can be used and how a researcher intends to use the data through additional terms that define intended research usage. DUO makes it possible to digitally encode consent terms so that data access decisions can be made by the AAI, allowing authenticated users to query and gain access to datasets pertaining to their research.

- Machine Readable Consent Guidance: provides instructions for researchers to integrate standard data sharing language into consent forms in a way that can be translated to a computable language, enabling researchers to search for datasets that have been consented to, for their research purposes.

6.5 Data analysis

6.5.1 Web interfaces for workflow execution and monitoring

Federated surveillance systems should use a web interface for workflow execution and monitoring. One such example is Workflow Execution Service (WES), an environment for running computational pipelines. WES makes it possible for data scientists and bioinformaticians to execute custom and best practices workflows in a way that is reproducible and highly scalable.

6.5.2 Registration and sharing of computational tools

Registries of computational tools should be used to support consistency in data processing to support analysis within federated surveillance systems. Examples of computational tool registries include:

- Dockstore: an open source platform for sharing reusable and scalable analytical tools and workflows;
- Tool Registry Service (TRS): a library of computational pipelines that can be executed by a WES. The TRS helps data scientists and bioinformaticians store versions of best practices and custom workflows and share them with collaborators.

6.5.3 Languages for writing reproducible workflows

Federated surveillance systems should use a specification for describing analysis workflows and tools in a way that makes them portable and scalable across a variety of software and hardware environments, from workstations to cluster, cloud and HPC environments. Examples include the following:

- Nextflow[3]: an open workflow management system that enables scalable and reproducible scientific workflows using software containers. It allows the adaptation of pipelines written in the most common scripting languages.
- Common Workflow Language (CWL): an open standard language for describing how to run command line tools and connect them to create workflows. Tools and workflows described using CWL are portable across a variety of platforms that support the CWL standards.
- Workflow Description Language (WDL): an open standard language for describing computational pipelines. The Workflow Description Language makes it possible for data scientists and bioinformaticians to define workflows that can be executed reproducibly in a variety of computational environments.
- Snakemake: an open workflow management system to create reproducible and scalable data analyses. Workflows are described via a human readable, Python based language. They can be seamlessly scaled to server, cluster, grid and cloud environments, without the need to modify the workflow definition.

7 Data linkage

7.1 Overview

Federated surveillance systems necessarily use pathogen genomic data and should be designed with these specific requirements in mind. While pathogen genomic data alone allow certain inferences (e.g. related cases and mutations leading to new variants), greater value can be obtained through combining and linking additional data sources like genomic or other ‘omics data, epidemiological and geographic data, clinical and demographic data including immunization or recovery data. As a result, federated networks can incorporate standards to facilitate sensitive data linkage to meet the intended objectives of the federated surveillance system.

[3] Nextflow is the registered trademark of Seqera Labs, SI. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of the product named.

7.2 Genomic and other ‘omics

Viral genome sequencing enables tracking of pathogenic mutations, transmission between hosts over time and across geographic regions and changes to the virus that can be more or less transmissible or lead to poorer health outcomes, or that can escape vaccines or therapeutics.

Host genome sequencing and other ‘omics data such as RNA-Seq facilitate improved understanding of the genetic reason why some individuals have a more severe response or poorer outcomes, leading to the development of biomarkers, diagnostics tools and therapeutics.

Wherever possible, read data (e.g. fastq) and associated metadata should be provided, enabling more comprehensive and accurate analysis, such as variant calling, when compared to assemblies.

7.3 Epidemiology

Information about the transmission context, including the geographic location and the reason for testing or sequencing (e.g. known contact or outbreak, travel, vaccine breakthrough infection) shall, where available, be provided to contextualize genomic data. Epidemiological data shall, where available, also include information about the source and location of exposure, such as the workplace, household, travel, community exposure, animal exposure, as well as any other contact investigation information (e.g. indoors vs outdoors, ventilation, community setting).

7.4 Medical records

Information from medical records, such as attributes of infected individuals as tracked by healthcare professionals, including vaccination (type, doses and dates) or past infection, treatments provided, outcomes such as symptoms and severity, and demographic aspects like age, comorbidities and exposure risks, shall be incorporated with genomic surveillance to facilitate deeper understanding of the impact of risk factors or interventions.