FINAL
DRAFT

TECHNICAL SPECIFICATION

ISO/DTS 8376

ISO/TC **215**/SC **1**

Secretariat: **KATS**

Voting begins on:
2023-09-15

Voting terminates on:
2023-11-10

Genomics informatics — Requirements for interoperable systems for genomic surveillance

Reference number
ISO/DTS 8376:2023(E)

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 before seen 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

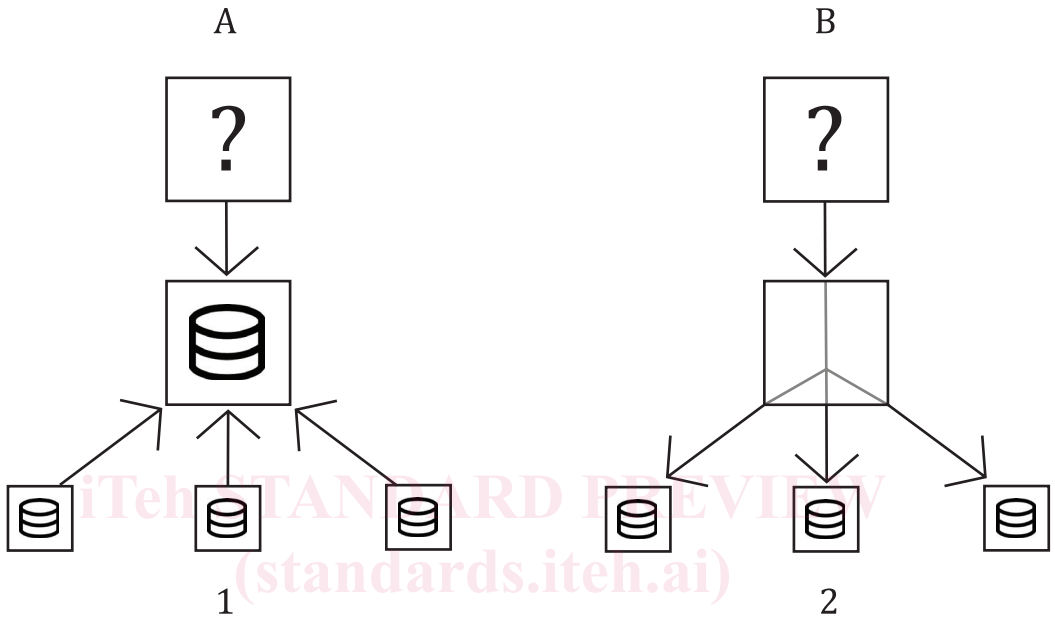
This document is intended to highlight the requirements for interoperability between networks, systems, and platforms required to share data within and between organizations and countries. It focuses on the technical, structural, and syntactic levels of interoperability, while acknowledging the importance 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.

---

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.

Further, data or insights facilitated by such tools can be consumed by individuals through mainstream media or research to understand the personal impact.

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 a number of 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(en), 3.2, modified — added "independent" and "distributed" to the definition]