GA4GH Data Sharing Lexicon

Background

Over the last two decades, advances in genetic and genomic technologies and their application are helping enable advances in human health. However, these potential benefits risk being undermined by challenges to genomic data sharing, which result in data remaining siloed within jurisdictions and disciplines. The GA4GH Data Sharing Lexicon has been developed in order to promote data sharing by facilitating a shared understanding of the terms which underpin the practice of data sharing for health and research purposes.This lexicon is intended to be used both by members of the Global Alliance for Genomics and Health (GA4GH) and researchers and institutions worldwide to promote an international community dedicated to advancing human health through genomic data.

The standards and policies developed by GA4GH are grounded in the Universal Declaration of Human Rights and aim to expand genomic data use within a human rights framework, highlighting that these technologies should be developed and used in ways that are equitable and responsible. Like other data sharing outputs, this work also builds on the Framework for Responsible Sharing of Genomic and Health Related-Data to accelerate progress in human health by establishing a shared global approach to responsible, broad and democratised use of genomic and related health data.

Purpose and scope

This revised GA4GH Data Sharing Lexicon supports international data sharing by promoting common/concordant terms within the GA4GH and across jurisdictions, disciplines and individual providers of health care and research. The scope of this Lexicon includes terms commonly encountered when sharing data for health or research use, but is not intended as a substitute for clinical, scientific or technical literature. Since the Lexicon is intended to be as generalisable as possible across jurisdictions, the terms also do not replicate legal/regulatory definitions in legal sources.

With these caveats in mind, and given that many of these terms are contested within and across disciplines, the definitions included in this Lexicon have been selected because they are accessible and succinct. However, they may not be appropriate for every context of data sharing.

Process

Where relevant, this lexicon builds on terminology used in GA4GH documents and policies, the most important of which are:

- The Framework for Responsible Sharing of Genomic and Health-Related Data
- Consent Policy v2
- Privacy and Security Policy v2
- Accountability Policy

It also incorporates terms already in use in existing GA4GH glossaries where these are relevant and appropriate.

Note to Users of this Lexicon

During the development of this Lexicon, we noted that some terms are related to one another either because they are used interchangeably when in fact their meanings are subtly different (e.g. cluster of terms relating to Data Identification), or because they describe sub categories of an overarching theme/category (e.g. terms describing different types of Consent decision or Technology). To guide the user we have grouped these related terms into clusters, which are colour coded for ease of reference:

- Consent
- Data identification
- Technology

Acknowledgements

The first version of the GA4GH Data Sharing Lexicon was developed by a working group, representing a wide spectrum of the law, security, bioethics, and genomics communities under the auspices of GA4GH's Regulatory and Ethics Work Stream. This revised version has been further developed by a small working group, and ratified by the larger GA4GH Regulatory and Ethics Workstream. We are grateful to members of the internal GA4GH group and the external expert group for their time and expertise.

- Internal group: Professor Yann Joly; Professor Dianne Nicol; Dr Diya Uberoi, Gemma Brown and Alison Hall
- External Expert Group: Professor Ainsley Newson for facilitating collaboration from the LINEAGE Project,[1] Jaime Guidry Auvil, David Bujold and Professor Heidi Beate Bentzen

[1] Funding Acknowledgements: "Ethical governance for clinical and genomic data" now known as LINEAGE: Law, Sociology and Ethics in Data Governance for Genomics, funded by the Commonwealth of Australia, Medical Research Future Fund, 2021 Genomics Health Futures, Mission Grant, grant MRF2015531.

Data Sharing Lexicon

Term (A-Z)	Definition
AI	The use of programming tools to enable machines to perform tasks that previously required human (or other biological) brainpower to accomplish.
Accountability	The obligation to explain and justify conduct.
Aggregate Data	Collated results of genomic data from many individuals included within a specific dataset.
Anonymization	The irreversible removal or alteration of identifiers from human individual, population data and associated metadata.
Anonymized Data	Data that have had personal identifiers irreversibly altered or removed so that individuals are unlikely to be identified, even with additional information.
Anonymous Data	Data that are not identifiable when collected.
Attribution	Identifying and/or acknowledging the source or author of an idea or written material.
Audit	A systematic review of practice to evaluate adherence to applicable laws and policies.
Big Data	Large and complex datasets typically combining multiple sources of information that can be analyzed through novel computational methods.
Biobank	An organized collection of human biological material and associated data which is stored, processed and searchable.
CARE Principles	An acronym summarising principles of data management for indigenous data governance including to promote Collective benefit, Authority to control, Responsibility and Ethics.
Clinical Test	A test to predict future disease, make a diagnosis, explain symptoms and guide clinical care or management.
Clinical Trial	A study involving human participants to investigate the efficacy and/or safety of one or more medicines, treatments or other therapeutic interventions.
Clinically Significant Finding	A finding that has the potential to improve [an individual's] medical prognosis, care or outcomes.
Cloud Computing	Distributed computing resources accessible over the internet that can include capabilities for processing and storing data.
Coding/ Pseudonymization	The act of replacing an identifier with a code for the purpose of avoiding direct identification of the participant, except by persons holding the key linking the code and identifier.
Cohort	A group of individuals identified by common characteristic(s) (e.g. demographic, exposure, illness).
Confidentiality	The ethical and legal obligation of an individual or organization to safeguard data by restricting access as authorized by law or by the data donor.

Conflict of Interest	One or more connections or interests (personal, social, financial or professional) that influence, or could be perceived to influence, professional integrity and independence.
Consent	Voluntary and informed expression of the will of a person or group, or their legal representative.*
Controlled/ Managed/ Restricted Access	Access to data that is subject to conditions and/or an approval process.
Conversational AI	A type of artificial intelligence that uses machine learning and other technologies to simulate human conversation.
Cross-border Data Transfer	The transfer of data between or across two or more independent governing jurisdictions.
Data	Facts, results or observations from which information could be derived.
Data Access Committee	A select group of individuals that review applications requesting access to sensitive data and authorize use in alignment with defined criteria.
Data Breach	The unauthorized collection, access, use, disclosure or release of data.
Data Curation	The process of selecting, annotating, maintaining, archiving and tracking data and any supporting files, objects, identifiers or metadata.
Data Custodian	An individual who is responsible for the technical environment and database structure.
Data Destruction	To take all necessary steps to ensure that data are no longer stored or able to be used.
Data Donor/Provider	An individual whose data have been collected, held, used or shared for research purposes with their consent.
Data Embargo	Defined period of time when data are unavailable for wider access.
Data Harmonization	The process of unifying or aligning certain policies, methodologies and approaches in order to achieve interoperability.
Data Linkage	The process by which records representing the same entity or individual are connected across two or more sources of data.
Data Localization	Legal/ ethical rules that aim to keep data within a jurisdiction that governs such data and efforts to comply with such rules.
Data Privacy	Maintaining data in a manner that ensures security of data files to prevent identification of data donors and allow their use in accordance with given consent.
Data Privacy Impact Assessment	A process designed to evaluate and minimize data privacy risks.
Data Protection Officer	An expert in data protection within an organization who ensures that an organization applies relevant laws, regulations, and guidelines protecting individuals' data.
Data Protections	The set of laws, policies, procedures and technologies that minimize intrusion into the privacy of individuals and associated populations, uphold confidentiality, maintain data security and penalize undue intrusions and/or breaches.
Data Security	The protection of the confidentiality, availability and integrity of data.

Data Sharing	Providing access to data to another agency, organization or person for an agreed purpose.
Data Sovereignty	The right of peoples to exercise ownership over data deriving from them, or pertaining to their members, knowledge systems, customs or territories, which is expressed through the creation, collection, access, analysis, interpretation, management, dissemination and reuse of those data.
Data Steward	An entity responsible for assuring the quality, integrity, use and access arrangements of data and metadata in a manner that is consistent with applicable law, institutional policy, individual permissions and the public good.
Data Subject	An individual, population or cohort whose data has been collected, generated, held, used or shared.
Data Transfer Agreement	An agreement between the provider and the recipient of data that describes conditions of transfer, use and disclosure.
Data Use	Use includes analysis, cleaning, reformatting, combining, or annotating the data or using the data for model training.
Data User	Individuals or organisations who are authorised to access or use data for an agreed purpose.
Data Visiting	A form of data sharing in which a user performs analysis on shared data within the provider's computing environment, whether through human or computational agents.
Database	A collection of data and information that are managed and stored in a systematic way to enable data analyses.
Dataset	A collection of data which may be part of a database.
De-identification	The removal or alteration of any direct identifiers that identify an individual or could, foreseeably, identify an individual in the future.
De-identified Data	Information that has had direct identifiers removed.
Disclosure Risk	The chance of private or confidential information being revealed about an individual.
Dynamic consent	A technology platform to enable ongoing engagement by participants.
Encryption	A mechanism of safeguarding stored data or information by making those data or information unreadable without access to the correct decryption method.
Ethical Guidelines	Guidance on decision-making based on accepted ethical principles and practice.
Ethics Review Committee/ IRB/ REC/ REB/ HREC	An independent committee for the ethical review and approval of research activities.
Exome Sequencing	Testing all the protein-coding regions of a person's genome.
FAIR Principles	An acronym summarising foundational principles of good data management that promote Findability of data, data Access, Interoperability and Reuse.
Federated Data Analysis	A form of data visiting that entails analysis across datasets from multiple providers without the data leaving each provider's respective computing environments.
GA4GH	The Global Alliance for Genomics and Health.

GA4GH Framework	The GA4GH *Framework for Responsible Sharing of Genomic and Health-Related Data* .
Gene	A portion of an individual's DNA.
Gene panels	Simultaneous testing of all genes associated with a particular set of clinical symptoms (phenotype) or disease.
Generative AI	The use of programming tools to generate new content in response to user prompts, that is indistinguishable from human-created data.
Genetic Discrimination	An individual or a group being negatively treated, unfairly profiled, or harmed, relative to the rest of the population, on the basis of actual or presumed genetic characteristics.
Genetic Test	A targeted test that examines a gene.
Genomic Test	A test that explores multiple genes such as gene panels, exome sequencing, and genome sequencing.
Genotype	The genetic make-up of an individual.
Governance	The process of policy making and management that guides and oversees health care and research in an ethically and legally compliant, consistent and structured manner.
Health Data	Data relating to the physical or mental wellbeing of an individual.
Identifiable Data	Data that alone or in combination with other data may reasonably be expected to identify an individual.
Identity and Access Management	A set of processes and supporting technologies that enable the creation, maintenance, use, and revocation of digital identity, often based on role.
Incidental Findings	A finding discovered in the course of clinical care or research concerning an individual, that is beyond the aims of the clinical care or research.
Individual Research Results	A finding discovered in the course of research concerning an individual that relates to the aims of the research.
Informed Consent	Consent given by a capable data subject (or legal representative) who has sufficient time, information and support to make an informed decision.
Interoperability	The technologies, processes and standards that allow data to be shared, accessed and used by multiple users and systems.
Key	A piece of data that an encryption algorithm uses to determine how to unscramble pseudonymized data.
Legacy Data or Biospecimens	Data (or biospecimens) previously collected for research or for clinical care, where new proposed uses may not be documented in a consent form.
Long-read Sequencing	A form of next generation sequencing in which long strands of DNA or RNA are sequenced without being broken up into smaller pieces.
Machine Learning	A form of AI enabling computers to learn from data and make decisions or predictions without being explicitly programmed to do so.
Material Transfer Agreement	A binding legal agreement between the provider and the recipient of materials that describes conditions of transfer, use and disclosure.
Medical/Health Record	A paper or electronic record created in the health care system which contains medical and health-related information about an individual and is used to record and support health care for that individual.

Metadata	Data about stored data including its properties, history, and versions.
Newborn Screening	Tests of newborn babies to detect and/or identify potential serious health conditions usually in the absence of clinical symptoms, to inform treatment or management options.
Newborn Testing	Diagnostic tests of newborn babies to detect signs and symptoms of serious health conditions.
Next Generation Sequencing	High throughput technologies which use sequencing to identify the order of nucleotides in entire genomes or targeted regions of DNA or RNA.
Open Data	Data that is available and accessible without restriction.
Opt-in consent	A consent mechanism where an active choice is required to participate.
Opt-out consent	A consent mechanism where consent is implied unless an active choice is expressed not to participate.
Organizational members	The organizations that are members of the GA4GH.
Personal Data	Data that alone or in combination with other data may identify an individual without requiring unreasonable identification efforts.
Personalized Medicine	Using an individual's genetic profile and other factors to guide and enable tailored/targeted medical decisions and interventions.
Phenotype	An individual's observable traits, which are determined by both genomic makeup (genotype) and environmental factors.
Polygenic Risk Score (PRS/PGS)	Estimating a person's inherited risk of a specific disease by analysing millions of common genetic variants associated with that condition, together with other genetic risk factors e.g. genetic ancestry.
Precision Medicine	Using a person's genes, environment and lifestyle to predict which intervention or treatment will work best for a particular patient.
Primary Data	Data that is generated or collected for a specific purpose or use.
Privacy	The right and freedom to control sensitive information about oneself.
Processing	Any operation on data, including collection, storage, use, dissemination and destruction.
Proportionate/Proportionality	Determining the justification for an action by balancing the objective, means, methods and potential consequences (including the risks and benefits) of that action.
Prospective data collection	Data that is yet to be collected.
Provenance metadata	Detailed information that documents the origin, history, and lineage of data, providing a comprehensive record of the data's lifecycle, including where it originated, how it has been processed or transformed, and where it has been transferred.
Pseudonymization/Coding	The act of replacing an identifier with a code for the purpose of avoiding direct identification of the participant, except by persons holding the key linking the code and identifier.
Pseudonymized Data	Data that may not be readily attributed to a specific data subject without the use of additional information, technical expertise or resource.

Public Engagement	An inclusive act involving a group of participants, ranging from active involvement in the development, management or governance of a project, to the provision of information and raising awareness of a project.
Quality	Assurance that data, biospecimens or processes meet specified standards.
Registered Access	Providing access to data contingent on a system of authentication and self-declaration.
Re-Identification	The ability to trace data back to an individual.
Remote Data Interrogation	A restricted type of data visiting where the user can query but not view raw data from the provider.
Replicability	The act of checking the consistency of results across studies which aim to answer the same scientific question using different datasets or different sources for the same data.
Retrospectively Collected Data	Data already collected for another purpose which may or may not include data consented for research.
Return of Results	Communication of research data or outcomes to an individual, designated health care provider, family member, or authorized representative.
Risk	The probability that an event, favorable or adverse, will occur within a defined time interval.
Safe Haven	A repository in which data are stored and accessed in ways that maintain their integrity and quality whilst meeting relevant ethical and legal controls on their use and dissemination.
Secondary Data	Data derived from alternate analysis of primary data.
Secondary Findings	Deliberately sought test results that are unrelated to the primary purpose for testing.
Secondary Use	Existing data which is used for an alternate purpose or by an alternate user following its primary use, such as clinical data being used for research.
Security	The process of protecting data from unauthorized access, use, change, disclosure, corruption or destruction.
Security Risk Assessment	An objective analysis of the effectiveness of the security controls that protect data.
Structured Data	Data that is organised and understandable/ interpretable by humans and machines.
Supervisory Authority	The public authority responsible for monitoring data privacy, data protection and data security laws within a given jurisdiction.
Surveillance	The systematic collection, monitoring and dissemination of health data to assist in planning, implementation and evaluation of a public health or health research intervention.
Sustainability	The ability to maintain or support a process over time through addressing economic, environmental and social considerations.
Synthetic Data	Artificially generated data created by algorithms or computer simulations which mimics real-world data.
Traceability	Verification of an item's history, location or application through documented records.

Training Data	Data used to train computers/ machine learning models to perform a specified task.
Transparency	The quality of actions being understood and accountable through data that have been made detectable and recognizable.
Trusted Research Environments	Highly secure and controlled computing environments that facilitate remote access to specified data.
Trusted Third Party	An individual or organization that safeguards access to information linking individuals to their data and biospecimens.
Trustworthy	Earning the trust of others through demonstrating integrity, honesty, competence and reliability.
Unstructured Data	Data lacking predefined structure or organisation which are not readily interpretable by computers or humans.
Validation	Evaluating the performance of a hypothesis, model, test or research finding using data/interventions not included in training or research.
Variants of Unknown Significance (VUS)	A genetic variant which has uncertain or unclear links to a health condition including where insufficient or contradictory evidence of a link exists.
Vulnerable Persons / Populations	Individuals or groups requiring special consideration, safeguards or protection because of the likelihood of them being denied individual, social or communal human rights.
Whole Genome Sequencing (WGS)	Form of next generation sequencing that reveals the complete DNA make-up of an organism, enabling better understanding of variations within and between species.
Withdrawal	The right for an individual to have data about them removed or destroyed or prevented from being used for research or clinical care.

*see consent tool kit, where there are multiple ways of describing consent.

Clusters of related terms:

Consent

- Dynamic Consent
- Informed consent
- Opt-in consent
- Opt-out consent

Data identification

- Anonymization
- Anonymized data
- Anonymous data
- Coding/Pseudonymization
- Data privacy
- De-identification

- De-identified data
- Identifiable data
- Identify and Access Management
- Pseudonymization/Coding
- Pseudonymized data
- Reidentification

Technology

- AI
- Big Data
- Cloud Computing
- Coding/Pseudonymization
- Conversational AI
- Exome Sequencing
- Generative AI
- Long-read Sequencing
- Machine Learning
- Next Generation Sequencing
- Polygenic Risk Score
- Whole Genome Sequencing