Version 1.0, 15 March 2016

Global Alliance for Genomics & Health
Data Sharing Lexicon

Preamble

The Global Alliance for Genomics and Health ("GA4GH") is an international, non-profit coalition of individuals and organizations working in healthcare, research, disease advocacy, life science, and information technology dedicated to improving human health by maximizing the potential of genomic medicine through effective and responsible data sharing. The sharing of genomic and health-related data for biomedical research is of key importance in ensuring continued progress in our understanding of human health and wellbeing. The challenges raised by international, collaborative research require a principled but nevertheless practical framework that brings together regulators, funders, patient groups, information technologists, industry, publishers, and research consortia to share principles about data exchange.

GA4GH's mission is to "accelerate progress in human health by helping to establish a common framework of harmonized approaches to enable effective and responsible sharing of genomic and clinical data, and by catalysing data sharing projects that drive and demonstrate the value of data sharing."

Sharing of genomic and health data is increasingly international, but must contend with discrepancies in the terms employed by applicable laws, ethics policies and regulatory systems. The **purpose** of this **GA4GH Data Sharing Lexicon** is to support international data sharing by promoting common/concordant terms within the GA4GH and across jurisdictions and research contexts with the ultimate aim of improving human health.

This **Data Sharing Lexicon** is a tool developed to promote data-sharing in general. It is not intended to focus on biobanking or the transfer or processing of tissues or samples. The terms included are predominantly derived from legal/regulatory sources but modified so that they are applicable to many jurisdictions. These terms are not intended to replicate definitions found in particular statutes, or within scientific and technical literatures. Additional resources can be found in the attached bibliography.

Where relevant, the Data Sharing 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
- Privacy and Security Policy
- Accountability Policy

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

Acknowledgements

This Policy is the result of the work of many people and committees. Developed under the auspices of the GA4GH's Regulatory and Ethics Working Group, the Policy was formulated by an international committee (Data Sharing Lexicon Task Team) representing a wide spectrum of the law, security, bioethics, genomics, and life science industry communities. Collaborative input was provided from individuals as well as biomedical, patient advocacy, and ethical, policy and legal organizations, committees, and projects from all regions of the world (for contributors, see the Data Sharing Lexicon home page).

Data Sharing Lexicon

Term	Definition
Accountability	The obligation to explain and justify conduct.
Anonymisation	The irreversible delinking of identifying information from associated data.
Audit	A systematic review to evaluate adherence to applicable laws and policies.
Big Data	Large and complex datasets typically combining multiple sources of information and analyzed through novel computational methods.
Biobank	An organized collection of human biological material and associated data which is stored, processed and searchable.
Clinical Trial	A study involving human participants to investigate the efficacy and/or safety of one or more medicines or other health-related interventions.
Cloud Computing	Shared computing resources accessible over the internet that can include capabilities for processing and storing data.
Coding/ Pseudonymisation	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) or studied over time using a common protocol.

Confidentiality	The ethical and legal obligation of an individual or organization to safeguard data or information by controlling 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 if incompetent, his/her legal representative.
Controlled/ Restricted Access	Access to data that is subject to conditions and an approval process.
Data	Observations, narratives or measurements that are assumed as the basis for further analysis, calculation or reasoning.
Data Access Committee	A committee that reviews and authorizes applications for data access and use.
Data Breach	The unauthorized collection, access, use, disclosure or release of data.
Data Curation	The process of selecting, annotating, maintaining, archiving and tracking data.
Data Donor	The individual whose data have been collected, held, used and shared.
Data Embargo	Defined period of time when data are unavailable for wider access.
Data Linkage	The process by which records representing the same entity or individual are linked across multiple data sources.
Data Protections	The set of laws, policies and procedures that aim to minimize intrusion into people's privacy, uphold confidentiality, and penalize undue intrusions and/or breaches.

Data Security	The protection of the confidentiality, availability and integrity of data.
Data Sharing	Extending access to data for the purpose of research or analyses.
Data Steward	An entity responsible for assuring the quality, integrity, and access arrangements of data and metadata in a manner that is consistent with applicable law, institutional policy, and individual permissions.
Data (or Material) Transfer Agreement	A binding legal agreement between the provider and the recipient of data (or materials) that sets forth conditions of transfer, use and disclosure.
Database	Data and information that are managed and stored in a systematic way to enable data analyses.
Dataset	A collection of data which may be a subset in a database.
De-identification	The removal or alteration of any data that identifies an individual or could, foreseeably, identify an individual in the future.
Destroy	To take all necessary steps to ensure that data are no longer stored or able to be used.
Disclosure	The revelation of confidential information about an individual.
Disclosure Risk	The probability of confidential information being revealed about an individual.
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	A framework to guide decision-making based on accepted ethical principles and practice.
Ethics review committee/ IRB/ REC/REB	An independent committee for the ethical review of research activities.

Governance	The process of policy making and management that guides and oversees research in a consistent and structured manner.
Harmonization	The process of unifying certain policies, methodologies and approaches in order to achieve interoperability.
Identifiable/ Personal 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.
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.
Information	Data that have already been interpreted, i.e. they have meaning in a specific context.
Legacy Data or Biospecimens	Data (or biospecimens) previously collected for research or for clinical care, where new proposed uses may not be covered.
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 that describe other data.
Open Data Access	Making data available without restriction.
Opt-in	A consent mechanism where an active choice is made to participate.
Opt-out	A consent mechanism where consent is implied unless an active choice is made not to participate.

Personal Data/ Identifiable Data	Data that alone or in combination with other data may reasonably be expected to identify an individual.
Privacy	The right and freedom to control access to information about oneself.
Pseudonymisation/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.
Public Domain	The body of knowledge and innovation in relation to which no person or other legal entity can establish or maintain proprietary interests.
Public Engagement	An inclusive act ranging from the active involvement of a population or sub-population in the development, management or governance of a project, to the provision of information and raising awareness of a project.
Quality	Conformity of data, biospecimens or processes with pre-established specifications appropriate to the purpose to which the data, biospecimens or processes will be put.
Registered Access	A system of authentication and self-declaration prior to providing access to data.
Re-Identification	The act of associating specific data or information within a dataset with an individual.
Return of Results	Communication of research results to an individual or a designated health care provider or family member.
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 uses	Using data or biospecimens in a way that differs from the original purpose for which they were generated or collected.

Supervisory Authority	The public authority (or authorities) in a given jurisdiction responsible for monitoring the application of law and administrative measures adopted pursuant to data privacy, data protection and data security.
Surveillance	The systematic collection, monitoring and dissemination of health data to assist in planning, implementation and evaluation of an action or intervention such as research or public health.
Traceability	The ability to verify the history, location, or application of an item, by means of documented recorded identification. For a biospecimen this pertains to any step of its handling, including donation, collection, processing, testing, storage, and disposition.
Trusted Third Party	An individual or organization that safeguards access to information linking individuals to their data and biospecimens.
Vulnerable Persons / Populations	Individuals or groups requiring special considerations and/or under the protection of governments, institutions or legal representatives including but not limited to children, the elderly, and those with mental health issues.

Appendix 1: Table of Concordance of Data Privacy and Security Terms[1]

(from GA4GH Privacy and Security Policy)

Spectrum of Identifiability

1 (Most identifiable)	2	3	4 (Least identifiable)
"Is or can be fully identifiable to everyone"	"Is unidentifiable to most, but remains re-identifiable to those with access to the key(s)"	"Is likely no longer identifiable to anyone"	"Never was identifiable"
• identified or identifiable • personal • nominative	• coded • key-coded pseudonymized • reversibly de-identified • linked anonymized • masked • encrypted	• anonymized • de-identified pseudonymized • irreversibly de-identified • non-identifiable • unidentifiable • unlinked anonymized	• anonymous[6]

Category 1: Identified/personal/nominative data are labelled with personal identifiers such as name or identification numbers. Data are directly traceable back to the Data Donor.

Category 2: Pseudonymization/coding/key-coding consists of replacing one attribute (typically a unique attribute) in a record by another. Pseudonymized/coded/key-coded data are labelled with at least one specific code and do not carry any personal identifiers, but an individual is still likely to be identified indirectly; accordingly, pseudonymization/coding/key-coding when used alone will not result in an anonymous data. Pseudonymization/coding/key-coding reduces the linkability of a dataset with the original identity of a Data Donor; as such, it is a useful security measure in genomic research, but it is *not* a method of anonymization.

Category 3: Anonymization is intended to prevent re-identification. Data must be processed in such a way that it can no longer be used to identify a Data Donor by using all the means likely reasonably to be used by person or entity. An important factor is that the processing *must* be irreversible to reasonable degree, i.e., anonymized data must not be traceable back to the Data Donor.

Category 4: Anonymous data are never labelled with personal identifiers when originally collected, nor is a coding key generated. Therefore, there is no potential to trace back Data to individual Data Donors. Anonymous data are of extremely limited utility in genomic research.

[1] Adapted from William W. Lowrance, *Learning from Experience Privacy and the Secondary Use of Data in Health Research* (London: Nuffield Trust, 2002) at 34; ICH, *Guidance for Industry: E15 Definitions for Genomic Biomarkers, Pharmacogenomics, Pharmacogenetics, Genomic Data and Sample Coding Categories* (April 2008); Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques; Knoppers BM, Saginur M. The Babel of genetic data terminology. Nature Biotechnology 2005; 23(8): 925-927.

Appendix 2: Additional Resources

- GA4GH, Privacy and Security Policy (2015), Appendix 1 "Glossary" and 2 "Table of Concordance of Data Privacy and Security Terms" [link]
- P3G, Biobank Lexicon [link]
- Presidential Commission for the Study of Bioethical Issues, Privacy and Progress in Whole Genome Sequencing (2012) "Appendix I: Glossary of Key Terms" [link]
- World Health Organisation, Standards and Operational Guidance for Ethics Review of Health-Related Research with Human Participants (2011) "Glossary" [link]
- BBMRI, Lexicon [link]
- Nuffield Council of Bioethics, The Collection, Linking and Use of Data in Biomedical Research and Health Care: Ethical Issues (2015) [link]
- OECD, Health Data Governance (2015) [link]
- Administrative Data Research Network (ADRN), Glossary, Definitions of Terms used in the Network (2015) [link]
- OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (2013) [link]
- Elliot M et al, Glossary on Statistical Disclosure Control (2009) [link]