GA4GH Machine-Readable Consent Guidance

How to Map Data Sharing Consent Language to the GA4GH Data Use Ontology

Current Challenge: When researchers draft a consent form and information sheet for a research project, they frequently start by obtaining a template provided by their Institutional Review Board (IRB) (also known as a Research Ethics Committee, Human Research Ethics Committee or Research Ethics Board) or other authority. While health research consent form templates are common, they do not always provide clear information about downstream data sharing; that is, they often do not say how the data generated during the study may be made available after the initial study to the broader research community for re-use, perhaps through repositories that were not involved in the drafting of the original consent form. Even where a detailed data sharing plan is provided, IRBs and participants may doubt that limits or conditions on data sharing will continue to be respected as data are made accessible to researchers around the world.

Standard Consent Language: Incorporating standard data sharing language into consent forms can benefit both participants and researchers. This includes standard descriptions of accessibility and use terms, and aims to ensure participants receive sufficient and clear information about how their data may be processed, shared, and re-used. Standard data sharing language also makes it easier for researchers and oversight bodies (such as data access committees or IRBs) to determine when and how data may be shared and re-used while respecting legal and ethical obligations towards participants. Standard clauses make data sharing more coherent and predictable, which is to everyone's benefit.

Machine-readable Consents: Another advantage of standard consent language is that it can be easily translated into a machine-readable format, i.e., language that computers can understand. Machine-readable consent language can be permanently attached to datasets as part of their metadata (descriptive data). Machine-readability of consents enables the research community to introduce software tools that:

1. Allow researchers to discover datasets for which consent has been provided their proposed use;
2. Allow data access committees to confirm if data access requests "match" the consent conditions associated with one or multiple datasets;
3. Reduce administrative burden by assisting researchers and IRBs to accurately interpret existing consent forms to determine if they permit data access and re-use; and
4. Reduce the risk of error or inconsistency resulting when IRBs or researchers misinterpret consent forms.

The GA4GH Data Use Ontology: An international standard for machine-readable consent in health research is the Global Alliance for Genomics and Health's (GA4GH) Data Use Ontology (DUO). The DUO is a structured, controlled vocabulary of data use terms that describe the scope of permitted research purposes for using a scientific resource (e.g., a dataset). The data use terms are based on reviews of international regulations and ethical guidelines, empirical studies of data sharing practices, and consensus deliberations. Scientific resources can be permanently tagged with DUO identifiers, so data use terms are preserved over time. The meaning of each DUO term is also stable and linked to a persistent webpage with a fixed description, definition, and examples of usage. The DUO was developed primarily, but not exclusively, to support the sharing of genomic and health-related data. It is the standard data

use vocabulary officially approved and maintained by the GA4GH, and has been implemented by the European Genome-Phenome Archive, the All Of Us Researcher Portal, NHLBI TOPMed, and the DUOS at the Broad Institute of MIT. Researchers drafting machine-readable consent forms should understand the meaning of the DUO data use terms and the basic relationships between them (illustrated in **Figure 1**). DUO terms are meant to be simple and easily understood by anyone. It is possible to select a combination of data use terms. Some combinations, however, would not make sense, such as General Research Use and Disease-Specific Use. The DUO terms may not map directly to specific legal categories in a particular jurisdiction.

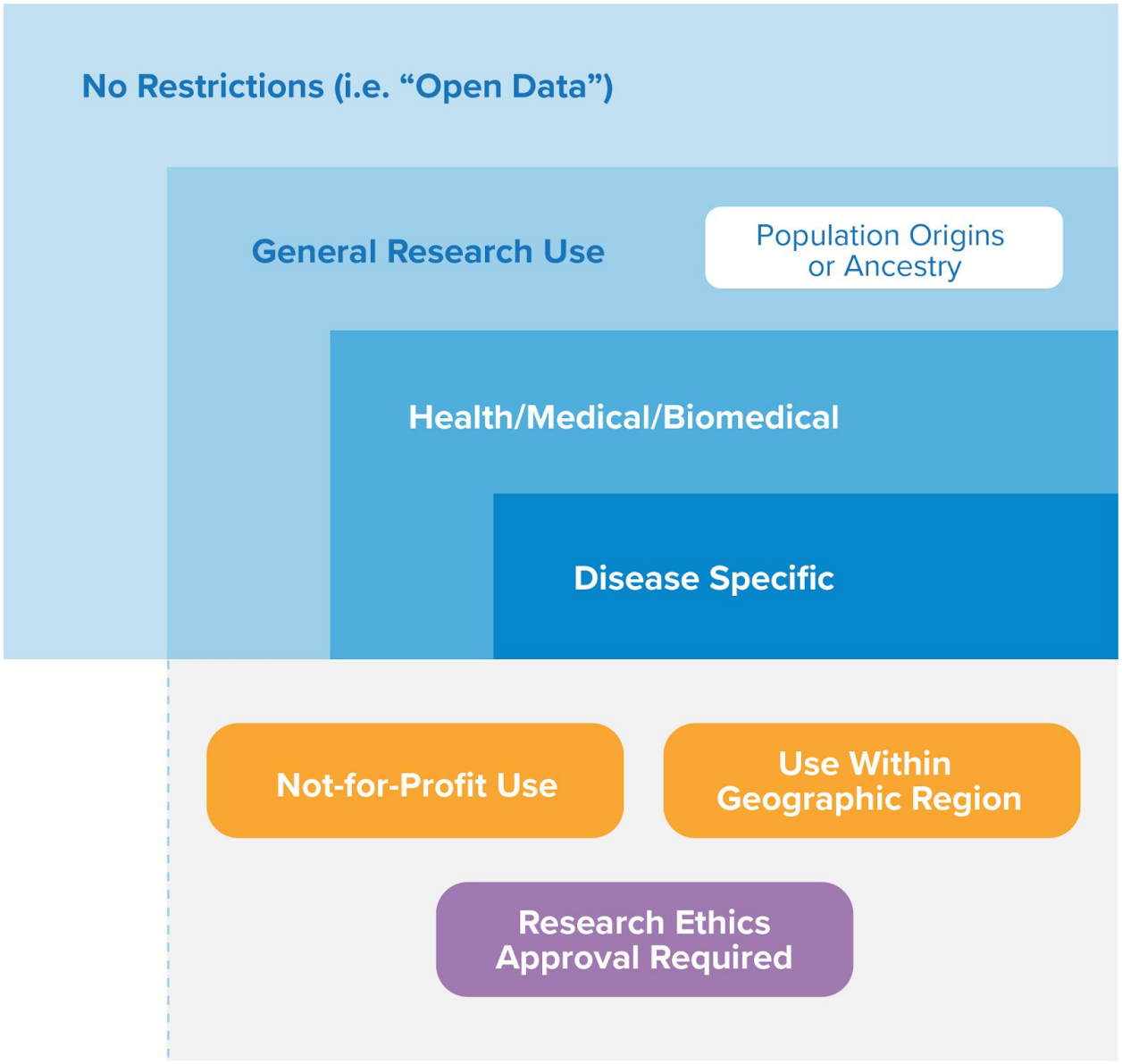
This guidance explains how to create a consent form that maps directly and unambiguously to the GA4GH DUO, which renders the consent machine-readable. We envisage a machine-readable consent form having three distinct elements:

1. A short consent clause providing a **summary** description of the data use term(s);
2. A detailed **explanation** of the meaning of the data use term(s), to ensure they are understood by the consented individual; and
3. An optional consent form **appendix** that unambiguously maps the consent language to specific data use term(s) in the GA4GH DUO.

Each consent clause summary and explanation below describes a data use term from the DUO. Data use terms are used to define the scope of permitted research purposes for which an individual's data may be shared and re-used. Consent forms must also include additional explanations to ensure participants are appropriately informed of the scope of permitted research purposes. Those drafting consent forms can select the terms most suitable to their regulatory and recruitment context. They also remain free to include data use terms not found in the DUO, although these terms will require review by humans.

Note that data sharing language in consent forms must address more than just the scope of permitted research purposes. To comply with applicable laws and to adequately inform individuals, consent forms also need to cover other information, such as what data types are being collected and shared, who may have access to the data, what security safeguards will be in place, etc. These additional aspects of data sharing consent are not addressed in this document. For more complete data sharing consent language, please refer to the **GA4GH Consent Clauses**. We encourage IRBs and policy makers who host and disseminate **consent form templates** to incorporate the consent clause summaries, explanations and accompanying guidance into their documentation.

Figure 1. Visual Summary of Relationships Between GA4GH Data Use Terms



Note that "No restrictions" encompasses "General Research Use", which in turn encompasses both "Health/Medical/Biomedical" and "Population Origins or Ancestry" Research. Technically, one could seek permission for both "Health/Medical/Biomedical" and "Population Origins or Ancestry" within a single consent form. Not-for-Profit Use Only and Use Within a Geographic Region (e.g., Country) can be added to limit a more general permission. Research Ethics Approval can be added as an additional requirement.

General Research Use (DUO:0000042)

Consent Form (CF) summary: "By participating in this study, I agree that my data may be used for different types of scientific research, including research related to non-disease traits."
Optional CF appendix: "This consent form represents the following GA4GH Data Use Ontology term: http://purl.obolibrary.org/obo/DUO_0000042."

This clause means participants' data may be used by the research team or other researchers for any type of scientific research purpose, which may not necessarily be health-related, including, but not limited to: clinical care, biological research, health/medical/biomedical research, socio-demographic, epidemiological and population and ancestry research. The exact nature of the research cannot be specified at the time of consent. This data will not be available for non-scientific research purposes, such as use to determine insurance eligibility. For illustration, general research purposes may include, but are not limited to:

- Research on any disease or condition, even if the research is on a disease very different from the disease being studied in the original project;
- Statistical methods research and development (e.g., development of software or algorithms) that may have applications to many different diseases or conditions;
- Research solely or primarily studying non-disease traits (e.g., intelligence; longevity; personality traits; socio-demographic); and
- Research relating to population structure, including research that involves the determination of allele frequencies in different populations or ancestries.

The scope of this category should be carefully explained to participants. Some types of research permitted under this category may pose risks of discrimination to communities or sub-populations, and may be objectionable to some participants. Furthermore, some jurisdictions may not permit consent for such broad research purposes. The risk that this broad term could be extended to objectionable research can be mitigated by limiting data access to qualified researchers, and by combining the term with Research Ethics Approval (see below).

Health/Medical/Biomedical Research Use (DUO:0000006)

CF summary clause: "By participating in this study, I agree that my data may be used for different types of health, medical, or biomedical research."
Optional CF appendix: "This consent form represents the following GA4GH Data Use Ontology term: http://purl.obolibrary.org/obo/DUO_0000006."

This means participants allow their data to be used for research on a variety of diseases and health conditions. This includes research where participants' data is used to understand how genetic changes affect the way a tissue, organ, or whole-body system works, but whose research purpose cannot be fully specified at the time of consent. Such research is often done in a hospital or research institute laboratory setting, but it can also be computer-based. These data would *not* be made available for research solely or primarily studying non-disease traits (e.g., intelligence; socio-demographic; personality traits), or population structure or ancestral origin. This term would not permit research with no clear relationship to understanding, preventing, diagnosing, or treating disease. The risk that this broad term could be extended to

objectionable research can be mitigated by limiting data access to qualified researchers and by combining it with Research Ethics Approval (see below).

Disease Specific Research Use (DUO:0000007)

CF summary clause: "By participating in this study, I agree that my data may be used only for research related to the following disease[s] [list name for disease(s)]."
Optional CF appendix: "This consent form represents the following GA4GH Data Use Ontology term: http://purl.obolibrary.org/obo/DUO_0000007. The name(s) for disease(s) are represented by the term(s) XXXXXXX selected from a standard disease ontology."

This means participants' data may be used to carry out research aiming to learn about, prevent, diagnose, or treat the disease(s) specified. In order to render this consent clause fully machine-readable, the disease name and use term must also be selected from a standard ontology, such as the Mondo Disease Ontology or the World Health Organization's International Classification of Diseases (ICD-11). NOTE: The GA4GH is still exploring how to capture disease categories in machine-readable format, at which point this guidance will be updated.

Population Origins or Ancestry Research Use (DUO:0000011)

CF summary clause: "By participating in this study, I agree that my data may be used for different types of population origins or ancestry research."
Optional CF appendix: "This consent form represents the following GA4GH Data Use Ontology term: DUO:0000011."

As a form of scientific research, population origins or ancestry research is encompassed within General Research Use. Nevertheless, it is considered separate from Health/Medical/Biomedical because research primarily concerned with population origins or ancestry research has a different risk profile (e.g., stigmatization of communities). This category may include the study of a genetic variation in a population or studying the traits of certain populations, such as the response to a medication in specific ethnic groups.

No Restriction on Use (i.e., Open Data) (DUO:0000004)

CF summary clause: "By participating in this study, I agree that my data may be placed in an open access database, meaning it can be accessed by anyone, including members of the public, and used for any purpose.
Optional CF appendix: "This consent form represents the following GA4GH Data Use Ontology term: http://purl.obolibrary.org/obo/DUO_0000004."

It is not common for rich individual-level data to be shared via open access, without restriction on use. Usually, only certain types of data are shared in this way (e.g., anonymized data), with appropriate permissions or consent provided. Some jurisdictions may prohibit or place strict limitations on such sharing. Researchers and IRBs should follow applicable laws (e.g., data

protection laws), as well as applicable data sharing and research ethics policies before proceeding with unrestricted data sharing.

Where unrestricted data sharing of rich individual-level data is allowed, participants need to be informed that their data may be used by anyone and for any purpose, and alerted to the risks that may be involved. This might include their full genomic data being uploaded to unrestricted online access repositories, known as "open datasets", to which any member of the public, not only researchers, might have access.

While this guidance document focuses on the sharing of rich individual-level genomic and health-related data, DUO categories can technically also be assigned to any type of dataset. For example, study results, aggregate statistics, and some variant data are typically published or otherwise shared in open access databases. Arguably, this knowledge should be free of restrictions on re-use. A DUO category of no restrictions could be applied to these types of data.

Research Ethics Approval Required (DUO:0000021)

CF summary clause: "My data will only be used for research projects approved by an Institutional Review Board."
Optional CF appendix: "This consent form represents the following GA4GH Data Use Ontology term: http://purl.obolibrary.org/obo/DUO_0000021."

The term "Institutional Review Board" may be replaced by a concordant term such as "Research Ethics Committee" or "Research Ethics Board". This consent clause should normally be combined with other data use terms that give permission to use data for a certain category of research (see Figure 1). It means that data can only be used (for the permitted research purposes) with the approval of such a committee. This condition would be satisfied by an approval by an appropriate research ethics committee in any country. Usually, a data access committee would be responsible for assessing the scope and conditions of the approval. It is a common regulatory requirement in many countries that permissions for broad consent to future research use of genomic and health-related data are subject to ongoing ethical oversight. There are exceptions, however. For example, this condition does not typically apply to the sharing of genomic and health-related data in the United States.

Not-for-Profit Use Only (DUO:0000018)

CF summary clause: "My data will only be used for non-commercial research purposes by non-profit organizations."
Optional CF appendix: "This consent form represents the following GA4GH Data Use Ontology term: http://purl.obolibrary.org/obo/DUO_0000018."

This data use term specifies that data cannot be used for commercial research purposes OR by for-profit organizations. This is therefore a highly restrictive term, as it could potentially prohibit data use for many research projects that have both a private and a public interest. There may be some particular research populations or communities that have low levels of trust in commercial parties, in which case introducing this term would be appropriate, otherwise the

participants or communities may withdraw permission for data sharing altogether. The DUO does not currently include terms that provide greater granularity, such as allowing for research without a direct commercial intent by a researcher at a for-profit organization.

Note that it is typically required that consent forms inform participants where data may be used for commercial research purposes or shared with commercial research organizations. In other words, consent forms should generally be explicit about the *absence* of commercial use restrictions, by using language such as the following: "*My data may be used for commercial or non-profit research purposes, including research by for-profit organizations."*

Use Within a Geographic Region (e.g., Country) (DUO: 0000022)

CF summary clause: "My data will only be used for research conducted within [name of specific countr-y/-ies or geographic region(s), e.g., European Union].
Optional CF appendix: "This consent form represents the following GA4GH Data Use Ontology term: http://purl.obolibrary.org/obo/DUO_0000022. The names(s) describing the countr-y/-ies or geographic region(s) represent(s) the ISO country code(s) ISO 3166-X:XX."

The GA4GH encourages international data sharing, and therefore cautions users about limiting data sharing to a particular country or region.This data use term is recommended for use only where necessary (e.g., legally required). This DUO term is mainly intended to enable data sharing within particular jurisdictions, namely countries or supra-national regions. The GA4GH is still in the process of identifying and endorsing a preferred geographic Ontology. The ISO 3166 standard – Codes for the representation of names of countries and their subdivisions - can be used in the interim, and we welcome suggestions if there is a better alternative.

Explanatory Notes

Defining the types of data that will be collected and shared
This guidance does not define the types of data referred to in the consent clauses above, because this will differ across contexts. It is the responsibility of those drafting consent forms to clearly define the types of data that will be collected and, subsequently, shared. The guidance generally envisages the sharing of rich genomic and health-related data collected or generated during participation in a research study, or as part of clinical care. This data may include information like gender, race, ethnicity, health status, vital signs, images (like x-rays or MRIs), family health history, medical records, or the results of tests or procedures.

Ensuring participants understand data use terms
The consent clauses provided above are only summary descriptions of data use terms, which define the scope of permitted research purposes for which data may be used. The clauses should not be considered complete descriptions of what is encompassed in the scope of a given data use term. Additional information needs to be included in consent form templates and other information to ensure participants understand the meaning of the consent clauses, based on the explanations provided above.

Addressing other aspects of data sharing consent beyond the scope of research purposes

The descriptions above only describe the scope of permitted research purposes for which data may be used. They do not address other key aspects of data sharing. For example, when there is a (even if very small) chance of re-identifying participants with this data, this risk should be explained to participants. Sharing and re-use of data may be subject to oversight, such as approval by a data access committee and/or IRB. Comprehensive data sharing consent language is available through the **GA4GH Consent Clauses**. The drafter must ensure that the additional data sharing explanations remain consistent with the clause or clauses selected from this guidance.

Adapting this guidance to different models of consent

This guidance also does not dictate a particular consent model or structure. A research project may adopt a uniform consent, or individuals can be permitted to choose their own data use terms, for example, through tick boxes or an electronic interface, although this may increase the complexity of capturing and managing data sharing consent conditions. In some cases, consent forms separate study participation from data sharing. In this case, these consent clauses may need to be modified to reflect this structure, for example by replacing “By participating in this study” with “By agreeing to share my genomic and health-related data”. With minor modifications, these template clauses could be adapted to health research beyond genomics, and to clinical contexts.

Optional Consent Appendix

When adding a machine-readable appendix, those drafting consent forms should ensure the appendix is accurate and consistent with the rest of the consent form. IRBs may be concerned that the appendix increases the length of documentation, risks creating inconsistencies, and may confuse participants. This can be addressed through careful design to ensure the appendix does not distract from the other consent language.

Including data use terms not found in DUO

Ideally, if a combination of data use terms can be selected that exhaustively describes the scope of permitted research purposes, the consent form will be fully machine-readable. The inclusion of additional data use terms is permissible, but a disclaimer in the appendix is then required that clarifies that the DUO terms selected are illustrative, but not exhaustive. In this case, the appendix could list both the machine-readable consent clauses and the additional data use terms in free text.

Related Standards

The GA4GH DUO primarily aims to define the scope of research purposes for which data may be used. The DUO can be combined with other standard ontologies to provide more comprehensive descriptions of compliance requirements. For example the W3C Data Privacy Vocabulary v0.1 can be used to describe all aspects of a data processing operation, which consists of the following categories: personal data, processing, processing method, processing purpose, legal basis for processing purpose and certain processing actions (such as a transfer from e.g. the EU to a third country), technical and organisational security measures, controller, user and recipient. Another related standard is the ISO/IEC 29184:2020 - Information technology — Online privacy notices and consent.