eTRIKS - Standards Starter Pack
Standards Guidelines
Release 1.0 - 29th June 2015

Authors (by alphabetical order)

Bratfalean, Dorina – CDISC Europe Foundation,
Braxenthaler, Michael – Roche,
Houston, Paul – CDISC Europe Foundation (*),
Munro, Robin – ID Business Solutions Limited,
Richard, Fabien – Centre National de la Recherche Scientifique,
Rocca-Serra, Philippe – Oxford e-Research Centre, University of Oxford (*),
Romacker, Martin – Roche,

(*) Correspondence to philippe.rocca-serra@oerc.ox.ac.uk or phouston@cdisc.org

Licence:

https://creativecommons.org/licenses/by-sa/4.0/
A Business Case for Standards in eTRIKS

Part 1. Introduction
1.1 eTRIKS mission and objectives
1.2 Document objective
1.3 Intended Audience
1.4 Standard Definition and Typology
   1.4.1 Definition of Standards:
   1.4.2 Typology of Standards
1.5 Purpose of Standards

Part 2. Procedure for standards selection and recommendation
2.1 Procedure outline
2.2 Attributes of standards
2.3. Standardization Bodies and Service Providers
2.4. Gaps in Standards
   2.4.1 Coverage gap in a domain covered by an existing standard
   2.4.2 Coverage gap in a domain not covered by standards
2.5 Changes, maintenance and updates to eTRIKS Standard Starter Pack

Part 3. Standards in data management
3.1 Principle of good annotation practice
   Example and Application: Procedure for selecting relevant standard given an eTRIKS dataset
3.2 Prospective data capture
3.4 Case study

Part 4. eTRIKS recommended resources
4.1 eTRIKS - Recommendations for Exchange Format for Clinical Study
   CDISC Standards
4.2 eTRIKS - Recommendations for Exchange Format for Non-Clinical Studies (Animal and in-vitro Studies)
   CDISC Standards
   Experimental Studies
4.3 eTRIKS - WP3 - Standard Starter Pack Recommendations for Database Resource Identification
   4.3.1 Resource Identification:
      4.3.1.1 Identification of Molecular Entities when reporting ‘omics’ data:
      4.3.1.2 Important Reagent Resources:
4.4 eTRIKS - Recommendations for Terminology Resources
   4.4.1 Content and Scope of the Document
   4.4.2 Selecting Terminologies
      4.4.2.1 Use Cases and Iterative Approach
      4.4.2.2 Selection Criteria
   4.4.3 Initial set of Core Terminologies
Organism, Organism Parts and Developmental Stages
Phenotype and Diseases
Pathology and Disease Specific Resources
Cellular entities
Molecular Entities
Assays and Technologies
Relations

4.4.4 Brokering Requests for New Terms
4.4.5 Open Portals and Tools
  4.4.5.1 Content and Browsing Resources
  4.4.5.2 Tools and APIs

4.5 eTRIKS-WP3 Starter-Pack Recommendations Exchange Format for Omics:

Part 5. Future work and roadmap

Appendix
  A.I. Glossary (terms and definitions)
    Person and Organization Roles
    Data Curation
    Data Labels and Controlled Terms
    Data Types and Levels
  A.II. eTRIKS Standards as available from BioSharing:
    A.III. tranSMART master tree

Bibliography:
A Business Case for Standards in eTRIKS

IMI eTRIKS project\(^1\) has released a set of documents aimed at project leaders and data managers alike to provide guidance and recommendations as to which standardization efforts can be relevant to them. The work carried out by eTRIKS is meant to be made available to all IMI projects\(^2\) to raise awareness about the present review work as well as to gain input from specific fields of translational research. Furthermore, eTRIKS aims to provide regular updates and releases on a quarterly basis, to incorporate additions and follow-ups on technology evolution and progress. eTRIKS information feeds (mailing list, website) will be used to relay these updates.

Data standards play an important role in managing and handling research data. On the one hand, regulatory agencies are increasingly mandating data standards for the submission of clinical research data and data sharing \(^3\)-\(^4\). On the other hand, data standards encourage the use of integrated metadata, which provides a solid foundation for systematically using, reusing, understanding, integrating, sharing, and exchanging research data.

Annotation standards such as MIAME guidelines\(^5\) or the Gene Ontology\(^6\) controlled vocabulary are becoming essential resources in modern molecular biology and computational biology. By defining how information is structured and what information is reported, standards, such as CDISC\(^7\) or ISA\(^8\), make it easier to distribute, disseminate and exchange information. They also allow scientific scrutiny to be exerted, a central activity in the life of scientists. There should be no barrier to data assessment and all stakeholders of the scientific endeavour must

\(^{1}\) "What is eTRIKS …… |." 2012. 6 Jun. 2015 <http://www.etriks.org/>
embrace efforts aiming at enhancing access to information so it can be efficiently mined and exploited.

Standards are developed to ensure scientific information is delivered consistently, efficiently and meaningfully to the benefit of the community. Patients will have greater confidence when their data is in an accepted standard format that it then can be analysed more easily and reused in the longer term. With the use of standards the analysis of aggregated study data becomes much more reliable and effective giving maximum opportunities for medical advances and new knowledge to come to light.

More broadly, the very low data comparability and reproducibility is a big issue also in Life Science\textsuperscript{9}, and this results in wasting significant amounts of resources in organizations worldwide, and, consequently, impairs/slowed down the scientific research and the development of new drugs and biomarkers for patients. Standardization of data, metadata and experimental/clinical reports in Life Science will effectively address that issue.

Part 1. Introduction

1.1 eTRIKS mission and objectives
eTRIKS should be a neutral and reference point for data management standards relevant to scientific research focusing on translational medicine in order to make the most of advances of animal model, in-vitro and clinical experimentation.

Among the goals of eTRIKS are:

- Standard harmonization for data annotation. Common list of eTRIKS-selected and recommended standards for data owners and curators
- Standard facilitation. “Bridge builder" between standards communities. Break the silos and facilitate communication between standard communities to drive out duplication and competing standards.
- Reporting standard creation. When not existing, leverage on the technological, medical and laboratory expertise across IMI consortia to develop common reporting standards.
- Standard adoption. Increase the adoption of standards by contributing to the development of annotation tools.
- Data preservation. Contribute to the development of eTRIKS repository that enables the preservation of standardized data through automatic standard updates
- Turning data into knowledge. Contribute to the development of eTRIKS metadata registry and semantic layer that enable smart data searches and inferences.

1.2 Document objective
This document aims to inform readers about eTRIKS guidelines and procedures dealing with data standards and stewardship of standards. eTRIKS strongly recommends eTRIKS
collaborators to follow these guidelines when applicable, in order to facilitate and increase data reusability, reproducibility, and preservation. The document is meant to help optimize annotation and enable *translational and Knowledge management* applications.

### 1.3 Intended Audience

The intended audience is:

- data producers (e.g. research scientists, clinicians, patients) to raise awareness in annotation practice,
- data curators to enable coordinated and agreed upon data cleanup and edition to eTRIKS annotation and curation guidelines,
- data managers in charge of establishing data management plans to guide them in choosing which data formats and terminologies to consider and rely on when collecting new study data, preferably in standard formats.
- software developers to guide development of submission and curation tools.
- knowledge engineers and terminology managers working on developing and supporting ontologies and data models to ensure resource alignment, semantic interoperability and convergence of terminologies and data standards.

### 1.4 Standard Definition and Typology

#### 1.4.1 Definition of Standards:

*Standards* are agreed-upon normative conventions defined by a community of users about a group of descriptive entities, and their combinations, specific to a domain and which facilitate information exchange and communication. They can be considered as a criterion or specification established by authority or consensus for 1) measuring performance or quality; 2) specifying conventions that support interchange of common materials and information (*for example, CDISC standards exist to support the exchange of clinical data, ISA to support exchange of omics data*). Standards may act at the syntactic or the semantic level; both are needed to support interoperability.

Standards should be identified by their name, their version number, the date of the last release, and, if available, a Unique Resource Identifier (URI).

*(See Section 2.2 for attributes of good standards)*

#### 1.4.2 Typology of Standards

Types of standards include the following:
1. **reporting requirements** also called **Minimum Information Guidelines (MIG)**; these define in non-formal ways the necessary and sufficient entities to describe a domain. eTRIKS-adopted or created MIG will specify which exchange formats and vocabulary standards are to be used. Those content standards ensure the exchange of meaning (semantics); they include data and metadata standards. Vocabularies are often treated separately, but they are a form of content standards. A standard may also refer to an integration profile, an implementation guide or a user guide.

2. **vocabularies**; these include a variety of terminologies, such as controlled vocabularies,–dictionaries/thesauri or ontologies that describe either entities, their data labels/names or their data values (i.e. text terms).

3. **exchange formats**; these are syntaxes defining formal ways to structure and organize groups of entities in order to form machine readable research objects, thereby allowing data exchanges between systems and/or organizations in general.

### 1.5 Purpose of Standards
Standards are developed to increase data interoperability, reproducibility, reusability. They also support traceability/provenance, automation and process improvement and preservation/archival of information/data. Three of these purposes are described in more details below.

**Interoperability:** To enable operational processes which underlie data exchange and sharing between different software systems.

**Reusability:** Conformance to standards ensures reliable and consistent description of information (both in structure and content), making it easier to develop robust software for exchanging data payload to be exploited by computational systems. Therefore, standards make data (and research objects) more usable, re-usable, and comparable across studies and/or organizations. Reusability is a central aspect of data preservation, working on the premises that dataset availability should allow meta analysis and discovery through data aggregation. Furthermore, good annotation standards lead to a higher reliability of meta-analysis results by better selecting data from different studies for those meta-analyses.

**Reproducibility:** Reporting standards enable to evaluate data quality, to ascertain solidity of claims and findings. They are therefore invaluable resources, as they allow information to be
assessed. On the one hand, reporting standards, by making key requirements explicit, allow for instance in the case of experimental information testing for confounding factors, thus enhancing reassessment and reproducibility. On the other hand, information provenance\textsuperscript{10} standards provide the means to records events to the data artefacts itself and the chain of custody associated with it. Both types contribute to good data stewardship.

**Long term preservation:** Data live beyond projects, consortia, or organizations. Standards allow for legacy data to be mobilized years after their creation, and compared with more recent or updated datasets. Standards ensure datasets are preserved in well documented, possibly self-describing, data structures.

### Part 2. Procedure for standards selection and recommendation

#### 2.1 Procedure outline

As recommended by the Standards Advisory Board (as of January 28\textsuperscript{th}, 2014), the selection and use of standards should be as neutral, objective, practical, and useful as possible. Information standards should be selected based on the available metadata. Practical applicability and sustainability of a standard rather than its completeness are preferred.

eTRIKS goal is to make recommendations of which standards should be used and in which domain. eTRIKS will demonstrate the benefits and applicability of the adoption of standards using practical examples of real use cases with supported projects. Over time the goal is to track the use and adoption of said standards using simple metrics, such as how many times they have been used in projects and how good the coverage was for the projects supported.

Where practical the following are used to assess whether to adopt a standard.

#### 2.2 Attributes of standards

Following is a list of attributes and criteria for selecting a good standard.

\<http://www.w3.org/2005/Incubator/prov/wiki/What_Is_Provenance>
They are not in order of priority.

Coverage: The standard addresses the domain with a sufficient number of concepts, term sets and metadata elements to meet the user's needs.

Depth and Breadth: The standard delivers at an adequate granularity level to address users needs and describing a study domain with accurate terms.

Relevant/Applicable: The standard is relevant to the goals of the project, study or data to which it is applied; it meets the intended purpose/use case

Necessary: The standard identifies elements and concepts, which must be described.

Available: The standard is freely available for eTRIKS, academic and non-profit organisations.

Pervasive: The standard is used worldwide and, preferably, across several organizations.

Authoritative: The standard is reliable, verified and accepted, based on a documented vetting procedure, preferably a consensus-based procedure by a standards development organization (SDO).

Quality: The standard is able to provide enough terms and associated metadata (e.g. name, label, definition, synonyms) as well as the relationships between terms (in case the standard is ontology)

Readable: The standard is available in human and machine-readable formats

Sustainable: The standard is viable and maintained by a recognized community or a sustained organization of good standing.

For each of these facets, evidence will be reviewed and used to assess the suitability of the standard for the purposes of eTRIKS.

As eTRIKS caters for many different disease areas, it is realised that conflicting interests will arise when selecting standards that cannot be expected to deal equally well with both specific and generic domain representations. The eTRIKS intent is to be practical and not prescriptive.
2.3. Standardization Bodies and Service Providers

Standardization activities are numerous and diverse, taking place in large organizations with industrial strength or at grass root level and academia or both. For historical reasons, many standardization initiatives started from and grew in specific domains of expertise (e.g. proteomics versus transcriptomics, regulatory studies versus research and exploratory studies). This state of affair results in overlapping and competing alternatives, fragmenting standardization efforts, and ultimately impairing integration of multi-type data.

As eTRIKS mission is to enable and ease integration of multi-type data, eTRIKS will build on the work and expertise of domain standards organizations and build an environment where each data type will be described by an eTRIKS-selected standard(s) (when it/they exist(s)).

Standards Development Organizations (SDO) include:
- W3C
- ISO
- CDISC
- HL7
- WHO
- OBO foundry

Vocabulary servers
- Bioportal
- NCI EVS
- Ontology Lookup Service
- LOV

Catalogue of Standards in Life Sciences:
- Biosharing

---

16 "The Open Biological and Biomedical Ontologies." 2006. 6 Jun. 2015 <http://www.obofoundry.org/>
2.4. Gaps in Standards

Two types of gaps in coverage can be found:

2.4.1 Coverage gap in a domain covered by an existing standard

In such a situation, study owners are aware of not only an eTRIKS-approved standard covering the domain of interest, but also a shortage of descriptors and values to accurately annotate their dataset. The central point here is the following: any eTRIKS recommended standard should provide a flexible framework supporting user-defined extensions. In CDISC SDTM or SEND standards, the Supplemental Qualifiers special purpose dataset model may be used to capture non-standard variables and their association to parent records in general-observation-class datasets (Events, Findings, Interventions) and Demographics (more on this topic in SDTM Implementation Guide Section 8.4.2). However, those should only be used if no coverage can be achieved by other more precise means available through CDISC domains. So CDISC documentation and training material should be consulted.

2.4.2 Coverage gap in a domain not covered by standards

This is often the case when new technologies emerge, when understanding of the error models is lacking and when field maturity is an issue making it difficult to standardize. The best advice in such a situation is to attempt to recycle existing module, principles in data management. The CDISC SDTM-Implementation Guide describes the overall process for creating a custom domain, which must be based on one of the three SDTM general observation classes. A custom domain may only be created if the data are different in nature and do not fit into an existing published domain.

Finally, direct contribution to standardization efforts could be made by joining development groups of SDOs or community efforts. When appropriate, a submission of a new CV term may also be logged to the relevant resources. To this end, the tables below supporting this document identify the respective issue trackers associated with each of the semantic artefacts.

For each, eTRIKS WP3 members will outline procedures intended to guide eTRIKS users in dealing with the situation in a principled manner. The main goal is to ensure request coordination and brokering by eTRIKS members and limit duplication and redundant efforts.

---

2.5 Changes, maintenance and updates to eTRIKS Standard Starter Pack

Science and technology are in constant evolution. As with anything, keeping abreast of those changes will be an essential part of eTRIKS Work Package 3. Therefore, it is essential that readers are aware that recommendations made about which data standards to use may change too. Disruptive technologies, both in the field of wet laboratory hardware but also in the field of computer science, computational biology and information technologies may be introduced and radically alter the way to handle specific data elements.

Conversely, substantial aspects of experimental science are not covered by broadly adopted standards, standards especially in the rapidly growing area of genomics and other -omics.

Standardization efforts can be slow to bring about actionable documents, meaning that users need to make do with the existing. Alternately, on-going efforts in specific area are known to exists and their output is announced (for instance, the various working groups in CDISC therapeutic areas publish roadmaps and calendar updates of their progress). For this reason, eTRIKS Work Package (WP) 3 participants will review the changing landscape of data standards and carry out revisions to our recommendations on a regular basis over the course of the eTRIKS project.

Versioning is also very important. The version of a standard should always be documented in any work utilizing standards for data collection, transport or reporting. In first approximation, clearly identifying the release version of the resource being used is a fundamental requirement. However, versioning could be foreseen at many different levels in particular if incremental updates to a standard exist (such as adding new synonyms to code list). Versioning may occur at the level of synonym sets, at the concept level, and at the level of all data and metadata elements making up a standard. A high level of granularity for versioning is required in validated environments.

Part 3. Standards in data management

3.1 Principle of good annotation practice

Many concepts should be standardized to enable cross-study queries and/or comparisons and achieve good query recall. Those queries can be performed:

---

• Within one given study class, e.g. when querying only clinical trials, or
• Across study classes, e.g. when querying clinical and in-vitro studies.

The latter holds most potential for insights or discoveries with relevance to Translational Medicine.

Therefore, we will prioritize our standardization effort on data labels and assays according to the following criteria and order:

a. The most commonly used data labels and their associated textual content across studies, such as (this is not an exclusive list): study protocol elements, study design, demographics, species, strains, organs/body parts, tissues/ primary cells, cell lines, virus, chemicals, peptides/proteins, RNAs (all kinds), genes, DNA variations, DNA modifications, vital signs, behavioral signs, structures/forms/colors, diseases, adverse events, interventions, medical history etc...

b. The data labels and their associated content (qualitative or quantitative values) of the most commonly used assays across studies, such as laboratory testing, gene expression microarray, RNAseq, SNP microarray, DNASeq.

c. In a given project the project-specific (those less commonly used) data labels and assays will be standardized according to the project time lines, following the basic procedure outlined earlier in the document.

The use of standards relies on the principles and basic rules of good annotation practice that are:

1. All the concepts (i.e. data labels and text content) are described by a Controlled Vocabulary Term (CVT) in-lieu of free-text. Concepts from legacy studies, medical comments, and observation notes are not replaced by CVTs but mapped to CVTs (principle of data provenance).
2. A CVT has a unique identifier issued by the associated authority responsible for maintaining the term.
3. Numerical values are converted in the International System (SI) of units\(^{25}\) while retaining the original values (principle of data provenance).
4. Derived data are collected with their primary data and algorithm or methodology used for the data derivation (principle of data provenance).

5. All measurements and observations obey to the principle of data provenance and are associated with the following concepts that answer the What, the Who, the When, the Where, the How and the Why:
  - What organization and/or individual perform them?
  - In what study class have they been performed?
  - For clinical studies, at what study activity identifier (ID) have they been performed?
  - Where (i.e. geographic location) have they been performed?
  - From what subject ID have they been performed?
  - From what specimen ID or part of the subject have they been performed?
  - When have they been performed or when has the specimen been collected (local time)?
  - What is measured or observed?
  - What assay has been used?
  - What biological material has been used by the assay? RNA, DNA, protein, serum etc...?

Example and Application: Procedure for selecting relevant standards given an eTRIKS dataset

Before starting the standard selection, the study owners have to define the investigation scope, the study(ies), the assays, and the variables that will be recorded in the eTRIKS platform. If several studies are recorded, then the workflow is used for each study.

The following steps for a curator to choose a suitable protocol / reporting / semantic / exchange standards for a study.

The workflow steps should be followed in the below described order.

A. Reporting standards
B. Vocabulary standards and standardized units
C. Exchange standards

3.2 Prospective data capture

Standards should be considered at the time of protocol and study design. Where possible data should be collected according to the chosen standards at the time of data generation and capture. To this end, eTRIKS WP3 starter pack recommends study data managers to create a ‘data management plan’ following the guidelines, which will be described in a series of “operational documents”.
3.3 Legacy data

Legacy data may be re-curated to conform to a given standard by the data curators. However, original data are always kept and mapped to CTs.

In either situation, dealing with retrospective or prospective data, a data validation plan (DVP) should be established prior to performing any modification on the submitted data. eTRIKS WP3 is currently working at creating specific documentation about this particular step.

3.4 Case study

One of the eTRIKS objectives is to show how and why the adoption and use of standards can benefit the downstream knowledge generation within and across projects. Initial experience gained from the U-BIOPRED project\textsuperscript{26} will be reported in another document.

---

Part 4. eTRIKS recommended resources

This section simply points to dedicated and specific documentations, which details further eTRIKS recommendation as to which standards may be used in the Data Management Plan.

4.1 eTRIKS - Recommendations for Exchange Format for Clinical Study

4.1.1 CDISC Standards

The CDISC suite of data standards have been designed to support various stages of the clinical research process while conforming to common research business processes and regulatory guidelines. Taken collectively, CDISC standards can streamline the medical research process, saving time and cost while improving quality. Use of data standards can increase the value and reusability of data while preserving meaning as data passes through various stages of the research process. The use of CDISC standards at project initiation has been found to save 70 - 90% of time and resources spent prior to first patient enrolled and approximately 75% of the non-patient participation time during the Study Conduct and Analysis stages. CDISC standards reap substantial benefits, qualitative and quantitative, during the entire research process for all types of research studies including academic, nutritional, device, outcomes and regulated research. Standards reduce complexity and generate a coherent data space.

<table>
<thead>
<tr>
<th>CDISC Standards</th>
<th>Uses/Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foundational Models</td>
<td></td>
</tr>
<tr>
<td>Protocol Representation Model (PRM), Study Design Model (SDM)</td>
<td>The Protocol Representation Model (PRM) is a BRIDG-based model and tools for representing standard clinical research protocol elements and relationships. The Study Design Model (SDM-XML) is an XML schema specification based on the Operational Data Model (ODM) for representing clinical study design, including structure, workflow and timing. PRM supports the interchange (re-use) of information standard to medical/clinical research protocols of any type. V1.0 supports study tracking and clinical trial registration (CTR) in clinicaltrials.gov, WHO or EudraCT; study design (arms, elements, epochs) and scheduled activities; eligibility criteria. In Unified Model Language (UML) format as a subset of BRIDG – spreadsheet and templates to ease use are in progress. A common problem with the typical protocol document is that it is not in a useful format for information management and re-use. The PRM is the foundation for a machine-readable protocol with such ‘re-use’ being one of</td>
</tr>
</tbody>
</table>

http://www.cdisc.org/protocol

the advantages as well as visibility and comprehensibility of the study design.

Clinical Trial data managers should primarily be concerned with obtaining study protocol in such format. When making cross project data comparisons, this summary information is the best way to understand the objectives of and background to the data collection. It enables to categorize studies, to make cross comparisons by identifying like data and the relationships between different datasets. When machine readable protocol representation is absent, one should be built by the data manager following the proposed standard representation.

**The PRM gives the added clinical research benefits of:**
- Increasing transparency of clinical research
- Adhering to study registry requirements
- Sending information to Ethics Committees
- Writing post study clinical reports
- Submission of trial summary info to regulators
- Machine readable search elements
- Avoid poor study designs and further costs and/or study re-runs.

<table>
<thead>
<tr>
<th>Clinical Data Acquisition Standards Harmonization (CDASH)</th>
<th>Clinical Data Acquisition Standards Harmonization is a specification describing basic data collection domains and variables for Case Report Form (CRF) data with standard question text, implementation guidelines, and best practices.</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.cdisc.org/cdash">http://www.cdisc.org/cdash</a></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Laboratory Data Model (LAB)</th>
<th>Specification describing standard content for the acquisition and interchange of clinical laboratory data between central labs and sponsors or contract research organizations (CROs). Vocabulary standard that facilitates exchange of clinical trial laboratory data between central laboratories and study sponsors, CROs or EDC vendors. The LAB model has an extension for microbiology and extensions for pharmacogenomics data.</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.cdisc.org/lab">http://www.cdisc.org/lab</a></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study Data Tabulation Model (SDTM)</th>
<th>Study Data Tabulation Model (SDTM) is the general model for representing study tabulation data used in clinical research. The SDTM Implementation Guide (IG) describes domains and variables for data from Human Clinical Trials for Drug Products and Biologics. SDTM is the standard for data tabulations from CRF data from multiple sites for a clinical study; it is the preferred method for providing data to the FDA for regulatory review. Collecting data in CDASH format can eliminate the need to map data to SDTM at the end of the clinical study process. Efficacy domains are in progress and are defined in the SDTM IG, as well as many described and available in the related Therapeutic</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.cdisc.org/sdtm">http://www.cdisc.org/sdtm</a></td>
<td></td>
</tr>
<tr>
<td><strong>Area User Guides. SDTM now also has a Pharmacogenomics (PGx) domain.</strong></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td></td>
</tr>
<tr>
<td><strong>Analysis Dataset Model (ADaM)</strong></td>
<td></td>
</tr>
<tr>
<td><a href="http://www.cdisc.org/adam">http://www.cdisc.org/adam</a></td>
<td></td>
</tr>
<tr>
<td>Analysis Data Model describes fundamental principles and standards for representing analysis datasets and metadata to support statistical analysis and also statistical regulatory reviews. It is the preferred method by FDA statistical reviewers for submitting research data. The ADaM Implementation Guide (IG) describes standard data structures, conventions and variables used with ADaM. A vocabulary standard for analysis datasets to support statistical analysis and also statistical regulatory reviews; preferred method for providing data for review by FDA statistical reviewers.</td>
<td></td>
</tr>
<tr>
<td><strong>Operational Data Model (ODM)</strong></td>
<td></td>
</tr>
<tr>
<td><a href="http://www.cdisc.org/odm">http://www.cdisc.org/odm</a></td>
<td></td>
</tr>
<tr>
<td>ODM is an XML transport standard that supports data acquisition and exchange of electronic CRF (eCRF) data (such as CDASH data); contains audit trail information per 21CFR11 and EMA eSource Guidance and serves for data archive in a manner independent of the data collection tool.</td>
<td></td>
</tr>
<tr>
<td><strong>Define-XML</strong></td>
<td></td>
</tr>
<tr>
<td><a href="http://www.cdisc.org/define-xml">http://www.cdisc.org/define-xml</a></td>
<td></td>
</tr>
<tr>
<td>The XML-based (ODM-based) standard referenced by FDA as the specification for the data definitions for CDISC SDTM, SEND and ADaM datasets and the current mechanism for providing eSubmissions metadata to FDA.</td>
<td></td>
</tr>
<tr>
<td><strong>Semantics</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Controlled Terminology</strong></td>
<td></td>
</tr>
<tr>
<td><a href="http://www.cdisc.org/terminology">http://www.cdisc.org/terminology</a></td>
<td></td>
</tr>
<tr>
<td>The controlled standard vocabulary and code sets for all of the CDISC models/standards; maintained openly and freely by NCI Enterprise Vocabulary Services (EVS: <a href="http://evs.nci.nih.gov/">http://evs.nci.nih.gov/</a>)).</td>
<td></td>
</tr>
<tr>
<td><strong>Specialty Area (SA) Standards</strong></td>
<td></td>
</tr>
<tr>
<td><a href="http://www.cdisc.org/therapeutic">http://www.cdisc.org/therapeutic</a></td>
<td></td>
</tr>
<tr>
<td>Various standards are now being developed to augment the basic CDISC standards that support safety data across essentially any protocol. These new standards are focused on specialty areas to support efficacy data (e.g. Alzheimer’s and Parkinson’s Diseases, Cardiovascular Disease, Diabetes, Tuberculosis) and also Imaging and Devices. These will add to existing domains for CDISC CDASH and SDTM, and Controlled Terminology.</td>
<td></td>
</tr>
<tr>
<td><strong>Glossary</strong></td>
<td></td>
</tr>
<tr>
<td><a href="http://www.cdisc.org/cdisc-glossary">http://www.cdisc.org/cdisc-glossary</a></td>
<td></td>
</tr>
<tr>
<td>Glossary with definitions of acronyms and terms commonly used in clinical research. Abbreviations and Acronyms also included.</td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>Details</td>
</tr>
<tr>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>Biomedical Research Integrated Domain Group (BRIDG) Model</td>
<td>Biomedical Research Integrated Domain Group (BRIDG) UML model of the semantics of protocol-driven clinical research. <a href="http://www.cdisc.org/bridg">http://www.cdisc.org/bridg</a></td>
</tr>
<tr>
<td>Clinical Outcome Assessment Instruments (Questionnaires)</td>
<td>SDTM Implementation Guide Supplements with annotated CRFs and Controlled Terminology for representing data from Clinical Outcome Assessments (COAs), Questionnaires, and Functional Tests commonly used in clinical studies. <a href="http://www.cdisc.org/ft-and-qt">http://www.cdisc.org/ft-and-qt</a></td>
</tr>
<tr>
<td>CDISC Shared Health and Research Electronic Library (SHARE)</td>
<td>CDISC Metadata Repository source for all CDISC standard metadata and terminology. <a href="http://www.cdisc.org/cdisc-share">http://www.cdisc.org/cdisc-share</a></td>
</tr>
</tbody>
</table>

**Therapeutic Area Standards**

<table>
<thead>
<tr>
<th>Description</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Therapeutic Area (TA) Standards</td>
<td>Various standards are now being developed to augment the basic CDISC standards that support safety data across essentially any protocol. These new standards are focused on specialty areas to support efficacy data (e.g. Alzheimer’s and Parkinson’s Diseases, Cardiovascular Disease, Diabetes, Tuberculosis) and also Imaging and Devices. These will add to existing domains for CDISC CDASH, SDTM, and Controlled Terminology. <a href="http://www.cdisc.org/therapeutic">http://www.cdisc.org/therapeutic</a></td>
</tr>
</tbody>
</table>

4.1.2 SPREC Guidelines for Solid and Fluid Samples:

In the context of clinical trial, it is critical to keep in mind issues related to human tissue and sample preservation and how preanalytical handling of the samples can impact the quality of biological signal derived from samples in downstream workflows. Therefore, eTRIKS WP3 needs to highlight the **Standard Preanalytical Coding for Biospecimens: Review and implementation of the Sample PREanalytical Code** (SPREC) guidelines produced by the International Society for Biological and Environmental Repositories (ISBER).

---

The guidelines, which start to gain momentum in the biobanking initiatives, define a coding system allowing for compact reporting of key collection, preanalytical processing, preservation and storage conditions for solid and fluid biological samples.
4.2 eTRIKS - Recommendations for Exchange Format for Non-Clinical Studies (Animal and *in-vitro* Studies)

4.2.1 CDISC Standards

<table>
<thead>
<tr>
<th>Standards Document</th>
<th>Uses/Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laboratory Data Model (LAB)</td>
<td>Vocabulary standard that facilitates exchange of clinical trial laboratory data between central laboratories and study sponsors, CROs or EDC vendors. The LAB model has an extension for microbiology and extensions for pharmacogenomics data.</td>
</tr>
<tr>
<td><a href="http://www.cdisc.org/lab">http://www.cdisc.org/lab</a></td>
<td></td>
</tr>
<tr>
<td>Standard for the Exchange of non-Clinical Data (SEND)</td>
<td>An extension of SDTM specifically developed for preclinical or non-clinical studies, e.g. toxicology.</td>
</tr>
<tr>
<td><a href="http://www.cdisc.org/send">http://www.cdisc.org/send</a></td>
<td></td>
</tr>
<tr>
<td>Controlled Terminology</td>
<td>The controlled standard vocabulary and code sets for all of the CDISC models/standards; maintained openly and freely by NCI Enterprise Vocabulary Services (EVS).</td>
</tr>
<tr>
<td><a href="http://www.cdisc.org/terminology">http://www.cdisc.org/terminology</a></td>
<td></td>
</tr>
<tr>
<td>Glossary</td>
<td>The CDISC dictionary of terms and their definitions related to the CDISC mission. Abbreviations and Acronyms also included.</td>
</tr>
<tr>
<td><a href="http://www.cdisc.org/cdisc-glossary">http://www.cdisc.org/cdisc-glossary</a></td>
<td></td>
</tr>
</tbody>
</table>

4.2.2 Experimental Studies

<table>
<thead>
<tr>
<th>Standards Document</th>
<th>Uses/Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Investigation Study Assay</td>
<td>‘Investigation’ (the project context), ‘Study’ (a unit of research) and ‘Assay’ (analytical measurement) Tabular format is a meta-format, built purposefully to manage diverse set of life science, environmental and biomedical experiments employing one or a combination of functional genomics technologies while ensuring data deposition to various key omic data repositories.</td>
</tr>
<tr>
<td><a href="http://isatab.sf.net">http://isatab.sf.net</a></td>
<td></td>
</tr>
<tr>
<td>Primary Data Format for Omics</td>
<td>The following link provides a complete overview of the existing format specifications available to support individual ‘omic like type of data.</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>eTRIKS-WP3-Standard-Starter-Pack-Recommandations-Exchange-Format-for-Omics</td>
</tr>
</tbody>
</table>
4.3 eTRIKS - WP3 - Standard Starter Pack Recommendations for Database Resource Identification

4.3.1 Resource Identification:

This is an integral part of the recommendations. Free text should be limited whenever possible and controlled metadata elements should be supplied instead, alongside with their associated identifier, the associated authority issuing it, without forgetting indicating the version of the database or semantic resource used.

The following section and specific documents will identify resources eTRIKS encourages submitters to rely on when preparing their submission in the case of retrospectives studies, or when planning data collection in the case of prospective studies.

If the submitters elect to follow eTRIKS advice, they will facilitate the curation tasks and speed up loading in the relevant tool while reducing operational cost. Should the submitters favour relying on resources outside those specified by eTRIKS, adherence to the resource identification requirements will be of help, leading to easier and more efficient mapping as eTRIKS curation team will be able to take advantage of mapping resources.

Free text terms can not be entirely avoided but controlled terminologies should always be preferred as used more efficiently by search and indexing software agents. In the absence of reliable or affordable natural language processing tools, enforcing controlled terms is a step to facilitate data integration.

4.3.1.1 Identification of Molecular Entities when reporting ‘omics’ data:
The following resources are recommended for tagging or linking entities of interest to database records. eTRIKS recommends using those resources and curation may be applied to align submission on those recommendations. We remind here that the purpose is to ensure annotation consistency, improve query recall and facilitate translational research use cases.

<table>
<thead>
<tr>
<th>Molecular Entity</th>
<th>Resource Name</th>
<th>Biosharing identifier</th>
<th>Resource URI</th>
<th>Resource Identifier pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small Molecules</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CHEBI</td>
<td>bsg-000039</td>
<td><a href="http://www.ebi.ac.uk/chebi/searchId.do?chebid=$id">http://www.ebi.ac.uk/chebi/searchId.do?chebid=$id</a></td>
<td>$id=^CHEBI:\d+$</td>
</tr>
<tr>
<td>---------</td>
<td>---------------</td>
<td>-----------------</td>
<td>--------------------------------------------------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>Drugs</td>
<td>DrugBank</td>
<td>biodbcore-000304</td>
<td><a href="http://www.drugbank.ca/drugs/$id">http://www.drugbank.ca/drugs/$id</a></td>
<td>$id=\text{DB[5]}$</td>
</tr>
<tr>
<td>Biopolymer</td>
<td>DNA</td>
<td>ensEMBL gene</td>
<td>biodbcore-000330</td>
<td><a href="http://www.ensembl.org/">http://www.ensembl.org/</a></td>
</tr>
<tr>
<td></td>
<td>messenger RNA</td>
<td>ensEMBL transcript</td>
<td>biodbcore-000330</td>
<td><a href="http://www.ensembl.org/">http://www.ensembl.org/</a></td>
</tr>
<tr>
<td></td>
<td>micro RNA</td>
<td>mirbase</td>
<td>biodbcore-000569</td>
<td><a href="http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=$id">http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=$id</a></td>
</tr>
<tr>
<td></td>
<td>Entrez Protein</td>
<td>biodbcore-000448</td>
<td><a href="http://www.ncbi.nlm.nih.gov/protein/$id">http://www.ncbi.nlm.nih.gov/protein/$id</a></td>
<td>$id=\text{\d+(\d+)}$</td>
</tr>
<tr>
<td>DNA variant (**)</td>
<td>SNP</td>
<td>NCBI dbSNP</td>
<td>biodbcore-000438</td>
<td><a href="http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=$id">http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=$id</a></td>
</tr>
<tr>
<td></td>
<td>Mutation</td>
<td>HGVS</td>
<td><a href="http://www.hgvs.org/disease-centered-central-mutation-databases">http://www.hgvs.org/disease-centered-central-mutation-databases</a></td>
<td></td>
</tr>
</tbody>
</table>

(*) WHOdrug is not freely available and its cost can be a major limitation for academic institutions.
Consider LRG-sequences now or in the future. (more information at: http://www.lrg-sequence.org/faq#faq_1)

4.3.1.2 Important Reagent Resources:

The table below lists major resources to be aware of when describing in-vitro based work. eTRIKS standard working group is aware of ongoing initiatives (e.g. cell line registry) and new versions of the eTRIKS Standard Starter Pack will reflect progress accordingly.

<table>
<thead>
<tr>
<th>Molecular Entity</th>
<th>Resource Name</th>
<th>Biosharing identifier</th>
<th>Resource URI</th>
<th>Resource Identifier pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>antibodies</td>
<td>antibody-registry</td>
<td>biodbcore-000182</td>
<td><a href="http://antibodyregistry.org/AB_Sid">http://antibodyregistry.org/AB_Sid</a></td>
<td>$id=^d{6}$</td>
</tr>
<tr>
<td>plasmids</td>
<td>addgene</td>
<td>biodbcore-000196</td>
<td><a href="http://www.addgene.org/Sid">www.addgene.org/Sid</a></td>
<td>$id=^d+$</td>
</tr>
<tr>
<td>cell lines</td>
<td>ATCC</td>
<td>biodbcore-00210</td>
<td><a href="http://www.lgcstandards-atcc.org/Products/All/$id.aspx">http://www.lgcstandards-atcc.org/Products/All/$id.aspx</a></td>
<td>$id=^d+$</td>
</tr>
</tbody>
</table>

4.4 eTRIKS - Recommendations for Terminology Resources

4.4.1 Content and Scope of the Document

This document provides a preliminary list of terminologies for clinical, lab data e.g. omics data and non-clinical data, animal data. Terminology is hereby used to refer to any
terminological artifact, e.g., controlled vocabulary, glossary, thesaurus, ontology. This document covers why terminologies are needed and how they have been selected. A list of resources providing browsing functionalities and web services access to the terminologies is also provided.

The scope of this document is to define a list of terminologies in order to inform: (i) the development of the starter pack in eTRIKS Work Package (WP) 3, (ii) curation activities in eTRIKS WP4, (iii) the implementation of the eTRIKS database and the search function (the ‘search app’) in eTRIKS WP2, and (iv) discussion at the IMI office.

To maximize dissemination and searchability of final list of eTRIKS-recommended terminologies, a view will be created in a dedicated page in the BioSharing portal (http://biosharing.org/standards/terminology_artifact).

4.4.2 Selecting Terminologies

4.4.2.1 Use Cases and Iterative Approach

1. The use and implementation of common terminologies will enable a normalization/harmonization of variable labels (data label) and allowed values (data term) when querying the eTRIKS database. Implementing use of common terminologies in the curation workflow will ensure consistency of the annotation across all studies.

2. The clusters of dependent annotations (related data label) also follows the eTRIKS Minimal Information Guidelines (MIGs), a set of core descriptors ensuring that a consistent breadth and depth of information is reported. Continuous feedback will be sought from eTRIKS WP2 and 4 and relevant users. The iterations will feedback into both MIGs and the terminology selections.

3. As part of this iterative process, the eTRIKS use cases and query cases will be documented in order to evaluate, revise and refine the set of terminologies and, where relevant, the associated selection criteria.

4.4.2.2 Selection Criteria

A set of widely accepted criteria for selecting terminologies (or other reporting standards) do not exists. However, the initial work by the Clinical and Translational
Science Awards’ (CTSA) Omics Data Standards Working Group and BioSharing\textsuperscript{29} has been used as starting point to define the eTRIKS criteria for selecting a terminology resource.

- **Exclusion criteria:**
  - absent licence or term of use (*indicator of usability*)
  - licences or terms of use with restrictions on redistribution and reuse (*avoiding any reuse restriction for non-profit organisations*)
  - absence of sufficient class metadata (*indicator of quality, for instance absence of term definition or absence of synonyms*)
  - absence of sustainability indicators nor sustainability of the organisation taking care of the resource
  - absence of term definitions

- **Inclusion criteria:**
  - scope and coverage meets the requirement of the concept identified by eTRIKS as priority target of harmonization (See Starter Pack document point 6.2.a)
  - unique URI, textual definition and IDs for each term
  - resources releases are versioned
  - size of resource (*indicator of coverage*)
  - number of classes and subclasses (indicator of depth)
  - number of terms having definitions and synonyms (indicator of richness)
  - presence of an help desk and contact point (*indicator of community support*)
  - presence of term submission tracker / issue tracker (*indicator of resource agility and capability to grow upon request*)
  - potential integrative nature of the resource by the provision of intra- and cross domain concepts and references (as *indicator of translational application potential*)
  - licensing information available (as *indicator of freedom to use*)
  - use of top level ontology (as *indicator of a resource built for generic use*)
  - pragmatism (as *indicator of actual, current real life practice*)
  - possibility of collaborating with eTRIKS: eTRIKS commit to “stamp” it as “recommended by eTRIKS” and be a portal for receiving users’ complaints/remarks that aim to fix or improve the terminology, while the resource organisation commits to fix or improve the terminology in brief delays (to be determined with the collaborating SDO)

---

\textsuperscript{29} “A sea of standards for -omics (‘genomics,’ ‘proteomics’ or ...” 2014. 8 Jun. 2015
<https://crowdcell.wordpress.com/2014/03/22/a-sea-of-standards-for-omics-data-sink-or-swim/>
4.4.3 Initial set of Core Terminologies

The terminologies have been organized by theme and scope. When possible, sections are organized in progression, from macroscopic scale (organism) to microscopic scale (molecular entities), and from general/generic (e.g. disease) to specialized/specific (e.g. infectious disease).

4.4.3.1 Organism, Organism Parts and Developmental Stages

<table>
<thead>
<tr>
<th>Scope</th>
<th>Name</th>
<th>Biosharing Identifier</th>
<th>File location</th>
<th>Top-Level Ontology</th>
<th>Licence</th>
<th>Issue Tracker URI</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organism</td>
<td>NCBITaxonomy</td>
<td>bsg-000154</td>
<td><a href="http://purl.obolibrary.org/obo/ncbitaxon.owl">http://purl.obolibrary.org/obo/ncbitaxon.owl</a></td>
<td>none specified</td>
<td>This ontology is made available via the UMLS. Users of all UMLS ontologies must abide by the terms of the UMLS license, available at <a href="https://uts.nlm.nih.gov/license.html">https://uts.nlm.nih.gov/license.html</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strain</td>
<td>Rat Strain Ontology</td>
<td>bsg-002625</td>
<td>ftp://rgd.mcw.edu/pub/ontology/rat_strain/</td>
<td>not available</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anatomy</td>
<td>(depends on organism)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mammalian Phenotype</td>
<td>MP</td>
<td>bsg-000129</td>
<td>ftp://ftpinformatics.jax.org/pub/reports/mp.owl</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.4.3.2 Phenotype and Diseases

<table>
<thead>
<tr>
<th>Scope</th>
<th>Name</th>
<th>Biosharing Identifier</th>
<th>File location</th>
<th>Top-Level Ontology</th>
<th>Licence</th>
<th>Issue Tracker URI</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thesaurus Term of Use</td>
<td>License</td>
<td>URL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------</td>
<td>---------</td>
<td>-----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SNOMED-CT</strong></td>
<td>not available</td>
<td>none</td>
<td><a href="http://www.ihtsdo.org/licensing/">http://www.ihtsdo.org/licensing/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ICD-10</strong></td>
<td>bsg-000274</td>
<td>login required</td>
<td><a href="http://www.who.int/about/copyright/en/">http://www.who.int/about/copyright/en/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Disease Ontology</strong></td>
<td>biodbcore-000025</td>
<td>BFO</td>
<td><a href="http://sourceforge.net/p/diseaseontology/feature-requests/">http://sourceforge.net/p/diseaseontology/feature-requests/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Infection Disease Ontology</strong></td>
<td>bsg-000095</td>
<td>BFO</td>
<td><a href="https://code.google.com/p/infectious-disease-ontology/issues/list">https://code.google.com/p/infectious-disease-ontology/issues/list</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Phenotype</strong></td>
<td>bsg-000131</td>
<td>BFO</td>
<td><a href="http://sourceforge.net/p/obo/human-phenotype-requests/">http://sourceforge.net/p/obo/human-phenotype-requests/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MedDRA</strong></td>
<td>bsg-002647</td>
<td>not available</td>
<td>Free for academic and other non-commercial uses. Commercial use of MedDRA requires obtaining a license from MSSO.</td>
<td><a href="https://mssotools.com/webcr/">https://mssotools.com/webcr/</a></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Login required
### 4.4.3.3 Pathology and Disease Specific Resources

<table>
<thead>
<tr>
<th>Scope</th>
<th>Name</th>
<th>Biosharing identifier</th>
<th>File location</th>
<th>Top-Level Ontology</th>
<th>Licence</th>
<th>Issue Tracker URI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Influenza</td>
<td>FLU</td>
<td>bsg-000094</td>
<td><a href="http://www.berkeleybop.org/ontologies/flu.owl">http://www.berkeleybop.org/ontologies/flu.owl</a></td>
<td>BFO</td>
<td>BSD license clause 4</td>
<td></td>
</tr>
<tr>
<td>Malaria</td>
<td>IDOMAL</td>
<td>bsg-000104</td>
<td><a href="http://www.berkeleybop.org/ontologies/idomal.owl">http://www.berkeleybop.org/ontologies/idomal.owl</a></td>
<td>BFO</td>
<td>not available</td>
<td></td>
</tr>
<tr>
<td>Rare disorder</td>
<td>ORDO</td>
<td>bsg-002716</td>
<td><a href="http://www.orphadata.org/data/ORDO/ordo_orphanet.owl.zip">http://www.orphadata.org/data/ORDO/ordo_orphanet.owl.zip</a></td>
<td>none</td>
<td>Attribution-NoDerivs 3.0 Unported</td>
<td></td>
</tr>
<tr>
<td>Alzheimer Disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autism spectrum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fanconi anemia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epilepsy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Immune disorder</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 4.4.3.4 Cellular entities

<table>
<thead>
<tr>
<th>Scope</th>
<th>Name</th>
<th>biosharing identifier</th>
<th>File location</th>
<th>Top-Level Ontology</th>
<th>Licence</th>
<th>Issue Tracker URI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td><a href="http://purl.obolibrary.org/obo/clo.obo">http://purl.obolibrary.org/obo/clo.obo</a></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell Molecular Phenotype Ontology</td>
<td>CMPO</td>
<td>not available</td>
<td><a href="https://github.com/EBI-SPOT/CMPO/tree/master/release">https://github.com/EBI-SPOT/CMPO/tree/master/release</a></td>
<td>BFO</td>
<td>Apache License version 2</td>
<td><a href="http://www.ebi.ac.uk/cmpo/submit">http://www.ebi.ac.uk/cmpo/submit</a></td>
</tr>
</tbody>
</table>
### 4.4.3.5 Molecular Entities

<table>
<thead>
<tr>
<th>Scope</th>
<th>Name</th>
<th>biosharing identifier</th>
<th>File location</th>
<th>Top-Level Ontology</th>
<th>Licence</th>
<th>Issue Tracker URI</th>
</tr>
</thead>
</table>

### 4.4.3.6 Assays and Technologies

<table>
<thead>
<tr>
<th>Scope</th>
<th>Name</th>
<th>biosharing identifier</th>
<th>File location</th>
<th>Top-Level Ontology</th>
<th>Licence</th>
<th>Issue Tracker URI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
<td>Vocabulary</td>
<td>Identifier</td>
<td>License</td>
<td>Notes</td>
<td>Licence Source</td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------------</td>
<td>------------</td>
<td>---------</td>
<td>-------</td>
<td>-----------------</td>
<td></td>
</tr>
<tr>
<td>Radiology</td>
<td>RADLex</td>
<td>bsg-002633</td>
<td>RadLex License Version 2.0</td>
<td><a href="http://radlex.org/suggest_term/index.cfm">http://radlex.org/suggest_term/index.cfm</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NMR Spectroscopy (instrument/acquisition parameter/spectrum related information)</td>
<td>NMR-CV</td>
<td>bsg-000564</td>
<td>BFO</td>
<td>Creative Commons Public Domain Mark 1.0</td>
<td><a href="https://github.com/nmrML/nmrML/issues?state=open">https://github.com/nmrML/nmrML/issues?state=open</a></td>
<td></td>
</tr>
<tr>
<td>Laboratory test (clinical context)</td>
<td>LOINC</td>
<td>bsg-000106</td>
<td>none specified</td>
<td>Mind the terms: <a href="https://uts.nlm.nih.gov/license.html">https://uts.nlm.nih.gov/license.html</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medical Imaging</td>
<td>DICOM [ISO 12052:2006]</td>
<td>bsg-000114</td>
<td>none specified</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.4.3.7 Relations
This section covers more advanced use cases of curation and annotation users may face. Besides identifying entities and concepts for annotation purposes, facts are commonly extracted from literature, expressed as statements and persisted to a knowledge base. An essential step in this process is the creation of these statements where 2 objects are linked via a relation. For example, **drug D inhibits enzyme E or radius bone part_of forearm**.

A few resource formally define relations, defining them in terms of domain and range, thus allowing input validation and reasoning.

While aware of these tasks being somewhat remote from day to day annotation, it is important to be familiar with the relational and semantic underpinnings of a number of terminology artefacts recommended or mentioned in the present documents. The entry below identifies one such important resource.

<table>
<thead>
<tr>
<th>Scope</th>
<th>Name</th>
<th>File location</th>
<th>Top-Level Ontology</th>
<th>Licence</th>
<th>Issue Tracker URI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relations</td>
<td>RO</td>
<td><a href="http://purl.obolibrary.org/obo/ro.owl">http://purl.obolibrary.org/obo/ro.owl</a></td>
<td>BFO</td>
<td>Creative Commons 3.0 BY</td>
<td><a href="https://code.google.com/p/obo-relations/issues/list">https://code.google.com/p/obo-relations/issues/list</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://purl.obolibrary.org/obo/ro.obo">http://purl.obolibrary.org/obo/ro.obo</a></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 4.4.4 Brokering Requests for New Terms

When a term or set of terms are not present in the terminology resources identified, WP3 will act as a broker to ensure the request is submitted to the appropriate resource. To facilitate this, WPs recommends user to submitting a term request using the following templates:

- single term request:

  logon to git.etriks.org with your eTRIKS LDAP credentials and register an issue at https://git.etriks.org/dashboard/issues

  tagging it with ‘Terminology’ label

  with the following fields supplied:

  Term Name:
  Term synonyms:
  Term textual definition:
  Term bibliographic evidence:
  Term submitter identification (name, institution,email):

  Resource targeted for term request:
the eHS team will be notified of the request upon submission.

- batch term request / programmatic handling:
  - WP3 can channel these requests by handling a template for batch submission
  - Batch class definition could be carried out using Ontomat on Google App in Google Spreadsheet: [http://goo.gl/9zsSSI](http://goo.gl/9zsSSI) according to templating procedure.

4.4.5 Open Portals and Tools

4.4.5.1 Content and Browsing Resources

The following terminologies portals allow browsing the resources and, in few cases, also offer useful annotation functionalities when implementing the eTRIKS terminologies in eTRIKS WP2 and WP4 activities and tools.

---


<table>
<thead>
<tr>
<th>Name</th>
<th>URL web interface</th>
<th>Supported Format</th>
<th>Programmatic Access</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCBO Bioportal</td>
<td><a href="http://bioportal.bioontology.org">http://bioportal.bioontology.org</a></td>
<td>OWL, OBO, RRF</td>
<td>yes</td>
<td>Most of it is under BSD license, parts of it is under the Eclipse Public License</td>
</tr>
<tr>
<td>Ontobee</td>
<td><a href="http://www.ontobee.org">http://www.ontobee.org</a></td>
<td>OWL</td>
<td>yes</td>
<td>Apache License, Version 2.0</td>
</tr>
<tr>
<td>EBI OLS</td>
<td><a href="http://www.ebi.ac.uk/ontology-lookup/">http://www.ebi.ac.uk/ontology-lookup/</a></td>
<td>OBO</td>
<td>yes</td>
<td>Apache License, Version 2.0</td>
</tr>
<tr>
<td>NCI EVS</td>
<td><a href="http://evs.nci.nih.gov">http://evs.nci.nih.gov</a></td>
<td>OWL, RRF</td>
<td>yes</td>
<td>not known</td>
</tr>
<tr>
<td>CDISC SHARE</td>
<td><a href="http://cdisc.org/cdisc-share">http://cdisc.org/cdisc-share</a></td>
<td>Excel, XML, RDF, OWL</td>
<td>all documents (PDF, XML, OWL can be made available for download)</td>
<td>not known</td>
</tr>
<tr>
<td>LOV</td>
<td><a href="http://lov.okfn.org/dataset/lov/">http://lov.okfn.org/dataset/lov/</a></td>
<td>RDF</td>
<td>yes</td>
<td>CC-by 3.0 Unported Licence</td>
</tr>
</tbody>
</table>

### 4.4.5.2 Tools and APIs

These are the commonly used API for manipulating terminology resources:

- Jena library: [https://jena.apache.org](https://jena.apache.org)
- OntoCAT: [http://www.ontocat.org](http://www.ontocat.org)
4.5 eTRIKS-WP3 Starter-Pack Recommendations Exchange Format for Omics:

The following table presents key reporting guidelines, exchange formats and terminologies associated to massive parallel molecular characterisation techniques, indicated in red. Fields of information with a blue header indicate supporting information allowing to classify the different laboratory techniques and their applications. The document also reports situations where no formal standard exists and where vendor format specification and instrument related files may act as *de facto* exchange format owing to their diffusion and acceptance as container for primary data.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>genetic variation</td>
<td>genome wide DNA variation profiling</td>
<td>DNA microarray</td>
<td>MIAME</td>
<td>Affymetrix</td>
<td>array design</td>
<td>.CDF file</td>
<td>&lt;none available&gt;</td>
<td>.CEL</td>
<td>.VCF</td>
</tr>
<tr>
<td>genome wide DNA variation profiling</td>
<td>DNA microarray</td>
<td>MIAME</td>
<td>Agilent</td>
<td>array design</td>
<td>.GAL</td>
<td>&lt;none available&gt;</td>
<td>agilent feature extraction .txt</td>
<td>.VCF</td>
<td></td>
</tr>
<tr>
<td>genome wide DNA variation profiling</td>
<td>DNA microarray</td>
<td>MIAME</td>
<td>Illumina</td>
<td>array design</td>
<td>.bpm file, .egt</td>
<td>&lt;none available&gt;</td>
<td>.idat</td>
<td>.VCF</td>
<td></td>
</tr>
<tr>
<td>targeted DNA variation profiling</td>
<td>DNA microarray</td>
<td>MIAME</td>
<td>&lt;miscellaneous&gt;</td>
<td>array design</td>
<td>.GAL</td>
<td>&lt;none available&gt;</td>
<td>export to .txt from instrument</td>
<td>.VCF</td>
<td></td>
</tr>
<tr>
<td>targeted DNA variation profiling</td>
<td>qRT-PCR</td>
<td>MIQE</td>
<td>Applied Biosystems</td>
<td>primer list</td>
<td>&lt;none available&gt;</td>
<td>RDML</td>
<td>export to .txt from instrument</td>
<td>.VCF</td>
<td></td>
</tr>
<tr>
<td>targeted DNA variation profiling</td>
<td>qRT-PCR</td>
<td>MIQE</td>
<td>Biorad</td>
<td>primer list</td>
<td>&lt;none available&gt;</td>
<td>RDML</td>
<td>export to .txt from instrument</td>
<td>.VCF</td>
<td></td>
</tr>
<tr>
<td>targeted DNA variation profiling</td>
<td>qRT-PCR</td>
<td>MIQE</td>
<td>Roche Applied Science</td>
<td>primer list</td>
<td>&lt;none available&gt;</td>
<td>RDML</td>
<td>export to .txt from instrument</td>
<td>.VCF</td>
<td></td>
</tr>
<tr>
<td>exome sequencing</td>
<td>nucleic acid sequencing</td>
<td>MINSEQE</td>
<td>Illumina</td>
<td>exon position list</td>
<td>not applicable</td>
<td>fastq</td>
<td>.BAM, .BigWIG, .BEDgraph</td>
<td></td>
<td></td>
</tr>
<tr>
<td>epigenetic modification</td>
<td>genome wide DNA methylation profiling</td>
<td>DNA microarray</td>
<td>MIAME</td>
<td>Affymetrix</td>
<td>array design</td>
<td>.CDF file</td>
<td>&lt;none available&gt;</td>
<td>.CEL</td>
<td>BAM,BigWIG, BEDgraph</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>----------------------------------------</td>
<td>----------------</td>
<td>-------</td>
<td>------------</td>
<td>-------------</td>
<td>-----------</td>
<td>-----------------</td>
<td>-----</td>
<td>-------------------</td>
</tr>
<tr>
<td>genome wide DNA methylation profiling</td>
<td>DNA microarray</td>
<td>MIAME</td>
<td>Illumina</td>
<td>array design</td>
<td>.bpm file, .egt</td>
<td>&lt;none available&gt;</td>
<td>.CEL</td>
<td>BAM,BigWIG, BEDgraph</td>
<td></td>
</tr>
<tr>
<td>genome wide DNA methylation profiling</td>
<td>DNA microarray</td>
<td>MIAME</td>
<td>Nimblegen</td>
<td>array design</td>
<td>.GFF</td>
<td>&lt;none available&gt;</td>
<td>.CEL</td>
<td>BAM,BigWIG, BEDgraph</td>
<td></td>
</tr>
<tr>
<td>targeted DNA methylation profiling</td>
<td>qRT-PCR</td>
<td>MIQE</td>
<td>Applied Biosystems</td>
<td>primer list</td>
<td>RDML</td>
<td>export to .txt from instrument</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>targeted DNA methylation profiling</td>
<td>qRT-PCR</td>
<td>MIQE</td>
<td>Biorad</td>
<td>primer list</td>
<td>RDML</td>
<td>export to .txt from instrument</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>targeted DNA methylation profiling</td>
<td>qRT-PCR</td>
<td>MIQE</td>
<td>Roche Applied Science</td>
<td>primer list</td>
<td>RDML</td>
<td>export to .txt from instrument</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>genome wide DNA methylation profiling</td>
<td>nucleic acid sequencing</td>
<td>MINSEQE</td>
<td>Illumina</td>
<td>not applicable</td>
<td>fastq</td>
<td>.CEL</td>
<td>BAM,BigWIG, BEDgraph</td>
<td></td>
<td></td>
</tr>
<tr>
<td>histone modification profiling</td>
<td>nucleic acid sequencing</td>
<td>MINSEQE</td>
<td>Illumina</td>
<td>not applicable</td>
<td>fastq</td>
<td>.CEL</td>
<td>BAM,BigWIG, BEDgraph</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chromatin occupancy profiling</td>
<td>nucleic acid sequencing</td>
<td>MINSEQE</td>
<td>Illumina</td>
<td>not applicable</td>
<td>fastq</td>
<td>.CEL</td>
<td>BAM,BigWIG, BEDgraph</td>
<td></td>
<td></td>
</tr>
<tr>
<td>transcription profiling</td>
<td>global transcription profiling</td>
<td>DNA microarray</td>
<td>MIAME</td>
<td>Affymetrix</td>
<td>array design</td>
<td>.CDF file</td>
<td>&lt;none available&gt;</td>
<td>.CEL</td>
<td>BAM,BigWIG, BEDgraph</td>
</tr>
<tr>
<td>global transcription profiling</td>
<td>DNA microarray</td>
<td>MIAME</td>
<td>Agilent</td>
<td>array design</td>
<td>.GAL</td>
<td>&lt;none available&gt;</td>
<td>.CEL</td>
<td>BAM,BigWIG, BEDgraph</td>
<td></td>
</tr>
<tr>
<td>global transcription profiling</td>
<td>DNA microarray</td>
<td>MIAME</td>
<td>Illumina</td>
<td>array design</td>
<td>.bpm file, .egt</td>
<td>&lt;none available&gt;</td>
<td>.CEL</td>
<td>BAM,BigWIG, BEDgraph</td>
<td></td>
</tr>
<tr>
<td>global transcription profiling</td>
<td>nucleic acid sequencing</td>
<td>MINSEQE</td>
<td>Illumina</td>
<td>not applicable</td>
<td>fastq</td>
<td>.CEL</td>
<td>BAM,BigWIG, BEDgraph</td>
<td></td>
<td></td>
</tr>
<tr>
<td>targeted transcription profiling</td>
<td>DNA microarray</td>
<td>MIAME</td>
<td>Applied</td>
<td>primer list</td>
<td>RDML</td>
<td>export to .txt from instrument</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>targeted qRT-PCR</td>
<td>MIQE</td>
<td>Applied</td>
<td>primer</td>
<td>RDML</td>
<td>export to .txt</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profiling Type</td>
<td>Technology</td>
<td>Instrument</td>
<td>List Availability</td>
<td>File Type</td>
<td>Other File Formats</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----------------------------</td>
<td>---------------------</td>
<td>-----------------------</td>
<td>----------------</td>
<td>-------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transcription profiling</td>
<td>qRT-PCR MIQE Biosystems</td>
<td>Roche Applied Science primer list</td>
<td>&lt;none available&gt; RDML</td>
<td>export to .txt from instrument</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targeted transcription profiling</td>
<td>qRT-PCR MIQE Biorad primer list</td>
<td>&lt;none available&gt; RDML</td>
<td>export to .txt from instrument</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>miRNA transcription profiling</td>
<td>nucleic acid sequencing MINSEQE Illumina</td>
<td>GTF file from miRBAS E</td>
<td>not applicable</td>
<td>fastq</td>
<td>.BAM, .BigWIG, .BEDgraph</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein profiling</td>
<td>mass spectrometry MIAPE</td>
<td>not applicable</td>
<td>mzML</td>
<td>mzIdentML</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targeted protein profiling</td>
<td>mass spectrometry MIAPE</td>
<td>protein list</td>
<td>mzML</td>
<td>mzIdentML</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targeted protein profiling</td>
<td>protein microarray MIAPE+ MIAME</td>
<td>protein list;array design</td>
<td>.GAL</td>
<td>&lt;none available&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tissue imaging</td>
<td>mass spectrometry MIAPE</td>
<td>.GAL</td>
<td>imzML</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metabolite profiling</td>
<td>mass spectrometry CIMR</td>
<td>Bruker</td>
<td>not applicable</td>
<td>mzML .netCDF</td>
<td>&lt;none available&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global metabolite profiling</td>
<td>NMR spectroscopy CIMR</td>
<td>Bruker</td>
<td>not applicable</td>
<td>NMR-ML .fid</td>
<td>&lt;none available&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global metabolite profiling</td>
<td>NMR spectroscopy CIMR</td>
<td>Bruker</td>
<td>not applicable</td>
<td>NMR-ML .acqus</td>
<td>&lt;none available&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targeted metabolite profiling</td>
<td>mass spectrometry CIMR</td>
<td>Bruker</td>
<td>metabolite list</td>
<td>mzML .netCDF</td>
<td>&lt;none available&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targeted metabolite profiling</td>
<td>NMR spectroscopy CIMR</td>
<td>Bruker</td>
<td>metabolite list</td>
<td>NMR-ML .fid</td>
<td>&lt;none available&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targeted metabolite profiling</td>
<td>NMR spectroscopy CIMR</td>
<td>Bruker</td>
<td>metabolite list</td>
<td>NMR-ML .acqus</td>
<td>&lt;none available&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global metabolite profiling</td>
<td>mass spectrometry CIMR</td>
<td>Agilent (Variant)</td>
<td>not applicable</td>
<td>mzML .netCDF</td>
<td>&lt;none available&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global metabolite profiling</td>
<td>mass spectrometry CIMR</td>
<td>Agilent (Variant)</td>
<td>not applicable</td>
<td>NMR-ML .fid</td>
<td>&lt;none available&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global metabolite profiling</td>
<td>NMR spectroscopy CIMR</td>
<td>Agilent (Variant)</td>
<td>not applicable</td>
<td>NMR-ML .propar</td>
<td>&lt;none available&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>Technology</td>
<td>Platform</td>
<td>List Available</td>
<td>File Formats</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>-----------------------------</td>
<td>-------------------</td>
<td>----------------</td>
<td>--------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targeted metabolite profiling</td>
<td>Mass spectrometry</td>
<td>CIMR, Agilent(Variant)</td>
<td>&lt;none available&gt;</td>
<td>mzML, .netCDF, &lt;none available&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targeted metabolite profiling</td>
<td>NMR spectroscopy</td>
<td>CIMR, Agilent(Variant)</td>
<td>&lt;none available&gt;</td>
<td>NMR-ML, .fid, &lt;none available&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targeted metabolite profiling</td>
<td>NMR spectroscopy</td>
<td>CIMR, Agilent(Variant)</td>
<td>&lt;none available&gt;</td>
<td>NMR-ML, .propar, &lt;none available&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Microbial diversity profiling</td>
<td>Nucleic acid sequencing</td>
<td>MiXs/MIEN, Illumina</td>
<td>&lt;none available&gt;</td>
<td>fastq, BAM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targeted microbial diversity profiling</td>
<td>Nucleic acid sequencing</td>
<td>MiXs/MIEN, Illumina</td>
<td>&lt;none available&gt;</td>
<td>fastq, BAM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targeted microbial diversity profiling</td>
<td>Nucleic acid sequencing</td>
<td>MiXs/MIEN, Roche Applied Science</td>
<td>&lt;none available&gt;</td>
<td>fastq, .sff, BAM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell characterization</td>
<td>Fluorescent activated cell sorting (FACS)</td>
<td>MIFlowCyt, Becton Dickinson</td>
<td>not applicable</td>
<td>.FCS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell sorting</td>
<td>Fluorescent activated cell sorting (FACS)</td>
<td>MIFlowCyt, EMD millipore</td>
<td>not applicable</td>
<td>.FCS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Part 5. Future work and roadmap

The present document can be viewed as a survey of the existing landscape of data exchange supporting standards in the field of life science relevant to translational medicine research. This is only a first step in the overall direction the eTRIKS project is advancing.

The goal is to deliver an environment to help and assist data managers in delivering more consistent and comparable datasets. To this end, eTRIKS WP3 intends to provide:

-a list of recommendations about relevant data standards to translational research (the eTRIKS standard starter pack). The eTRIKS Standard Starter Pack will undergo updates according to a 3 to 6 month release frequency to reflect the evolution and progress of standardization initiatives. For instance, CDISC releases Therapeutic Areas (TA) standards regularly as disease domains reach maturity. In the field of genomics, the Global Alliance for Genomic Health\(^\text{32}\) is spearheading a new initiative to devise programmatic means to data exchange

-a set of operational guidelines, meaning clear procedure for creating ‘data management plans’ and ‘data validation plans’. eTRIKS WP3 expects to release these documents in the fourth quarter of 2015 (15Q4)

-a set of use-cases, user requirements that will be used to draft the functional specifications for a curation infrastructure as several needs have been identified such as an eTRIKS metadata registry which would:

1. store eTRIKS vetted terminology artefact
2. store eTRIKS vetted representation of data format
3. store collections of value sets specific to public or IMI studies curated by the eTRIKS curation team. While the list of variables collected in the study could be queried by all, the actual individual, subject level value would be access-controlled in order to preserve any intellectual property.

A graphical overview of roadmap is presented below.

The colored cones indicate planned releases and milestones. However, the output is not confined to single point releases and documents. Training materials, examples and tutorials will be posted from eTRIKS portal as they are developed.
Appendix

A.I. Glossary (terms and definitions)

Organizations and Consortia

· *eTRIKS* refers to the eTRIKS consortium.

· *CDISC* stands for Clinical Data Interchange Standards Consortium.

· *TCGA* stands for The Cancer Genome Atlas.\(^3^3\)

· *SI units* refer to the International System (SI) of units

· *tranSMART Foundation* ([www.transmartfoundation.org](http://www.transmartfoundation.org)) is an organization looking after the tranSMART software.

Person and Organization Roles

· *A study owner* is the legal person (natural or judicial) who is responsible for authorizing the access and/or the use of data from a study.

· *A collaborator* is a study owner who 1) gives the right of handling the data of a study to eTRIKS, and 2) follows eTRIKS guidelines, where applicable.

Data Curation

· *Data curator* is someone who performs data curation, namely a group of management activities required to ensure long-term research data preservation such that data are available for reuse and evaluation. These management activities consist in harmonizing annotation, cleaning, converting, standardizing, and formatting data to ensure consistency, increase recall and enable cross study comparison.

· *Curated data* are data for which the values, the labels, the formats, and the provenances follow the curation rules and conventions defined by eTRIKS.

---

Data Labels and Controlled Terms

- **Data labels** (also called variables in data management) are descriptions of data (often names; in a table they are column headers)

- **Data Dictionary** is a flat list of terms whose labels and definitions are agreed upon

- **Controlled Terminology** is a tree of terms whose labels and definitions are agreed upon and which are organized in a hierarchical structure.

- **A Reference Ontology** is a semantic resource developed to represent formally a domain of Science, defining entities, their properties and relation with respect to other entities. The Gene Ontology\(^{34}\) is a reference ontology for defining gene function, molecular process and biological component while Human Phenotype Ontology\(^{35}\) is a reference ontology for the description of human disorders.

- **An Application Ontology** is a semantic resource developed specifically to answer uses cases and specific tasks defined by a focused software application such as user interface. Application Ontology often combines controlled vocabulary terms from various ‘reference’ resources (i.e. reference ontologies) by mixing and matching in an \textit{ad-hoc} fashion (in the worst of cases), or according to principled way (for instances by combining reference ontologies sharing the same development practices). Application Ontologies requires constant synchronization with Parents/Source artefacts, something which can be achieved through software agents but places infrastructure demands. EFO, The experimental Factor Ontology\(^{36}\), is an application ontology specifically developed for EMBL-EBI ArrayExpress needs.

- **A Controlled Vocabulary Term (CVT)** is a term that belongs to a terminology, a dictionary, or an ontology for which an authoritative textual definition exists (complemented by a formal definition for ontologies).

- **An eTRIKS Controlled Vocabulary Term (eCVT)** is a unique CVT in the eTRIKS CVT library, and has a corresponding identifier and the associated standard source.

- The eCVT library contains all the eCVT used by eTRIKS in eTRIKS.

- **eTRIKS data labels** are eCVT.

---

\(^{36}\) “Experimental Factor Ontology < EMBL-EBI.” 2009. 8 Jun. 2015 <http://www.ebi.ac.uk/efo>
Standardized data are either eCVT or numerical values converted to International System (SI) of units.

Data Types and Levels

- **Metadata** provide descriptive and provenance information about data.

- **Primary data (Level 1 Data)** according to The Cancer Genome Atlas (TCGA) classification ([https://tcga-data.nci.nih.gov/tcga/tcgaDataType.jsp](https://tcga-data.nci.nih.gov/tcga/tcgaDataType.jsp), also known as “raw data”) are assay results that have not been processed/transformed, and are either measurements or observations.

- **Derived data (Level 2 Data)** according to TCGA classification) are data that are calculated from, or given according to, several primary or derived data. Treatment responses are derived data: they are assigned according to primary data.

  Example 1. A treated patient with a tumor size (primary data) above an arbitrary threshold is considered as “non-responder” (derived data).

  Example 2. Ages are derived data calculated from the birth and study starting dates (primary data).

- **Interpreted data (Level 3 Data)** according to TCGA classification) are data that result from the interpretation of Level 1 or 2 Data by using reference data.

  Example. In a microarray, normalized intensity values associated with a probe set IDs are level 2 data, while the gene names associated with the probe set IDs are level 3 data.

- **Reference data** provide information from biological databases and resources (e.g. gene annotation of a microarray probe set; SNP location in the genome and their mapping to genes).

Investigation, Study and Observations, Assays and Measurements

- A study is a central unit containing information on subjects under study and its characteristics. A study has associated assays.

- A study class is defined according to the nature (type) of subjects (i.e. human, non-human animal, cell, virus) under study.
A clinical study is a type of study where study subjects are human subjects.

A preclinical study is a type of study where study subjects are animals, tissues, or cells.

An investigation or project is a collection of related studies.

A subject is the living entity or organism under study, and can be a human, a non-human animal, a cell, or a virus.

An assay is a measurement process performed either on a subject or on material derived from the subject. Assay results are findings.

Measurements are quantitative data of an assay and have a numerical value.

Observations are qualitative data of an assay result, and do not have a numerical value.

An image is an observation, while its signal levels are measurements.

An ‘omic’ assay is a molecular biology techniques that enables simultaneous measurement of a large collection of molecular entities (transcripts, protein, small molecules). An ‘omic’ profiling may be “targeted” (meaning a limited number of known entities are assayed, such as in ELISA, Luminex or RT-PCR multiplex panel) or may be “untargeted” (meaning any entity in a given molecular class may be measured (such as in pan-genome microarrays, RNA-Seq)

TranSMART:

TranSMART (TM) is the data warehouse that eTRIKS will contribute to develop in order to enable data hosting, sustainability, visualization and analysis.

A tranSMART concept tree refers to the overall organisation and representation of the study concepts in the TranSMART User Interface (UI) (see an example of a tranSMART concept tree in Annex III).
A.II. eTRIKS Standards as available from BioSharing:

BioSharing (www.biosharing.org) is an open source initiative aiming at providing an up-to-date overview of the standards landscape in the life science. Besides various advanced search and filtering features, the registry offers communities to present the set of resources they rely on for their data management needs. The following figure illustrates how eTRIKS may use the Biosharing website to further broadcast and publicize technical recommendations.

Figure 1. The eTRIKS view of relevant standards as available from Biosharing website.
https://www.biosharing.org/collection/5?q=&view=table
A.III. tranSMART master tree

First pass of a recommended hierarchy to use for tranSMART data explorer
Bibliography

STANDARDS


Ann Marie Martin, Nathalie Seigneuret, Ferran Sanz,and Michel Goldman; Data Standards are needed to move Translational Medicine forward; (2012)Transl Med 3.2:: 2161-1025 (url: http://dx.doi.org/10.4172/2161-1025.1000119)


**ONTOLOGIES:**

Haber, Margaret W et al. "Controlled terminology for clinical research: a collaboration between CDISC and NCI enterprise vocabulary services." Drug information journal 41.3 (2007): 405-412.


**VOCABULARY SERVERS**


TRANSMART
