European Translational Information and Knowledge Management Services

eTRIKS Deliverable report
Grant agreement no. 115446
Deliverable 3.4
Final standards report – manual of guidelines/recommendations for standards in eTRIKS v2

Due date of deliverable: Month 28
Actual submission date: Month 30

Dissemination Level		
PU	Public	xx
PP	Restricted to other programme participants (including Commission Services)	
RE	Restricted to a group specified by the consortium (including Commission Services)	
CO	Confidential, only for members of the consortium (including Commission Services)	

DELIVERABLE INFORMATION

Project	
Project acronym:	eTRIKS
Project full title:	European Translational Information and Knowledge Management Services
Grant agreement no.:	**115446**
Document	
Deliverable number:	3.4
Deliverable title:	Final standards report – manual of guidelines/recommendations for standards in eTRIKS v2
Deliverable version:	1
Due date of deliverable:	January 31st 2015
Actual submission date:	March 31st 2015
Leader:	
Editors:	
Authors:	Philippe Rocca-Serra, Fabien Richard, Dorina Bratfalean, Chris Marshall
Reviewers:	Michael Braxenthaler, Paul Houston, Robin Munro, Trevor Garrett
Participating beneficiaries:	
Work Package no.:	3
Work Package title:	Standards Research and Coordination
Work Package leader:	Michael Braxenthaler, Paul Houston
Work Package participants:	
Estimated person-months for deliverable:	6 pm
Nature:	Document
Version:	1
Draft/Final:	Draft
No of pages (including cover):	
Keywords:	

Purpose

As set out in the Description of Work, WP3 aims to identify, develop and adopt data standards and controlled terminologies for translational researchers using the eTRIKS Platform and in the wider community. The consistent application of standards to a scientific discipline is a prerequisite for meaningful exchange of information and combining of data from multiple sources.

In this document, we provide a consolidated review of current standards and controlled terminologies available to translational researchers, a procedure for identifying good quality data standards and terminologies and recommendations for use of these standards by all organisations planning to use the eTRIKS Platform. The content have been reviewed by the eTRIKS Standards Advisory Board and forms part of the wider offering of the eTRIKS project to the translational research community through the eTRIKS web site.

www.etriks.org

The standards set out in this document should be treated as a living document, the current version of which represents a snapshot of current standards and terminologies that are valid at the time of publication. It should be recognized that these might, and indeed are expected to change over the course of the delivery of the eTRIKS project. The current version is available on the etriks website referenced above. The expectation is that we will operate a quarterly cycle of review and update for this document.

Intended audience

The readership of this document is assumed to be familiar with eTRIKS and its overall aims, including being aware of the work completed to date with respect to the tranSMART for eTRIKS software.

Contents

A Business Case for Standards in eTRIKS

IMI eTRIKS project 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 to raise awareness about the present review work as well as to gain input from specific fields of translational research. Furthemore, 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.

Regulatory agencies now mandate data deposition and data sharing. This means relying on communication standards. Data management is evolving and the landscape around these activities is moving forward. Scientists and data managers can no longer afford to ignore these evolutions. Standardization efforts will not go away and ignoring them will only add to the 'growing pains' of large research projects.

Annotation standards such as MIAME guidelines or the Gene Ontology controlled vocabulary are no longer challenged as major contributions, however imperfect, to the birth and establishment of essential resources in moder molecular biology and computational biology. By regularizing how information is structured and reported, standards, such as CDISC or ISA, 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 stake holders of the scientific endeavour must embrace efforts aiming at enhancing access to information so it can be efficiently mined and exploited.

Standards are not developed by insular groups dedicated to creating more red-tape entrapping 'real scientists'. Standards are developed to ensure scientific information is delivered ***consistently***, ***efficiently*** and ***meaningfully*** to the benefit of the community. Building such infrastructures do not occur overnight. It requires investments from all parties and also the appreciation from funding agencies and stakeholders that data management is a new, essential activity befalling on researchers. Data management has associated costs. These should be properly evaluated and considered by funding agencies when supporting research efforts. Conversely, as funding agencies increasingly require data management plans to be supplied in grant application, it is obvious that the constraints of data management are being appreciated. Therefore, instead of being seen as a burden, standardization efforts and standards should be in fact perceived as unique helping tools to enhance impact of the work carried out by scientists. Data production and data dissemination can now be acknowledged by modern means and will become one of the ways to monitor scientific and research output, In this context, standards and standard compliance is only a means to an end.

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 as well as ontologies and data models for the eTRIKS repository.

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 specific to a domain and which facilitate information exchange and communication. They can 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 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***; define in non-formal ways the necessary and sufficient entities to describe a domain. 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 standards 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.

4. ***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

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 data exchange, sharing and operational process 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 data to be evaluated, ascertain solidity of claims and findings, thus assist in allowing reproducibility. Reporting standards, by making key requirements explicit allow testing for confounding factors, thus enhancing reassessment and reproducibility.

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 Jan 28th 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. They are not in order of priority:

- Coverage: The standard addresses the domain adequately to meet the users' needs.
- 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.
- Depth/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)
- Depth and Breadth: The standard delivers at an adequate granularity level to address users needs and describing a study domain with accurate terms.
- 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)
- 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, 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), includes:

- ISO
- CDISC
- HL7
- WHO
- IHE
- OBO foundry

Vocabulary servers

- Bioportal
- NCI EVS
- Ontology Lookup Service

- LOV

Catalogue of Standards in Life Sciences:

- Biosharing

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 to accurately annotate their dataset. The central point here is the following: any eTRIKS approved standard should have a clearly identified capability for handling and supporting users' requests.

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.

Finally, direct contribution to standardization efforts could be made by joining development groups of SDOs or community efforts.

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, ongoing 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

http://blogs.fda.gov/fdavoice/index.php/tag/coalition-for-accelerating-standards-and-therapies-cfast/#sthash.bLKDDn44.dpuf")

For this reason, eTRIKS Work Package 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 program.

Note: include a section detailing how to deal with versioning and related issues/ DB mentions how industry deals with legacy studies -> reliance on contemporary standards and guidelines, not new ones)

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.

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 class, e.g. when querying clinical and in-vitro studies.

The latter holds most potential for insights or discoveries with relevancy 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, RNA (all kinds), genes, DNA variations, DNA modifications, vital signs, behavioral signs, structures/forms/colors, diseases, adverse events, interventions, medications.
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 (*CT) in-lieu of free-text. Concepts from legacy studies, medical comments, and observation notes are not replaced by CTs but mapped to CTs (principle of data provenance).
2. A CT 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 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, and the How:

    - What organization and/or individual perform them?
    - In what study class have they been performed?
    - For clinical studies, at what study activity 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, …?

Example and Application: Procedure for selecting relevant standard 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 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 will be reported elsewhere.

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

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 bring order to complexity.*

CDISC Standards	**Uses/Value**
Foundational Models	
Protocol Representation Model (PRM), Study Design Model (SDM) http://www.cdisc.org/protocol The PRM toolkit gives 30 basic concepts essential for all protocols and is more easily understandable than the full UML model.	The Protocol Representation Model 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. The 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 the advantages as well as visibility and comprehensibility of the study design. For eTRIKS project data this should be the first point with which a data manager should be concerned with. If a PRM does not exist then one should be built from the protocol information. When making cross project data comparisons this summary information is the best way to understand the objectives and background to the data and thus categorize the studies,

	being able to make cross comparisons by identifying like data and the relationships between different data sets. **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.
Clinical Data Acquisition Standards Harmonization (CDASH) http://www.cdisc.org/cdash	Clinical Data Acquisition Standards Harmonization is a specification describing basic data collection domains and variables for CRF data with standard question text, implementation guidelines, and best practices.
Laboratory Data Model (LAB) http://www.cdisc.org/lab	Specification describing standard content for the acquisition and interchange of clinical laboratory data between central labs and sponsors or 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.
Study Data Tabulation Model (SDTM) http://www.cdisc.org/sdtm	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 Area User Guides. SDTM now also has a Pharmacogenomics (PGx) domain.
Standard for the Exchange of non-Clinical Data (SEND) http://www.cdisc.org/send	An extension of SDTM specifically developed for pre-clinical or non-clinical studies, e.g. toxicology.
Analysis Dataset Model	Analysis Data Model describes fundamental principles and standards for

(ADaM) http://www.cdisc.org/adam	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.
Operational Data Model (ODM) http://www.cdisc.org/odm	ODM is an XML transport standard that supports data acquisition and exchange of 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.
Define-XML http://www.cdisc.org/define-xml	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.
Semantics	
Controlled Terminology http://www.cdisc.org/terminology	The controlled standard vocabulary and codesets for all of the CDISC models/standards; maintained openly and freely by NCI Enterprise Vocabulary Services (EVS).
Specialty Area (SA) Standards http://www.cdisc.org/therapeutic	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.
Glossary http://www.cdisc.org/cdisc-glossary	Glossary with definitions of acronyms and terms commonly used in clinical research. Abbreviations and Acronyms also included.
Biomedical Research Integrated Domain Group (BRIDG) Model http://www.cdisc.org/bridg	Biomedical Research Integrated Domain Group (BRIDG) UML model of the semantics of protocol-driven clinical research.

Clinical Outcome Assessment Instruments (Questionnaires) http://www.cdisc.org/ft-and-qt	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.
CDISC Shared Health and Research Electronic Library (SHARE) http://www.cdisc.org/cdisc-share	CDISC Metadata Repository source for all CDISC standard metadata and terminology.
Therapeutic Area Standards	
Therapeutic Area (TA) Standards http://www.cdisc.org/therapeutic	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.

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 starts to gain momentum in the biobanking initiatives, defines 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)

CDISC Standards

Standards Document	**Uses/Value**
Laboratory Data Model (LAB)	Vocabulary standard that facilitates exchange of clinical trial laboratory data between central laboratories and study sponsors,

http://www.cdisc.org/lab	CROs or EDC vendors. The LAB model has an extension for microbiology and extensions for pharmacogenomics data.
Standard for the Exchange of non-Clinical Data (SEND) http://www.cdisc.org/send	An extension of SDTM specifically developed for pre-clinical or non-clinical studies, e.g. toxicology.
Controlled Terminology http://www.cdisc.org/terminology	The controlled standard vocabulary and codesets for all of the CDISC models/standards; maintained openly and freely by NCI Enterprise Vocabulary Services (EVS).
Glossary http://www.cdisc.org/cdisc-glossary	The CDISC dictionary of terms and their definitions related to the CDISC mission. Abbreviations and Acronyms also included.

Experimental Studies

Standards Document	**Uses/Value**
Investigation Study Assay http://.isatab.sf.net	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.
Primary Data Format for Omics	eTRIKS-WP3-Standard-Starter-Pack-Recommandations-Exchange-Format-for-Omics

4.3 eTRIKS - WP3 - Standard Starter Pack Recommendations for Database Resource Identification

4.3.I Resource Identification:

This is an integral part of the recommendations. Free text should be limited whenever possible and metadata elements should be associated with an identifier, the authority.resource issuing it and the version of the database.resource.

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

In so doing, submitters will facilitate the curation tasks and if they elect to follow eTRIKS advice 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 can not be entirely avoided but placing tokens of information in a metadata framework is a step towards data integration.

i. 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.

Molecular Entity	Resource Name	Resource URI	Resource Identifier pattern	Resource Name	Resource URI	Resource Identifier pattern
Small Molecules						
Metabolites	Pubchem (biodbcore-000455)	http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=$id	$id=^\d+$	CHEBI	http://www.ebi.ac.uk/chebi/searchId.do?chebiId=$id	^CHEBI:\d+$
Lipids	Lipid Maps (biodbcore-000559)	http://www.lipidmaps.org/data/get_lm_lipids_dbgif.php?LM_ID=$id	^LM(FA|GL|GP|SP|ST|PR|SL|PK)[0-9]{4}([0-9a-zA-Z]{4,6})?$			
Drugs	DrugBank (biodbcore-000304)	http://www.drugbank.ca/drugs/$id	^DB\d{5}$	*WHOdrug (*)*		
Biopolymer						
DNA	ensEMBL gene (biodbcore-000330)	http://www.ensembl.org/	$id=ENSG\d+$	Entrez Gene (aka NCBI Gene)	http://www.ncbi.nlm.nih.gov/gene/$id	^\d+$
messenger RNA	ensEMBL	http://www.en	$id=ENST\d+$			

	transcript (biodbcore-000330)	sembl.org/				
micro RNA	mirbase (biodbcore-000569)	http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=$id	MI\d{7}			
Protein	Uniprot (biodbcore-000544)	http://www.uniprot.org	^([A-N,R-Z][0-9]([A-Z][A-Z, 0-9][A-Z, 0-9][0-9]){1,2})|([O,P,Q][0-9][A-Z, 0-9][A-Z, 0-9][A-Z, 0-9][0-9])(\.\d+)?$	Entrez Protein	http://www.ncbi.nlm.nih.gov/protein/$id	^(\w+\d+(\.\d+)?)|(NP_\d+)$
DNA variant (**)						
SNP	NCBI dbSNP (biodbcore-000438)	http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=$id	^rs\d+$	HGVS	www.hgvs.org (***)	
Structural Variation	NCBI dbVar (biodbcore-000463)					

(*)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)
(***) incorrect uri supplied leading to an ambiguity, hence recommendation is to rely on dbSNP and dbVAR until resolution.

ii Important Reagent Resources:

Molecular Entity	Resource Name	Resource URI	Resource Identifier pattern	Biosharing/ biodbcore identifier		
antibodies	antibody-registry	http://antibodyregistry.org/AB_$id	^\d+{6}$	biodbcore-000182		
plasmids	addgene	www.addgene.org/$id	^'\d+$	biodbcore-000196		
cell lines	ATCC	http://www.lgcstandards-atcc.org/Products/All/$id.aspx	^'\d+$	biodbcore-000210		

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.

This scope of this document is to define a list of terminologies to inform: (i) the development of the starter pack in WP3, (ii) curation activities in WP4, (iii) the implementation of the eTRIKS database and the search function (the 'search app') in 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 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 (http://jamia.bmj.com/content/early/2013/10/03/amiajnl-2013-002066.long) has been used as starting point top define the eTRIKS criteria for excluding and/or including 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 (*as indicator of translational application potential*)
  - licensing information available (*as indicator of freedom to use*)
  - use of 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 (one month after receipt?)

4.4.3 Initial set of Core Terminologies

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

Organism, Organism Parts and Developmental Stages

Scope	Name	File location	Top-Level Ontology	Licence	Issue Tracker URI	Comment
Organism	NCBITaxonomy	http://purl.obolibrary.org/obo/ncbitaxon.owl	none specified	This ontology is made available via the UMLS. Users of all UMLS ontologies must abide by the terms of the UMLS license, available at https://uts.nlm.nih.gov/license.html		
Strain	Rat Strain Ontology	http://data.bioontology.org/ontologies/RS/submissions/46/download?apikey=4ea81d74-8960-4525-810b-fa1baab576ff				
Anatomy	(depends on organism)					
Vertebrate Anatomy	UBERON	http://purl.obolibrary.org/obo/uberon/ext.owl http://purl.obolibrary.org/obo/uberon/ext.obo	BFO	CC-by 3.0 Unported Licence	https://github.com/obophenotype/uberon/issues	*Integrative Resource engineered to go across species*
Mouse Anatomy	MA					
Mouse Phenotype	MPO					

Phenotype and Diseases

Scope	Name	File location	Top-Level Ontology	Licence	Issue Tracker URI
Pathology/Disease (generic)					
	SNOMED-CT			http://www.ihtsdo.org/licensing/	
	NCI thesauru	http://evs.nci.nih.gov/ftp1/N		http://evs.nci.nih.gov/ftp1/NCI_Thes	

	s	CI_Thesaurus		aurus/ThesaurusTermsofUse.htm	
	ICD-10		login required [http://apps.who.int/classifications/apps/icd/ClassificationDownloadNR/login.aspx?ReturnUrl=%2fclassifications%2fapps%2ficd%2fClassificationDownload%2fdefault.aspx]	http://www.who.int/about/copyright/en/	
	UMLS			http://www.nlm.nih.gov/databases/umls.html	
	Disease Ontology	http://purl.obolibrary.org/obo/doid.owl	BFO	CC-by 3.0 Unported Licence	http://sourceforge.net/p/diseaseontology/feature-requests/
	Infection Disease Ontology	https://code.google.com/p/infectious-disease-ontology/source/browse/trunk/src/ontology/ido-core/ido-main.owl	BFO	most probably: CC-by 3.0 Unported Licence	https://code.google.com/p/infectious-disease-ontology/issues/list
Phenotype	Human Phenotype Ontology	http://compbio.charite.de/hudson/job/hpo/lastStableBuild/	BFO	most probably: CC-by 3.0 Unported Licence	http://sourceforge.net/p/obo/human-phenotype-requests/
	PATO		BFO		http://sourceforge.net/p/obo/phenotypic-quality-pato-requests/
	MedDRA			This ontology is freely accessible on this site for academic and other non-commercial uses. Users anticipating any commercial use of MedDRA must contact the MSSO to obtain a license.	https://mssotools.com/webcr/ Login required

Pathology and Disease Specific Resources

Scope	Name	File location	Top-Level Ontology	Licence	Issue Tracker URI
Influenza	FLU		BFO		
Malaria	IDOMAL		BFO		
Dengue Fever	IDODEN		BFO		
Alzheimer Disease					
Autism spectrum					
Fanconi anemia					
Epilepsy					
Immune disorder					
Rare disorder	ORDO				

Cellular entities

Scope	Name	File location	Top-Level Ontology	Licence	Issue Tracker URI
Cell	CL	http://purl.obolibrary.org/obo/cl.owl http://purl.obolibrary.org/obo/cl.obo	BFO	most probably: CC-by 3.0 Unported Licence	https://code.google.com/p/cell-ontology/issues/list
Cell Lines	CLO	http://clo-ontology.googlecode.com/svn/trunk/src/ontology/clo.owl	BFO	most probably: CC-by 3.0 Unported Licence	https://code.google.com/p/clo-ontology/issues/list
Cell Molecular Phenotype Ontology	CMPO	https://github.com/EBISPOT/CMPO/tree/master/release	BFO		

Molecular Entities

Scope	Name	File location	Top-Level Ontology	Licence	Issue Tracker URI
Chemicals and Small Molecules	CHEBI	http://ftp.ebi.ac.uk/chebi.owl http://ftp.ebi.ac.uk/chebi.obo	BFO	most probably: CC-by 3.0 Unported Licence	http://sourceforge.net/p/chebi/annotation-issues/
Drug	National Drug File			https://uts.nlm.nih.gov/license.html	
Gene Function, Molecular Component, Biological Process	GO	http://purl.obolibrary.org/obo/go.obo http://purl.obolibrary.org/obo/go.owl	BFO	CC-by 4.0 Unported License	http://sourceforge.net/p/geneontology/ontology-requests/
Protein/peptide	PRO	http://ftp.pir.georgetown.edu/pro.obo	BFO	CC-by 3.0 Unported Licence	

Assays and Technologies

Scope	Name	File location	Top-Level Ontology	Licence	Issue Tracker URI
Sample Processing/Reagents/Instruments Assay Definition	OBI	http://svn.code.sf.net/p/obi/code/releases/2014-03-29/obi.owl	BFO	CC-by 3.0 Unported Licence	http://sourceforge.net/p/obi/obi-terms/
Biological screening assays and their results including high-throughput screening (HTS)	BAO	http://www.bioassayontology.org/bao/bao_complete_bfo_dev.owl	BFO	CC-by 3.0 Unported Licence	
Experimental Design, Statistical Methods and Statistical Measures	STATO	https://raw.githubusercontent.com/ISA-tools/stato/dev/src/ontology/stato.owl	BFO	CC-by 3.0 Unported Licence	https://github.com/ISA-tools/stato/issues?state=open
Radiology	RADLex				
Mass Spectrometry (instrument/acquisition parameter/spectrum related	PSI-MS	http://psidev.cvs.sourceforge.net/viewvc/psidev/psi/psi-ms/mzML/con	none specified	CC-by 3.0 Unported Licence	https://lists.sourceforge.net/lists/listinfo/psidev-vocab

information)		trolledVocabulary/psi-ms.obo (No OWL file)			
NMR Spectroscopy (instrument/acquisition parameter/spectrum related information)	NMR-CV	http://nmrml.org/cv/v1.0.rc1/nmrCV.owl	BFO	Creative Commons Public Domain Mark 1.0	https://github.com/nmrML/nmrML/issues?state=open
Laboratory test	**LOINC**	LOINC and RELMA Complete Download File (All Formats Included)	none specified	https://uts.nlm.nih.gov/license.html	wait for Bron 's feedback regarding CDISC lab test descriptors to handle/avoid overlap with LOINC coverage
Medical Imaging	DICOM				

Relations

Scope	Name	File location	Top-Level Ontology	Licence	Issue Tracker URI
Relations	RO	http://purl.obolibrary.org/obo/ro.owl http://purl.obolibrary.org/obo/ro.obo		Creative Commons 3.0 BY	https://code.google.com/p/obo-relations/issues/list

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:

send a mail to <email address to be determine: WP3 term tracker>
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:

- batch term request / programmatic handling:
  - WP3 can channel these request by handling a template for batch submission
  - Class definition could be carried out using Ontomaton Google App in Google Spreadsheet: http://goo.gl/9zsSSI

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 offers annotation functionalities, useful when implementing the eTRIKS terminologies in WP2 and WP4 activities and tools.

Name	URL web interface	Supported Format	Programmatic Access	License
NCBO Bioportal	http://bioportal.bioontology.org	OWL,OBO,RRF	yes	Most of it is under BSD license, parts of it is under the Eclipse Public License
Ontobee	http://www.ontobee.org	OWL	yes	Apache License, Version 2.0
EBI OLS	http://www.ebi.ac.uk/ontology-lookup/	OBO	yes	Apache License, Version 2.0
NCI EVS	http://evs.nci.nih.gov	OWL, RRF	yes	not known
CDISC SHARE	http://cdisc.org/cdisc-share	Excel,XML,RDF,OWL	?	not known
LOV	http://lov.okfn.org/dataset/lov/	RDF	yes	CC-by 3.0 Unported Licence

4.4.5.2 Tools and APIs

These are the commonly used API for manipulating terminology resources:

- Jena library: https://jena.apache.org
- OWLAPI: http://owlapi.sourceforge.net
- OntoCAT: http://www.ontocat.org

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

The following table present 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 report situations where no formal standard exists and where vendor format specification and instrument related files may act as *de factor* exchange format owing to their diffusion and acceptance as container for primary data.

Measurement Category		Technology	Reporting Guideline	Manufacturer	Probe Design	Probe Design File (Annotation File)	Standard Format [Primary Data]	Primary Data Vendor File Format	Standard Format [Derived Data File]
genetic variation	genome wide DNA variation profiling	DNA microarray	**MIAME**	Affymetrix	array design	.CDF file	*<none available>*	.CEL	**.VCF**
	genome wide DNA variation profiling	DNA microarray	**MIAME**	Agilent	array design	.GAL	*<none available>*	agilent feature extraction .txt	**.VCF**
	genome wide DNA variation profiling	DNA microarray	**MIAME**	Illumina	array design	.bpm file, .egt	*<none available>*	.idat	**.VCF**
	targeted DNA variation profiling	DNA microarray	**MIAME**	<miscellaneous>	array design	.GAL	*<none available>*	export to .txt from instrument	**.VCF**
	targeted DNA variation profiling	qRT-PCR	**MIQE**	Applied Biosystems	primer list	*<none available>*	**RDML**	export to .txt from instrument	**.VCF**
	targeted DNA variation profiling	qRT-PCR	**MIQE**	Biorad	primer list	*<none available>*	**RDML**	export to .txt from instrument	**.VCF**
	targeted DNA variation profiling	qRT-PCR	**MIQE**	Roche Applied Science	primer list	*<none available>*	**RDML**	export to .txt from instrument	**.VCF**
	exome sequencing	nucleic acid sequencing	**MINSEQE**	Illumina	exon position list	not applicable	**fastq**		**BAM,BED,BigWIG,BEDgraph**
epigenetic modification	genome wide DNA methylation profiling	DNA microarray	**MIAME**	Affymetrix	array design	.CDF file	*<none available>*	.CEL	**BAM,BED,BigWIG,BEDgraph**
	genome wide DNA	DNA microarray	**MIAME**	Illumina	array design	.bpm file, .egt	*<none available>*	.idat	**BAM,BED,BigWIG,BEDgraph**

	methylation profiling								
	genome wide DNA methylation profiling	DNA microarray	**MIAME**	Nimblegen	array design	.GFF	*<none available>*	.idat	**BAM,BED,BigWIG,BEDgraph**
	targeted DNA methylation profiling	qRT-PCR	**MIQE**	Applied Biosystems	primer list	*<none available>*	**RDML**	export to .txt from instrument	
	targeted DNA methylation profiling	qRT-PCR	**MIQE**	Biorad	primer list	*<none available>*	**RDML**	export to .txt from instrument	
	targeted DNA methylation profiling	qRT-PCR	**MIQE**	Roche Applied Science	primer list	*<none available>*	**RDML**	export to .txt from instrument	
	genome wide DNA methylation profiling	nucleic acid sequencing	**MINSEQE**	Illumina		not applicable	**fastq**		**BAM,BED,BigWIG,BEDgraph**
	histone modification profiling	nucleic acid sequencing	**MINSEQE**	Illumina		not applicable	**fastq**		**BAM,BED,BigWIG,BEDgraph**
	chromatin occupancy profiling	nucleic acid sequencing	**MINSEQE**	Illumina		not applicable	**fastq**		**BAM,BED,BigWIG,BEDgraph**
transcription profiling	global transcription profiling	DNA microarray	**MIAME**	Affymetrix	array design	.CDF file	*<none available>*	.CEL	
	global transcription profiling	DNA microarray	**MIAME**	Agilent	array design	.GAL	*<none available>*		
	global transcription profiling	DNA microarray	**MIAME**	Illumina	array design	.bpm file, .egt	*<none available>*	.idat	
	global transcription profiling	nucleic acid sequencing	**MINSEQE**	Illumina		not applicable	**fastq**		**BAM,BED,BigWIG,BEDgraph**
	targeted transcription profiling	DNA microarray	**MIAME**		array design	.CDF;.GAL	*<none available>*		
	targeted transcription profiling	qRT-PCR	**MIQE**	Applied Biosystems	primer list	<none available>	**RDML**	export to .txt from instrument	
	targeted transcription profiling	qRT-PCR	**MIQE**	Roche Applied Science	primer list	<none available>	**RDML**	export to .txt from instrument	
	targeted transcripti	qRT-PCR	**MIQE**	Biorad	primer list	<none available>	**RDML**	export to .txt from	

	on profiling							instrument	
	miRNA transcription profiling	nucleic acid sequencing	**MINSEQE**	Illumina	GTF file from miRBASE	not applicable	**fastq**		**BAM,BED,BigWIG,BEDgraph**
protein profiling	global protein profiling	mass spectrometry	**MIAPE**			not applicable	**mzML**		**mzIdentML**
	targeted protein profiling	mass spectrometry	**MIAPE**		protein list	<none available>	**mzML**		**mzIdentML**
	targeted protein profiling	protein microarray	**MIAPE + MIAME**		protein list;array design	.GAL	*<none available>*		
	tissue imaging	mass spectrometry	**MIAPE**			.GAL	**imzML**		
metabolite profiling	global metabolite profiling	mass spectrometry	**CIMR**	Bruker		not applicable	**mzML**	.netCDF	*<none available>*
	global metabolite profiling	NMR spectroscopy	**CIMR**	Bruker		not applicable	**NMR-ML**	.fid	*<none available>*
	global metabolite profiling	NMR spectroscopy	**CIMR**	Bruker		not applicable	**NMR-ML**	.acqus	*<none available>*
	targeted metabolite profiling	mass spectrometry	**CIMR**	Bruker	metabolite list	<none available>	**mzML**	.netCDF	*<none available>*
	targeted metabolite profiling	NMR spectroscopy	**CIMR**	Bruker	metabolite list	<none available>	**NMR-ML**	.fid	*<none available>*
	targeted metabolite profiling	NMR spectroscopy	**CIMR**	Bruker	metabolite list	<none available>	**NMR-ML**	.acqus	*<none available>*
	global metabolite profiling	mass spectrometry	**CIMR**	Agilent(Variant)		not applicable	**mzML**	.netCDF	*<none available>*
	global metabolite profiling	mass spectrometry	**CIMR**	Agilent(Variant)		not applicable	**NMR-ML**	.fid	*<none available>*
	global metabolite profiling	NMR spectroscopy	**CIMR**	Agilent(Variant)		not applicable	**NMR-ML**	.propar	*<none available>*
	targeted metabolite profiling	mass spectrometry	**CIMR**	Agilent(Variant)	metabolite list	<none available>	**mzML**	.netCDF	*<none available>*
	targeted metabolite profiling	NMR spectroscopy	**CIMR**	Agilent(Variant)	metabolite list	<none available>	**NMR-ML**	.fid	*<none available>*
	targeted metabolite profiling	NMR spectroscopy	**CIMR**	Agilent(Variant)	metabolite list	<none available>	**NMR-ML**	.propar	*<none available>*

microbial diversity profiling	global microbial diversity profiling	nucleic acid sequencing	**MiXs/MIMARS/MIENS**	Illumina		<none available>	**fastq**		**.BAM**
	targeted microbial diversity profiling	nucleic acid sequencing	**MiXs/MIMARS/MIENS**	Illumina	primer list	<none available>	**fastq**		**.BAM**
	targeted microbial diversity profiling	nucleic acid sequencing	**MiXs/MIMARS/MIENS**	Roche Applied Science	primer list	<none available>	**fastq**	.sff	**.BAM**
cell characterization	cell counting	fluorescent activated cell sorting (FACS)	**MIFlowCyt**	Becton Dickinson	protein list	not applicable	**.FCS**		
	cell sorting	fluorescent activated cell sorting (FACS)	**MIFlowCyt**	EMD millipore	protein list	not applicable	**.FCS**		

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)

- Set of operational guidelines, meaning clear procedure for creating 'data management plans' and 'data validation plans'. This draft of this documents are already quite advanced and eTRIKS WP3 expects to offer a release in the first quarter of 2015 (15Q1)

- 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 eTRISK 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 of eTRIKS studies or IMI studies curated by eTRIKS curation team. While the element of the value-sets could be queried by all, the actual value sets would be accessed controlled in order to preserve any intellectual property.

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.
- *SI units* refer to the International System (SI) of units
- *tranSMART Foundation* (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 label and definition are agree upon
- *Controlled Terminology* is a tree of terms whose label and definition are agree 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 is a reference ontology for defining gene function, molecular process and biological component while Human Phenotype Ontology 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 *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, is an application ontology specifically developed for EMBL-EBI ArrayExpress needs.

. A *Controlled Vocabulary Term (CT)* 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 (eCT)* is a unique CT in the eTRIKS CT library, and has a corresponding identifier and the associated standard source.

· The *eCT library* contains all the eCT used by eTRIKS in eTRIKS.

· *eTRIKS data labels* are eCT.

· *Standardized data* are either eCT 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, 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 subject (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 pre*clinical study* is a type of study where study subjects are animals or 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 all 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 pangenome microarrays, RNA-Seq)

TranSMART:

· *TranSMART* ™ is the data warehouse that eTRIKS will contribute to develop in order to enable data hosting, sustainability, visualization and analysis. Hereafter, TM refers to the TM instance of eTRIKS, unless specified differently.

· 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 Annexes).

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.

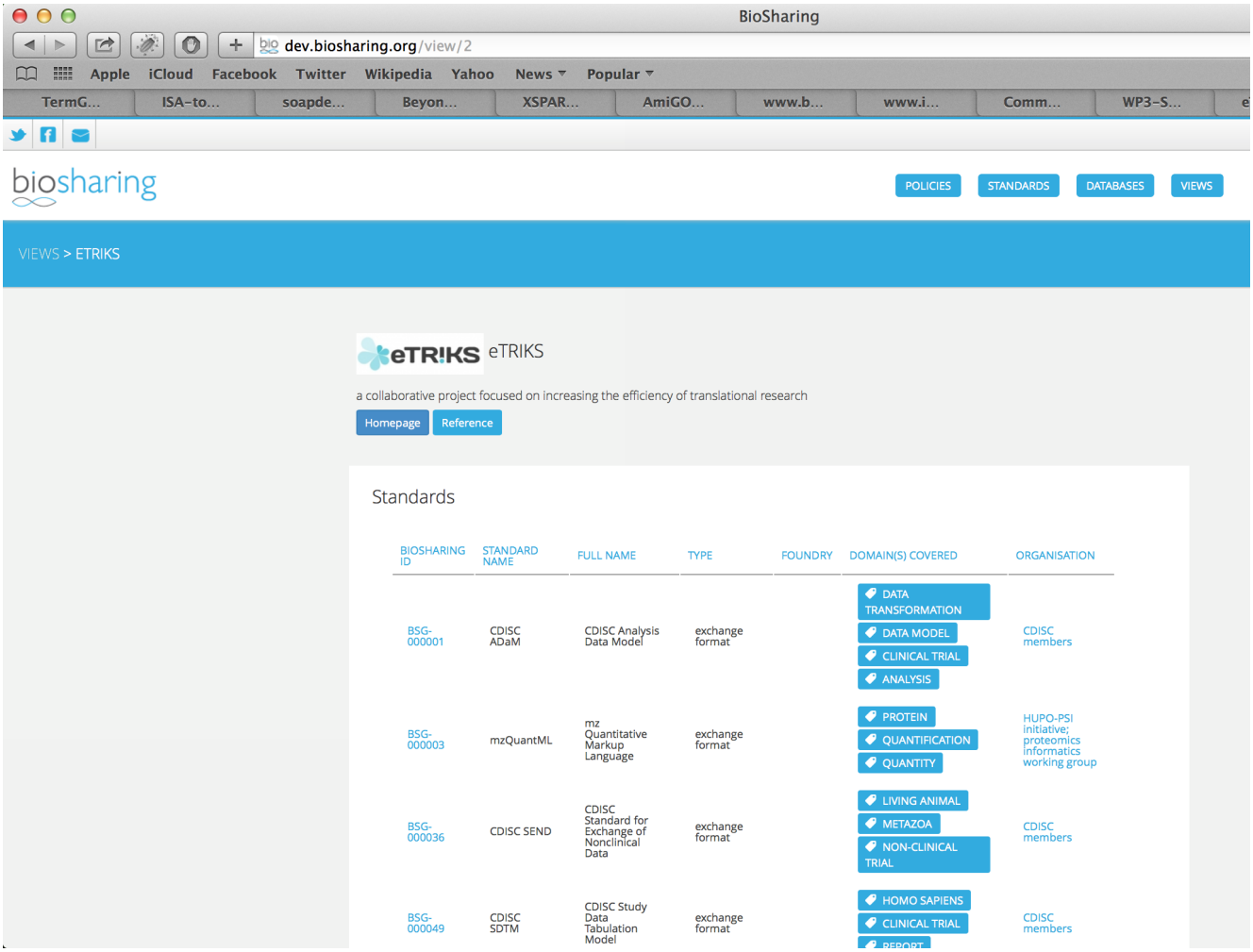

figure1: the eTRIKS view of relevant standards as available from Biosharing website.
http://www.biosharing.org/view/5

A.III. tranSMART master tree

First pass of a recommended hierarchy to use with tranSMART data explorer is provided as part of D3.5

A.IV. Standards currently in use by IMI projects

A review of the initial standards used by supported eTRIKS projects is provided in D3.2