eTRIKS -Standards Starter Pack
Standards Guidelines
Release 1.1 - 25th April 2016

Authors (by alphabetical order)

Bratfalean, Dorina – CDISC Europe Foundation,
Braxenthaler, Michael – Roche Innovation Center New York,
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 Innovation Center Basel,
Sansone, Susanna-Assunta – Oxford e-Research Centre, University of Oxford.

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

Version History

Version	Date	Who	Role	Notes
1.0	11 February 2015	Philippe Rocca-Serra, Fabien Richard, Dorina Bratfalean	Draft creator and maintainer	Draft for internal use
1.1	25 April 2016	Philippe Rocca-Serra	maintainer	Public release

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 may be relevant to them. The work carried out by eTRIKS is meant to be made available to all IMI projects[2] to raise awareness as well as to gain input from specific fields of translational research and further development and refinement of data standards. Furthermore, eTRIKS aims to provide regular updates and releases, to incorporate additions and follow-ups on technology evolution and progress in standardization initiatives. 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 discovering, retrieving, understanding, integrating, disseminating, exchanging and reusing research data.

Annotation resources such as MIAME guidelines[5] or the Gene Ontology[6] controlled vocabularies have become 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 access, distribute, disseminate and exchange

[1] "What is eTRIKS…… |." 2012. 6 Jun. 2015 <http://www.etriks.org/>
[2] "Innovative Medicines Initiative: Home | IMI." 2007. 6 Jun. 2015 <http://www.imi.europa.eu/>
[3] "NIH Data Sharing Policy - National Institutes of Health." 2002. 6 Jun. 2015 <http://grants.nih.gov/grants/policy/data_sharing/>
[4] "An essential guide to open access for Wellcome Trust ..." 2011. 8 Jun. 2015 <http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/WTVM050569.pdf>
[5] Brazma, Alvis et al. "Minimum information about a microarray experiment (MIAME)—toward standards for microarray data." *Nature genetics* 29.4 (2001): 365-371.
[6] Ashburner, Michael et al. "Gene Ontology: tool for the unification of biology." *Nature genetics* 25.1 (2000): 25-29.
[7] "CDISC | Strength Through Collaboration." 6 Jun. 2015 <http://www.cdisc.org/>
[8] "ISA Tools." 2005. 6 Jun. 2015 <http://www.isa-tools.org/>

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 embrace efforts aiming at enhancing access to information so it can be efficiently mined, analyzed and exploited.

To understand the impact of the standardization process, one should consider the roadmap set out by the European Medicines Agency for implementing a series of ISO Standards for the identification of medicinal products. The process is now widely known as the IDMP [9]. It is an extremely significant milestone in standardization process as, starting from July 2016, Commission Implementing Regulation (EU) No 520/2012External link icon (articles 25 and 26) obliges Member States, marketing-authorisation holders and EMA to make use of the terminologies defined in ISO IDMP standards. In practice, this means that submission to EMA, will be made using the HL7 SPL format based on the ISO IDMP standards, ISO IDMP technical specifications and HL7 common product model, a combination of standardization tools designed to supercede the eXtended EudraVigilance Product Report Message (XEVPRM) format.

Standards are developed to ensure scientific information is represented ***consistently***, ***efficiently*** and ***meaningfully*** to the benefit of the community. It is expected that scientists and science stakeholders will have greater confidence when standard compliant datasets are available, ensuring data analysis and reuse in the longer term are made more straightforward. 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 in Life Science[10], and this results in wasting significant amounts of resources in organizations worldwide, and, consequently, impairs/slows down the scientific research and the development of new drugs and biomarkers for patients. The lack of adequate reporting standards adversely affects the overall quality of available data. Therefore efforts in standardization of data, metadata and experimental/clinical reports in Life Science represent significant endeavour at rectifying this issue.

With the present document, IMI eTRIKS aims to bring about endorsement of the **FAIR principles** for data, that is to make data **'Findable, Accessible, Interoperable and Reusable'**, as

[9] "Implementation of the ISO IDMP standards - European ..." 2015. 14 Mar. 2016 <http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000645.jsp&mid=WC0b01ac058078fbe2>

[10] Mobley, A. "A Survey on Data Reproducibility in Cancer Research ..." 2013. <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0063221>

outlined by the Force 11 group[11]. These principles are being adopted by a growing number of stakeholders, from publishers (e.g. NPG Scientific Data (http://www.nature.com/sdata/about/principles), to Funders and Repositories (e.g. Dryad[12], Figshare[13]) to support data publication.

Finally, The entire set fo eTRIKS recommended resources can be accessed in a browseable, searcheable form from the Biosharing catalogue of standards and resources from https://biosharing.org/collection/5?q=.

[11] "The FAIR Data Principles - FOR COMMENT | FORCE11." 2014. 6 Mar. 2016 <https://www.force11.org/group/fairgroup/fairprinciples>
[12] "Dryad Digital Repository - Dryad." 2008. 10 Mar. 2016 <http://datadryad.org/>
[13] "figshare - credit for all your research." 2012. 10 Mar. 2016 <https://figshare.com/>

Part 1. Introduction

1.1 eTRIKS mission and objectives

eTRIKS aims to provide a 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. Recommendations are needed to provide guidance in wide array of specifications available, produced both by academics and standard development organizations (SDO).

Among the goals of eTRIKS are :

- Standard harmonization for data annotation. Common list of eTRIKS-selected and recommended standards for data owners, curators and consumers.
- 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 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.
- **data quality officers** responsible for ensuring procedures are adhered to and data meet expected grade.
- **data curators** to enable coordinated and agreed upon data cleanup and edition to eTRIKS annotation and curation guidelines,
- **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 commonly shared and accepted 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 and/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 Uniform 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, usually in non-formal ways, the necessary and sufficient entities to describe a domain. eTRIKS-adopted or created MIGs will specify which exchange formats and vocabulary standards are to be used. Those content standards ensure the information exchange based on community shared meaning (semantics); they include data and metadata standards. Vocabularies are often treated separately, but they are a form of content standards and are a prerequisite to support semantic alignment. A standard may also refer to an integration profile, an implementation guide or a user guide***.**

2. ***vocabularies***; these include a variety of terminologies, possibly multi-lingual, such as controlled vocabularies,—dictionaries/thesauri or ontologies that describe their 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 major purposes are described in more details below. They are therefore essential elements for ensuring delivering **FAIR datasets**.

Accessibility & 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. The notion of accessibility includes issues related to data protection and patient privacy. Therefore, information governing access permission, patient consent and encryption need to be described in standardized ways.

Interoperability: To enable operational processes which underlie data exchange and sharing between different software systems. Two distinct facets of interoperability need to be addressed simultaneously to reach efficiency. One dimension covers the syntactic alignment,

which can be viewed as a more technical layer. The second dimension concerns itself with the meaningfulness of interoperation, something designated as 'semantic interoperability'.

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[14] 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.

Part 2. Procedure for standards selection and recommendation

2.1 Procedure outline

As recommended by the eTRIKS Standards Advisory Board (as of January 28th, 2014), the selection and use of standards should be as 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

[14] "What Is Provenance - XG Provenance Wiki." 2010. 19 Jun. 2015 <http://www.w3.org/2005/Incubator/prov/wiki/What_Is_Provenance>

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 criteria are used to assess whether to adopt a standard, and which one where competing standards exists.

2.2 Attributes of standards

Following is a list of attributes and criteria for selecting a standard suitable for use by eTRIKS.

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.

Relevance/Applicability: 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

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

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

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

Authority: 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) . When ontologies are concerned, it means assessing the nature and accuracy of relationships between terms needs to be assessed.

Readability: The standard is available in human and machine readable formats. Hence, availability of resources if format such as RDF, OWL, SKOS can be used as metric.

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

Note: the order is not an indication of importance.

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 influences 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 but the current recommendations are not prescriptive. It is therefore left to data managers to decide on which resources to used and provide all the necessary details.

2.3 Versioning of Standards

Standards are artefacts resulting from activities by dedicated bodies, acting in a specific realm to deliver normative documents. This process is often iterative, dealing with use cases depending on pragmatic prioritization and obeying release cycles. It is of paramount important to understand that Standards evolve and adapt to new needs and therefore often undergo alteration, extensions and incremental modifications that warrant the release of updated normative specifications on a regular or ad-hoc basis. Therefore, Standards should always be identified by their name, their version number, the date of the latest release, and, if available, a Unique Resource Identifier (URI). 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.

2.4. 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) that have been considered so far, include:

- W3C[15]
- ISO[16]
- CDISC[17]
- HL7[18]
- WHO[19]
- OBO foundry[20]

Vocabulary servers

- US National Center for Biomedical Ontologies: Bioportal[21]
- US National Cancer Institute Enterprise Vocabulary Services: NCI EVS[22]
- US Unified Medical Language System: UMLS Terminology Server (requires license agreement and account)
- EMBL-EBI Ontology Lookup Service[23]
- Open Knowledge Foundation Linked Open Vocabulary: LOV[24]
- CDISC SHARE API (under development): http://www.cdisc.org/cdisc-share

[15] "World Wide Web Consortium (W3C)." 19 Jun. 2015 <http://www.w3.org/>
[16] "ISO Standards - ISO." 2012. 6 Jun. 2015 <http://www.iso.org/iso/home/standards.htm>
[17] "CDISC | Strength Through Collaboration." 6 Jun. 2015 <http://www.cdisc.org/>
[18] "HL7." 6 Jun. 2015 <http://www.hl7.org/>
[19] "World Health Organization: WHO." 6 Jun. 2015 <http://www.who.int/>
[20] "The Open Biological and Biomedical Ontologies." 2006. 6 Jun. 2015 <http://www.obofoundry.org/>
[21] "Welcome to the NCBO BioPortal | NCBO BioPortal." 2008. 6 Jun. 2015 <http://bioportal.bioontology.org/>
[22] "Welcome to EVS — EVS." 2006. 6 Jun. 2015 <http://evs.nci.nih.gov/>
[23] Côté, Richard G et al. "The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries." *BMC bioinformatics* 7.1 (2006): 97.
[24] "Linked Open Vocabularies." 2012. 8 Jun. 2015 <http://lov.okfn.org/>

Catalogue of Standards and Resources in Life Sciences:

- Biosharing[25]

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[26]). 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 [27].

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.

---

[25] "BioSharing: policies, standards and communication in ..." 2011. 8 Jun. 2015 <http://precedings.nature.com/collections/biosharing>

[26] "Study Data Tabulation Model (SDTM) | CDISC." 2009. 19 Jun. 2015 <http://www.cdisc.org/sdtm>

[27] "CDISC Training Campus." 2012. 19 Jun. 2015 <http://cdisc.trainingcampus.net/>

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 [28])

For this reason, eTRIKS Standards Work Package 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.

Part 3. Standards in data management

3.1 Standards for Data Security, Data Privacy and Compliance with Ethical Guidelines.

While the main focus of the eTRIKS Standards Starter Pack aims at documenting and advising content and annotation standards, the object of the present section is to draw attention to normative guidelines and procedures dedicated to ensure proper handling of clinical data.

3.1.1 Data Access and Security

When it comes to patient and clinical data, the Information Security Standard ISO 27000 family (http://www.iso.org/iso/home/standards/management-standards/iso27001.htm) should be considered as a reference point to establish an industry strength implementation of secured data access and data encryption. The standards specify reliance on two-factor authentication processes for every interaction session with the system. It also specifies appropriate measures for data access logging and audit.

[28] "Coalition For Accelerating Standards and Therapies (CFAST)." 9 Jun. 2015 <http://blogs.fda.gov/fdavoice/index.php/tag/coalition-for-accelerating-standards-and-therapies-cfast/>

3.1.2 Data Privacy and Anonymization

To preserve patient privacy and effectively remove the risk of patient re-identification, data managers must be aware of the latest recommendations for anonymization. Recommendations by BioMedbridges [29] [30] and eTRIKS WP7 [31] on ethics should be used as reference documentation.

3.1.3 Patient consent and Ethical use of research information.

Whether clinical trial data or electronic health records are involved, it is essential for data managers to clearly identify and document the modalities of recording consent information provided by patients as well as the extent of usage the patients have agreed to. As noted before, there are widespread disparities between countries in Europe and across the world on that topic. One can therefore only refer to current normative documents produced by standardization bodies or regulatories agencies. For the former, the HL7 V3 Security and Privacy Ontology, Release 1[32] FHIR, an effort towards a patient friendly, machine readable consent description, should be considered. For the latter, the US Food and Drug Administration maintains significant documentation and the 21CFR50.20[33] general requirements provide a framework but also the IMI eTRIKS code of practice on secondary use of medical data[34] which has been widely adopted in Europe and US.

3.2 Principles of good annotation practice

In order to enable within study consistency and ultimately, cross-study queries and/or comparisons and achieve good query recall, many concepts need to be standardized. 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:

---

[29] Sariyar, M. "Sharing and Reuse of Sensitive Data and Samples ..." 2015. <http://www.ncbi.nlm.nih.gov/pubmed/26186169>

[30] "Supporting researchers sharing sensitive data: identifying ..." 2016. 11 Mar. 2016 <https://www.biomedbridges.eu/supporting-researchers-sharing-sensitive-data-identifying-requirements>

[31] Bahr, A. "Code of practice on secondary use of medical data in ..." 2015. <http://idpl.oxfordjournals.org/content/early/2015/09/19/idpl.ipv018.abstract>

[32] "HL7 Version 3 Standard: Security and Privacy Ontology ..." 2014. 10 Mar. 2016 <http://www.hl7.org/implement/standards/product_brief.cfm?product_id=348>

[33] "Search for FDA Guidance Documents > A Guide to Informed ..." 2009. 6 Mar. 2016 <http://www.fda.gov/RegulatoryInformation/Guidances/ucm126431.htm>

[34] "Code of Practice | eTRIKS." 2015. 10 Mar. 2016 <https://www.etriks.org/code-of-practice/>

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, RNA seq, 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[35] 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?

[35] Bureau International des Poids et Mesures, Commission électrotechnique internationale, and Organisation internationale de normalisation. *Guide to the Expression of Uncertainty in Measurement*. International Organization for Standardization, 1995.

- 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 separately for each study.
The following steps guide a curator towards choosing the most 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

3.3 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" and as now mandated by IMI2 and H2020 programs[36].

3.4 Retrospective data capture and 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 Controlled Vocabulary Terms (CVT).

[36] "Guidelines on Data Management in Horizon 2020." 2016. 14 Mar. 2016 <https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf>

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.

3.5 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[37] will be reported in another document.

[37] Bel, EH et al. "Unbiased Biomarkers for the Prediction of Respiratory Disease Outcome (U-BIOPRED) Consortium, Consensus Generation. Diagnosis and definition of severe refractory asthma: an international consensus statement from the Innovative Medicine Initiative (IMI)." *Thorax* 66.10 (2011): 910-7.

Part 4. eTRIKS recommended resources

This section points to dedicated and specific documents and specifications that detail further eTRIKS recommendations as to which standards may be used in Data Management Plans (for instance, as described in H2020 data management document)

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 [38]. 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*.

CDISC Standards	Uses/Value	Application
Foundational Models		
Protocol Representation Model (PRM), Study Design Model (SDM) http://www.cdisc.org/protocol	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	Planning

[38] "Business Case for Standards | CDISC." 2010. 22 Jun. 2015 <http://www.cdisc.org/business-case>

The PRM toolkit gives 30 basic concepts essential for all protocols and is more easily understandable than the full UML model.	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 reuse. 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. 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.	
Clinical Data Acquisition Standards Harmonization (CDASH)	Clinical Data Acquisition Standards Harmonization is a specification data collection domains and variables for Case Report Form (CRF) data with standard question text, implementation guidelines, and best practices.	Planning/Data Collection

http://www.cdisc.org/cdash		
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 contract research organizations (CROs). Vocabulary standard that facilitates exchange of clinical trial laboratory data between central laboratories and study sponsors, CROs or EDC vendors.	Planning/Data Collection
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 model and its implementation guidelines 3.2 highlight new specific finding domains: extensions for microbiology specimen, microbiology susceptibility and pharmacogenomics data.	Data Tabulation / Preparation to Preservation (and Regulatory Submission if needed)
Analysis Dataset Model (ADaM) http://www.cdisc.org/adam	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.	Data Tabulation / Preparation to Preservation (and Regulatory Submission if needed)
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	Preparation to Preservation (and Regulatory

	mechanism for providing eSubmissions metadata to FDA.	Submission if needed)
Study Design in XML http://www.cdisc.org/study-trial-design	The CDISC Study Design Model in XML (SDM-XML) version 1.0 allows organizations to provide rigorous, machine-readable, interchangeable descriptions of the designs of their clinical studies, including treatment plans, eligibility and times and events. As an extension to the existing CDISC Operational Data Model (ODM) specification, SDM-XML affords implementers the ease of leveraging existing ODM concepts and re-using existing ODM definitions. SDM-XML defines three key sub-modules – Structure, Workflow, and Timing – permitting various levels of detail in any representation of a clinical study design, while allowing a high degree of authoring flexibility	Preparation to Preservation (and Regulatory Submission if needed)
Dataset-XML http://www.cdisc.org/dataset-xml	Dataset-XML, released for comment under the name "StudyDataSet-XML", and renamed to avoid confusion with the CDISC SDS team, is a new standard used to exchange study datasets in an XML format. Dataset-XML supports the interchange of tabular data for clinical research applications using ODM-based XML technologies. The Dataset-XML model is based on the CDISC Operational Data Model (ODM) standard and should follow the metadata structure defined in the CDISC Define-XML standard.	Preparation to Preservation (and Regulatory Submission if needed)
Semantics		
Controlled Terminology http://www.cdisc.org/terminology	The controlled standard vocabulary and code sets for all of the CDISC models/standards; maintained openly and freely by NCI Enterprise Vocabulary Services (EVS: http://evs.nci.nih.gov/)).	Annotation
Glossary http://www.cdisc.org/cdisc-glossary	Glossary with definitions of acronyms and terms commonly used in clinical research. Abbreviations and Acronyms also included.	Annotation

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.	Annotation
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.	Annotation
Specialty 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..	Preparation to Preservation (and Regulatory Submission if needed)
Medical Devices http://www.cdisc.org/devices	The Study Data Tabulation Model Guide for Medical Devices (SDTMIG-MD) v.1.0 defines recommended standards for the submission of data from clinical trials in which medical devices are used. The document includes seven new domains, developed by a team comprised of medical device experts, CDISC experts, and the FDA (CDRH and CBER), and represents years of work by the members of the CDISC Medical Device team. Training on these seven new domains has been incorporated into the standard SDTM training available through CDISC	Preparation to Preservation (and Regulatory Submission if needed)
Pharmacogenomics/Genetics (PGx) http://www.cdisc.org/pharmacogenomics-genetics	Version 1.0 Provisional of the SDTM Implementation Guide: Pharmacogenomics/Genetics (SDTMIG-PGx), published 2015-06-01, describes standards to guide the organization, structure and format of gene-related tabulation datasets submitted as part of a product application to a regulatory agency. SDTMIG-PGx is being released for provisional use, since it introduces new	Preparation to Preservation (and Regulatory Submission if needed)

	variables and constructs that will be added to the forthcoming SDTM v1.5, and to allow operational testing and evaluation of this new standard by the CDISC user community.	
Specialty Areas http://www.cdisc.org/specialty-areas	Solution Kits for Specialty Areas describe how to use CDISC Foundational Standards to represent content for specific types of trials or certain broad categories of data. For example, the SDTM Implementation Guides for Medical Devices and Pharmacogenomics/Genetics both began as new specialty area projects before being formally released as independent standards. Some of these specialty areas are very similar to therapeutic area data standards, but they are not governed under the CFAST initiative so are currently listed separately (though in some cases them may evolve into CFAST projects at a later date.	Preparation to Preservation (and Regulatory Submission if needed)
Questionnaires, Ratings and Scales (QRS) http://www.cdisc.org/qrs	Questionnaires are named, stand-alone measures designed to provide an assessment of a concept. Questionnaires have a defined standard structure, format, and content; consist of conceptually related items that are typically scored; and have documented methods for administration and analysis. Questionnaires consist of defined questions with a defined set of potential answers. Most often, questionnaires have as their primary purpose the generation of a quantitative statistic to assess a qualitative concept. CDISC publishes standard Questionnaires, Ratings and Scales (QRS) products, which include SDTM Annotated CRFs and Supplements to the SDTMIG along with the related controlled terminology.	Regulatory Submission / Tabulation

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)[39] 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.

[39] Lehmann, Sabine et al. "Standard preanalytical coding for biospecimens: Review and implementation of the Sample PREanalytical Code (SPREC)." *Biopreservation and biobanking* 10.4 (2012): 366-374.

4.2 eTRIKS - Recommendations for Exchange Format for Non-Clinical Studies (Animal and *in-vitro* Studies)

4.2.1 CDISC Standards for non-clinical studies

Standards Document	Uses/Value
Laboratory Data Model (LAB) http://www.cdisc.org/lab	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.
Standard for the Exchange of non-Clinical Data (SEND) http://www.cdisc.org/send	An extension of SDTM specifically developed for preclinical or non-clinical studies, e.g. toxicology.
Controlled Terminology http://www.cdisc.org/terminology	The controlled standard vocabulary and code sets 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.

4.2.2 Non-Regulatory Standards for Research Studies

Standards Document	Uses/Value
Investigation Study Assay	'Investigation' (the overall project context which may group several studies), 'Study' (a defined research experiment why may use several

http://.isatab.sf.net	different types of assays) and 'Assay' (sets of data acquisition events) 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.
Primary Data Format for Omics	The following link provides a complete overview of the existing format specifications available to support individual 'omic like type of data. Section 4.5 eTRIKS-WP3-Standard-Starter-Pack-Recommandations-Exchange-Format-for-Omics

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 prefered 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.

Molecular Entity	Resource Name	Biosharing identifier	Resource URI	Resource Identifier pattern	Comment
Small Molecules					
Metabolites	**Pubchem**	**biodbcore-000455**	http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=$id	$id=^\d+$	
	CHEBI	**bsg-000039**	http://www.ebi.ac.uk/chebi/searchId.do?chebiId=$id	$id=^CHEBI:\d+$	
Lipids	**Lipid Maps**	**biodbcore-000559**	http://www.lipidmaps.org/data/get_lm_lipids_dbgif.php?LM_ID=$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	$id=^DB\d{5}$	
	WHOdrug (*)	Not available	http://www.umc-products.com/DynPage.aspx?id=73588&mn1=1107&mn2=1139		(*)WHOdrug is not freely available and its cost can be a major limitation for academic institutions.
Biopolymer					
DNA	**ensEMBL gene**	**biodbcore-000330**	http://www.ensembl.org/	$id=ENSG\d+$	
	Entrez Gene (aka NCBI Gene)	**biodbcore-000449**	http://www.ncbi.nlm.nih.gov/gene/$id	$id=^\d+$	
messenger RNA	**ensEMBL transcript**	**biodbcore-000330**	http://www.ensembl.org/	$id=ENST\d+$	
microRNA	**mirbase**	**biodbcore-000569**	http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=$id	$id=MI\d{7}	
Protein	**Uniprot**	**biodbcore-000544**	http://www.uniprot.org	$id=^([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	**biodbcore-000448**	http://www.ncbi.nlm.nih.gov/protein/$id	$id=^(\w+\d+(\.\d+)?)\|(NP_\d+)$	
DNA variant (**)					
SNP	**NCBI dbSNP**	**biodbcore-000438**	http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=$id	$id=^rs\d+$	Human Genome Variation Guidelines for annotation and nomenclature (http://www.hgvs.org/mutnomen/) (used by CDISC PGX extension)
Structural Variation	**NCBI ClinVar**	**biodbcore-000739**	http://www.ncbi.nlm.nih.gov/clinvar/		Human Genome Variation Guidelines for annotation and nomenclature (http://www.hgvs.org/mutnomen/) (used by CDISC PGX extension)

(**) Consider Locus Reference Genomic (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.

Molecular Entity	Resource Name	Biosharing identifier	Resource URI	Resource Identifier pattern
antibodies	antibody-registry	**biodbcore-000182**	http://antibodyregistry.org/AB_$id	$id=^\d+{6}$
Plasmids and vectors	addgene	**biodbcore-000196**	www.addgene.org/$id	$id=^'\d+$

cell lines	ATCC	biodbcore-000210	http://www.lgcstandards-atcc.org/Products/All/$id.aspx	$id=^'\d+$

4.3.1.3 Important Resources for Describing Medical Devices

In March 2016, the United States Foods and Drug Administration, in collaboration with the US National Library of Medicine released the 'accessGUDID' database of medical devices, with the aim of providing a consistent and standard way to identify medical devices throughout their distribution and use by health care providers and patients.

Entity	Resource Name	Biosharing identifier	Resource URI	Resource Identifier pattern
Medical Device	GUDID	biodbcore-000748	https://accessgudid.nlm.nih.gov/devices/search?query=$id	$id=^\d+{14}$

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[40] 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)

[40] "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/>

- 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)

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

Scope	Name	Biosharing Identifier	File location	Top-Level Ontology	Licence	Issue Tracker URI	Comment
Organism	NCBITaxonomy	bsg-000154	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		

Vertebrate Anatomy	UBERON	bsg-000016	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*
Strain	Rat Strain Ontology	bsg-002625	ftp://rgd.mcw.edu/pub/ontology/rat_strain/	none specified	not available		Species specific resource

4.4.3.2 Phenotype and Diseases

The following table summarizes major, generic and well established semantic resources for which constitute central elements for describing and designated pathologies and their signs. The same convention is used to layout the information, with a specific highlight on **licensing** terms as well as **regulatory requirements**. These are deemed critical information about the resources, of relevance when defining data management plans.

Scope	Name	Biosharing identifier	File location	Top-Level Ontology	Licence	Issue Tracker URI	Comment
Pathology Disease (generic)							
	NCI thesaurus	bsg-000154	http://evs.nci.nih.gov/ftp1/NCI_Thesaurus	none specified	http://evs.nci.nih.gov/ftp1/NCI_Thesaurus/ThesaurusTermsofUse.htm		**Use Mandated by FDA for regulatory submissions**
	SNOMED-CT	bsg-000098	not available	none specified	http://www.ihtsdo.org/licensing/ Refer to this page for the full details of terms of use and fees: http://www.ihtsdo.org/snomed-ct/get-snomed-ct		**Use Recommended by FDA for regulatory submissions**
	ICD-10	bsg-000274	**login required** [http://apps.who.int/classifications/apps/icd/ClassificationDownloadNR/login.aspx?ReturnUrl=%2fclassifications%2fapps%2ficd%2fClassificationDownload%2fdefault.aspx]	none specified	http://www.who.int/about/licensing/classifications/en/ If your organization is planning to use WHO classifications for non-commercial or research purposes, then you may qualify for a licence for non-commercial research use, register here: http://goo.gl/4ueZpI For commercial use, register here: http://goo.gl/2D7H1s		

	UMLS	not available	not available	none specified	http://www.nlm.nih.gov/databases/umls.html		
	Disease Ontology	biodbcore-000025	http://purl.obolibrary.org/obo/doid.owl	BFO	CC-by 3.0 Unported Licence	http://sourceforge.net/p/diseaseontology/feature-requests/	
	Infection Disease Ontology	bsg-000095	https://code.google.com/p/infectious-disease-ontology/source/browse/trunk/src/ontology/ido-core/ido-main.owl	BFO	CC-by 3.0 Unported Licence	https://code.google.com/p/infectious-disease-ontology/issues/list	
Phenotype	Human Phenotype Ontology	bsg-000131	http://compbio.charite.de/hudson/job/hpo/lastStableBuild/	BFO	CC-by 3.0 Unported Licence	https://github.com/obophenotype/human-phenotype-ontology/issues/	
	Mammalian Phenotype	bsg-000129	ftp://ftp.informatics.jax.org/pub/reports/mp.owl			https://github.com/obophenotype/mammalian-phenotype-ontology/issues	
	Phenotypic Quality Ontology (PATO)	bsg-000134	https://raw.githubusercontent.com/obophenotype/pato/master/pato.owl	BFO		https://github.com/pato-ontology/pato/issues/	
	MedDRA	bsg-002647	not available		Free for academic and other non-commercial uses. Commercial use of MedDRA requires obtaining a license from MSSO.	https://msstools.com/webcr/ Login required	**Use Mandated by FDA for regulatory submissions**

4.4.3.3 Pathology and Disease Specific Resources

The following table lists several terminology resources specifically focused around a particular pathology, from pathogen induced disorders to orphan disease, often of genetic origin.

The content of this component will evolve in line with the progress of development occurring in the CDISC Therapeutic Area domains.

Scope	Name	Biosharing identifier	File location	Top-Level Ontology	Licence	Issue Tracker URI
Influenza	FLU	bsg-000094	http://www.berkeleybop.org/ontologies/flu.owl	BFO	BSD license clause 4	
Malaria	IDOMAL	bsg-000104	http://www.berkeleybop.org/ontologies/idomal.owl	BFO	not available	
Rare disorder	ORDO	bsg-002716	http://www.orphadata.org/data/ORDO/ordo_orphanet.owl.zip	none	Attribution-NoDerivs 3.0 Unported	

Note: This section dedicated to specific disease area will be expanded as eTRIKS and IMI projects come together as well as in accordance to progress made under CDISC Therapeutic areas according to the CFAST (http://www.cdisc.org/cfast) initiative.

4.4.3.4 Cellular entities

Non-clinical studies make heavy use of cellular systems. The rise of cell based therapies will only reinforce the need for reliable identification and description to the cells, their origins, their genotypic and phenotypic properties. The following semantic artefacts provide central resources. It should be however noted that these are under active development and important coverage gaps exist. Those will be addressed following user requests and as the communities coalesce in cooperative organizations.

Scope	Name	biosharing identifier	File location	Top-Level Ontology	Licence	Issue Tracker URI
Cell	CL	bsg-000009	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	bsg-002627	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
	Cellosaurus	**http://web.expasy.org/cellosaurus/**	ftp://ftp.expasy.org/databases/cellosaurus	None specified	None specified	
Cell Molecular Phenotype Ontology	CMPO	not available	https://github.com/EBISPOT/CMPO/tree/master/release	BFO	Apache License version 2	http://www.ebi.ac.uk/cmpo/submit

4.4.3.5 Molecular Entities

The following section is dedicated to identifying major terminology efforts aimed at providing controlled terms for denoting *molecular entities and their properties*, from structure to function or effects on biological systems, as well as their classification and use for pharmacology and therapeutic applications.
The resources listing below encompass naturally occurring molecules as well as man-made, synthetic molecules, from small molecules to macromolecules.
The table below highlight semantic resources and complements the table found in section 4.3.1.1 , which listed databases of molecular entities. Records in those databases may be annotated with the terminologies presented below.

Scope	Name	biosharing identifier	File location	Top-Level Ontology	Licence	Issue Tracker URI	Commnet
Unique Ingredient Identifier	UNIII		http://www.fda.gov/ForIndustry/DataStandards/SubstanceRegistrationSystem-UniqueIngredientIdentifierUNII/default.htm	*none*			**Use Mandated by FDA for regulatory submissions**
Chemicals and Small Molecules	CHEBI	**bsg-000039**	http://ftp.ebi.ac.uk/chebi.owl http://ftp.ebi.ac.uk/chebi.obo	BFO	most probably: CC-by 3.0 Unported Licence	https://github.com/ebi-chebi/ChEBI/issues	
Drug	National Drug File	bsg-002592	http://www.pbm.va.gov/NationalFormula	*none*	**Mind the terms:** https://uts.nlm.nih.		

			ry.asp		gov/license.html		
Gene Function, Molecular Component, Biological Process	GO	bsg-000089	http://purl.obolibrary.org/obo/go.obo http://purl.obolibrary.org/obo/go.owl	BFO	CC-by 4.0 Unported License	https://github.com/geneontology/go-ontology/issues/	
Protein/peptide	PRO	bsg-000139	http://ftp.pir.georgetown.edu/pro.obo	BFO	CC-by 3.0 Unported Licence	http://sourceforge.net/p/pro-obo/term-requests/	

4.4.3.6 Assays and Technologies

Biological signals are acquired through a range of techniques, each requiring specific instruments, settings, data processing and quantitation description. The following table aggregates community vetted and regulatory agency approved (for some) resources assembling terminologies to describe unambiguously analytical and experimental techniques used in research.

Scope	Name	biosharing identifier	File location	Top-Level Ontology	Licence	Issue Tracker URI	Comment
Therapeutic Device	**GUDID**						
Laboratory test (clinical context)	**LOINC**	bsg-000106	LOINC and RELMA Complete Download File (All Formats Included)	none specified	**Mind the terms:** https://uts.nlm.nih.gov/license.html		**Use Mandated by FDA for regulatory submissions**
Sample Processing/Reagents/Instruments **Assay Definition (non-clinical assay,experimental test)**	OBI	bsg-000070	http://purl.obolibrary.org/obo/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) (non-clinical, in-vitro)	BAO	bsg-002687	http://www.bioassayontology.org/bao/bao_complete_bfo_dev.owl	BFO	CC-by 3.0 Unported Licence	https://github.com/BioAssayOntology/BAO/issues	
Experimental Design, Statistical Methods and Statistical Measures	STATO	bsg-000548	http://purl.obolibrary.org/obo/stato.owl	BFO	CC-by 3.0 Unported Licence	https://github.com/ISA-tools/stato/issues?state=open	
Radiology	RADLex	bsg-002633	http://data.b	none specified	RadLex License	http://radlex.org/su	

			ioontology.org/ontologies/RADLEX/submissions/31/download?apikey=8b5b7825-538d-40e0-9e9e-5ab9274a9aeb		Version 2.0 Open source/ free use	ggest_term/index.cfm	
Mass Spectrometry (instrument/acquisition parameter/spectrum related information)	PSI-MS	bsg-000068	http://psidev.cvs.sourceforge.net/viewvc/psidev/psi/psi-ms/mzML/controlledVocabulary/psi-ms.obo (No OWL file)	none specified	CC-by 3.0 Unported Licence	https://lists.sourceforge.net/lists/listinfo/psidev-vocab	
NMR Spectroscopy (instrument/acquisition parameter/spectrum related information)	NMR-CV	bsg-000564	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	
Medical Imaging	**DICOM [ISO 12052:2006]**	bsg-000114	http://medical.nema.org/medical/dicom/current/output/pdf/part06.pdf	none specified			
Experimental Factor Ontology	**EFO**	bsg-000082	http://www.ebi.ac.uk/efo/efo.owl	BFO	www.apache.org/licenses/LICENSE-2.0	https://www.ebi.ac.uk/panda/jira/secure/CreateIssue!default.jspa	Application Ontology

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.

Scope	Name	Biosharing identifier	File location	Top-Level Ontology	Licence	Issue Tracker URI
Relations	RO	bsg-000144	http://purl.obolibrary.org/obo/ro.owl	BFO	Creative Commons 3.0 BY	https://github.com/oborel/obo-relations/issues

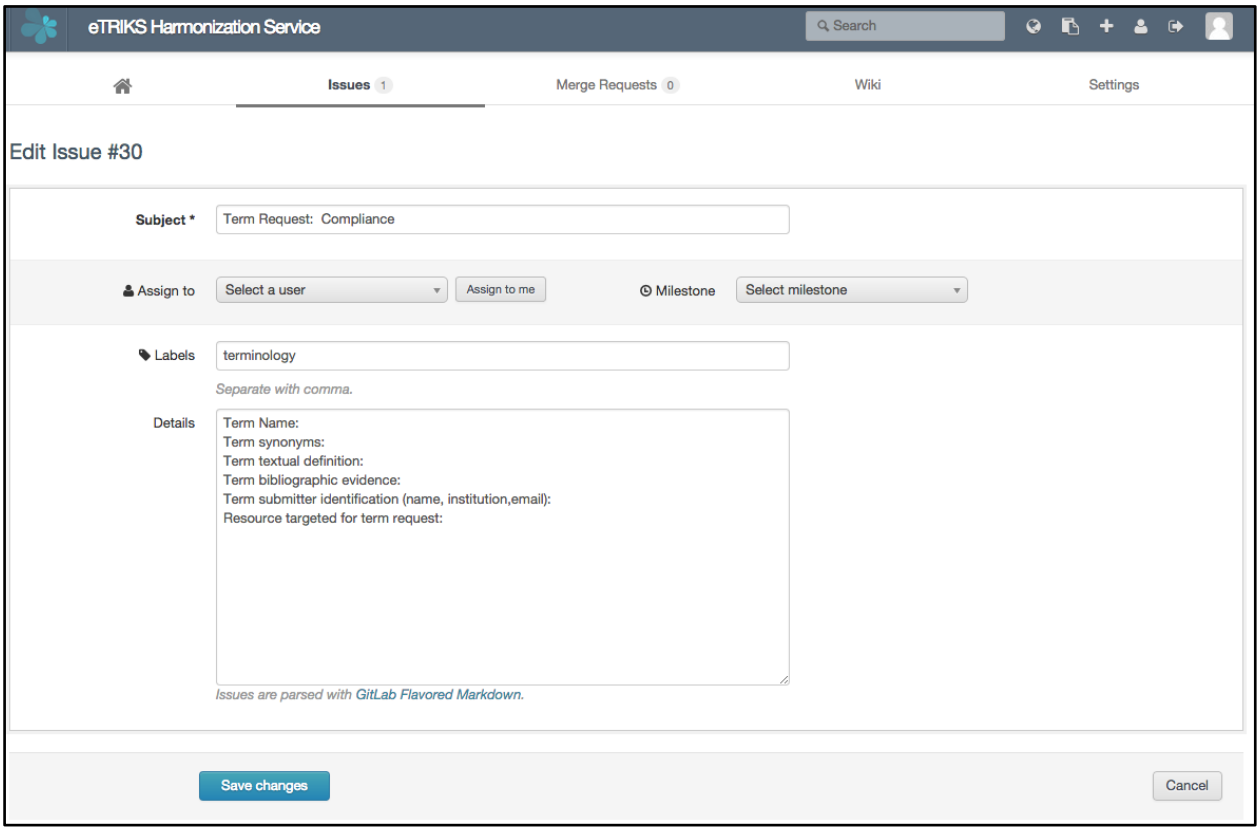
4.4.4 Brokering Requests for New Terms

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 eTRIKS Standards 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 Ontomaton Google App[41] in Google Spreadsheet: http://goo.gl/9zsSSI according to templating procedure[42].

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.

[41] Maguire, Eamonn et al. "OntoMaton: a Bioportal powered ontology widget for Google Spreadsheets." *Bioinformatics* (2012): bts718.
[42] Rocca-Serra, Philippe et al. "Overcoming the ontology enrichment bottleneck with quick term templates." *Applied Ontology* 6.1 (2011): 13-22.

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
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	all documents (PDF, XML, OWL can be made available for download)	not known
Ontobee	http://www.ontobee.org	OWL	yes	Apache License, Version 2.0
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 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.

Measurement Category	Assay Name	Technology	Reporting Guideline	Maker	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,BigWIG, BEDgraph
epigenetic modification	genome wide DNA methylation profiling	DNA microarray	MIAME	Affymetrix	array design	.CDF file	***<none available>***	.CEL	.BAM,BigWIG, BEDgraph
	genome wide DNA methylation profiling	DNA microarray	MIAME	Illumina	array design	.bpm file, .egt	***<none available>***	.idat	.BAM,BigWIG, BEDgraph
	genome wide DNA methylation profiling	DNA microarray	MIAME	Nimblegen	array design	.GFF	***<none available>***	.idat	.BAM,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,BigWIG, BEDgraph
	histone modification profiling	nucleic acid sequencing	MINSEQE	Illumina		not applicable	fastq		.BAM,BigWIG, BEDgraph
	chromatin occupancy profiling	nucleic acid sequencing	MINSEQE	Illumina		not applicable	fastq		.BAM,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,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 transcription profiling	qRT-PCR	MIQE	Biorad	primer list	<none available>	RDML	export to .txt from instrument	
	miRNA transcription profiling	nucleic acid sequencing	MINSEQE	Illumina	GTF file from miRBASE	not applicable	fastq		.BAM,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 spectroscropy	CIMR	Bruker		not applicable	NMR-ML	.fid	*<none available>*
	global metabolite profiling	NMR spectroscropy	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 spectroscropy	CIMR	Bruker	metabolite list	<none available>	NMR-ML	.fid	*<none available>*
	targeted metabolite profiling	NMR spectroscropy	CIMR	Bruker	metabolite list	<none available>	NMR-ML	.acqus	*<none available>*
	global metabolite	mass spectrometry	CIMR	Agilent(Variant)		not applicable	mzML	.netCDF	*<none available>*

	profiling								
	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/MIENS	Illumina		<none available>	fastq		.BAM
	targeted microbial diversity profiling	nucleic acid sequencing	MiXs/MIENS	Illumina	primer list	<none available>	fastq		.BAM
	targeted microbial diversity profiling	nucleic acid sequencing	MiXs/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:

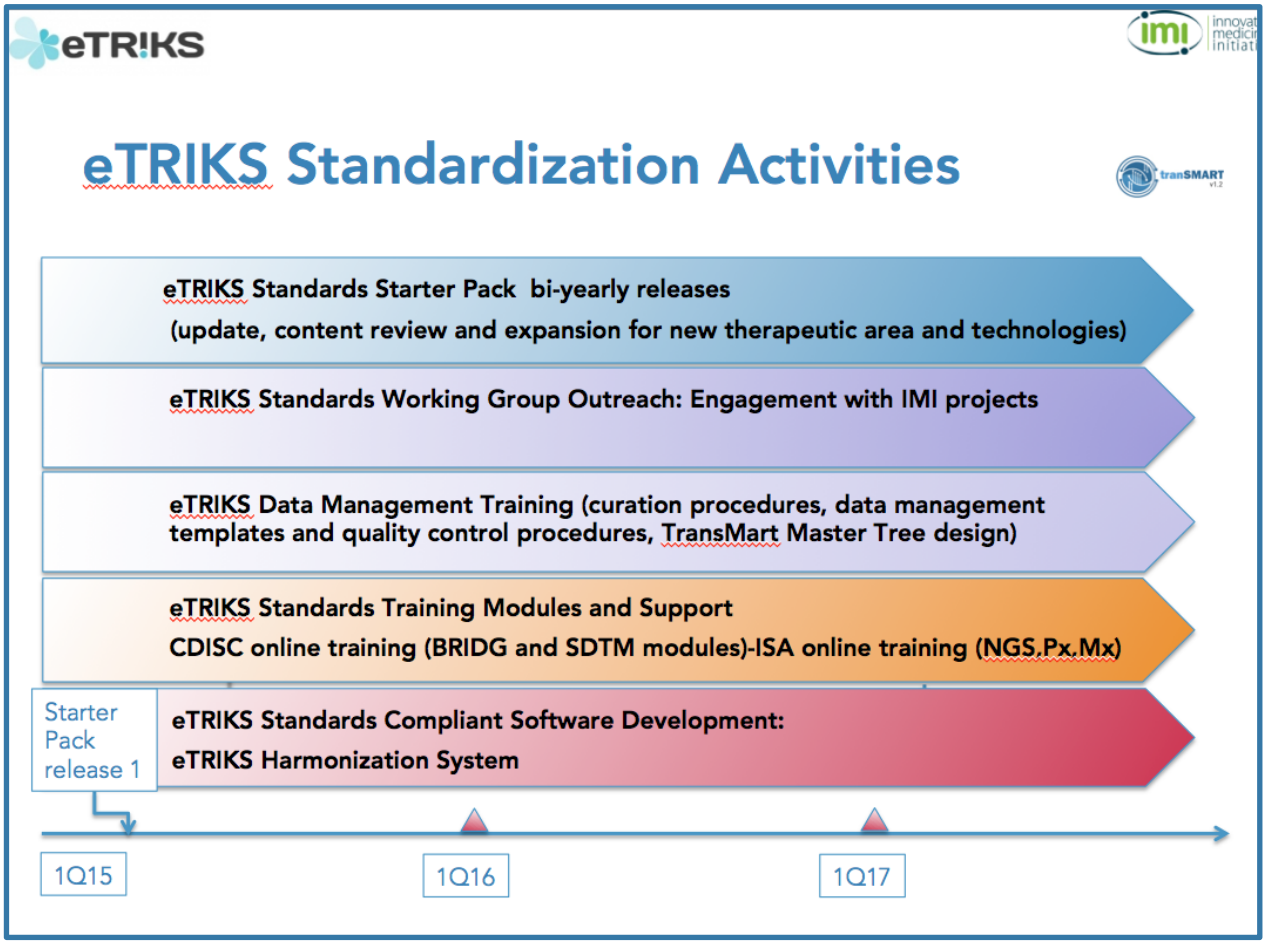
- a list of recommendations about relevant data standards to translational research (the eTRIKS Standard Starter Pack). The eTRIKS Standard Starter Pack will undergo annual updates to reflect the evolution and progress of standardization initiatives. For instance, CDISC releases Therapeutic Areas (TA) standards regularly as disease domains reach maturity. In the field of genomics, the Global Alliance for Genomic Health[43] 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 and 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.

[43] "Global Alliance for Genomics and Health: Home." 2014. 19 Jun. 2015 <http://genomicsandhealth.org/>

A graphical overview of the roadmap is presented below.



The coloured 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.[44]
- *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.

[44] "Home - The Cancer Genome Atlas - Cancer Genome - TCGA." 2005. 8 Jun. 2015 <http://cancergenome.nih.gov/>

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[45] is a reference ontology for defining gene function, molecular process and biological component while Human Phenotype Ontology[46] 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[47], 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.

[45] "Gene Ontology Consortium." 2002. 8 Jun. 2015 <http://geneontology.org/>
[46] "The Human Phenotype Ontology." 2008. 8 Jun. 2015 <http://www.human-phenotype-ontology.org/>
[47] "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, 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 pre*clinical 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.

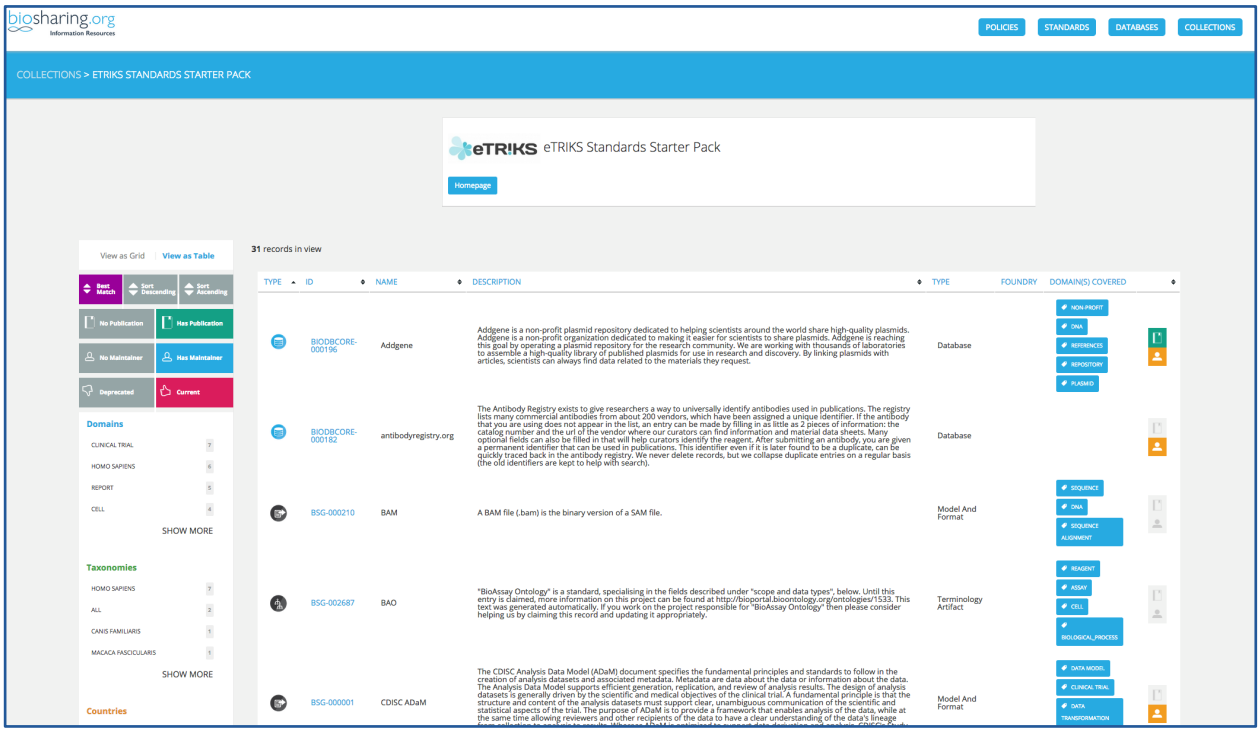
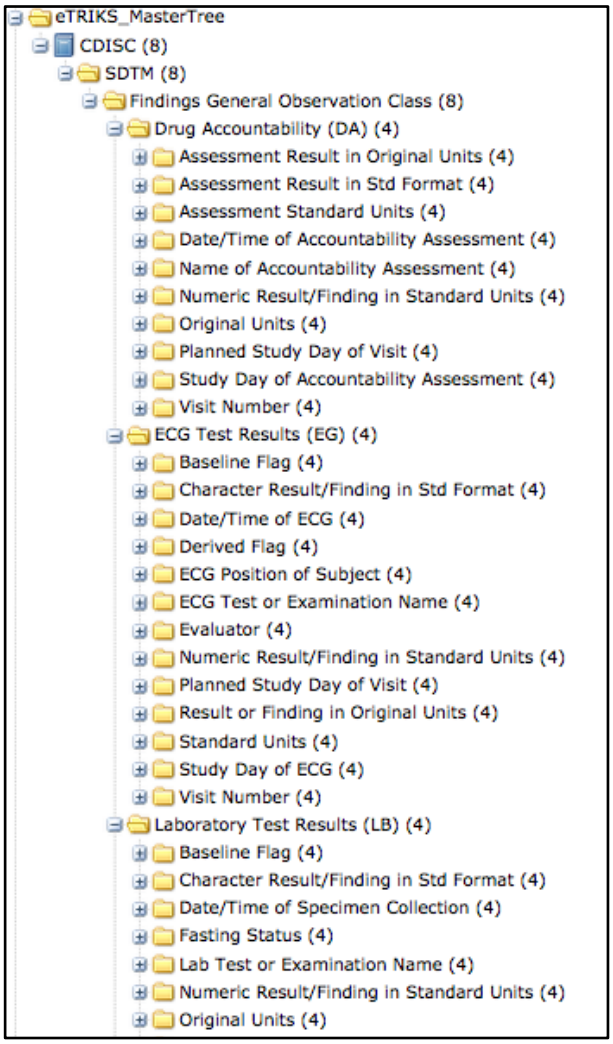


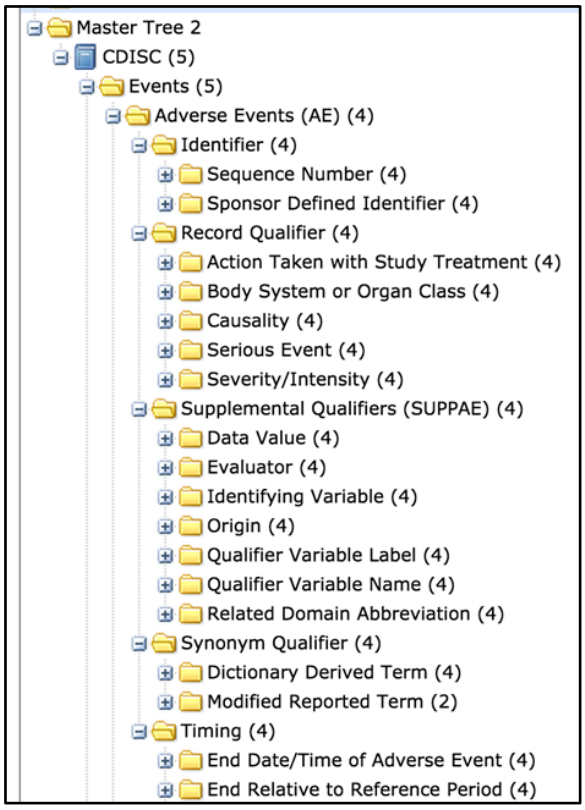
Figure 1. The eTRIKS view of relevant standards as available from Biosharing website.
https://www.biosharing.org/collection/5?q=&view=table

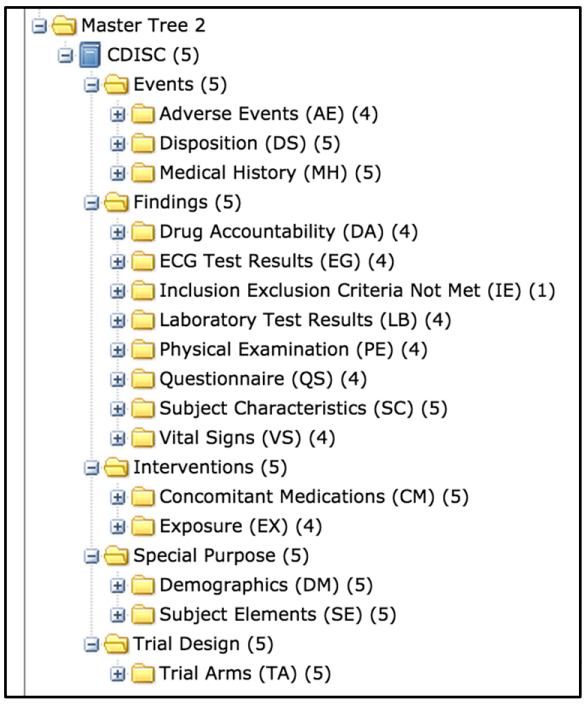
A.III. tranSMART master tree

Left pane: First pass of a recommended hierarchy to use for tranSMART data explorer

Right pane: **2 views of a CDISC SDTM based TransMART master tree data explorer.** The lower pane expands the 'CDISC STDM Adverse Event Domain showing possible descriptors used in the CDISC Tabulation format.







Bibliography

STANDARDS

Kubick, Wayne R, Stephen Ruberg, and Edward Helton. "Toward a comprehensive CDISC submission data standard." *Drug information journal* 41.3 (2007): 373-382.

Souza, Tammy, Rebecca Kush, and Julie P Evans. "Global clinical data interchange standards are here!." *Drug discovery today* 12.3 (2007): 174-181.

Rebecca D. Krush; Current status and future scope of CDISC standards; Clinical Journal; Clinical Data Interchange Standards Consortium, October, (2012).(url: http://www.cdisc.org/system/files/all/article/application/pdf/current_status___future_scope_of_cdisc_standards_kush.pdf)

Andrea Vadakin, Rebecca D. Krush; CDISC standards and Innovations; Clinical Evaluation 40, Suppl XXXI; 217-28, (2012) (url: http://www.cdisc.org/system/files/all/article/application/pdf/cdisc_standards_and_innovations_vadakin__kush.pdf)

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)

Dolin, Robert H et al. "HL7 clinical document architecture, release 2." *Journal of the American Medical Informatics Association* 13.1 (2006): 30-39.

Sansone, Susanna-Assunta et al. "Toward interoperable bioscience data." *Nature genetics* 44.2 (2012): 121-126.

Rocca-Serra, Philippe et al. "ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level." *Bioinformatics* 26.18 (2010): 2354-2356.

Leinonen, Rasko, Hideaki Sugawara, and Martin Shumway. "The sequence read archive." *Nucleic acids research* 39 (Database issue) (2011): D19–D21.

Martens, Lennart et al. "mzML—a community standard for mass spectrometry data." *Molecular & Cellular Proteomics* 10.1 (2011): R110. 000133.

Sansone, Susanna-Assunta et al. "Agenda of the “BioSharing workshop–Bringing catalogues of bio-resource and standards together”." 713 (2011).

Gardner, Daniel et al. "The neuroscience information framework: a data and knowledge environment for neuroscience." *Neuroinformatics* 6.3 (2008): 149-160.

Larson, Stephen D, and Maryann E Martone. "NeuroLex. org: an online framework for neuroscience knowledge." *Frontiers in neuroinformatics* 7 (2013).

Bandrowski, A et al. "Exploring mammalian brain connectivity using NeuroLex." *Front. Neuroinform. Conference Abstract: 5th INCF Congress of Neuroinformatics. doi: 10.3389/conf. fninf* 2014.

Mobley, A.; Linder, S. K.; Braeuer, R.; Ellis, L. M.; Zwelling, L. (2013). "A Survey on Data Reproducibility in Cancer Research Provides Insights into Our Limited Ability to Translate Findings from the Laboratory to the Clinic". In Arakawa, Hirofumi. PLoS ONE 8 (5): e63221. doi:10.1371/journal.pone.0063221. PMC 3655010

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.

McDonald, Clement J et al. "LOINC, a universal standard for identifying laboratory observations: a 5-year update." *Clinical chemistry* 49.4 (2003): 624-633.

Smith, Barry et al. "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration." *Nature biotechnology* 25.11 (2007): 1251-1255.

Mungall, Christopher J et al. "Uberon, an integrative multi-species anatomy ontology." *Genome Biol* 13.1 (2012): R5.

Degtyarenko, Kirill et al. "ChEBI: a database and ontology for chemical entities of biological interest." *Nucleic acids research* 36.suppl 1 (2008): D344-D350.

Natale, Darren A et al. "The Protein Ontology: a structured representation of protein forms and complexes." *Nucleic acids research* 39.suppl 1 (2011): D539-D545.

Bard, Jonathan, Seung Y Rhee, and Michael Ashburner. "An ontology for cell types." *Genome biology* 6.2 (2005): R21.

Sarntivijai, Sirarat et al. "Cell Line Ontology: Redesigning the Cell Line Knowledge base to Aid Integrative Translational Informatics." *ICBO* 833 (2011).

Meehan, Terrence F et al. "Logical development of the cell ontology." *BMC bioinformatics* 12.1 (2011): 6.

Robinson, Peter N et al. "The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease." *The American Journal of Human Genetics* 83.5 (2008): 610-615.

Ashburner, Michael et al. "Gene Ontology: tool for the unification of biology." *Nature genetics* 25.1 (2000): 25-29.

Botstein, David et al. "Gene Ontology: tool for the unification of biology." *Nat Genet* 25.1 (2000): 25-29.

Brinkman, Ryan R et al. "Modeling biomedical experimental processes with OBI." *J. Biomedical Semantics* 1.S-1 (2010): S7.

Schürer, Stephan C et al. "BioAssay ontology annotations facilitate cross-analysis of diverse high-throughput screening data sets." *Journal of biomolecular screening* 16.4 (2011): 415-426.

Malone, James et al. "Modeling sample variables with an Experimental Factor Ontology." *Bioinformatics* 26.8 (2010): 1112-1118.

Rustici, Gabriella et al. "ArrayExpress update—trends in database growth and links to data analysis tools." *Nucleic acids research* 41.D1 (2013): D987-D990.

Gkoutos, Georgios V, Paul N Schofield, and Robert Hoehndorf. "The Units Ontology: a tool for integrating units of measurement in science." *Database* 2012 (2012): bas033.

Brown, Elliot G, Louise Wood, and Sue Wood. "The medical dictionary for regulatory activities (MedDRA)." *Drug Safety* 20.2 (1999): 109-117.

Wishart, David S et al. "DrugBank: a comprehensive resource for in silico drug discovery and exploration." *Nucleic acids research* 34.suppl 1 (2006): D668-D672.

Kong, Jun et al. "Integrative, multimodal analysis of glioblastoma using TCGA molecular data, pathology images, and clinical outcomes." *Biomedical Engineering, IEEE Transactions on* 58.12 (2011): 3469-3474.

Brennan, Cameron W et al. "The somatic genomic landscape of glioblastoma." *Cell* 155.2 (2013): 462-477.

Taylor, Chris F et al. "Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project." *Nature biotechnology* 26.8 (2008): 889-896.

Brazma, Alvis et al. "Minimum information about a microarray experiment (MIAME)—toward standards for microarray data." *Nature genetics* 29.4 (2001): 365-371.

Taylor, Chris F et al. "The minimum information about a proteomics experiment (MIAPE)." *Nature biotechnology* 25.8 (2007): 887-893.

Bustin, Stephen A et al. "MIQE precis: Practical implementation of minimum standard guidelines for fluorescence-based quantitative real-time PCR experiments." *BMC molecular biology* 11.1 (2010): 74.

Lee, Jamie A et al. "MIFlowCyt: the minimum information about a Flow Cytometry Experiment." *Cytometry Part A* 73.10 (2008): 926-930.

Yilmaz, Pelin et al. "Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications." *Nature biotechnology* 29.5 (2011): 415-420.

Tenenbaum, Jessica D, Susanna-Assunta Sansone, and Melissa Haendel. "A sea of standards for omics data: sink or swim?" *Journal of the American Medical Informatics Association* 21.2 (2014): 200-203.

Lehmann, Sabine et al. "Standard preanalytical coding for biospecimens: Review and implementation of the Sample PREanalytical Code (SPREC)." *Biopreservation and biobanking* 10.4 (2012): 366-374.

VOCABULARY SERVERS

Whetzel, Patricia L et al. "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications." *Nucleic acids research* 39.suppl 2 (2011): W541-W545.

Maguire, Eamonn et al. "OntoMaton: a Bioportal powered ontology widget for Google Spreadsheets." *Bioinformatics* 29(4) (2013): 525-7.

Côté, Richard G et al. "The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries." *BMC bioinformatics* 7.1 (2006): 97.

Xiang, Zuoshuang et al. "Ontobee: A Linked Data Server and Browser for Ontology Terms." *ICBO* 26 Jul. 2011. (url: http://www.ontobee.org/Ontobee_ICBO-2011_Proceeding.pdf)

De Coronado, Sherri et al. "NCI Thesaurus: using science-based terminology to integrate cancer research results." *Medinfo* 11.Pt 1 (2004): 33-7.

Adamusiak, Tomasz et al. "OntoCAT--simple ontology search and integration in Java, R and REST/JavaScript." *BMC bioinformatics* 12.1 (2011): 218.

Vasilevsky, Nicole A et al. "On the reproducibility of science: unique identification of research resources in the biomedical literature." *PeerJ* 1 (2013): e148.

TRANSMART

Szalma, Sándor et al. "Effective knowledge management in translational medicine." *Journal of translational medicine* 8.1 (2010): 68.

Athey, Brian D et al. "tranSMART: an open source and community-driven informatics and data sharing platform for clinical and translational research." *AMIA Summits on Translational Science Proceedings* 2013 (2013): 6.