European Translational Information and Knowledge Management Services

eTRIKS Deliverable report

Grant agreement no. 115446

Deliverable 3.5

Manual of guidelines and recommendations for adoption of standards in eTRIKS v3

Due date of deliverable: Month 30

Actual submission date: Month 31

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

DELIVERABLE INFORMATION

Project	
Project acronym:	eTRIKS
Project full title:	European Translational Information and Knowledge Management Services
Grant agreement no.:	115446
Document	
Deliverable number:	D3.5
Deliverable title:	Manual of guidelines and recommendations for adoption of standards in eTRIKS v3
Deliverable version:	1
Due date of deliverable:	Month 30
Actual submission date:	Month 31
Leader:	
Editors:	
Authors:	Philippe Rocca-Serra, Fabien Richard, Dorina Bratfalean
Reviewers:	Serge Eiffes, Paul Houston, Chris Marshall
Participating beneficiaries:	
Work Package no.:	3
Work Package title:	Standards Research and Coordination
Work Package leader:	Michael Braxenthaler, Paul Houston
Work Package participants:	
Estimated person-months for deliverable:	3 pm
Nature:	Document
Version:	1
Draft/Final:	Final
No of pages (including cover):	30
Keywords:	Guidelines, recommendations, Data Standards

Purpose

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

This document describes the input from WP3 Standards group to the design of the eTRIKS v3.0 Platform. It describes the three principal requirements for eTRIKS v3.0 in order for it to support the implementation of consistent data standards for data curators and uploaders.

Intended audience

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

Contents

Overview

Three significant issues are outstanding to support the consistent and compatible uploading of data into tranSMART that is an essential prerequisite for sharing data within or between projects.

1) eTRIKS MetaData Registry

   A mechanism to ensure that consistent terms are used for entities, and where this is not possible due to the restrictions of the original data set, a mechanism is available to align terms for identical entities with different names.

2) eTRIKS Data Curation Guidelines

   Guidelines to ensure that the MetaData Registry and Master Data Tree mechanisms are consistently applied to all data as it is curated.

3) eTRIKS tranSMART Master Data Tree

   A mechanism to guide the placing of data entities on the tranSMART data tree to ensure the data tree can be quickly and easily compared between studies and navigated by researchers unfamiliar with the specific data set.

eTRIKS MetaData Registry - Use Cases and Functional Requirements

Background:

The work of eTRIKS work package 3 on standards identified a gap in the infrastructure being built to manage IMI clinical and non-clinical studies. Namely, the absence of a dedicated architecture supporting a 'Data as a Service' approach, to enable vocabulary lookup, metadata curation and data validation.

Therefore, eTRIKS work package 3 decided to prepare a use-case and functional requirements document necessary to drive the development and establishment of an eTRIKS MetaData Registry (eMDR).

eTRIKS WP3 considers this piece of infrastructure essential to the success of standards implementation and data harmonization in the eTRIKS project and therefore requires corresponding investment to deliver the capability. Failing to do so would severely impair possibilities of effective progress towards better cross study searching, exploration and analysis.

Use cases:

1. Serve study variables and controlled terminologies vetted by eTRIKS

- Allow data annotation tools to access controlled vocabulary terms (or Coding Terms (CT))
- Allow curators to perform entity recognition and tagging
- Allow curators and users to log CT request

2. Create standard compliant data submission templates

   - Devise tabular data acquisition template aligned with MIGs
   - Ensure good practice in data capture

3. Create 'study metadata signature' database for eTRIKS and IMI submissions:
   A 'study metadata signature' is a study-specific set of free text terms (variables or their contents) that are mapped to CTs.

   Note: The access to the 'study metadata signature' will be restricted to study owners, while the access to individual mappings of free text terms to CTs will be publically available.

Task/Functional Requirements

A. Core Function / Requirements

1. **Metadata Model:**
   - Reliance on standard compatible metadata models (ISO-11179 Metadata Registry, W3C-RDF) or define a metamodel for metadata in order to allow nature, function, position and dependencies associated to individual tokens of metadata.

2. **Vocabulary Server:**
   - Support metadata loading from the following resources
     i. Models/representations (e.g. CDISC model, BRIDG model (http://www.bridgmodel.org) as Protocol-driven research and its associated regulatory artifacts
     ii. Ontologies expressed in OBO format, OWL, RIF, SKOS format
     iii. XSD supporting xml based standards
     iv. ISA configurations as MIGS for omics techniques

3. **Metadata Library with access control and user based restriction**
   - Grant access to all individual metadata tokens
   - Restrict access of study-specific metadata sets to study owner until made public. This last functionality enables the creation of a 'study fingerprint' that remains under the ownership of the study submitter unless specified otherwise.

B. Key Capabilities consuming data from the eTRIKS MetaData Registry

1. **Annotation and Term Tagging Tool**

    - Similar to NCBO Annotator, enable the creation of 'term mapping files".

2. **Extract Transform Load Transformation file creation**

    - Similar to ISA Mapping tool, to facilitate the creation of curation files, holding submitter labels and values, as well as their mapping to Standard compliant variables and any business rule applied for cleaning.

3. **Curation File Repository**

    - Allow storage of "term mapping files", files resulting from the mapping from submitted variables to eTRIKS vetted terminologies along with the substitution rules. NOTE: this is by product of the 'data curation process' and should be preserved for provenance purpose, quality control evaluation but also as starting templates/stubs for future curation elements.

4. **Data Submission Template Generator**

    - Based on study type, eTRIKS MIGs, generate template for data input templates.

5. **Formalization of 'minimal information guidelines**

    - Anchoring those to well established laboratory and bioassay terminologies (SIO, BAO, OBI...). In practise this is a difficult capability to achieve, but successfully implemented, it will allow capturing of consistent ontology patterns that can be used for data compliance validation.

6. **Study Fingerprint Database**

    - Study metadata signature database for eTRIKS and IMI submissions: A 'study metadata signature' is a study-specific set of free text terms (variables or their contents) that are mapped to CTs.

The diagram presented in Figure 1 illustrates the overview of Meta Data Registry mechanism

Overview:

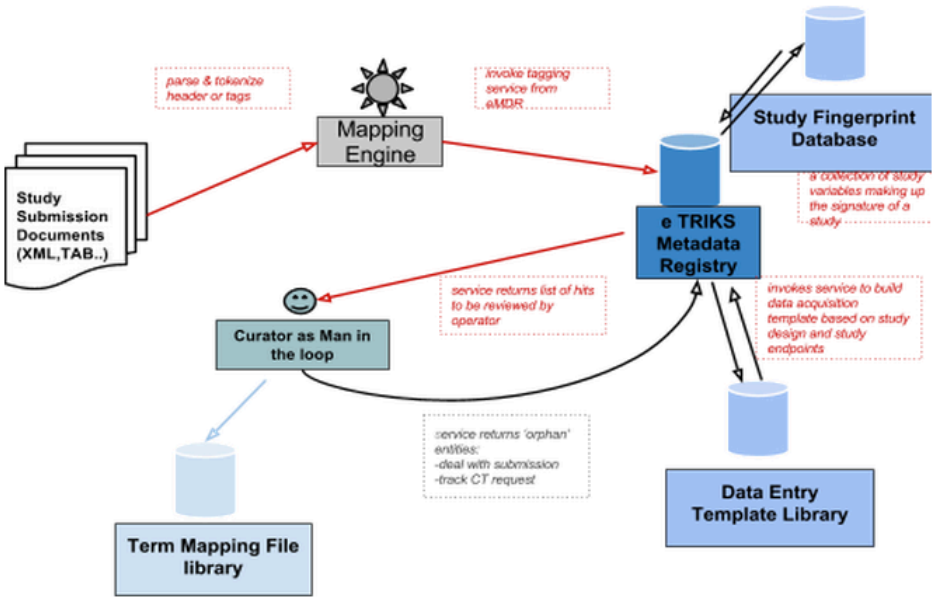

Figure 1. Overview of Meta Data Registry

eTRIKS Data Curation Guidelines

Purpose of this section

The objective of this section is to outline the eTRIKS principles and recommendations on the good practices of data annotation as well as the curation process that eTRIKS is putting in place as part of eTRIKS v3.0.

The goal is to deliver an operational document that is straightforward to read and navigate for collaborators and curators.

eTRIKS strongly recommends to follow its guidelines whenever possible to facilitate data persistence and integration into eTRIKS platform.

This is organized into the following sections:

- Section A: provides definitions and meanings of terms related to the data curation process. The document uses only the term meanings described in this section. The readers should not consider other possible meanings that are not described in that section in order to avoid misunderstanding.

- Section B: provides a stepwise description of the eTRIKS operational curation process for a quick and easy read.

- Section C: details the step-by-step operational procedures, showing inputs and outputs at each step as well as examples.

- Section D: provide a stepwise description of the eTRIKS tranSMART Master Tree data as a CDISC fashion.

A. Terms and definitions

Data curation is a group of information management activities needed to ensure long-term preservation of **research data thus enabling their mining, integration and analyses across projects**. **These activities, also known as extract, transform, load (ETL) activities, deal with cleansing, conversion, standardization, and formatting data.**

Table 1 below presents the acronyms and definitions which will be dedicated as vocabulary terms used within the eTRIKS curation processes.

Table 1eTRIKS acronym & definition

Full name	**Acronym/ Abbreviation**	**Definition**
European Translational Information and Knowledge Management Services	eTRIKS	eTRIKS, an IMI consortium, provides a platform and services that enable data integration and translational medicine.
tranSMART	TM	A data warehouse that eTRIKS will contribute to develop in order to enable data hosting, visualization and analysis. Hereafter, TM refers to the TM instance of eTRIKS, unless specified otherwise.
Extract, Transform, Load	ETL	A process in data warehousing that Extracts data from outside sources, Transforms them to fit operational needs, which can include quality levels, and Loads it into a data warehouse (e.g. TM).
electronic data capture	EDC	A tool for clinical data collection in which the data manager is able to build clinical database by creating, developing and testing electronic case report forms (eCRFs).
electronic case report forms	eCRFs	An electronic document designed to record all of the protocol-required information to be reported to the sponsor on each trial subject.
user interface*	UI	In information technology, the **user interface**(UI) is everything designed into an information device with which a human being may interact -- including display screen, keyboard, mouse, light pen, the appearance of a desktop, illuminated characters, help messages, and how an application program or a Web site invites
Investigation Study Assay	ISA	ISA is format specification for reporting multi-omics experiments. More about the project can be found at htttp://www.isa-tools.org
Minimum Information Guideline*	MIG	MIGs are prepared by standards and development team (WP3) of eTRIKS project (if they do not exist already), and show what minimal information is required to fully describe the provenance of data. (disambiguation note: Homonyme warning :do not confuse Model Implementation Guidelines as found for CDISC documents)

The Cancer Genome Atlas	TCGA	The Cancer Genome Atlas (TCGA) is a comprehensive and coordinated effort to accelerate our understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing. TCGA is a joint effort of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), two of the 27 Institutes and Centers of the National Institutes of Health, U.S. Department of Health and Human Services.
International System*	SI	A complete, coherent system of units used for scientific work, in which the fundamental quantities are length, time, electric current, temperature, luminous intensity, amount of substance, and mass.
Uniform Resource Name*	URN	URN is a persistent, globally unique name assigned to an object. In contrast to a URL, which changes whenever the location of an object changes, a URN has no location dependence and therefore a longer lifetime. This is realized by using a naming service which in most cases will provide a mapping from URNs to URLs. Thus, even if the URL of an object changes, its URN remains the same, since only the object's entry in the naming service has to be updated.
Uniform Resource Locator*	URL	Is a reference to a **resource** that specifies the location of the **resource** on a computer network and a mechanism for retrieving it. A URL is a specific type of **uniform resource** identifier.
Clinical Data Interchange Standards Consortium	CDISC	CDISC is a global, open, multidisciplinary, non-profit organization that has established standards to support the acquisition, exchange, submission and archive of clinical research data and metadata. The CDISC mission is to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare. CDISC standards are vendor-neutral, platform-independent and freely available via the CDISC website
TranSMART Master Tree*	TM MT	A reference tree built to represent and structure data, based on the data domains defined by CDISC, in order to accommodate consistently the persistence and display of studies relevant to translational research in the TranSMART platform (see an example of a TM tree in Annexes).
Quality Control	QC	Operational techniques and activities undertaken within the quality assurance system to verify that the requirements for quality of the trial-related activities have been fulfilled
Standard Operating Procedure*	SOP	A detailed instruction written to achieve uniformity of the performance of a specific function.

Study data management plan	SDMP	A document which describes procedures for organizing and controlling research data per study including activities and responsibilities for collection and curation.
Good Clinical Practice	GCP	Good clinical practice (GCP) is an international quality standard which follows the International Conference on Harmonisation (ICH) of GCP guidelines. GCP enforces tight guidelines on ethical aspects of a clinical study.

Table 2 below presents the listed terms and definitions used in curation processes

Table 2 eTRIKS operational procedure

Term		**Definition**
WorkFlow diagram		A graphic meant to depict the steps or activities that constitute a process
Clinical protocol		A document that describes the objective(s), design, methodology, and statistical considerations of a study Clinical protocols are protocols covering study design with human subjects. In animal or in-vitro studies (preclinical context), the study design is analogous to clinical protocol in human studies. In other words, both describe how to conduct the study.
Laboratory protocol*		Experimental protocols are specifications detailing laboratory processing (for example mRNA extraction).
Protocol amendment *		A written description of a change (or changes) to, or formal clarification of, a protocol.
Specification document*		A document that states the requirements to which a given product or service should conform
Statistical Analysis Plan	SAP	The aim of the Statistical Analysis Plan is to minimise bias by clearly stating the proposed methods of dealing with protocol deviators, early withdrawals, missing data, and the way(s) in which anticipated analysis problems will be handled as well as many other possible issues. The Statistical Analysis Plan will usually include sample layouts for tables and listings to be produced. Therefore preparation of a Statistical Analysis Plan is a key component in the conduct of a rigorous clinical trial and requires a statistician with both formal statistical training and significant experience in the pharmaceutical industry

Standard		A unique syntactic or semantic specifications of the representations of data, provided by a standard resource/organization (e.g. ICD-10, Gene Ontology, International System of Units, DICOM etc…). Standards regarding terminology are commonly called controlled vocabulary terms (CT), or preferred terms.
Variable Name (Data Label)		A variable name is a shorthand for designating a dimension whose values are observed or measured. examples: "AGE" is the name of a variable that describes a set of numerical values denoting the interval of time elapsed since birth. "SEX" is the name of a variable that describes the value of gender and which can assume the values or "male", "female).
Data element		
eTRIKS variable name (eTRIKS data label)		*eTRIKS variable name* are unique and standardized names according to eTRIKS-selected terminology resources.
*Entity provenance information**		Information about entities, activities, and people involved in producing an entity (e.g. data, standards), and can be used to assess the quality, reliability or trustworthiness of that entity
*Standard provenance information**		The provenance information of a standard describes the origin a standard comes from. Information such as its name and its version number or, if the version number does not exist, the date of its last release at the time the standard source is used. When a standard source is a web resource the Uniform Resource Name (URN), the uniform resource locator (URL), and its version describe the standard source.
Source information / source data		All information / data that are produced by/for the study (e.g. findings, events, protocol, studyl design). Source data are contained in source documents such as original records or certified copies of original records
Curated Data		Data are defined as curated by eTRIKS when their value, variable, format, and provenance follow the curation rules and conventions defined by eTRIKS
Primary Data		Also called raw data. Assay results that have not been processed/transformed, and are either measurements or observations. This is *Level 1 Data* according to The Cancer Genome Atlas (TCGA) classification (https://tcga-data.nci.nih.gov/tcga/tcgaDataType.jsp)

Derived Data		Data calculated from, or given according to, several primary or derived data. When they come from one subject or one sample, this is *Level 2 Data* according to TCGA classification. Example:Ages are derived data calculated from the birth and visit dates (primary data)). Second example: discretization of answers to a questionaire. for instance if number of of cigarette per day >10 , classify as 'heavy smoker'.
*Interpreted data**		Data that result from the interpretation of *Level 1 or 2 Data* (primary or derived data, respectively) by using reference data. This is *Level 3 Data* according to TCGA classification. 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. a microarray probe set gene annotation; SNP location in the genome and their mapping to genes).
Resource		A resource is an asset such as a specification or a terminology from which elements can be drawn for the purpose of data standardization and curation
Standardized data		Either data that replace and correspond to the original and non-standardized data, or numerical values that are converted in the International System (SI) of units.
Metadata		Provide information about the data, including structural information (how the data is organized) and descriptive information (what the data is about and how it was collected for instance)
*Subject identification code**		Refers to a unique identifier assigned by the investigator to each study/trial subject to protect the subject's identity and to be used in lieu of the subject's name when the investigator reports adverse events and/or other study/trial related data
Metadata registry		A software infrastructure to manage (create, maintain, map, deprecate) metadata and serves as a resource for annotation tools or EDC systems.
eTRIKS standard library		All standards used by eTRIKS

Study		A *study* is a central unit containing information on subjects and study characteristics. The term "study" covers any type of study such as interventional (e.g. drug development trials) and observational studies (e.g. patient stratification, identification of prognostic or diagnostic biomarkers). In wet laboratories the term "experiment" is homologous to "study" in clinical environment. A study has associated assays (adapted from the ISA definition: http://www.isa-tools.org).
Study site*		The location(s) where study/trial-related activities are actually conducted
Subject		A living biological entity (e.g. individuals, animals, or cells) that is under study. Example 1: patients under treatment in an interventional study. Example 2: not treated patients in an observable study. Example 3: animals under treatment in an interventional and toxicological study. Example 4: cells under treatment in an interventional and in vitro study.
Assay		An *assay* is a test that is performed either on material taken from a subject or on the whole initial subject, and aims to obtain quantitative and/or qualitative data (adapted from the ISA definition: http://www.isa-tools.org). Examples of assays: blood pressure test; microarray on cultured cells is an assay (attention: the cell culture is not part of the assay); immunochemistry on colon tissue; number of brain Gd+ lesions by Magnetic Resonance Imaging.
*Quantitative data**		Quantitative data do have a numerical value. The concentration of blood glucose is a quantitative result
*Qualitative data**		Qualitative data do not have a numerical value. The color of the eye is a qualitative result.
Observation (as understood and used in data tabulation process)		A set of qualitative or quantitative results for a given patient (on one row if data is organized by patient or on a set of rows if the data is organized by grouping criteria (also known as domain as in CDISC)
eTRIKS study owner*		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 under eTRIKS consortium.
eTRIKS collaborator		A *collaborator* is a study owner who agrees 1) to provide eTRIKS with data from a study and 2) to follow eTRIKS guidelines for concept tree design and data curation, where applicable.

Data curation		Is a term used to indicate management activities required to maintain research data long-term such that it is available for reuse and preservation In an IT (information technology) context, data curation is roughly synonymous with data lifecycle management(DLM)
Data lifecycle management	DLM	Is a policy-based approach to managing the flow of an information system's data throughout its life cycle
Data Management		Describes procedures for organizing and controlling research data
Data management plan		Template documents containing the study data management plan
Data analytics	DA	Is the science of examining raw data with the purpose of drawing conclusions about that information. In computing, data quality is the reliability and application efficiency of data, particularly when kept in a data warehouse.
Data aggregation		Any process in which information is gathered and expressed in a summary form, for purposes such as statistical analysis.
Data cleaning		Also called data cleansing or data scrubbing, is the process of cleaning up data in a database that is incorrect, incomplete, or duplicated.
Data collection		A systematic approach to gathering information from a variety of sources to get a complete and accurate picture of an area of interest
Tabulation		The systematic arrangement of the clinical data in columns or rows as rational tables
Systematized Nomenclature of Medicine -- Clinical Terms	SNOMED CT	A standardized, multilingual vocabulary of clinical terminology that is used by physicians and other health care providers for the electronic exchange of clinical health information

B. Operational Procedure

Table 3 below describes the steps in the eTRIKS operational procedure for clinical and non-clinical data that ensures high quality data standardization. The table is organized as follows:

1st- column: names the procedure step
2nd- column: indicates the chronological order of the procedure step
3rd- column: defines the action taken at the procedure step
4th- column: describes input information required for doing the action
5th- column: describes output information resulting from the action
6th- column: provides examples of the action
7th column: provides references of supplementary documentation

Note: The order of curation steps described in the below table may be changed according to the study design and/or the type of data source to be curated. In any case these changes will be reported in the document 'Standard Plan of Data Curation'.

Table 3 below is showing the operational curation processes.

Table 3eTRIKS operational procedure

Name of the procedure step	Step order	Definition of the action taken at that step	Input information required for the step action	Output resulting from the step action	Links to supplementary documentation
START COLLECTION PROCESSES (STUDY DESIGN & CASE REPORT FORMS & DATA FILES) ACTIVITIES					
Triaging	0	The process of identifying the main research theme and study protocol/ study design	1. Prospective studies. The medical writer plans to write the clinical or/and pre-clinical protocols. eTRIKS curation team (Standard Team) focuses on identifying a set of data elements suitable for a computational representation of structured protocols. eTRIKS curation team reviews and gives advice for implementation of standards and control terminology. 2. Ongoing studies or longitudinal studies. eTRIKS curation team may give advice to write amendments of protocols if applicable. 3. Retrospective studies eTRIKS curation team needs to obtain access right of studies documents .	a list of keywords. a list of user data elements: ● variable, ● questions, ● data types, ● frequencies All these elements are based on the main objectives of the protocols. Note: The data elements should be suitable for a computational representation of structured protocols.	CDISC online training: For any collaborator: http://www.cdisc.org/education-and-events For eTRIKS members: http://cdisc.trainingcampus.net Study data management plan: https://docs.google.com/document/d/151MGFo_qwxe-HWzu7HQNwmMxBWCjfmhnSufsN3rVwSs/edit
Data collection process	1	The process of supporting the creation/development , update/ testing/production of data acquisition forms (e.g. Case report forms, eCRF, ISA tools configuration) in compliance with eTRIKS recommendations.	1. Prospective studies. eTRIKS curation team gives advice for the development of electronic case report forms (eCRFs) and the implementation of validation rules and standards variable (items) 2. Ongoing studies or longitudinal studies. eTRIKS curation team give advice for update of eCRFs if applicable 3. Retrospective studies eTRIKS curation team needs to get familiar with the concept of EDC system and/or data collection processes.	If prospective studies, then: ● design of database in EDC system ● eCRFs, the structured data acquisition forms If ongoing or retrospective studies, then ● a set of user variables ● a set of user validation rules ● an issue tracker	CDISC online training: For any collaborator: http://www.cdisc.org/education-and-events For eTRIKS members: http://cdisc.trainingcampus.net Study data management plan: https://docs.google.com/document/d/151MGFo_qwxe-HWzu7HQNwmMxBWCjfmhnSufsN3rVwSs/edit Data standards translational strategy Standards workflow: https://drive.google.com/drive/#folders/0B_CayBYdTcllanZYcWJVOGw4WG8/0B9d3MkxIKuPGMmFXUWVRVjdfUkE/0B9d3MkxIKuPGRFQzM

					011X2M3T0
data collection variable ingestion (iterative process with built-in quality control)	2	The process of deconvolution/guessing user submitted variables and variables attributes values, validation/business rules descriptions	1. Prospective studies. 2. Ongoing or longitudinal studies. 3. Retrospective studies ● eCRFs, the structured data acquisition forms ● a set of user variables ● a set of user validation rule ● an issue tracker	a set of requirements from curator to submitter a set of curator annotated validation rules an issue tracker identifier a list of set of user data elements which could cover the following: ● variable, ● label, ● questions (definition), ● data types, ● data formats ● data size ● frequencies, ● validation rules ● user internal code ● user standard dictionary used ● explicit user study design ● other data element sources than EDC system	Meta Data Registry-Functional-Requirements Study data management plan: https://docs.google.com/document/d/151MGFo_qwxe-HWzu7HQNwmMxBWCjfmhnSufsN3rVwSs/edit Data standards translational strategy Standards workflow: https://drive.google.com/drive/#folders/0B_CayBYdTcllanZYcWJVOGw4WG8/0B9d3MkxIKuPGMmFXUWVRVjdfUkE/0B9d3MkxIKuPGRFQzM011X2M3T0
START ETL (EXTRACT/TRANSFORM/LOAD) ACTIVITIES					
EXTRACTION					
data transfer	3	the process of sending the data from the study owner to the eTRIKS curation team	1. Prospective studies. 2. Ongoing or longitudinal studies. 3. Retrospective studies Endpoints of Transfer: emission point/reception point Protocol of transfer: (e-mail, FTP, CD, DVD, etc.), Format of transfer: (CSV, ASCII files, XML files, etc.) Security of transfer: Frequency of transfer: Integrity of transfer: (MD5/SHA1 checksum)	● a collection of datasets files an issue tracker identifier Consolidation of datasets from EDC system or other sources including the list of set of user data elements.	Study data management plan: https://docs.google.com/document/d/151MGFo_qwxe-HWzu7HQNwmMxBWCjfmhnSufsN3rVwSs/edit Data transfer procedure: https://docs.google.com/document/d/1ROlZKLsLby8NrOhohlvaLLOzmqF6gITKOhVDUMMJU_I/edit
TRANSFORMATION					
data cleaning	4	The process of cleaning data during collection process.	Consolidation of datasets from EDC system and/or other sources including the list of set of user data elements.	Clean datasets and validation rules a set of specifications for data quality problems including:	Study data cleaning plan:

				● single source problem : data entry errors: Misspelling, redundancy/duplicates, contradictory values ● multi sources problems overlapping, aggregations , inconsistent data	
data annotation by standards	5	The process of associating verified user variables to controlled terminologies chosen from eTRIKS recommended standards (models or terminologies)	Clean datasets and/or other sources including the list of set of user data elements.	● a set of specifications for data annotated including: ● selection rule of parameters key for metadata structure ● a set of Standards domains ● a set of standard variable associated with user variables ● a set of user control terminology ●a set of eTRIKS control terminology ●a set of specification for derived data ● transformation rules ● a set of Laboratory model specifications Note: LB model includes: routine clinical lab, preclinical, omics, cell, in vitro model..etc ● a set of specification for possibly to combine multiple nomenclatures where possible and then fill in the gaps with non-standard naming. ● a set of specification for developing the Master Tree User data elements and eTRIKS convention for upgrade of Meta Data	Study data management plan: https://docs.google.com/document/d/151MGFo_qwxe-HWzu7HQNwmMxBWCjfmhnSufsN3rVwSs/edit Meta Data Registry-Functional-Requirements Data standards translational strategy Standards workflow: https://drive.google.com/drive/#folders/0B_CayBYdTcllanZYcWJVOGw4WG8/0B9d3MkxIKuPGMmFXUWVRVjdfUkE/0B9d3MkxIKuPGRFQzM011X2M3T0 Data annotated processes https://docs.google.com/document/d/1kUb2PCjFlmYr_EuJVp06NsfjeRjj1yqnuNW4mReedrg/edit
domain tabulation processes (data structure transformation)	6	The process of transforming user data structure into standard compliant ones	a collection of data sets files a set of specifications for data annotated including: ● selection rule of parameters key for metadata structure ● a set of Standards domains rules ● a set of standard variable rules associated with user	a set of syntactically standard compliant a collection of standard datasets as metadata files Standards metadata structure includes: ● Tabulated data sets into the Standards	Study data management plan: https://docs.google.com/document/d/151MGFo_qwxe-HWzu7HQNwmMxBWCjfmhnSufsN3rVwSs/edit Meta Data Registry-Functional-Requirements

			variables ● a set of user supplied documents ● transformation rules	domains ● Draft tabulated datasets NOTE: metadata files kept content value from data sources	Data standards translational strategy Standards workflow: https://drive.google.com/drive/#folders/0B_CayBYdTcllanZYcWJVOGw4WG8/0B9d3MkxIKuPGMmFXUWVRVjdfUkE/0B9d3MkxIKuPGRFQzM011X2M3T0
transformation verification (quality control for the process execution)	7	The process of standardization follow by the quality control step to ascertain the transition between data structure conforms with specifications	Standards metadata structure includes: ● Tabulate data sets into the Standards domains ● Draft tabulate datasets a set of specifications for data annotated including: ● Set of standard control terminology ● Set of eTRIKS control terminology ● Specification for derived data ● Transformation rules a set of specification for possibly to combine multiple nomenclatures where possible and then fill in the gaps with non-standard naming.	a set of verified , syntactically standard compliant data structures/documents the new tabulated metadata including standards control terminology and derived data is deliver	Study data management plan: https://docs.google.com/document/d/151MGFo_qwxe-HWzu7HQNwmMxBWCjfmhnSufsN3rVwSs/edit Meta Data Registry-Functional-Requirements Data standards translational strategy Standards workflow: https://drive.google.com/drive/#folders/0B_CayBYdTcllanZYcWJVOGw4WG8/0B9d3MkxIKuPGMmFXUWVRVjdfUkE/0B9d3MkxIKuPGRFQzM011X2M3T0
data “cleaning” (quality control step for content)	8	The process of cleaning performing of set of rules for detecting missing values, detecting data type errors or out of bound values.	The new tabulated metadata, including standards control terminology and derived data a set of specifications for solving data quality problems including: ● single source problem : data entry errors: Misspelling, redundancy/duplicates, contradictory values ● multi source problems overlapping, aggregations, inconsistent data a set of specification for possibly to combine multiple nomenclatures where possible and then fill in the gaps with non-standard naming. a set of content validation	a set of cleaned, and verified syntactically standard compliant data structures/documents the new tabulated clean metadata	Study data management plan: https://docs.google.com/document/d/151MGFo_qwxe-HWzu7HQNwmMxBWCjfmhnSufsN3rVwSs/edit Meta Data Registry-Functional-Requirements Data standards translational strategy Standards workflow: https://drive.google.com/drive/#folders/0B_CayBYdTcllanZYcWJVOGw4WG8/0B9d3MkxIKuPGMmFXUWVRVjdfUkE/0B9d3MkxIKuPGRFQzM011X2M3T0

			rules a set of verified, syntax standard compliant data structures/documents		
data 'harmonization'	9	The process of harmonization by enactment of a set of rules aimed at ensuring that variable categorical value labels are from controlled terminologies and used consistently within a given submission This process is following the validation step.	the clean tabulated metadata from several sources with and within eTRIKS projects. a set of Laboratory model specifications Note: LB model includes: routine clinical lab, preclinical, omics, cell, in vitro model, etc... a set of content terminology substitution rules a set of cleaned, and verified, syntactically standard compliant data structures/documents	a set of cleaned, and verified, syntactically and locally semantically standard compliant data structures/documents the new tabulated harmonized and standardized metadata is deliver	Study data management plan: https://docs.google.com/document/d/151MGFo_qwxe-HWzu7HQNwmMxBWCjfmhnSufsN3rVwSs/edit Meta Data Registry-Functional-Requirements Data standards translational strategy Standards workflow: https://drive.google.com/drive/#folders/0B_CayBYdTcllanZYcWJVOGw4WG8/0B9d3MkxIKuPGMmFXUWVRVjdfUkE/0B9d3MkxIKuPGRFQzM011X2M3T0 Study Lab model:
LOADING					
data integration/ persistence/loading	10	The process of creating of TransMART study Master Tree Data and loading the study in TransMART database.	1.Prospective studies. 2.Ongoing or longitudinal studies. 3.Retrospective studies the new tabulated, harmonized and standardized metadata a set of specification to develop the Master Tree	a new entry in TransMART repository Master Tree per study Consolidated metadata TranSMART database	tranSMART Mastre Tree Data:

C. Data Curation workflow

The diagram below illustrates the overview of the curation processes including set up, planning and the implementation steps.

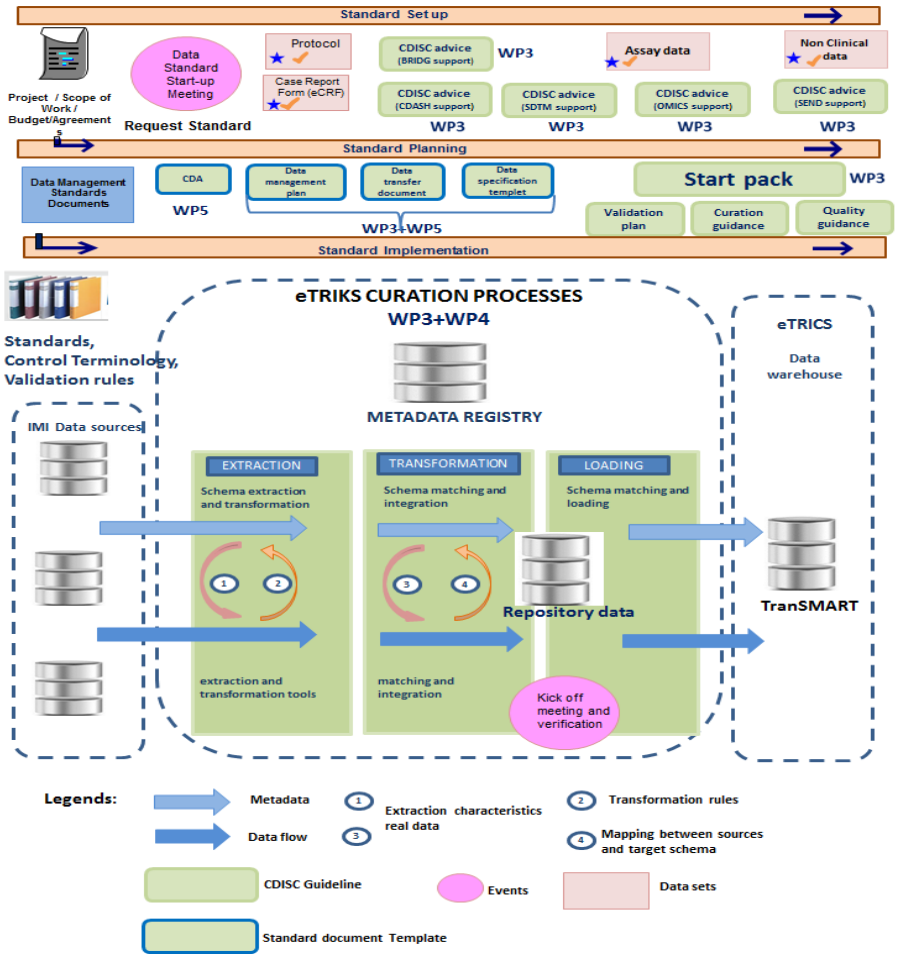

Figure 2 Study data management workflow

The diagram below illustrates the overview of the ETL curation workflow showing all the steps of data curation operational process.

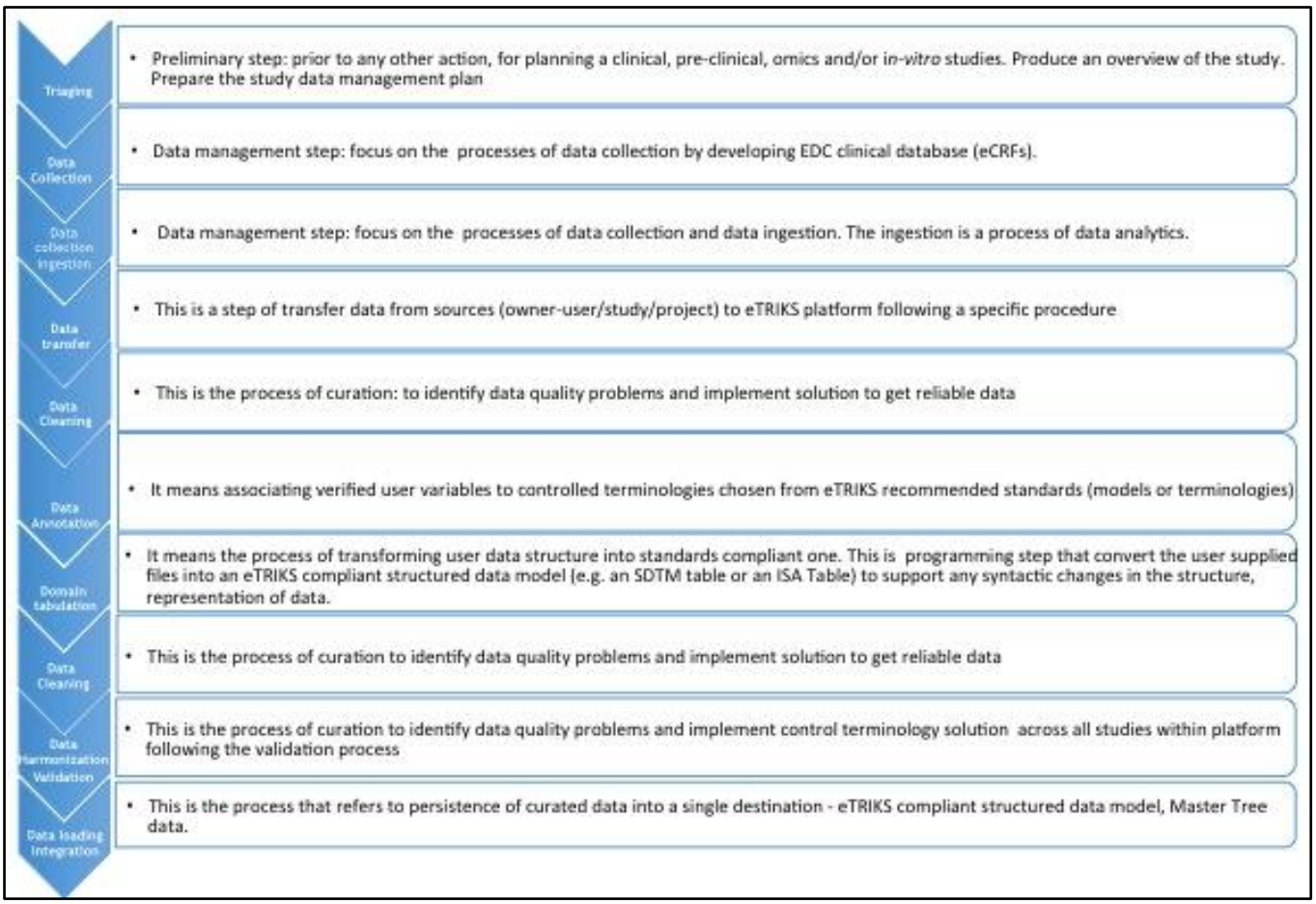

Figure 3 Study ETL workflow

eTRIKS tranSMART Master Data Tree

Overview Master Concept Tree

This section will show how data mapping to the TransMART "ontology tree" (*TM tree*) can be executed and presented consistently with the CDISC data standards.

NOTE: Bear in mind that the notion of TransMART "ontology tree" (TM tree) is unrelated to the notion of ontology as understood in the semantic web sense. The TransMART "ontology".
The concept of the Master Tree of an investigation should derive from the eTRIKS master concept tree in order to ease the users' reading and comprehension and to improve cross-study comparability. The eTRIKS master tree represents domains (i.e. CDISC definition) that are common to all investigations (master branches) such as demographics, adverse events, etc. This domain representation is adapted from the CDISC data categorization.

The CDISC data standards define 'data set structures' whereas the 'Master tree' is a way of presenting data points. The mapping of IMI project data to a generic MT structure will naturally have to make some bends e.g. to leave out CDISC required information simply because it is not available.

A concept map presented in the diagram below illustrates how data from a specific study can be mapped to a tree structure against CDISC standards. If an appropriate CDISC Therapeutic Area Standard does not exist then the procedure described in D3.4 and the eTRIKS "Standards Starter Pack" should be followed to work with CDISC to meet the demand.

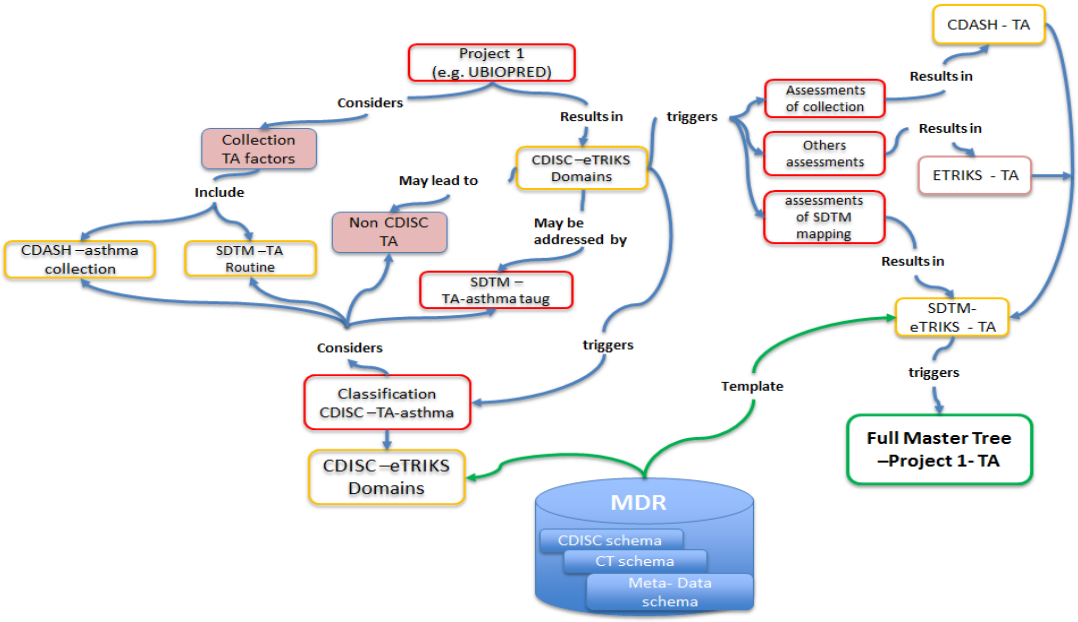

Figure 4 Concept map -TranSMART Master Tree

Requirements and workflows

1. Data label tree

a. The designed concept tree shall:

- Comply with the definition of Master Tree
- Provide a structure to the data in a way that the collaborators want to visualize them in the TransMart UI
- Be approved by both eTRIKS and collaborators before proceeding further.

b. To understand the investigation in detail the collaborators shall provide eTRIKS with a detailed description for the following points:

- The goal of the investigation;

  The clinical and/or laboratory study design;
- The assays, the measurements, and the observations;
- The analytical plans (i. e. what data will be analyzed? What analyses will be performed?);
- The conventions (e.g. the reference time point is the baseline time point).

c. To design the concept tree collaborators and eTRIKS shall first define:

- A complete list of domains (as defined by CDISC) and data labels,
- The data status:
  - "Mandatory": missing data are not accepted;
  - "Optional": missing data are accepted.

d. Based on the provided information, eTRIKS will propose a concept tree to the collaborators.

e. The concept tree could be graphically shown in order to facilitate comprehension and discussions between eTRIKS and collaborators.

f. Based on the 1st version of the concept tree eTRIKS will provide a 1st version of the column mapping file that will be used for the data loading into TM.

g. If, as a result of data curation for the study, the concept tree requires modification, this will be reviewed and agreed with the collaborators.

The workflow eTRIKS traSMART Master tree is described in the diagram below from figure 5

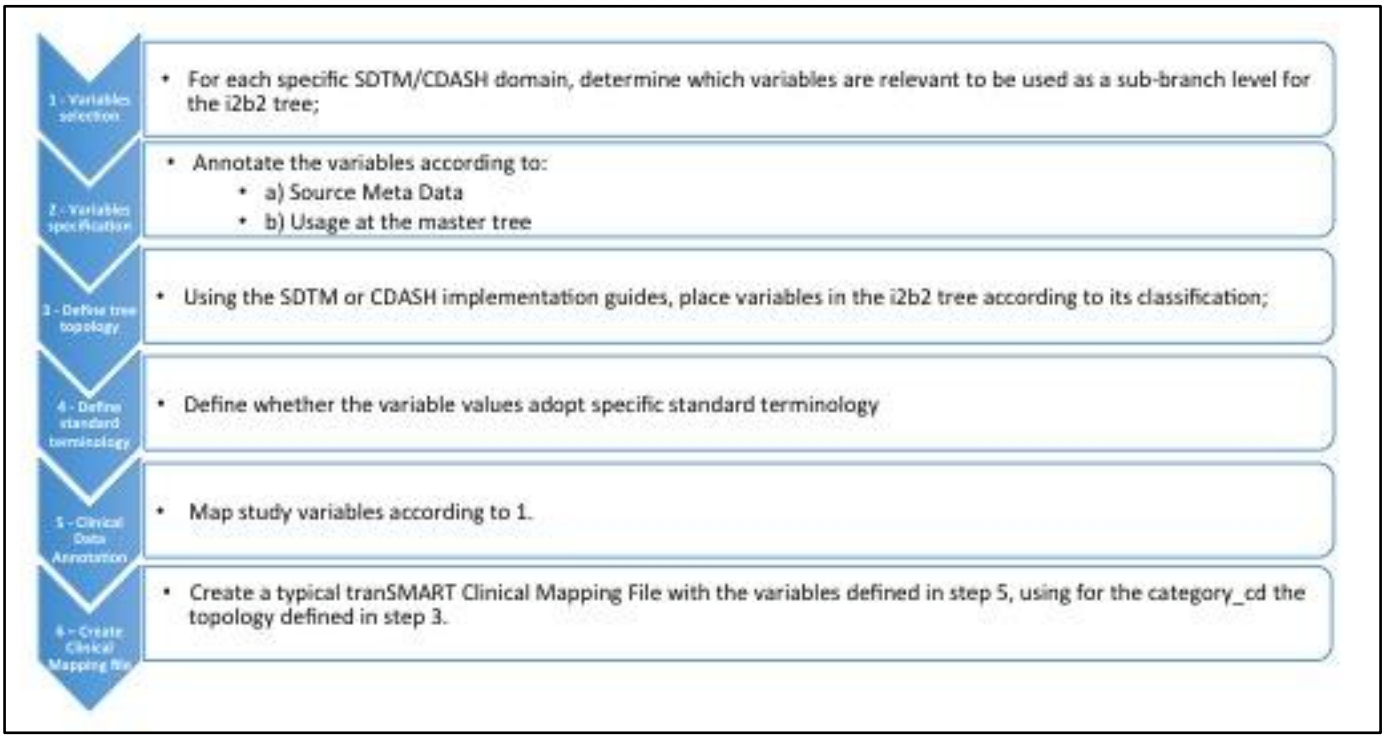

Figure 5 Master Tree Data flow

An example of a tranSMART concept tree

The *Master tree* is the user's guide to data points available in the TransMart data base. Hence, the tree should be understood as a guide to data point selection. The data tree should not be read as a regular data *file* explorer since the leaves on the tree are not full data sets but the value level of an analysis parameter; e.g.

figure 6 below shows the branch and leaf design of the master tree

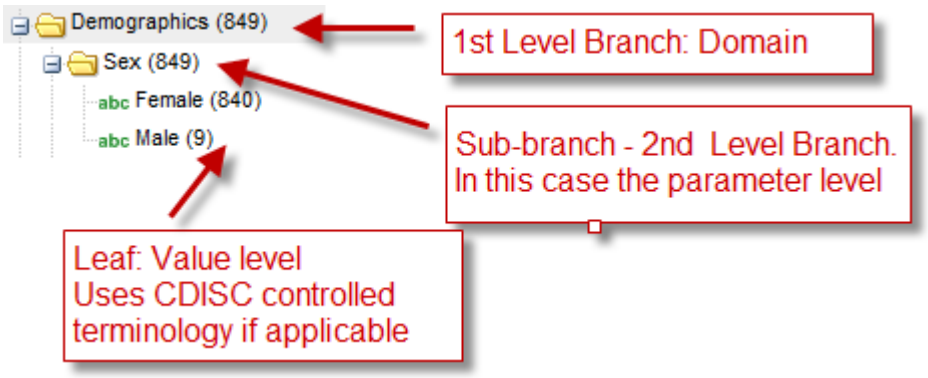

Figure 6 Master Tree Branch and Leaf Example

Table 4
eTRIKS SDTM -DM example data structure

TranSMART Tree	Folder Level	Folder Level	Not shown in tree	Folder Level	Leaf Value
SDTM variable name -->	**STUDYID**	**DOMAIN**	**USUBJID**	**Sex**	
Row 1		DM	AAAAA		Male
Row 2		DM	BBBBB		Female

Table 4 above presents the example SDTM Demographics Data Set Structures for the tree structure presented in Figure 6 (folder level STUDYID omitted from figure)

The following linked document is used for generating the TransMART master tree from a tabulated list of CDISC data elements:

https://docs.google.com/spreadsheets/d/1pocBl_GfOP9lshOgGOGZubx7QL9Zt7Iu_2yzsJAPM5s/edit#gid=2023004721

The Master Tree is displayed as is presented in the figure 7

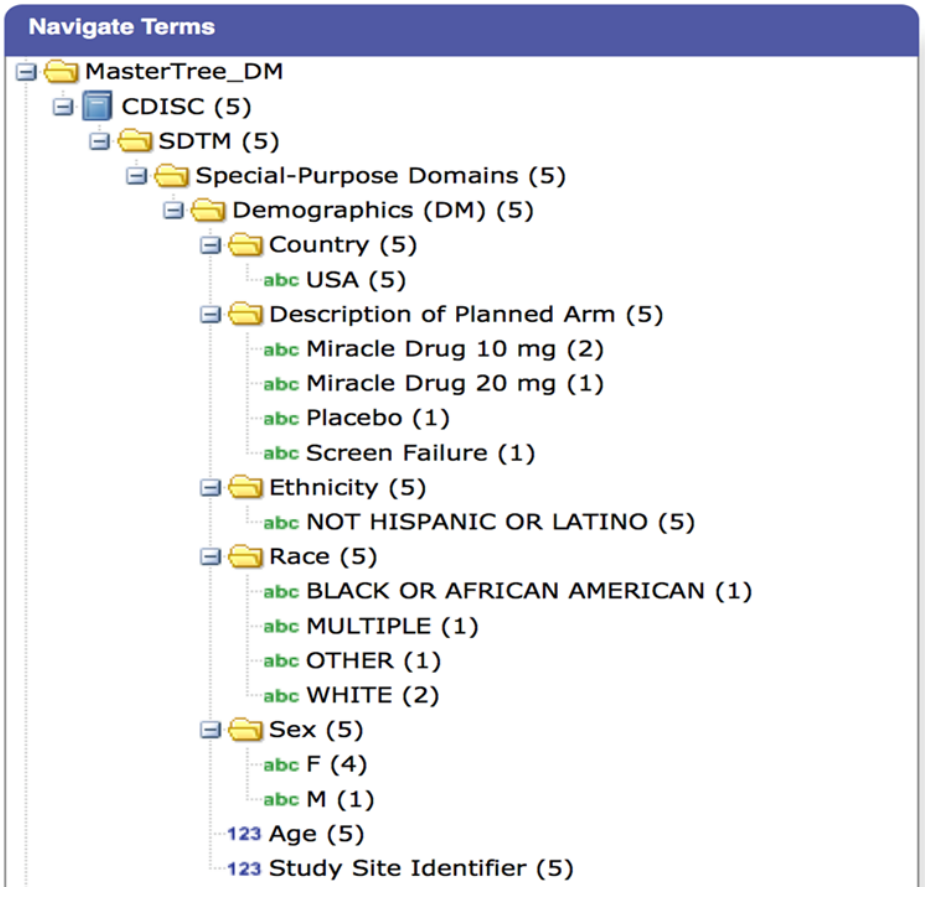

Figure 7 TranSMART Master Tree-CDISC Demographic

ISA model (from http://www.isa-tools.org)

ISA-Tab is a syntax specification for managing functional genomics studies, including DNA microarray based transcriptomics, next generation sequencing applied to expression, gene regulation and epigenomic but also targeted or untargeted metabolite profiling using mass spectroscopy and NMR spectroscopy. ISA grammar allows the definition of workflows describing the actions affecting biospecimen but also data generated by the various assays supported by the specifications. ISA grammar allows controlled terminology annotation and coding of biospecimen characteristics, protocol parameters and experimental design variables, thus promoting good practice for data sharing and data distribution. Furthermore, ISA-Tab specification have been developed with awareness of other standards, in particular in the field of clinical and toxicological areas. Hence, it documents how to reference CDISC SEND USUBJID, thereby affording meshing with the clinical world if need be.

ISA-Tab grammar also provides structure and consistency to the reporting of level 1 and level 2 data (raw data and normalized data, respectively).

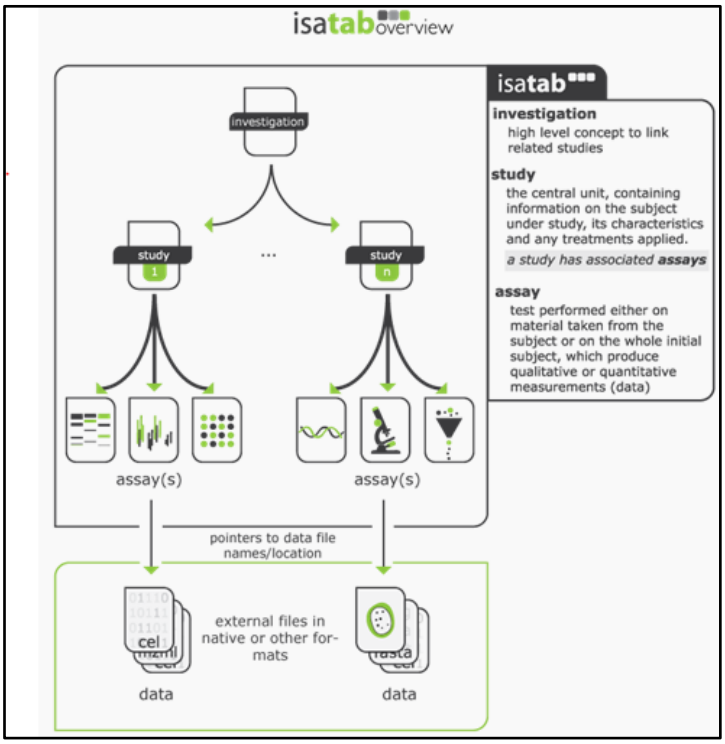

Figure 8. An overview of the structure of an ISA data archive. 3 metadata file types, investigation, study and assay referencing level 1 and level 2 data matrices.

In support on ISA-Tab based data reporting, ISA team has developped a suite of tools and assay specific configurations (see Figure 9). The former count an editor, a validator, a converter and web application database bundle, completed with an R package. The latter can be viewed as annotation templates implementing curation guidelines in line with major public data repositories (e.g NCBI Sequence Read Archive & EMBL-EBI European Nucleotide Archive, EMBL-EBI PRIDE proteomics repository and EMBL-EBI Metabolights database). For each of these configurations, key annotation elements are embedded, reminding and guiding end-users in the data preservation process. ISA configurations can be augmented or amended in order to match annotation and curation policies agreed in a project.

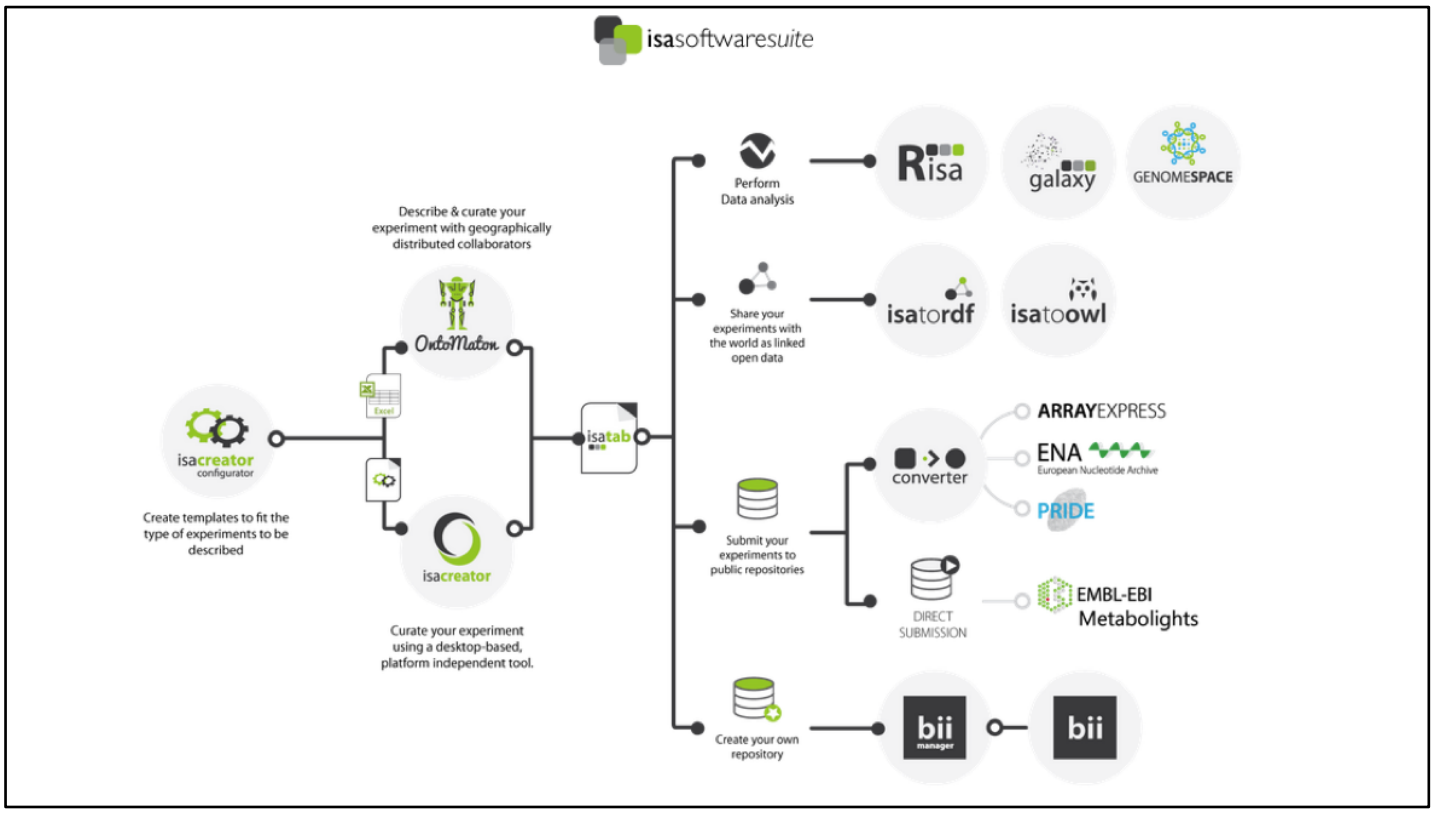

Figure 9: An overview of ecosystem of tools supporting ISA syntax, from creation to analysis, conversion and persistence.

ISAcreator, the standalone editor for creating or curating ISA-Tab documents operates in 3 distinct modes: 1/ a spreadsheet editor, with embedded terminology lookup service, relying on NCBO Bioportal API. 2/ a mapping / import tool capable of ETL from existing, third party spreadsheets or table to ensure formatting to ISA-Tab specifications. 3/ a wizard creation mode, relying of the principles of factorial experimental design to automate the creation of ISA-Tab data document. This is particularly useful in either retrospective cases to rapidly get curators of the ground or in prospective mode, when scientists are planning their experiments and devise data management plans. Figure 10 provides screenshots of some of the modes for illustration.

Tabular definition grants ease of use and rendering of information. Various XSLT transformations have been devised for EBI/NCBI Short Read Archive xml format, Bruker and Biocrates AG export XML formats as well GEO MiniML. This means that public datasetd can easily be made available to the eTRIKS curation activities.

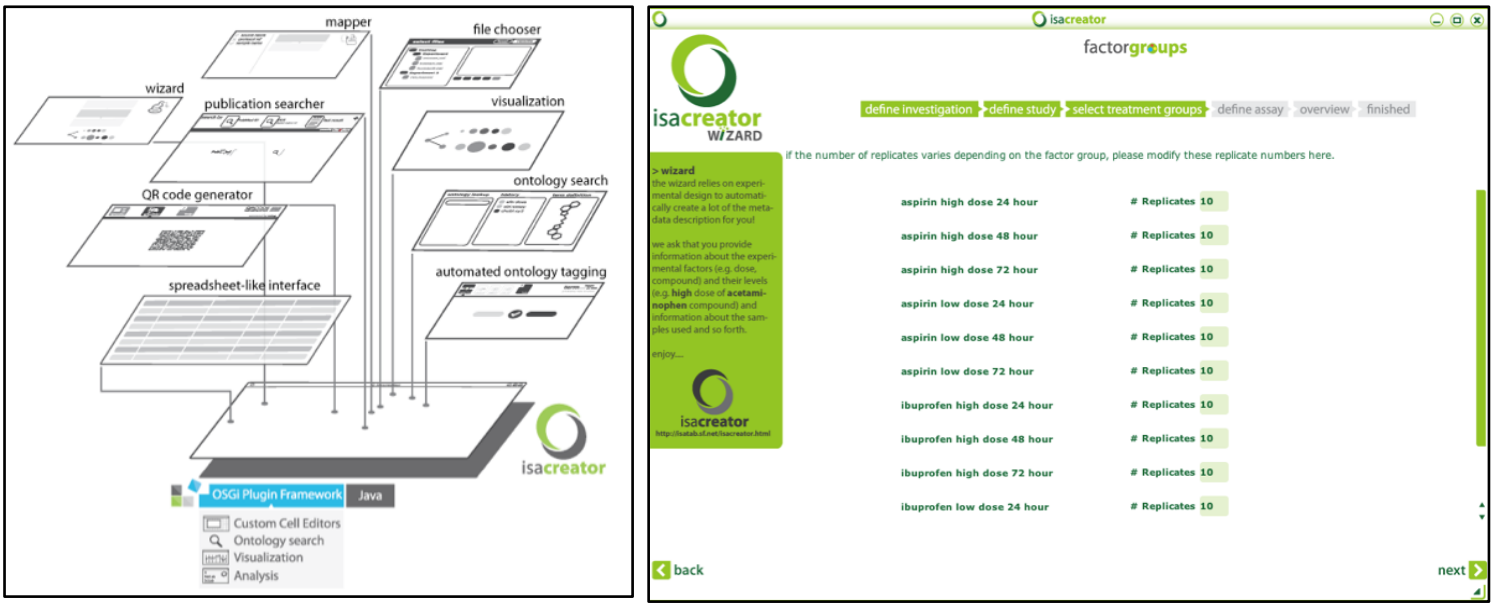

Figure 10: ISAcreator main features on the right panel. A view of the study design wizard in ISAcreator used to speed up ISAarchive creation.

In keeping with the data management and curation guidelines, the main principles, described in the first sections of the present document, need to be applied when using ISA syntax for data preservation. Hence, training programs will be put in place.