European Translational Information and Knowledge Management Services

eTRIKS Deliverable report

Grant agreement no. 115446

D4.4 - Data Curation Guide

Due date of deliverable: September 2013

Actual submission data: November 2013

Dissemination Level		
PU	Public	PU
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:	D4.4
Deliverable title:	Data Curation Guide
Deliverable version:	1.1
Due date of deliverable:	September 30th 2013
Actual submission date:	November 2013
Leaders:	Fabien Richard, Manfred Hendlich, and Reinhard Schneider
Editors:	
Authors:	Fabien Richard
Reviewers:	Serge Eifes, Maria Biryukov, Adriano Barbosa, Venkata Satagopam, Wie Gu, Reinhard Schneider, Manfred Hendlich, Ioannis Pandis, David Johnson, Ibrahim Emam, David Henderson, Antigoni Elefsinioti, Chris Marshall, Paul Dodson, Angus McAllister, Anthony Rowe
Participating beneficiaries:	
Work Package no.:	WP4
Work Package title:	Analytics Research & Content Curation
Work Package leader:	Fabien Richard, Manfred Hendlich, and Reinhard Schneider
Work Package participants:	
Estimated person-months for deliverable:	1
Nature:	
Version:	1.1
Draft/Final:	Final
No of pages (including cover):	
Keywords:	Curation, standards, data tree, tranSMART

Table of Contents

Data Tree Design and Data Curation

eTRIKS Guidelines and Procedures

Version 1.1 - November 2013

This document aims to inform collaborators about eTRIKS guidelines and procedures regarding data tree design and data curation. eTRIKS strongly recommends collaborators to follow these guidelines when applicable in order to facilitate and speed-up the work of data curation and loading into tranSMART.

I. Terms and definitions

- *eTRIKS* refers to the eTRIKS consortium.
- *Data curation* is a group of management activities that are required to **maintain research data long-term such that data are available for reuse and preservation (i.e. data sustainability)**. **These management activities consist of cleansing, converting, standardizing, and formatting data.** To enable data sharing and reusability, all data must be curated by using the same rules and conventions.
- *Curated data*. Data are defined as curated by eTRIKS when their value, label (also called variable), format, and provenance follow the curation rules and conventions defined by eTRIKS.
- *tranSMART* (*TM*) is the data warehouse that eTRIKS will contribute to develop in order to enable data hosting, sustainability, visualization and analysis. Hereafter, TM refers to the TM instance of eTRIKS, unless specified differently.
- A *data tree* refers to the overall structure and representation of the study data in the TM User Interface (UI) (see an example of a TM data tree in Annexes).
- 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 agrees 1) to provide eTRIKS with data from a study and 2) to follow eTRIKS guidelines for data tree design and data curation, where applicable.
- An *investigation* or project is a detailed inquiry or systematic examination of one to several studies (adapted from the ISA definition: http://www.isa-tools.org) (see ISA model in Annexes).
- A *study* is a central unit containing information on subjects under study and its characteristics. A study has associated assays (adapted from the ISA definition: http://www.isa-tools.org).
- A *clinical trial* or a laboratory experiment is a type of study.

- An *assay* is a test performed either on material taken from the subject or on the whole initial subject. Assay results are measurements and/or observations (adapted from the ISA definition: http://www.isa-tools.org).
- *Measurements* are quantitative data of an assay result and has a numerical value.
- *Observations* are qualitative data of an assay result, and does not have a numerical value.
- An *image* is an observation, while its signal levels are measurements.
- *Primary data* (also called raw data) are 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* 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 visit dates (primary data)). When they come from one subject or one sample, this is *Level 2 Data* according to TCGA classification.
- *Processed data* are derived data (e. g. normalization of microarray data). When they come from one subject or one sample, this is *Level 2 Data* according to TCGA classification.
- *Interpreted data* are data that result from the interpretation of *Level 1 or 2 Data* 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).
- *Metadata* provide information about one or more aspects of the data.
- *Data labels* (also called *variables* in data management) are descriptions of data (often names; in a table they are column headers).
- *Standards* are unique syntactic and 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 themselves are commonly called vocabulary controlled terms, or preferred terms.
- *Standardized data* are 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. Data labels can also be standardized.
- A *standard source* describes the origin from which the standard comes. It is defined by 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.

- A *dictionary* or *ontology* is a standard source.
- The *eTRIKS standard library* describes all standards used by eTRIKS in eTRIKS, including the eTRIKS data labels.
- *eTRIKS data label* are unique, standardized, and eTRIKS-compliant data labels.
- *Minimum Information Guidelines* (*MIG*) show what minimal information is required to fully describe the provenance of data. As a minimal requirement, the following questions must be answered:

    1. What organization generates, provides, and/or stores data hosted in TM?
    2. What is the subject ID used for generating these data? What organism/species type does the studied subject belong to?
    3. Where does the sample or the measurement/observation come from? Tissue name? Cell name? Body part name?...
    4. What biological material is used? RNA, DNA, protein, …?
    5. When is the measurement/observation done?
    6. How is the measurement/observation done? What assay/measurement/observation, technology, and/or methodology is/are used?

    Other questions may be needed according to the data type and/or assay.
- *CDISC* stands for Clinical Data Interchange Standards Consortium.
- *TCGA* stands for The Cancer Genome Atlas.
- *SI units* refer to the International System (SI) of units.
- *ETL* refers to 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).

II. Principles and guidelines

1. The data tree of an investigation should derive from the eTRIKS master data tree in order to ease the users' reading and comprehension. The eTRIKS master tree[1] represents domains (i.e a collection of observations on a particular topic (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. Therefore, the master branches should be represented in the same way between studies within an investigation, and between investigations;

[1] eTRIKS master data tree will be described in the next version of the document.

2. When an investigation encompasses several studies, study-specific data trees should be derived from this investigation data tree. Therefore, the branches of the investigation data tree that are common between the study data trees but are not master branches should be defined;
3. eTRIKS is responsible for quality level and harmonization of data annotation between TM instances. Therefore, quality level and harmonization of data annotation must be checked and approved by eTRIKS WP4 curators before data are loaded into collaborators' or eTRIKS TM instance. The collaborators can curate their data themselves, or request eTRIKS WP4 to curate the data with/for them;
4. Data and metadata required by MIG (provided in the next versions of the document) are mandatory, where applicable. When quality level of data annotation is below the minimal level defined by the MIG, data are flagged as "below the minimal information level";
5. Data status (access-restricted or public) is mandatory. eTRIKS annotates it based on collaborator's input. Basic data such as sex, age, ethnicity, study treatment name, disease name, the names of measurements and observations should be publicly available in order to let users perform basic queries on restricted-access studies (assuming that patient consents authorize that sharing);
6. Primary and derived data are hosted in eTRIKS;
7. Capturing the provenance of derived data (i.e. the primary data and the methodology used for generating derived data) is mandatory. If the values of primary data used for calculating derived data are not available or are not required by MIG, capturing the primary data labels and methodology remains mandatory;
8. To allow cross-study comparison, data hosted in TM are normalized according to an eTRIKS-predefined normalization methodology;
9. Units used in TM are SI units. Therefore, original measurements with non-SI units are converted into measurements with SI units;
10. A subjective and/or arbitrary observation should be flagged as "not supported observation" in TM UI when it is not supported by any measurement. When an observation is supported by one or more measurements, the measurements must be hosted in TM;
11. Capturing data types and levels (as defined by TCGA) is mandatory;
12. eTRIKS selects and uses standard sources that are freely available at least for not-for-profit organizations;
13. eTRIKS follows the standard orthogonality rule in order to enable data harmonization: data, and labels must correspond to a single standard. Two or more different standards cannot be applied to the same data or data label;
14. eTRIKS follows the standard granularity rule in order to standardize data in the most detailed way. Example: if the original term is "left heart ventricle" for the data label "tissue", then the standard could be "heart/left ventricle" and not "heart". This enables the user to search/analyze studies by selecting either "heart" or "left ventricle";

15. Data are standardized by using eTRIKS-selected standard sources. When a standard does not exist in any standard source, eTRIKS creates a standard and adds it to the eTRIKS standard library. The eTRIKS standard then is used in TM until a suitable international standard can replace it;
16. eTRIKS uses the CDISC-recommended standard sources when they are compliant with the points 13 and 14. If not, then eTRIKS uses other standard sources;
17. eTRIKS uses only the CDISC terminology for standardizing data labels. If a CDISC standard does not exist for an original data label, then eTRIKS creates a standard compliant with the naming rules of the CDISC terminology;
18. Capturing the origin of data standards (i.e. its standard source and its version) is mandatory;
19. Reference data sources used in TM must be public and sustained by international organizations;
20. eTRIKS keeps the provenance of data during the curation process. Thus, original and intermediate source data files that are generated along the curation process are all stored in the eTRIKS repository.
21. Acronyms are not accepted when they are not standards. Full names must be captured in the TM databases. If the full name to be captured is too long, then a short and explicit name should be captured in the TM databases and linked to the full name in a searchable eTRIKS glossary.

III. Requirements and workflows

A. Data tree

1. The designed data tree shall:
    a. comply with *Principle 1*,
    b. provide a structure to the data in a way that the collaborators want to visualize them in the TM UI,
    c. be approved by both eTRIKS and collaborators before proceeding further.

2. To understand the investigation in detail the collaborators shall provide eTRIKS with a detailed description for the following points:
    a. the goal of the investigation;
    b. the clinical and/or laboratory study design;
    c. the assays, the measurements, and the observations;
    d. the analytical plans (i. e. what data will be analyzed? What analyses will be performed?);
    e. the conventions (e.g. the reference time point is the baseline time point).

3. To design the data tree collaborators and eTRIKS shall first define:
    a. a complete list of domains (as defined by CDISC) and data labels,
    b. the data status:
        - "Mandatory": missing data are not accepted;
        - "Optional": missing data are accepted.

4. Based on the provided information, eTRIKS will propose a data tree to the collaborators.

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

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

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

B. Data Curation

1. The data curation procedure shall comply with the above principles.

2. Before starting data curation the following points should/must be addressed:
    a. A non-disclosure agreement (NDA) and/or a material transfer agreement (MTA) must be signed between eTRIKS and collaborators, and as such cover all the data that the collaborators want eTRIKS to curate and/or host in TM.
    b. Data pseudonymization or anonymization must be performed by the collaborators, and approved by eTRIKS work package 7.
    c. Collaborators should provide eTRIKS with the following information:
        - Information listed in the previous section "data tree";
        - Study status (retrospective/prospective study);
        - Number of source data file formats. It depends on the data types, the assays, and the data collection sites, and impacts eTRIKS resources and workload;
        - Representative examples of source data files for each source data file format;
        - Names and versions of the standard sources used by the collaborators;
        - Provenance of the derived data (as described in *Principle 7*);
        - Metadata required for unit conversion and mapping (e.g. conversion of g/l to mM requires the molecular weight of the molecule; or 1 = Male, 2 = Female, -9 = unknown, …);

  - Detailed status of data access rights for the different data parts of a study/investigation (i.e. “public” or “restricted”);
  - If “restricted” is selected, collaborators will provide a list of users to whom access may be granted.
  - Names and versions of the reference data sources used by collaborators for interpreting study data.

d. Define the business rules of the data cleansing with collaborators. The business rules will be used for automatic quality assurance.

e. Collaborators and eTRIKS agree on the source data file format for each assay.

3. The eTRIKS data curation process has the following steps (see workflow diagram in Annexes):

   a. Check that:

      - there are appropriate agreements in place (MTA and/or NDA)
      - data have been been pseudonymised or anonymised, and the pseudonymisation or anonymisation methodology has been approved eTRIKS work package 7;
      - the data and metadata required by MIG are provided;
      - the metadata required for unit conversion and mapping are provided;
      - the collaborator-selected reference data and standard sources are eTRIKS-compliant as defined by work package 3;
      - the data type, level, and provenance are provided;
      - the data labels of the study are synonyms with eTRIKS data labels stored in the eTRIKS standard library. If no corresponding eTRIKS data label exists in the eTRIKS standard library, a new eTRIKS data label is created by eTRIKS and added it in the eTRIKS standard library.

   b. Setting/configuration

      - Define standard sources (for data and reference data) if not provided by collaborators, or if collaborator-selected standard sources are not eTRIKS-compliant.
      - Write the data label mapping file according to the source data files.
      - Configure the cleansing engine according to business rules.
      - Configure the unit converter engine.
      - Configure the standardization engine according to standard sources and data labels/columns headers that will be used.

- Write the scripts that transform the standard source data files into the standard format files that are used by the ETL scripts for data loading into TM.
- Write ETL scripts when new types and/or formats of the curated data and metadata are loaded into TM.
- Write the subject-sample and column mapping files that are used by ETL scripts.

c. Curation

  i. Map the original data labels to the standard data labels by using the data label word mapping file. Generate the 1-standard source data files.

  ii. Consolidate data by grouping data from various 1-standard source data files into one 2-standard source data file. High dimension data are grouped into one 2High-standard source data file.

  iii. Clean data by using the cleansing engine, and saved in 3-standard or 3High-standard source data files. Missing or inconsistent data that the cleansing engine fails to clean are flagged and corrected manually by eTRIKS curators, or collaborators. Data that cannot be corrected are rejected. Data consistency is checked according to the defined business rules (see point 2-b-iv).

  iv. Convert data with original units into data with SI units by using the converter engine, and save them in 4-standard or 4High-standard source data files.

  v. Standardize data by using the standardization engine, and save them in 5-standard or 5High-standard source data files.

  vi. Manual checking and correction of what the standardization engine fails to standardize. Save checked/corrected data in 6-standard or 6High-standard source data files.

  vii. Process primary and checked data (i.e. getting Level 2 Data from Level 1 Data; e.g. microarray probeset intensity level normalization, calculation of derived data), and save them in 7-standard or 7High-standard source data files. Methodologies of data processing are saved in eTRIKS repository.

d. Transformation and loading

  i. Transform standard source data files into standard format files that are used by ETL scripts to load data into TM.

  ii. Load data and metadata into TM development server by using ETL procedure.

e. Validation

   i. The curation and loading process is tested on a subset of the study data. Each curation step is validated by checking the intermediate files.

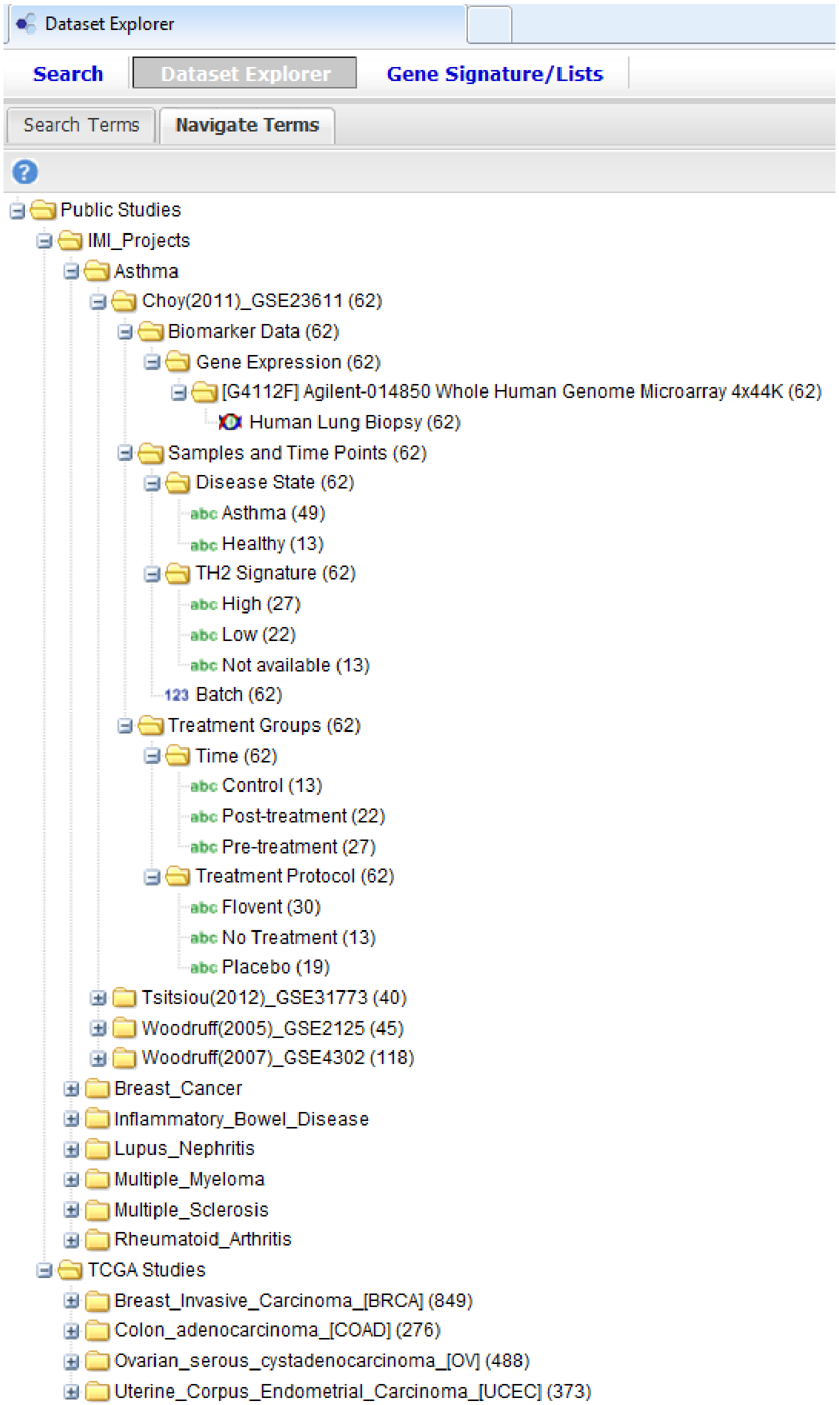
   ii. Check that the visualization of the data subset in the TM UI is correct, and satisfies collaborators. If not, change the data tree and the column mapping file accordingly.

   iii. Check that the summary statistics results are correct.

   iv. Load the whole study dataset into the TM production server following collaborators' approval.

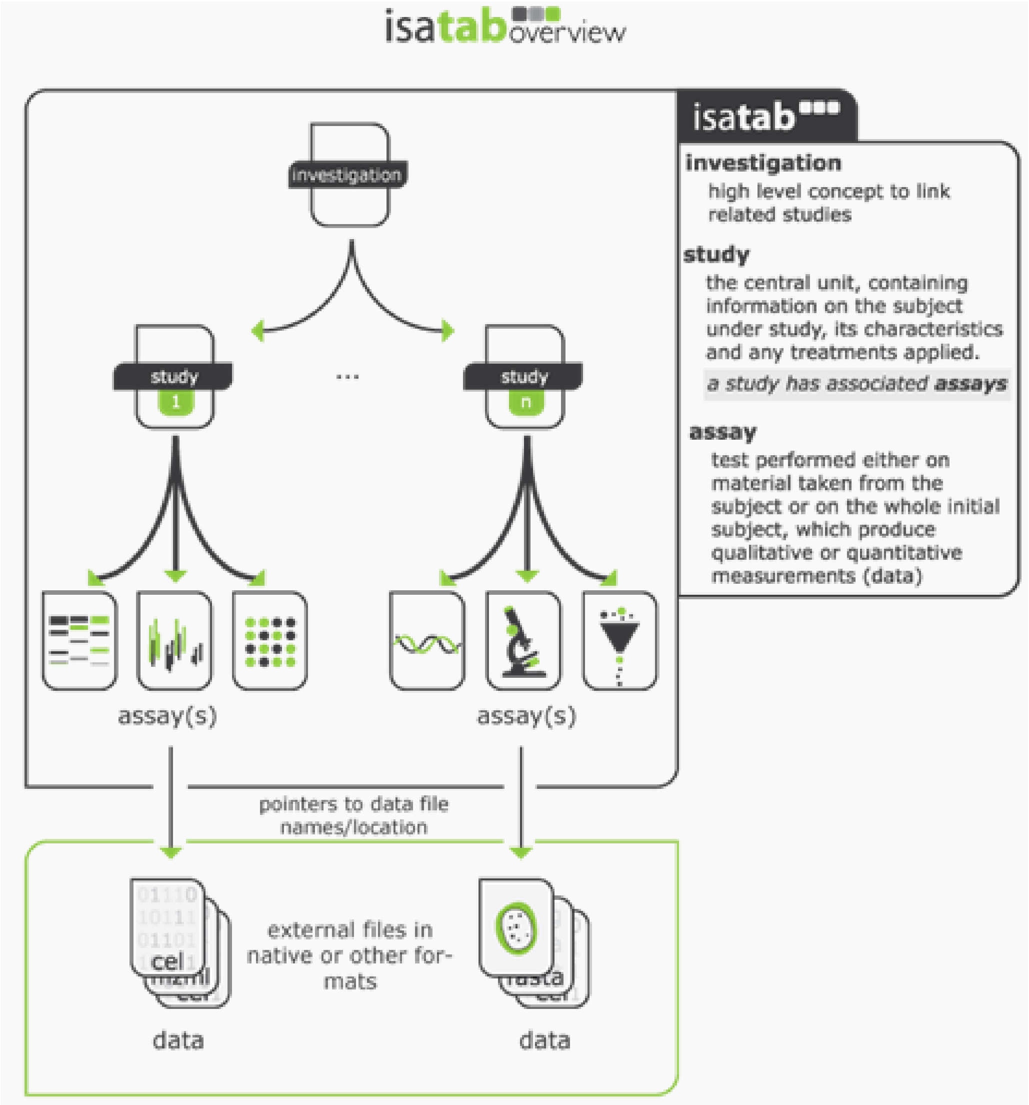
IV. Annexes

A. An example of a tranSMART data tree



- Public Studies
  - IMI_Projects
    - Asthma
      - Choy(2011)_GSE23611 (62)
        - Biomarker Data (62)
          - Gene Expression (62)
            - [G4112F] Agilent-014850 Whole Human Genome Microarray 4x44K (62)
              - Human Lung Biopsy (62)
        - Samples and Time Points (62)
          - Disease State (62)
            - abc Asthma (49)
            - abc Healthy (13)
          - TH2 Signature (62)
            - abc High (27)
            - abc Low (22)
            - abc Not available (13)
          - 123 Batch (62)
        - Treatment Groups (62)
          - Time (62)
            - abc Control (13)
            - abc Post-treatment (22)
            - abc Pre-treatment (27)
          - Treatment Protocol (62)
            - abc Flovent (30)
            - abc No Treatment (13)
            - abc Placebo (19)
      - Tsitsiou(2012)_GSE31773 (40)
      - Woodruff(2005)_GSE2125 (45)
      - Woodruff(2007)_GSE4302 (118)
    - Breast_Cancer
    - Inflammatory_Bowel_Disease
    - Lupus_Nephritis
    - Multiple_Myeloma
    - Multiple_Sclerosis
    - Rheumatoid_Arthritis
  - TCGA Studies
    - Breast_Invasive_Carcinoma_[BRCA] (849)
    - Colon_adenocarcinoma_[COAD] (276)
    - Ovarian_serous_cystadenocarcinoma_[OV] (488)
    - Uterine_Corpus_Endometrial_Carcinoma_[UCEC] (373)

B. ISA model



investigation
high level concept to link related studies

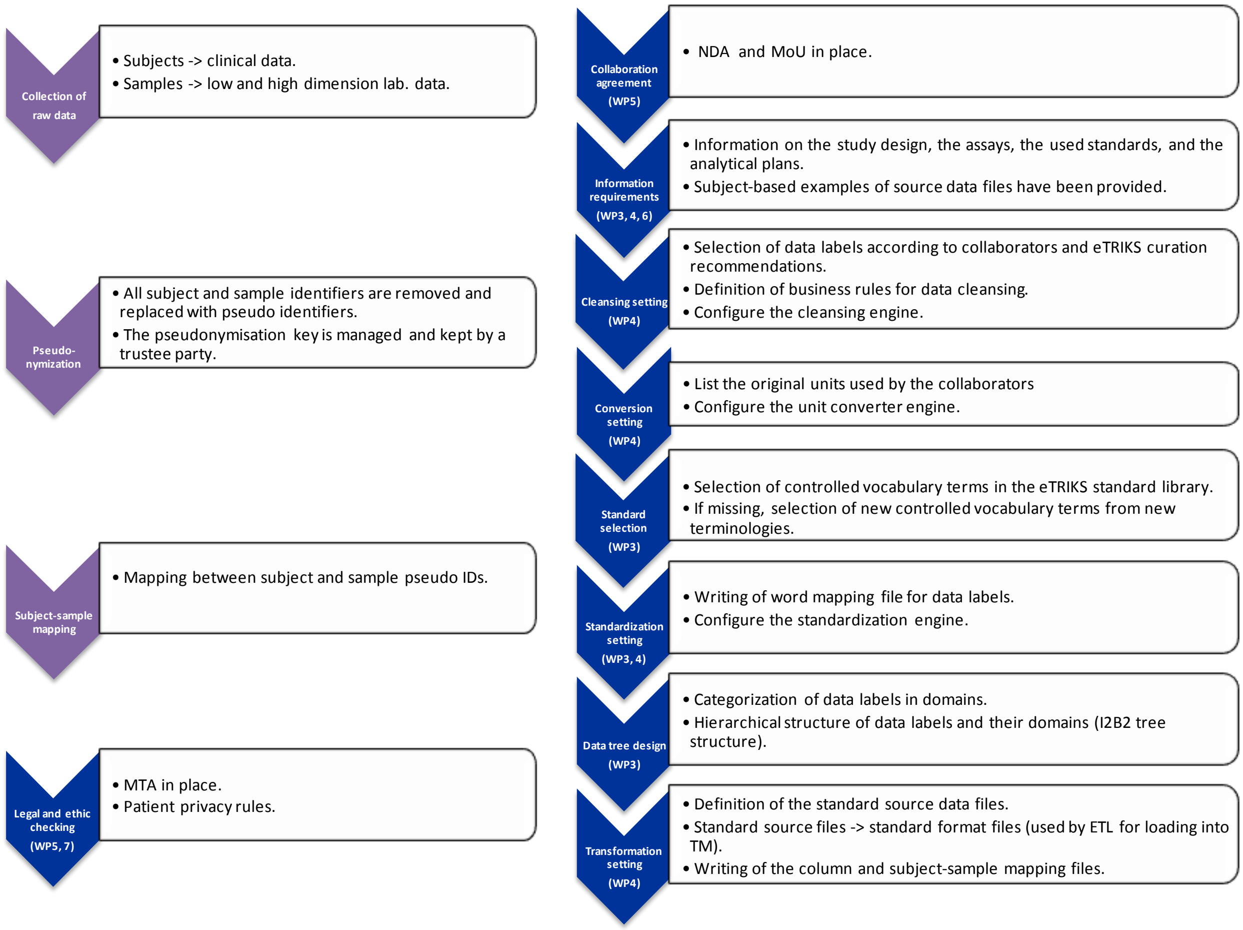
study
the central unit, containing information on the subject under study, its characteristics and any treatments applied.

a study has associated ***assays***

assay
test performed either on material taken from the subject or on the whole initial subject, which produce qualitative or quantitative measurements (data)

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

C. Curation workflow



Collection of raw data

- Subjects -> clinical data.
- Samples -> low and high dimension lab. data.

Pseudo-nymization

- All subject and sample identifiers are removed and replaced with pseudo identifiers.
- The pseudonymisation key is managed and kept by a trustee party.

Subject-sample mapping

- Mapping between subject and sample pseudo IDs.

Legal and ethic checking (WP5, 7)

- MTA in place.
- Patient privacy rules.

Collaboration agreement (WP5)

- NDA and MoU in place.

Information requirements (WP3, 4, 6)

- Information on the study design, the assays, the used standards, and the analytical plans.
- Subject-based examples of source data files have been provided.

Cleansing setting (WP4)

- Selection of data labels according to collaborators and eTRIKS curation recommendations.
- Definition of business rules for data cleansing.
- Configure the cleansing engine.

Conversion setting (WP4)

- List the original units used by the collaborators
- Configure the unit converter engine.

Standard selection (WP3)

- Selection of controlled vocabulary terms in the eTRIKS standard library.
- If missing, selection of new controlled vocabulary terms from new terminologies.

Standardization setting (WP3, 4)

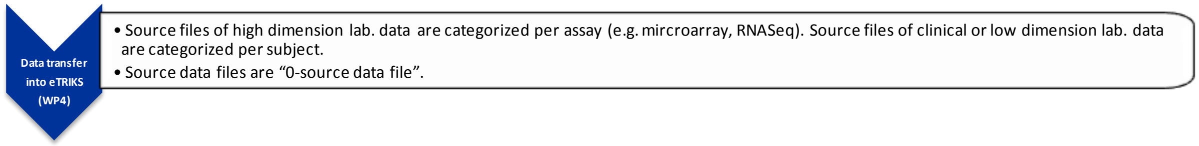
- Writing of word mapping file for data labels.
- Configure the standardization engine.

Data tree design (WP3)

- Categorization of data labels in domains.
- Hierarchical structure of data labels and their domains (I2B2 tree structure).

Transformation setting (WP4)

- Definition of the standard source data files.
- Standard source files -> standard format files (used by ETL for loading into TM).
- Writing of the column and subject-sample mapping files.



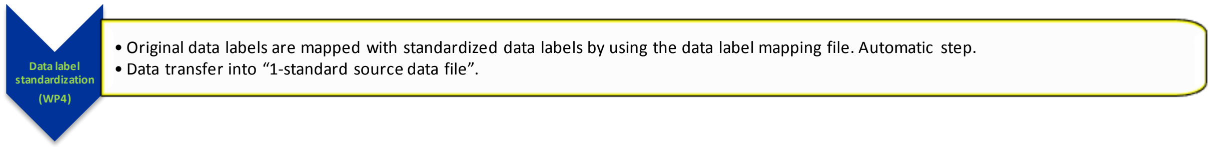
Data transfer into eTRIKS (WP4)

- Source files of high dimension lab. data are categorized per assay (e.g. mircroarray, RNASeq). Source files of clinical or low dimension lab. data are categorized per subject.
- Source data files are "0-source data file".

0-Source data files

Patient ID	Sample ID	age	sex	height	weight	glucose level
12345	12345	23	2	5 feet 9 inches	150lb	0.8g/l
67789	9876	46	male	182 cm	95kg	1g/l

Patient ID	Sample ID	age	sex	height	weight	glucose level
97531	54321	32	M	171 cm	76kg	0.97g/l
19283	63197	95	female	1.66 m	60kg	0.78g/l



Data label standardization (WP4)

- Original data labels are mapped with standardized data labels by using the data label mapping file. Automatic step.
- Data transfer into "1-standard source data file".

1-Standard source data files after data label mapping

Patient ID	Sample ID	age	sex	height	weight	glucose level
PATIENT ID	SAMPLE ID	OR-AGE	OR-SEX	OR-HEIGHT	OR-MASS	OR-GLYCEMIA
12345	12345	23	2	5 feet 9 inches	150lb	0.8g/l
67789	9876	46	male	182 cm	95kg	1g/l

Patient ID	Sample ID	age	sex	height	weight	glucose level
PATIENT ID	SAMPLE ID	OR-AGE	OR-SEX	OR-HEIGHT	OR-MASS	OR-GLYCEMIA
97531	54321	32	M	171 cm	76kg	0.97g/l
19283	63197	95	female	1.66 m	60kg	0.78g/l

Data consolidation (WP4)

- If source files of clinical data or low dimension lab. data come from several sites, data are combined into one file “2-standard source data file”.
- Source files of high dimension data per assay and sample are combined into one file to form a matrix sample x measurements that is saved in “2High-standard source data file”.

2-Standard source data files after consolidation

	Patient ID	Sample ID	age	sex	height	weight	glucose level
eTRIKS FILE #	PATIENT ID	SAMPLE ID	OR-AGE	OR-SEX	OR-HEIGHT	OR-MASS	OR-GLYCEMIA
xxxx	12345	12345	23	2	5 feet 9 inches	150lb	0.8g/l
xxxx	67789	9876	46	male	182 cm	95kg	1g/l
yyyy	97531	54321	32	M	171 cm	76kg	0.97g/l
yyyy	19283	63197	95	female	1.66 m	60kg	0.78g/l

Data cleansing (WP4)

- Missing data or data with incorrect data types (e.g. a string while numerical value is expected) are flagged for manual checking.
- Manual correction is done. If not possible, data are sent back to collaborators for completion/correction.
- Cleaned low and high dimension data are saved in “3-standard source data file” or “3High-standard source data file”, respectively

3-Standard source data files after data cleansing

	Patient ID	Sample ID	age			sex	height			weight			glucose level		
eTRIKS FILE #	PATIENT ID	SAMPLE ID	OR-AGE	DC-AGE	DC-AGE-UNITS	OR-SEX	OR-HEIGHT	DC-HEIGHT	DC-HEIGH-UNITS	OR-MASS	DC-MASS	DC-MASS-UNITS	OR-GLYCEMIA	DC-GLYCEMIA	DC-GLYCEMIA-UNITS
xxxx	12345	12345	23	23	YEAR	2	5 feet 9 inches	5-9	feet-inches	150lb	150	Lb	0.8g/l	0.8	$g.l^{-1}$
xxxx	67789	9876	46	46	YEAR	male	182 cm	182	cm	95kg	95	kg	1g/l	1	$g.l^{-1}$
yyyy	97531	54321	32	32	YEAR	M	171 cm	171	cm	76kg	76	kg	0.97g/l	0.97	$g.l^{-1}$
yyyy	19283	63197	95	95	YEAR	female	1.66 m	1.66	m	60kg	60	kg	0.78g/l	0.78	$g.l^{-1}$

Data conversion (WP4)

- Data with original units are converted into data with SI units.
- Converted low and high dimension data are saved in “4-standard source data file” or “4High-standard source data file”, respectively .

4-Standard source data files after data conversion

	Patient ID	Sample ID	age			sex	height					weight					glucose level		
eTRIKS FILE #	PATIENT ID	SAMPLE ID	OR-AGE	DC-AGE	DC-AGE-UNITS	OR-SEX	OR-HEIGHT	DC-HEIGHT	DC-HEIGH-UNITS	CV-HEIGHT	CV-HEIGH-UNITS	OR-MASS	DC-MASS	DC-MASS-UNITS	CV-MASS	CV-MASS-UNITS	OR-GLYCEMIA	DC-GLYCEMIA	DC-GLYCEMIA-UNITS
xxxx	12345	12345	23	23	YEAR	2	5 feet 9 inches	5-9	feet-inches	175.26	cm	150lb	150	LB	68.04	kg	0.8g/l	0.8	$g.l^{-1}$
xxxx	67789	9876	46	46	YEAR	male	182 cm	182	cm	182	cm	95kg	95	kg	95	kg	1g/l	1	$g.l^{-1}$
yyyy	97531	54321	32	32	YEAR	M	171 cm	171	cm	171	cm	76kg	76	kg	76	kg	0.97g/l	0.97	$g.l^{-1}$
yyyy	19283	63197	95	95	YEAR	female	1.66 m	1.66	m	166	cm	60kg	60	kg	60	kg	0.78g/l	0.78	$g.l^{-1}$

Content standardization (WP4)

- The standardization engine tries to match free text terms to controlled vocabulary terms, and saves the controlled vocabularyterms and their eTRIKS IDs in “5-standard source data file” or “5High-standard source data file”.
- An eTRIKS term ID is unique and corresponds to its controlled vocabulary term, its standard source, its synonyms in theeTRIKS standard library.

Manual validation of standardization (WP4)

- Curators validate the automatic standardization. When the standardization engine fails to recognize a term, the latter is flagged to the curators who will do the matching manually. Collaborator support may be needed to understand the term meaning.
- New manual matches are saved in the eTRIKS standard library. Learning process of the standardization engine.
- Validated data are saved in “6-standard source data file” or “6High-standard source data file”.

5-Standard source data files after content standardization

	Patient ID	Sample ID	age			sex			height					weight					glucose level		
eTRIKS FILE #	PATIENT ID	SAMPLE ID	OR-AGE	DC-AGE	DC-AGE-UNITS	OR-SEX	ST-SEX	SEX-eTRIKS_ID	OR-HEIGHT	DC-HEIGHT	DC-HEIGH-UNITS	CV-HEIGHT	CV-HEIGH-UNITS	OR-MASS	DC-MASS	DC-MASS-UNITS	CV-MASS	CV-MASS-UNITS	OR-GLYCEMIA	DC-GLYCEMIA	DC-GLYCEMIA-UNITS
xxxx	12345	12345	23	23	YEAR	2	F	1234	5 feet 9 inches	5-9	feet-inches	175.26	cm	150lb	150	LB	68.04	kg	0.8g/l	0.8	$g.l^{-1}$
xxxx	67789	9876	46	46	YEAR	male	M	5678	182 cm	182	cm	182	cm	95kg	95	kg	95	kg	1g/l	1	$g.l^{-1}$
yyyy	97531	54321	32	32	YEAR	M	M	5678	171 cm	171	cm	171	cm	76kg	76	kg	76	kg	0.97g/l	0.97	$g.l^{-1}$
yyyy	19283	63197	95	95	YEAR	female	F	1234	1.66 m	1.66	m	166	cm	60kg	60	kg	60	kg	0.78g/l	0.78	$g.l^{-1}$

Raw data Process (WP4)

- Derived data are calculated from primary lab. or clinical data.
- Derived low and high dimension data are saved in “7-standard source data file” and “7High-standard source data file”, respectively.

6-Standard source data files after data process

	Patient ID	Sample ID	age			sex			height					weight					glucose level				
eTRIKS FILE #	PATIENT ID	SAMPLE ID	OR-AGE	DC-AGE	DC-AGE-UNITS	OR-SEX	ST-SEX	SEX-eTRIKS_ID	OR-HEIGHT	DC-HEIGHT	DC-HEIGH-UNITS	CV-HEIGHT	CV-HEIGH-UNITS	OR-MASS	DC-MASS	DC-MASS-UNITS	CV-MASS	CV-MASS-UNITS	OR-GLYCEMIA	DC-GLYCEMIA	DC-GLYCEMIA-UNITS	BODY MASS INDEX	BODY MASS INDEX-UNITS
xxxx	12345	12345	23	23	YEAR	2	F	1234	5 feet 9 inches	5-9	feet-inches	175.26	cm	150lb	150	LB	68.04	kg	0.8g/l	0.8	$g.l^{-1}$	22.2	kg.m-2
xxxx	67789	9876	46	46	YEAR	male	M	5678	182 cm	182	cm	182	cm	95kg	95	kg	95	kg	1g/l	1	$g.l^{-1}$	28.7	kg.m-2
yyyy	97531	54321	32	32	YEAR	M	M	5678	171 cm	171	cm	171	cm	76kg	76	kg	76	kg	0.97g/l	0.97	$g.l^{-1}$	26.0	kg.m-2
yyyy	19283	63197	95	95	YEAR	female	F	1234	1.66 m	1.66	m	166	cm	60kg	60	kg	60	kg	0.78g/l	0.78	$g.l^{-1}$	21.8	kg.m-2

Data transformation (WP4)

- “7-standard source data file” or “7High-standard source data file” are transformed into standard format files that are used by ETL scripts to load data into TM.

Data loading into TM (WP4)

- ETL scripts are run to load data into TM.
- Data representation is checked in the TM UI .

Color codes:

- **Writing.** Green : fully automated step; orange: semi-automated step; white: manual step
- **Arrows.** Purple :collaborator task; blue: eTRIKS task