European Translational Information and Knowledge Management Services

eTRIKS Deliverable report

Grant agreement no. 115446

D4.3 Data Provenance Guidelines

Due date of deliverable: Month 9

Actual submission date: Month 23

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:	D4.3
Deliverable title:	Data Provenance Process and Management Guidelines
Deliverable version:	
Due date of deliverable:	31st May 2014
Actual submission date:	
Leader:	Reinhard Schneider and Manfred Hendlich
Editors:	
Authors:	Venkata Satagopam and Manfred Hendlich
Reviewers:	Leila El Hadjam and Chris Marshall
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:	
Nature:	
Version:	
Draft/Final:	Final
No of pages (including cover):	
Keywords:	

Purpose of this document:

The purpose of this document is to define business rules for capturing data provenance information for data sets processed and/or hosted by eTRIKS personnel in the eTRIKS technical environment. The objective is to precisely record the origin of data sets ("data owner", version, and the workflow detailing the curation process and to make every dataset submitted to eTRIKS uniquely identifiable and traceable. In order to achieve this goal, we are prescribing the following guidelines.

Out of scope:

- The definition a technical solution for managing provenance information.
- Definition of rules for project partners that get only technical support from eTRIKS (e.g. tranSMART installation, training etc.) and that do not submit data to eTRIKS.
- The tracking of analysis workflows within the tranSMART environment.
- Backup plan for whole tranSMART including data provenance.

Referenced documents:

- D7.1 - Data Ethics Requirements and Framework design-part 1 and 2.
- The Code of Practice on secondary use of medical data in scientific research projects.
- eTRIKS Security Package.
- eTRIKS Global Glossary Document.
- Material Transfer Agreement (MTA).
- Task Plan.

Process:

In eTRIKS each dataset (i.e. outcome of an investigation) will have a unique provenance ID (eTRIKS ID). These provenance IDs describe the provenance of all data files as well as the methods used to curate that data via tranSMART web interface. The Fig.1 illustrates the provenance workflow and stakeholders.

1. Once the memorandum of understanding (MOU) and non-disclosure agreement (NDA), material transfer agreement (MTA) has been signed between eTRIKS and the study owner and hosting infrastructure is provided, the study owner should move the data to the eTRIKS'LandingZone'.

2. Study owner defines the time lines, version numbers, user groups and roles, who is data administrator, curator, auditor etc according to the predefined task plan. This will enable who has access to which data set.

3. eTRIKS'LandingZone' is a file system located either at CC-IN2P3 or other pre-defined location.

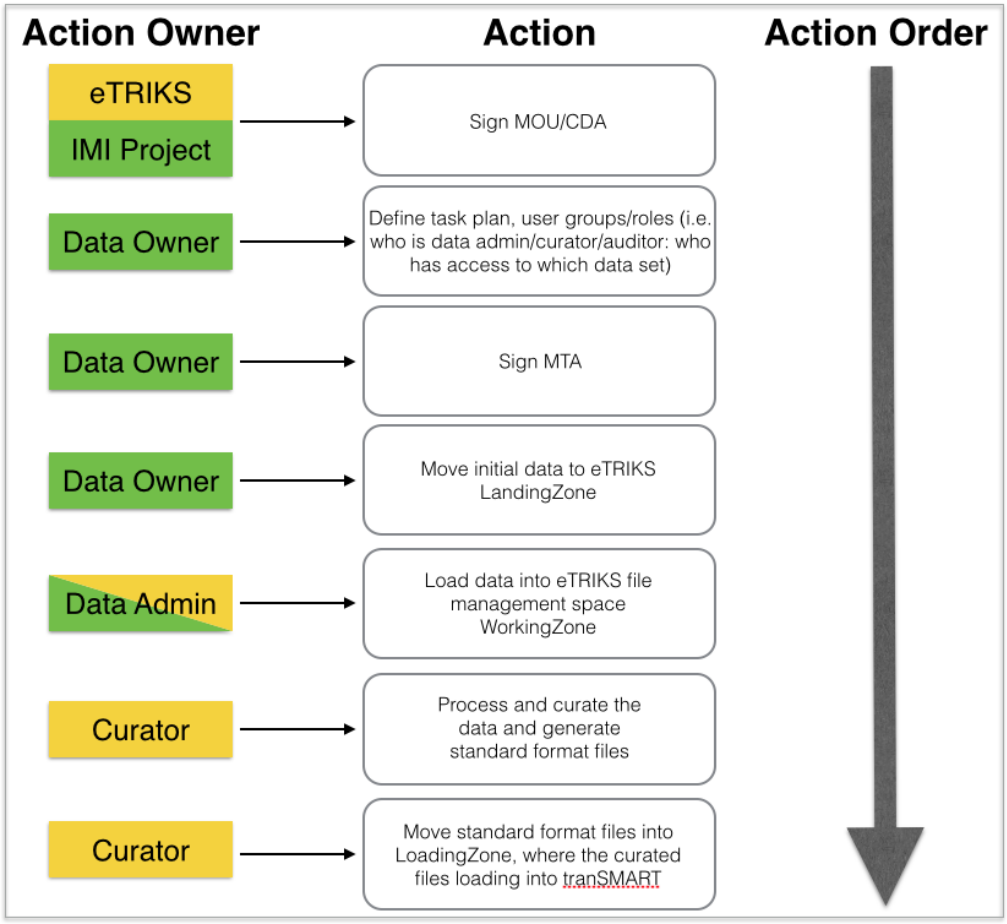

Figure 1: The diagram describes the workflow of different activities and responsibilities with-in data provenance framework

4. The initial data from the 'LandingZone' will be loaded into the eTRIKS file management space - 'WorkingZone'(e.g. MongoDB/GridFS, Hadoop, Network File System (NFS) etc.) along with version information. This environment, then becomes focal point for the primary data.

5. Each study in the eTRIKS will be represented with a unique name (e.g.: IMI project name - ABIRISK, OncoTrack etc.). On the repository file system a directory will be created with project name. The provenance of initial data and curated data will be under this high level directory in different sub-folders.Three sub-folders 'LandingZone','WorkingZone', 'LoadingZone' will be created under the project name directory. All the sub-directories and files under these zones will be accessible only by authorized users that are defined by study owner. 'LandingZone' is the place holder for the initial data deposited by the study owner. 'WorkingZone', where curation will be done and intermediate and curated files will be produced. The curated files will be moved to 'LoadingZone' where loading to the tranSMART will takes place using ETL scripts.

   A project can have different study locations;next sub-folder after above mentioned zones is the 'StudyLocation', for example KarolinskaInstitutet (KI), Technical

University Munich (TUM) etc. Some projects like for example ABIRISK working on more than one disease/scientific domain - Multiple Sclerosis (MS), Hemophilia A (HA), rheumatoid arthritis (RA) and inflammatory bowel diseases (IBD), next sub-directory will be 'StudyDomain', underneath will be 'VersionNumber' with timestamp.

Under the 'VersionNumber' folder, one dedicated folder for each data-type (clinical, omics (gene-expression, genomics, proteomics, metabolomics, lipidomics and so on), imaging etc.) where the files will be located.

6. The sub-directory structure of all the three zones ('LandingZone', 'WorkingZone' and 'LoadingZone') will be the same, unless until the data is pooled from all the study locations. In such case the 'StudyLocation' folder will skipped under ''WorkingZone' and 'LoadingZone'.In addition to above directory structure, each data-type in 'WorkingZone' contains 'README' file provided by the curation team with provenance information like name and e-mail of the study owner, study location. It also provides a workflow detailing the curation process and analysis methods used. In addition 'LoadingZone' will contains a'ETLScripts' folder under each data-type. One can achieve similar provenance using MongoDB/GridFS or Hadoop.
   This 'README' file will help in presenting curation and provenance details on tranSMART web interface.

7. In order to handle the different versions of the data coming from the same study, version numbering using the study date will be used ('Version ID')
   Version_<date in yyyymmdd format>
   e.g.: Version_20130822

8. The curation team in WP4 accesses the data from the above-mentioned 'WorkingZone' either from database or repository file system in order to process and curate. During this curation process few intermediates files will be generated, it is worth to archive these intermediate files. Generated mapping files (e.g. column mapping files, sample subject mapping files, word mapping files, etc.) in the 'WorkingZone' that generated by the curator should be located in the same folder. In case data will be pooled, mapping files should contain information of the absolute path of the data files. The final curated data will be moved to 'LoadingZone'that will be loaded into the designated tranSMART server.

9. The directory structure of the a study (e.g: ABIRISK) data provenance in eTRIKS framework is shown in the below picture

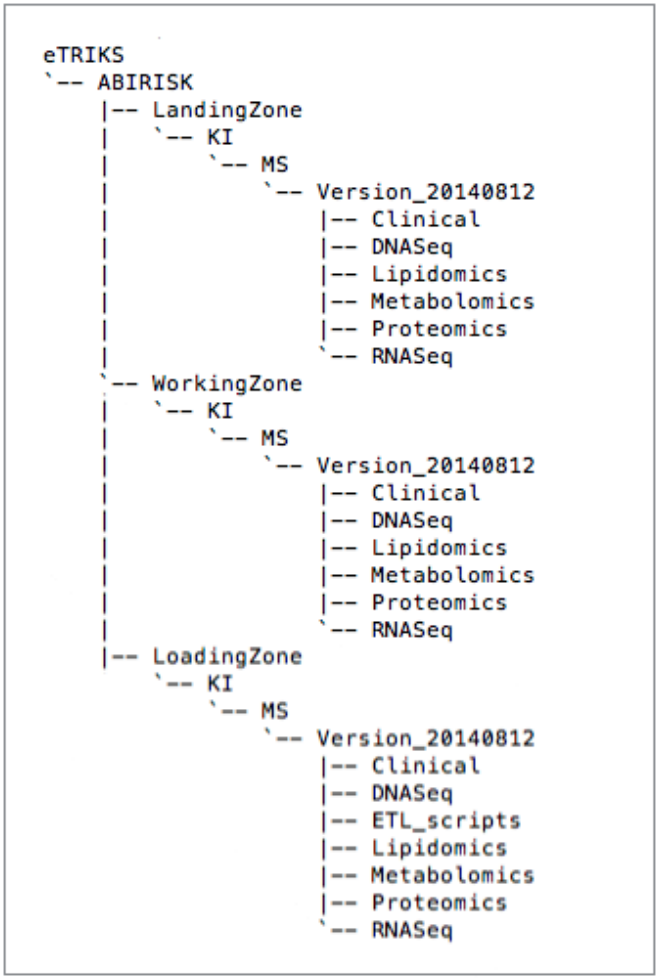

Figure 2: Example directory structure of a study data provenance in eTRIKS framework

10. Each data type (clinical, genomics, proteomics, metabolomics, lipidomicsetc) in a project as well as related mapping files will be assigned with an unique 'eTRIKS ID' and is simply by concatenation of all the folder names with underscore '_', an example is shown below. The project name, data type, location of the files either in the repository file system or database, eTRIKS ID will be recorded in a relational database and will help in access the data provenance from tranSMART or other relevant application by an authorized user.

    Example of eTRIKS ID for initial clinical data from ABIRISK Multiple Sclerosis is 'eTRIKS_ABIRISK_LandingZone_KI_MS_Version20140812_Clinical'

11. In eTRIKS framework the hosting infrastructure is provided by CC-IN2P3 and is mirrored by University of Luxembourg (UL) and Imperial College London (ICL).

12. Access to the initial data and curated data will be made available via tranSMART (TM) to the authorized users with the help of eTRIKS ID.

13. The eTRIKS IDs will be assigned by the eTRIKS data manager by following these guidelines.

14. eTRIKS data provenance framework will meet current legal and ethical requirements.

Responsibility for assigning IDs and for managing provenance:

It is the responsibility of the WP4 Luxemburg team to assign unique IDs to data sets to ensure traceability, to manage abbreviation for data owners, data types, and to ensure overall compliance with these rules. These abbreviations will be documented in eTRIKS Global Glossary Document.

Technical storage and visualization of the data provenance in the eTRIKS technical environment (tranSMART) will be the responsibly of WP2. Details will be described elsewhere.