European Translational Information and Knowledge Management Services

eTRIKS Deliverable report

Grant agreement no. 115446

Deliverable D5.8
eTRIKS best practices for service creation and delivery

Due date of deliverable: 30th September 2017

Actual submission date: 14h September 2017

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)	

Project	
Project acronym:	eTRIKS
Project full title:	European Translational Information and Knowledge Management Services
Grant agreement no.:	**115446**
Document	
Deliverable number:	D5.8
Deliverable title:	eTRIKS Best practices for service creation and delivery
Deliverable version:	1.0
Due date of deliverable:	30th September 2017
Actual submission date:	14th September 2017
Leader:	Chris Marshall, Scott Wagers
Editors:	Neil Fitch,
Authors:	Chris Marshall, Neil Fitch, Paul Peeters, Wei Gu, Francisco Bonachela Capdevilla, Gino Marchetti
Reviewers:	Jay Bergeron,
Participating beneficiaries:	BioSci, CNRS, ICL, Oxford, UL
Work Package no.:	5
Work Package title:	Governance and Business Model
Work Package leader:	Jay Bergeron, Scott Wagers
Work Package participants:	Jay Bergeron, Scott Wagers, David Henderson, Yi-Ke Guo, Charles Auffray, Francisco Bonachela Capdevilla
Estimated person-months for deliverable:	2
Nature:	Report (PU)
Version:	1
Draft/Final:	Final
No of pages (including cover):	44
Keywords:	Best practice, services, eTRIKS, Bioinformatics, tranSMART

Table of contents

3

Executive Summary

eTRIKS has developed a range of best practices to advantage the discipline of translational research knowledge management. This document describes the most significant contributions to knowledge management provided by eTRIKS. The specific practices are described in finer detail in the referenced documents and reports. The IMI mid-term reviewers recommended that eTRIKS collate best practices that emerged from the experiences of eTRIKS personnel as a formal deliverable.

Inputs and Outputs from related deliverables

See specific entries for each best practice

Description of work achieved

A wide range of practices have been developed over the course of the eTRIKS project to aid in productive translational research data and knowledge management. This document will highlight the following of these best practices.

- MetaData catalogue
- Open source software development
- Asset management
- Knowledge hosting
- Data security
- Open source software training
- Data privacy training
- Knowledge management platform design
- Information standards
- Bug reporting
- Deployment of experimental research software
- Legal agreements
- Knowledge curation
- Project Engagement

Each of the sections below is a short self-contained description of one best practice developed by eTRIKS.

Contributions to this deliverable come from personnel involved in all eTRIKS workpackages and represent cumulative learnings over the full course of the project.

Content of deliverable

Good practice guides follow, as per the report index:

Good Practice for developing open access software

Opportunity
There are many cases (Apache, Linux, Eclipse, etc.) in which open source software communities, in leveraging the capabilities of a wide breadth of contributors, have produced high quality sharable software. TranSMART version 1.0, the initial release of the base open source/license software product from which the eTRIKS platform was developed, required considerable effort to enhance functionality, security, stability and usability to serve as a productive platform for eTRIKS clients. With a community quality management system for tranSMART lacking at the start of the eTRIKS project, Work Packages 1, 2 and 6 needed to develop baseline quality practices to promote client success.

Challenges
eTRIKS faced several challenges with regard to the tranSMART open source development strategy, including:

1. tranSMART was commercially built for a single company and there was little community organization prior to its open license release.
2. Coordination of multiple different development/user groups having different priorities and development capabilities requires an open but controlled development environment that did not exist.
3. The technology stack of the initial tranSMART version included a commercially-licensed database management system that needed to be replaced.
4. eTRIKS developers were strongly skilled and motivated but limited in number.
5. Academic groups were resistant, as can be expected, to performing development activities that were not readily publishable. Quality management activities such as testing, defect resolution, and code refactoring are difficult to publish in peer-reviewed journals.
6. Traditional project management practices, including central "Command and Control" monitoring and delivery structures, may be challenged in a community development setting.
7. There was no commercial blueprint to guide the development of eTRIKS. Many successful open source/license products are developed based on existing commercial solutions having demonstrable prior success in serving clients.

Roles

Requirements Team
A small group of individuals drawn from eTRIKS partners and supported projects tasked with creating and prioritising client requirements.

Requirements manager
Manages, maintains and distributes the consolidated requirements list including the status of agreed development activities.

Developers
Academic, pharmaceutical and contract developers working on tranSMART development.

Academic and pharmaceutical testers who evaluate software, report defects and raise issues regarding usability, and functionality.

Approach
eTRIKS created repeatable quality practices that included.
1. Elicite and harmonize software requirements across all eTRIKS stakeholders (and often include requirements pertinent to the broader community). These requirements were refreshed for each successive version of the product.
2. Design, build and unit-test eTRIKS-specific deliveries prior to contributing code to the wider community.
3. Qualify and release as eTRIKS-branded products, outside of community releases, as necessary to support immediate needs of eTRIKS clients.
4. Participate in community-wide project delivery teams, to share code and design, integration and testing resources, leading to major community releases that can then be provided as core eTRIKS branded versions.
5. Qualify and deploy the major version releases on the eTRIKS Public Server prior to updating eTRIKS project instances.

Activities

Build a single prioritised requirements list
- Consult and interview a wide a range of eTRIKS (and broader community) stakeholders representing IMI projects (and other pertinent public private partnerships), pharmaceutical companies and software developers having tranSMART experience.
- Based on the elicitation above, create a consolidated list of requirements.
    o The consolidated requirements are also used as a communication tool to report progress against development targets.
- Prioritise the requirements based on overall value to the community, immediacy of client need, and anticipated ease of implementation, including any dependency on efforts conducted in parallel (a value vs. risk approach to prioritization).
- Personnel from Work Package 6 maintain the requirements, perform follow up elicitation as warranted and monitor progress of development against requirements.

Community Challenges with respect to scoping
Following the release of tranSMART v.2 eTRIKS undertook a 6-month requirements effort with the tranSMART Foundation to elicit and consolidat requirements from the larger community. Although a dozen organizations (including eTRIKS representing eTRIKS' current and potential clients) contributed to the consolidated requirements only two organizations, in addition to eTRIKS, were willing to commit resources against these requirements. The disparity of interest vs. resource (freeriding) substantially delayed development effort and the inability to share resourcing lead to individual development efforts.

The community approach to software development depends on a breadth of people and organizations willing to contribute resources. Having a critical mass of contribution within the tranSMART community appeared highly likely early in the program (pre eTRIKS v2/tranSMART v1.2) although community support diminished substantially following tranSMART v1.2. The potential for community support should be considered carefully for

any new program that intends to, or depends upon, leveraging a diverse set of participants. For tranSMART, development efforts that were most effective (including work to produce the tranSMART 1.0 by Janssen) were performed in isolation and provided to the community when complete. In addition to community uncertainty, as tranSMART use is marked by diversity of use cases, clients and data attributes, reuse of custom developed features proved problematic in many cases. Requirements that appear to be aligned during community scoping become specialized under the scrutiny of detailed scoping and execution. This tendency, which appears to be pertinent to tranSMART, can impede the delivery for individual clients as implementations of aggregate requirements risk being incomplete and/or unsuitable for use cases associated with individual clients or projects.

Synergistic development approaches

- Hackathons
    - Rapid advances in development can be achieved by bringing together academic and industry developers in multi-day "Hackathon" events focussed on specific development challenges such as integrating multiple research versions. Examples include the integration of I2B2 and tranSMART conducted in 2013, the Integration of the Harvard Sample Explorer tool conducted in 2013, the integration of Galaxy workflows conducted in 2014 and Hackathons in anticipation of tranSMART 2.0.
    - To improve the odds of success, hackathons ideally have:
        - Specified goals that can be reasonably achieved within the allotted time frame
        - A participant group sufficient to achieve the goals (not too small nor too large)
        - Experienced Developers who can lead and monitor the activities of the Hackathon
        - Developers who are prepared to use compatible development and code repository environments prior to the start of the hackathon.
        - By experience, the 2013 I2B2 and Sample Explorer integrations were development efforts that were not reasonably achievable as Hackathons. The tranSMART 2.0 hackathons were successful in the sense that face to face efforts expedited development through focussed team effort. The Galaxy hackathon was successful in preparing developers, who were inexperienced with respect to tranSMART, to use the API available at the time.
        - Successful hackathons have specific general goals and are designed to best attain the intended goals:
            - Creation of new features, refactoring, integration
                - Dependent on scoping and delivery expectations as well as developer experience and preparation.
            - Introduction for new developers.
                - Dependent on tractable but compelling problems
                - Assign experienced developers to serve in roles of guides, trainers, reviewers
            - Face to face interaction
                - Same as creation of new features above, and…
                    - Participation limited based on ability and importance to delivering the intended outcome.

- - Participants have prior experience working effectively together as a remote project team before assuming substantial expense in appropriating a venue, travel and accommodations.
- eTRIKS Labs
  - eTRIKS Labs are a set of eTRIKS branded tools that are hosted, both pre and post production, on a dedicated computing environment. This environment was prepared for eTRIKS developers to enable development of new individual applications, features and services for eTRIKS customers and have these applications positioned for robust agile evaluation.
  - An agile environment allowing users to access new pre-production platform elements for ad hoc testing and feedback.
  - A promotional environment for eTRIKS applications and services as such application and services matured.

- tranSMART Foundation
  - The tranSMART Foundation (tMF) was set up to coordinate development and release activities for the tranSMART community
  - eTRIKS personnel contributed substantially to release activities for tranSMART versions 1.x, and 16.x (which have corresponding eTRIKS version releases) and 17.1, which will be used as a base system during the eTRIKS extension
  - Three eTRIKS pharmaceutical partners also heavily funded the tMF as members providing annual payments as well as via major funders of the tranSMART 17.1 initiative.
  - Notes
    - Early on in the eTRIKS project the tMF and eTRIKS visions were, in part, redundant. This led to difficulty, and concern, with respect to the eTRIKS Strategic Advisors in differentiating each initiative.
    - The tMF provided coordination/management with some direct technical development/release services. However, eTRIKS provided most of the technical integration services for releases 16.2 and 17.1.
    - eTRIKS resources applied to tMF-branded efforts constitute a substantial portion of the tangible outcomes of the tMF.
  - Insights with regard to partnering, by IMI projects, with open community organizations
    - Ensure that expectations for promotion and credit assignment are understood, agreed upon and acceptable
    - Ensure that product licensing, copyright policies and policies concerning intellectual property are in place and acceptable
    - Ensure responsibilities and activities are clear and acceptably differentiated. Ensure any competitive activities are acknowledged and, if partnership remains possible, competitive activities are pursued ethically and in a manner consistent with applicable legal statutes.
    - Ensure that resource management responsibilities are clear. Management of collaboration resources by the open community organization should be per explicit agreement. Collaboration management must intervene in cases in which external entities assume,

or have the perception of assuming, inappropriate direction over collaboration personnel.

Testing
- Supported projects and eTRIKS partners assumed quality control responsibilities and applied resources to participate in testing.
- Where possible, experienced test teams were leveraged to expedite the writing of test plans and the execution of test cases. For eTRIKS, Sanofi, Pfizer and other pharmaceutical partners having in-house tranSMART experience contributed personnel resources to expedite testing.
- Through the eTRIKS Labs model new functionality could be tested and assessed independently of the main public version of eTRIKS before incorporation.
- Testing processes and training experiences were brought together to develop a suite of automated tests to speed the quality assurance process for new versions of the eTRIKS platform.

Documentation and control
- GitHub provides a flexible accessible environment in which multiple development activities can be coordinated and merged.
- A centralized code versioning environment is essential when development activities are being carried out asynchronously by several groups and allows each group to operate independently of other groups until a new release is planned.

Comments
There are clearly cases in which developer communities have produced highly robust software leading to substantial distribution. Nevertheless, challenges abound for satisfying clients through community software production. The size and diversity of both software and client communities, similarity and precedence of need across software clients and the potential to drive economy via shared development impact the potential of open source/license efforts. Programs that intend to leverage community efforts must consider the structure and approach to community production very carefully and monitor progress.

Good Practice for Project Engagement

Opportunity

Establish and nurture productive relationships with potential partners of the client project and plan how eTRIKS can help meet their knowledge management issues.

Challenges

- Projects are often unfamiliar with the activities of eTRIKS
- Projects are not consistent in how they approach knowledge management
- It can be unclear how knowledge management is governed within a project
- Knowledge management challenges faced by projects are highly variable
- Knowledge management resources available internally in a project are highly variable
- There is often a significant under appreciation of the resource required in a project to deliver good knowledge management
- Project needs change over time both due to new circumstances and better understanding.

Approach

It is crucial to understand the information needs of the project in the context of the science they are trying to carry out. Engagement is best mediated through trained scientists as well as information specialists. Attention must be given to the capabilities within the project to ensure that eTRIKS effort is value added and project informaticians understand that eTRIKS will productively supplement their roles rather than compete with, or otherwise impede, their duties. Active engagement, with expectations being set for the engagement process, is the key to success.

Activities

First Contact

eTRIKS maintains close interaction with new and established projects through a network of contacts within consortia partners and awarding bodies, but initial contact can come from unexpected sources within or outside the process. It is the role of everyone in eTRIKS to be aware of possible engagement opportunities and to alert the Engagement and Outreach team of prospective clients as quickly as possible. A contact name and email address should be obtained and as much general information as possible on the aims and background of the project. It is essential that first contacts are followed up promptly.

Exploratory Meeting

The first step is to establish a teleconference as quickly as possible between the eTRIKS Engagement and Outreach team and the prospect's Information Scientists, or equivalent, accountable for data and knowledge management. The Project Coordinator should also be invited as they will likely have a responsibility to approve any work between the projects. Before the meeting the contact in the project should be sent the eTRIKS website address so they can begin to understand the scope of what eTRIKS can offer and the eTRIKS Engagement and Outreach team should seek as much information as possible about the project.

- Understand the aims and drivers of the project

The meeting should focus on the aims of the project and any particular information challenges they foresee. eTRIKS staff should familiarize themselves with the prospective project and devise a plan to elicit useful information prior to the meeting. The intent is to drive beyond the situational (what is the project about, what partners are involved, etc.) to specific concerns and problem that could be advantaged by an engagement with eTRIKS.

- Explain the aims of eTRIKS

As part of the same meeting the eTRIKS representatives discuss the aims and capabilities of eTRIKS. Some of the challenges that have been encountered in other projects should also be described if these appear relevant. Each service area of eTRIKS should be introduced.

Use the project information checklist to gauge the likely scale of the engagement. The outcome of the meeting should be a high-level view of the areas in which eTRIKS can complement the project's own resources and a plan for further engagement as necessary.

- Project information checklist
    a. *Scientific questions*
    b. *Problems in data management*
    c. *Project use cases - functionality needed*
    d. *Data Management Plan*
    e. *Data types*
    f. *Is there a need for public data?*
    g. *Project data becoming public?*
    h. *Data curation needs*
    i. *Data sustainability plans*
    j. *Is there a dedicated data manager within the project?*
    k. *System deployment options*
    l. *Special security needs?*
    m. *CDA/MTA/others will be needed?*
    n. *Project resources available*
    o. *Data is anonymised or personal?*

Establish contact people and processes
If additional discussions following the exploratory meeting are desired then an eTRIKS Account Manager should be identified to work with a counterpart in the client project to plan and drive the subsequent activities.

Identify areas where eTRIKS can contribute to the project
The scale of the requested engagement can be anything from access to good practise guidance to full project support including data hosting. At an early stage the scale of the likely engagement needs to be established. For light engagements resource within eTRIKS and the client project are likely to be readily available. More extensive engagements will require dedicated resource and will need to be agreed by the eTRIKS Executive Committee and the equivalent governing body in the client project.

Plan the engagement
As described above, the engagement may be very small or very large. A plan should be created identifying those activities required of eTRIKS by the project. This may require teleconferences and face to face meetings between members of the project teams. The plan should identify what resources may be required from each eTRIKS Work Package.

For a small scale engagement this need be no more than a list of resources and how to access them. For larger engagements it may form part of a legal agreement between the projects. It should be recognised that the plan is not immutable and should be reviewed regularly.

Set up monitoring and management of the engagement
- Add the project to the list of eTRIKS supported projects.
- Create an entry in the eTRIKS CRM Smartsheet with the details of the contact people, a link to the project and the required activities.
- For larger scale engagements (data hosting, systems development, extensive curation) a specific teleconference series should be created to manage the activities.
- Significant changes or developments in the project or engagement should be recorded in the CRM and reported to WP6.
- Operational issues should be raised at the weekly eTRIKS Operations TC.

Roles

Initial Contact
In principle this can be almost anyone in the client project, in eTRIKS or in IMI. They will have at least some awareness of both eTRIKS and the client project and can see a possibility for beneficial engagement.

Project coordinator
Within the client project, the coordinator needs to take a role in managing the initial contact as this will establish the engagement as a potentially valuable activity for the project.

Engagement and Outreach team
Through Work Package 6 eTRIKS maintains a group of information and knowledge scientists able to understand the needs of a project and align the needs to the capabilities and assets of eTRIKS.

Project information scientists
The project information scientists, data managers, informaticians and similar roles will most directly benefit from an engagement with eTRIKS. However, they also may view eTRIKS as a threat to their roles. It is essential that a good early relationship be established between these groups and eTRIKS to understand the role of eTRIKS as a facilitator rather than a replacement of the project's own capabilities.

Account Manager
Accountable within eTRIKS for the engagement with the project. The ideal situation would be for the Account Manager to be a participant is in both project consortia. The Account Manger should be skilled in the traditional activities of a Business Analyst and have sufficient knowledge of the science and information areas to lead discussions between the projects.

eTRIKS work package contacts
eTRIKS delivers benefits through its internal Work Package structure. Each of the work packages should maintain a capability to respond to requests from supported projects coordinated through a WP contact in cooperation with the Account Manager.

Good Practice for Asset Maintenance

Opportunity

1. The impact will tremendously increase when assets are maintained and sustained
2. Best practice implementation will lead to precompetitive advantage and clear step actions
3. Cross consortium expertise and stakeholder engagement
4. Avoidance of redundant efforts that diminish the potential of public and private funding
5. Better understanding of the disease mechanism and develop strategies to treatment

Challenges

- Maintenance of data after project funding ends is often an issue that is considered late in the project.
- Funding research infrastructure to support software and data assets after project termination is difficult and requires planning and post-project funding.
- Promotion of assets diminishes once a project ends leading to a subsequent of decrease of awareness and use.
- Assets may be applied differently across stakeholders, including public-private partnership projects, pharmaceutical companies, SME's, IMI and academics.
- Licensing and other legal and ethical barriers to asset use.
- Enthusiasm for community projects can diminish. Collaborative ventures require new participants, and possibly turn-over of personnel, to maintain community momentum.

Approach

"The use of business model canvasses as outlined in figure 1 are a viable means to develop the elements of an asset maintenance plan. Having all the requisites together in one canvass is a very effective mechanism to promote sustainability strategies which require input across a diverse stakeholder group (pharmaceutical, academic and SME stakeholders).

Asset Maintenance Canvas

What obstacle (s) could be expected?	What problem (s) does its solve?	Unique Value Proposition	Location?	Supported by?
	Demands for potential customers?	End users?	How is the asset funded?	Promotion:
Rough Costs for maintenance		Road map for future development?		

Creative Commons license - derived from the Business Model Canvas as developed by www.biosciconsulting.com

Activities

Asset identification

All Work Packages as part of their 4th year activities were required to identify all assets they had created that would have lasting value beyond the end of the eTRIKS project. The lists were combined into a single Asset Register. For each asset a primary responsible person and organisation was identified

Component model - eTRIKS Lab development

eTRIKS is built on a core central technology with added plug-in components. The plug-in components developed by eTRIKS are branded eTRIKS Labs. Each lab is the product of a single eTRIKS partner and as such there is clear line of sight to the group who are tasked with maintaining it.

Interviews

Interviews were conducted with all asset owners to understand how and where the asset would be maintained. The asset maintenance canvas (see above) was completed for each

Report

The series of interviews was drawn up into a single report that documents for each asset how and where it will be maintained.

Asset	Maintenance
eTRIKS HS	ICL
eTRIKS AE	ICL
SNF	EISBM
WGCNA	EISBM
SmartR	tranSMART
Disease Maps Lab	EISBM
Disease Networks	EISBM
HiDome	tranSMART
Federated Data Integration	DSI
ImmunoMap	DSI
Play Decide	BioSci
Standards starter pack	OXFORD
Data Catalogue	UL
Training	UL
Galaxy	UL

Table 1: Summary of the asset maintenance interviews (as show in appendix table 1a-d) showing which institute will house and maintain the asset.

References
eTRIKS website – www.etriks.org
websites of eTRIKS partners
Offer insights best practices for each asset in the data value chain
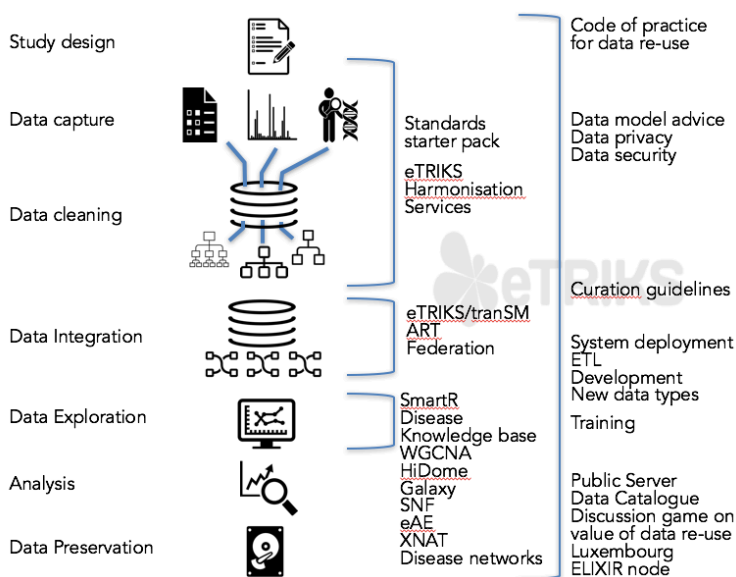
Fig 1 Component map of eTRIKS assets in the data value chain

Good Practice for Data Hosting for supported projects

Opportunity

Some projects are unable to offer a supported platform from which their translational researchers can explore the related clinical and high dimensional molecular data sets collected or created by the project. eTRIKS is able to offer a tailored hosted service to provide secure access at no cost to a secure tranSMART instance for the supported project.

Challenges

- The skills needed to set up, optimise and maintain a tranSMART instance are rare
- The correct form of legal agreement needs to be established before data can be transferred from an IMI supported project
- Data must be correctly curated to be usable in tranSMART, data curation is a complex task in its own right
- Well managed security control is essential to protect the data. Security is a joint responsibility of the project and eTRIKS.

Approach

eTRIKS provides a secure cloud-based environment in which translational research data can be shared and explored within a project. A common hardware platform is maintained at CC-IN2P3.

This core hardware platform is provided to supported projects through the Open Stack Cloud Management Platform using Apache TomCat and HTTP Server, solR and Postgres. All components are open source through GPL or similar, ensuring there are no licensing costs to supported projects.

Each supported project has a fully independent instance of the eTRIKS Knowledge Management Platform (KMP) and data manipulation area that is maintained in isolation from all other eTRIKS instances. Management of the instance and associated study data is jointly carried out by the eTRIKS

Activities

Agree the data in scope

Not all data are necessarily in scope. Those data sets that are in scope need to be checked for suitability.

Set up legal agreement

A suitable legal agreement must be set up between the project and eTRIKS. As the activities will include data hosting the agreement needs to be based on the eTRIKS MTA template. It is possible to begin work ahead of the full signing of the agreement. Data review and curation and setting up a test environment can all be done with just a CDA.

Create a dedicated eTRIKS KMP instance

WP1 are tasked with maintaining the online services for eTRIKS. The WP1 leaders agree and allocate resource as necessary to create and maintain the environment required by the project. The setup includes a production environment and a test environment.

Curate and load the data

Raw data is uploaded via SFTP to a cinder block storage volume.

The processes described in the "eTRIKS good practise for data curation" is used to prepare the data for hosting. This involves curation expertise from WP4 working with the project scientists through the Account Manager.

Alternatively the project may choose to curate the data themselves outside the eTRIKS environment. In this case the curated data are uploaded and transferred into the project test environment directly.

All data transfers to and storage on the eTRIKS platform is encrypted.

Test the eTRIKS KMP platform

Before release to the full project a subset of project researchers are invited to test the platform to ensure that it is operational and the data is available. Any necessary changes are managed through the WP1 and WP4 teams as necessary

Access control

Generally the project agreement will require that data should be shared within the project but not outside the project.

A project data Czar (PDC) is identified with the role of ensuring only those people who The data sets are made available through the eTRIKS Portal but in a restricted environment with full username/password controlled access.

Access is requested through the eTRIKS portal. The request is passed to the PDC for approval and access is only granted if both the PDC and the eTRIKS administrators agree that access is appropriate. The PDC has access to an administration dashboard linked to the project data space in which user accounts can be managed. New users must set up and regularly change strong passwords.

Annual reviews of users are carried out to ensure that all accounts still require access to the data.

If the data set contains any data identified as personal data then before access is granted a new user must complete the Data Privacy Training programme.

Maintenance

When new versions of the eTRIKS KMP are created the service team liaise with the Account Manager and the PDC to agree the timelines for testing and upgrading the Project instance. It is strongly encouraged to upgrade the instance as quickly as possible for security and supportability reasons.

Roles

WP1 Service experts

WP1 maintain the public and project eTRIKS service environments.

WP4 curation experts

WP4 curation team are experts in creating high performance eTRIKS Postrgresql data bases from clinical and omics data sets.

Account Manager

An eTRIKS resource that maintains the relationship between a project and eTRIKS

Project Coordinator

The leader of the supported project with accountability for making decisions on the

Project data Czar

A specific role within a project with data management knowledge to provide interface between the project users and the eTRIKS KM platform

Project researchers

Scientists in a project who which to explore their data

References

D1.2 Hosting and Management of the eTRIKS KM Platform
D1.7 Publication of the current hosting model, hosting requirements and support for eTRIKS supported projects
D5.7 Legal access guidelines
D7.6 Security Review

Good Practice for Open source data security

Opportunity
eTRIKS is mandated to manage data on behalf of other IMI projects.   These data may include personal data as defined by Directive 95/46/EC and the General Data Protection Regulation (GDPR) which will be introduced in May 2018. As such it is necessary to manage the access and security of the support projects' data to an appropriate level.

Challenges
tranSMART was designed as an in-house system for a single pharmaceutical company. As such, it relied on the company firewall for much of its security layer. This was insufficient both technically and administratively for a publicly accessible platform managing personal data.

The security provisions in tranSMART had never been fully tested by an experienced system security analyst

Approach
Security for the eTRIKS version of tranSMART was enhanced through system development, system testing and access control management.

The eTRIKS public instance contains only public and fully anonymised data sets and it was decided as a matter of policy that the eTRIKS public instance should be available to anyone. As such there is no access authentication required, but the IP address of all access attempts and the use of system features are recorded.

Activities

System security design
LDAP layer added to manage system access for all non-public instances of eTRIKS KMP. Anti-virus software is deployed on all eTRIKS KMP components with intruder alerting to the System administrator in the case of any suspicious activity.

eTRIKS curates and performs ETL services on behalf of the supported projects. The data preparation and staging areas are further protected from access by a firewall and are only accessible by data curators and ETL engineers. Thus only a copy of the final processed data is available to end users.

System security review
A full security review was conducted by an experienced Security Review Officer from GSK. The review looked at the threat profiles for the data and reviewed the processes and systems in place to protect the data. The review looked at all aspects of access to the eTRIKS KMP including protection against malicious and inadvertent threats. This was reported through Deliverable D7.6.

Access control
All support project instances of eTRIKS KMP are protected by LDAP security layer. In order to be granted access to the instance an applicant must apply through the Project Data Czar. The project data czar, seconded by the Account Manager for the project recommends either allowing or refusing access for the applicant. This is then approved by the System administrator and access is granted. It is mandatory for all new users to complete data privacy training before their account is activated.

All use of the system is logged and monitored.

Access control maintenance

A list of users with access to the project instance is maintained.   The Project data czar notifies the System administrator in the event that anyone no longer requires access to the data and their access is removed.

An annual review of users is also carried out to ensure no individual is inadvertently overlooked.

Roles

Security review officer

Leads the review of the security layer in eTRIKS

Project data czar

Vets application for access to a project's eTRIKS instance.

System administrator

Ultimately responsible for the smooth and secure working of the eTRIKS KMP. Monitors all accesses to the eTRIKS KMP

Account Manager

Seconds the Project data czar and liaises between the project and eTRIKS for example in arranging data privacy training

References

EU Directive D95/46/EC
General Data Protection Regulation (GDPR) 2018
D7.6 Compliance assessment of the Security of the eTRIKS Platform

Good Practice for Data Privacy Training

Opportunity
Provide data handling stakeholders with sufficient information to be able to determine safe practice and when further expert advice is necessary.

Challenges
- Stakeholders or projects hold and process data in different countries, with different rules
- It is not always clear who is responsible for data handling at the project level
- Data is processed in different ways in different projects
- European law and country level laws have a number of grey areas, often related to being able to define 'reasonable' against a particular circumstance or event
- The knowledge of potential trainees about data privacy is variable
- The knowledge of potential trainees about how data is processed within their projects, by other project partners, is variable
- Data privacy law, including regulations and the question of what is anonymised data, is a moving target

Approach
eTRIKS identified the need for data privacy training after interactions with client projects that had demonstrated that many project members had a limited understanding of data privacy issues. The intention was to raise the level of understanding of the data privacy issues, so enabling safe data re-use and sharing in order to make the most of eTRIKS tools and services. A peer to peer approach was used with eTRIKS project members undertaking the training for colleagues in client projects.

Activities

Creating content
eTRIKS used a course outline drawn up by a data privacy expert, focused on the re-use of data in research projects. This was then formed into slides accompanied by a script. By defining the question to be answered the content could be refined towards more specific issues relating to that question.

In drawing sources to build the training, eTRIKS was able to use the code of re-use of data developed across different projects and organisations. This source document provided a clear direction for the training in the wide field. Further sources included wording from the regulations themselves and case studies to demonstrate the risks in data privacy and the steps that can be taken to reduce these risks. Having someone familiar with or expert in the field is important in being able to identify the right sources and to review the training material, although the training can be delivered by non-experts with some data privacy understanding.

The training was designed as entry level information and was split into 4 sections:
- Why is data privacy important?
- What is personal data?
- Personal data protection principles
- Developing our own guidance

- Conclusion

Identifying trainees

Identifying potential trainees was carried out through using existing project structures. This was felt to be the best way of both getting to the people that need the information and of ensuring that client projects consider and identify who processes data within their project. In doing this eTRIKS utilised established communication avenues set up by the eTRIKS Engagement and Outreach team. The projects were identified as 'supported' by eTRIKS and therefore the target audience was these projects.

Promotions

Even with a defined audience who nominally have a need for a no cost service, it is still important to describe the benefit and purpose of the training. This was done in communications with the client projects, highlighting that attendees need to be able to recognise data privacy issues and that the course is designed to provide that basic understanding to be better able to undertake their work.

Delivering the training

The training slideset was accompanied by a script, which allowed the trainer to deliver the training in a well-paced and considered manner. Due to the geographical spread of the eTRIKS client projects the training was delivered online, through a screen share conference call system (gotomeeting/gotowebinar).

The group responsible for ethics within eTRIKS composed the data privacy training and presenters were drawn from this group, who had the necessary knowledge to answer most questions posed in the sessions.

The training sessions were offered for free and were recorded, allowing people the ability to re-view the session.

Future contact

Although eTRIKS was not resourced to offer a question and answer service, representatives from client projects were able to ask follow up questions and the answers or signposts from the questions were provided by email.

To enable the training to have a longer life span, a training video was produced. This followed a refined and updated version of the slides and script and was made available online via the eTRIKS website.

Roles

Trainer

The trainers were people from within the WP team who had particular knowledge and expertise in data privacy, and included a data privacy expert and a data scientist who has held an IMI project co-ordinator position. The trainers were also responsible for building the training content.

Training co-ordinator

A central role in the eTRIKS project, the co-ordinator helps to facilitate and promote training sessions for the client projects.

Review team members

Team members were drawn from the WP responsible for ethics (WP7). The team reviewed the training slides at various iterations during the project, including during the twice monthly WP calls.

Outreach and engagement team members

Members of the outreach and engagement WP (WP6) provided communications lines through which to reach potential trainees.

References

Code of practice on secondary use of medical data in European scientific research projects Anne Bahr & Irene Schlünder - International Data Privacy Law 2015 - *doi: 10.1093/idpl/ipv018 (free access)*

European Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data (24 October 1995)

Italian Data Protection agency website :
http://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/1544575

Int J Law Info Tech (2010) 18 (4): 356-378.
Google case: Google Italy vs. Associazione Vivi Down
(http://ijlit.oxfordjournals.org/content/18/4/356.full)

New York Times: April 21 2010: Indian Tribe wins fight to limit research of its DNA

Good Practice for tranSMART training

Opportunity
Any system as complex as tranSMART requires effort to understand and use to its full potential. Training provides an effective way of sharing the deep knowledge of the eTRIKS KMP and tranSMART garnered by the eTRIKS partners over the 5 years of the project. By building and offering a suite of training aids, eTRIKS provides a lasting valuable legacy for all tranSMART users.

Challenges
- The depth of knowledge of tranSMART (and bioinformatics) is highly varied among translational researchers, even within a single organisation
- Trainees derive vastly more from being trained on data they recognise but this requires that the trainer is familiar with the data and has ethical and legal access to it
- It is difficult to bring together people from widespread projects for a face to face training session
- To be of value training must be regularly updated to reflect the latest software version

Approach
Since the beginning of the project, eTRIKS has recognised the pivotal role of training in developing the user community for eTRIKS supported projects. To this end we have created a panel of trainers and a training coordinator to offer training services on all aspects of the eTRIKS platform. As the project nears its conclusion we have consolidated this knowledge through the Training content team to provide a maintainable asset.

Activities

Consistent
Training as an activity is recognised as of ubiquitous value across eTRIKS, WP1, 2, 3, 4, 6 and 7 all have roles or identified activities in the eTRIKS DoW around training. To maintain consistent quality this is coordinated through WP6 as part of the engagement activities.
By creating a Training Coordinator role and a cadre of trained trainers, eTRIKS is able to coordinate and maintain the consistency of the delivered training
A training content team drawn from main partners has been established to create and prepare appropriate training materials that can be easily maintained as part of the asset legacy of eTRIKS.

Tailored content
Training has been developed and delivered on
Using the eTRIKS KMP version of tranSMART
- Aimed at all scientists using the eTRIKS KMP
- Focussed on how tranSMART works and the use of analysis tools to address typical translational research use cases
- Can be tailored to the skill level of the audience
- Can be delivered using a project's own data (or to randomised data with a similar composition)

Curation of data
- Aimed at scientists in charge of project data curation and management

- o Shows how to prepare data and load it into a tranSMART instance
- o Details appropriate use of standards
- o Provides basic skills of trouble shooting in tranSMART data sets

Data privacy
- o For all scientists working with personal data
- o Explains the risks and considerations when working with personal medical data
- o Not intended as legal advice

Trainer training
- o For eTRIKS partners wishing to deliver eTRIKS training
- o Covers the standard end-user eTRIKS training and how best to deliver it
- o No formal certification, but a register of trained trainers is maintained

Platform and service administration
- o For managers and systems administrators of distributed instances of the eTRIKS KMP
- o Installation, maintenance and administration of the full eTRIKS software suite and supporting tools

Flexible delivery

It was quickly recognised that no one delivery method would match the training needs of all eTRIKS supported projects. To overcome this eTRIKS offers several forms of training

Collective training
- o class room - organized in a venue of one of the three academics partners (CC-IN2P3, Lyon; University of Luxembourg, Esch-sur-Alzette; Imperial College, London)
- o workshop - organized around a conference or satellite meetings

Individual training and Self-learning
- o virtual meeting technology
- o video and regular Webinars
- o training blog system

Customized training
- o class room - dedicated to a pharmaceutical company
- o class room - dedicated to a specific project

Sustainable value

To maintain the value of the training materials a training content team consisting of experienced trainers from several eTRIKS partners was brought together to plan the materials.

The outcomes of this planning are:
- • A consistent training slide set with training notes
- • A series of webinars covering typical translational research use cases
- • A series of webinars on new features of tranSMART developed through eTRIKS
- • A full webinar of Privacy Training
- • Space on the eTRIKS website to assist trainees in finding training resources

All these will remain available as resources to end users and trainers after the completion of the eTRIKS project

Roles

Training coordinator
A role in WP6 in charge of managing and reporting on all eTRIKS training activities

Trainers
A group of 3 – 5 individuals trained and able to provide eTRIKS training

Training content team
The trainers augmented with systems administrators tasked with creating a lasting eTRIKS training legacy

Systems administrators
Manage the set up and security of training instances and advise on the training content team

Account managers
Responsible for advertising and identifying training opportunities within supported projects

References
eTRIKS website training pages - https://www.etriks.org/etriks-training/
D6.6 Final Training Curriculum

Good Practice for Standards Recommendation

Opportunity

Standards are essential for the management of data in a FAIR (Findable, Accessible, Interoperable, Reusable) way.   Without standards it is impossible to know if two entities can be compared within or between studies.   The need for good data standards within the eTRIKS Curation Team pointed the way to a wider dissemination of our standards knowledge. With the increased access to study data envisioned by IMI it is imperative that guidance on standards is available to projects.

Challenges

Creation and maintenance of standards is a complex task with few dedicated resources
Translational research (TR) data are complex with multiple modalities needing to be managed in a single study or between studies
Standards for TR entities are spread across multiple domain specific areas – it is not easy to find all the relevant standards
Some standards are better than others within a single domain
Standards are still evolving for cutting edge modalities such as GWAS

Approach

Bring together as wide a group of experts as possible to identify the best standards in each area of TRD. Use this knowledge to provide an "index" of good standards in TR – the Standards Starter Pack which is maintained by regular updates

Activities

Consult widely within eTRIKS and in the broader standards community

A Standards Advisory Board was created to pool and call on as wide a range of expertise in standards deployment as possible.
Engage deep knowledge of clinical and omics data standards through CDISC and Oxford University partners in eTRIKS
Engage with external organisations working in the Standards area,such as GH4GA

Develop the Standards Starter Pack (SSP)

The combined knowledge was used to create the first draft of the SSP. This is divided into two main sections, one covering the need and theory of standards use, the other containing a detailed list for each entity encountered in translational research data sets of those standards that eTRIKS believes to be robust and generally applicable.
Users are able to

Maintain the SSP

The SSP is maintained by annual reviews and updates led by the standards coordinator the team at Oxford University but consulting widely in the standards community identified through the StAB at the beginning of the process.

Roles

Standards Coordinator
Charged with managing the SSP, resolving any conflict in standards and releasing the annual update to the SSP.

StAB

Standards Advisory Board consisting of expert from eTRIKS and external organisations. The StAB members provide expert input and review the SSP before each update.

References

The eTRIKS Standards Starter Pack

http://www.nature.com/articles/sdata201618

Good Practice for IMI MetaData Catalogue

Opportunity

The IMI Data Catalogue centralizes ongoing and past IMI project level metadata with a focus to expose general information regarding research studies.

Translational research scientists want to know what projects are underway (or have been undertaken), corresponding descriptions of motives and methods and how to access study data, if possible.

This Data Catalogue:
1. Promotes the awareness of medical projects, especially IMI and public private partnerships.
2. Preserves a directory of completed research projects with refernces to stakeholders and data.

Challenges

The database could eventually manage a large number of study descriptions and curation is required to remove redundancies and ensure the quality of each project entry.
Promoting the Data Catalogue to increase use.
Determining the appropriate amount of information to include in an entry, enough to be informative without burdensome to those providing the information.
Engaging project contacts in a manner that encourages their participation in providing data for the catalogue.

Approach

A set of interviews were performed to understand the potential and feasibility of the catalog, including what information would be valued and how the catalog could be deployed. An initial version of the catalogue was developed at the UL using the open source CKAN cataloging application. With the application in place, project leads in the IMI were contacted and encouraged to provide information to create catalog entries corresponding to their project.

Interviews:
Interviews were conducted with the following projects
- DRIVE-AB
- ADAPT-SMART
- BTCURE
- EBiSC
- ADVANCE
- iABC

- AETIONOMY
- eTOX
- APPROACH
- ENABLE
- EPAD

General impression:

Most project coordinators were aware of the FAIR concept and supportive of providing their data.

Template building

Each project received a template requesting pertinent information for the catalog. The template required ~30 minutes to complete and was delivered by BioSci Consulting. IMI leadership supported the outreach campaign by providing IMI contacts and encouraging IMI project coordinators to participate in registering their project to the catalogue. Over 100 projects were contacted and, at the time of this writing, thirteen projects have provided catalogue entries. It has been proposed that the template be made mandatory for new IMI projects and be completed before project launch.

Potential limitations

There are a variety of efforts aimed at creating project catalogues underway. Differentiation and consolidation across these efforts is likely needed but will be problematic.
Post-consortium promotion and maintenance is an issue.
Not all projects are interested in registering.

Recommendations:

Extend/ levels of study descriptions with mandatory and recommended details to promote comprehensive project registrations.

Roles

eTRIKS WP6

Started the metadata repository idea and in collaboration with UL/ELIXIR created the template based on the interviews report. Coordinated the entire project in collaboration with ELIXIR UL.

ELIXIR/UL

Developed and maintained the catalogue application.

IMI

IMI office promoted the catalog concept to projects.
FAIRIFICATION call (fall 2018) will substantially elevate the awareness of the importance of the catalogue.

References

Slide deck pitch distributed and publically available on eTRIKS and ELIXIR/UL website.
Awareness building via media channels and on IMI website
Use case scenario's Preservation— Oncotrack/ELIXIR
http://datacatalog.elixir-luxembourg.org/ckan/about
http://datacatalog-dev.elixir-luxembourg.org/limesurvey/?r=survey/index&sid=779596

Good Practice for bug reporting and resolution

Opportunity
All software has errors and inadequacies. An effective mechanism for allowing users to report these and for skilled staff to investigate and resolve them improves the system and the confidence of the users.

Challenges
- eTRIKS is a complex knowledge management environment with multiple potential components. Issues can be difficult to trace to their cause
- User groups are not always using the same version
- Dialogues between issue reporters and resolves need to be maintained
- Some issues are "features" and need to be recorded as change requests not bug fixes
- Maintaining knowledge of past issues.

Approach
We have put in place a bug reporting and tracking system eTRS (https://cc-usersupport.in2p3.fr/otrs/customer.pl) based on the commercial OTRS ticketing system (https://www.otrs.com/). Triage of issues ensures they are passed to the most likely group to resolve. Issues are raised at weekly Operations teleconferences where Operations staff and developers work together to identify and resolve issues.

Activities

Report an issue – raise a ticket
Anyone using eTRIKS KM Platform has access to the "Report a Bug" facility in the Utilities tab. This directs the user to the eTRS system where detail of the issue can be recorded. eTRS is username/password protected but any eTRIKS user can request an account to report issues. The issue reporter is asked on which version of the system they have encountered the issue. A ticket is raised and a unique identification number allocated to it.

Allocate issue to likely resolver
Tickets are triaged by a small team of eTRS administrators. This group has good knowledge of the overall system and is able to resolve some issues immediately. Others they allocate to the most likely person who can resolve it. The potential resolver is notified and is able to communicate with the ticket raiser through eTRS.

Raise at Operations TC
For issues that cannot be immediately resolved an Operations TC is run weekly at which all outstanding tickets are reviewed and progress monitored. The Operations team includes representatives from all work packages and groups contributing to eTRIKS. Additional eTRIKS participants can be invited if needed to look at specific issues.

Maintain dialogue with issue reporter
eTRS has a facility to allow email communication between the issue raiser and resolvers. The advantage of using this over direct email is that a record is kept of the communications to aid future ticket resolution. Requests for further details or for the issue reporter to test the possible solution can be made, and their responses captured.

Resolve or mark as a future request

Resolution can take many forms and may take many weeks or even months to reach. Resolved tickets are closed. Those that cannot be resolved in the current version of the system are marked as future requirements and the closed. The Issue reporter is informed of the closure of a ticket and has the option to reopen it if they believe the issue is not resolved. In the event that a ticket has lain dormant for several months with no response from the issue reporter it may be closed by the eTRS administrators if they believe the issue has been resolved.

Maintain awareness of common issues

Some issues occur frequently. For these cases preserving the dialogue between previous issue raisers and resolvers allows the eTRS administrators to quickly recognise duplicate issues and provide quick resolutions. Issue reporters can also review past entries.

Roles

Issue reporter

Anyone using an eTRIKS supported version of tranSMART

eTRS administrator

Maintain the eTRS system and make initial triage review of new tickets

eTRIKS Operations team

Group meeting weekly to review outstanding tickets

Issue resolvers

Often the developers of system components

Good Practice for eTRIKS Labs deployment

Opportunity

Novel bioinformatics software functionality can be difficult to develop from the initial idea to a fully working prototype. Research benefits from early feedback to help guide the idea to a useable state. The eTRIKS Labs initiative provides an environment in which bioinformatics software ideas can be shared and developed.

Challenges

- It is difficult for many students to provide access to their software
- Setting up the first usable version
- Getting an environment that can be generally used
- Ensuring adequate protection for the IP of the idea
- Encouraging use of the Lab.

Approach

Once an informatics idea has reached the point that it can be shown, eTRIKS provides a managed environment in which the software can be shared with potential users to gain their feedback. The eTRIKS Labs environment is open to public access and users are encouraged to feed back ideas to the Lab creator. By deploying the software outside the developing institution we are able to encourage and verify the general usability of the software. A team of infrastructure experts in WP1 work with the Lab owner to deploy the software, this approach reduces the need for the developer to be deeply knowledgeable about all the components needed in the back-end of their application.

Activities

Create a description of the Lab
Before acceptance a minimum standard of documentation in require in the form of a "One Pager" description. The Lab Owner provides key information about the purpose, installation requirements, dependencies and contact point for the Lab.

Approve Lab for deployment
The eTRIKS Labs Work-team are accountable for ensuring the Lab is of suitable standard to deploy. They work with the Lab Owner to ensure that lab provides useful functionality, is documented adequately for public use and the resources are available to make the Lab available.

Liaise with Work Package 1 to ensure Lab is usable
WP1 are accountable for the eTRIKS Public Server and are skilled in deploying the infrastructure needed for bioinformatics software. WP1 maintain a Labs environment into which new eTRIKS Labs can be added with minimum disruption.
Working with the Lab Owner, the WP1 team provides a mechanism for deploying the Lab as a robust, supportable software tool with adequate infrastructure that does not violate good software deployment principles.

Created deployable instance
Deployment may require addition of new back-end functionality. This will be discussed and deployed as necessary. Deployment may also require access to third party software or data

bases in which case the eTRIKS Labs Work team will work with the Lab owner to secure suitable rights are obtained for the third part components. Creation of a DOCKER or similar file to allow users to download the software may also be undertaken.

Attach disclaimer
All the software deployed in the eTRIKS Labs environment in experimental and cannot be guaranteed to work or to be entirely secure. We therefore add a disclaimer to all Labs before they are made publicly available.

Disclaimer

The eTRIKS Labs modules provided via this website are offered to the public as freely available resources, solely for non-commercial research use. Some aspects of these experimental modules may still be under development, and are made available with no warranties about the completeness, reliability, accuracy and security of any of the software packages or the services. Please bear this in mind, especially if you wish to analyse personal and/or confidential data. Some of the services provided use public data and/or software sources that are available through third parties on the internet – these are identified by links provided under 'Acknowledgements' on the respective module pages. If you use or re-distribute these public data or software for any purposes, you are responsible for adhering to the license requirements of these public data sources.

Advertise the Lab
A link to the new lab is provided on the eTRIKS Portal page and the etriks.org website. Also the website runs a banner advertisement highlighting the new Lab. The Lab Owner is encouraged to create a webinar presentation and short videos that can be further used to advertise the work.

Beyond the Lab
Lab software that is shown to be stable and valued by users "graduates" to the core eTRIKS Knowledge Management Platform and becomes part of the regular eTRIKS maintenance and update cycle.

Roles

Lab Owner
The researcher who has created the lab software

eTRIKS WP1
eTRIKS service operations group

eTRIKS Labs Work-team
A small team of eTRIKS partners from WP1, 2 and 6 tasked with encouraging and vetting eTRIKS Labs ideas

eTRIKS Website owners
Web masters for the eTRIKS Portal and .org websites

Good Practice for Legal Agreements for Data Management

Opportunity

eTRIKS is tasked with providing data management capabilities to other IMI projects. That necessarily requires eTRIKS partners to interact with the data assets of the supported projects. Whenever non-public data is exchanged between a project and eTRIKS a legal agreement must be in place to protect the researchers and study participants. Developing these agreements from scratch is a time complex and time consuming activity. eTRIKS has developed a process and templates to streamline the work.

Challenges

- Depending on the nature of the data and the nature of the desired engagement between eTRIKS and the project different forms of agreement are required
- Existing legal agreement templates focus on material exchange not data exchange
- Because projects are not legal entities agreements need to be signed by all partners involved in the data exchange activities
- Legal support is rarely factored into project resourcing
- There are no standard IMI cross-project agreement templates
- Data are Foreground for an IMI project and therefore their IP status must be protected.
- Legal representatives are often busy, elusive and hard to engage in dialogue.

Approach

We have created a series of templates and a workflow to guide discussions to the best fit for the particular engagement. Experience has shown that minimising the number of partners that need to sign the agreement to only those engaged in the data activities greatly streamlines the process. Some rewording of the template will always be necessary due to the diverse nature of projects and the engagements they need.

Each agreement is divided into two parts. One section covers the legal agreement between the parties and the other section covers the scientific and practical activities. The templates are set up such that it is possible to change the scientific and practical activities in the light of evolving circumstances without the need to resign the whole agreement.

Activities

Recognise that a legal agreement is required
Whenever there is a need to look at or interact with data from a supported project the need for a legal agreement between eTRIKS and the project needs to be considered. Typically this will be brought to the attention of the eTRIKS Legal team by the Account Manager for the project.

Understand the needs of the engagement
The major question to consider is to what extend TRIKS will be working with the data from the project. Specifically it needs to be considered whether eTRIKS will be in any way processing the data, including curating or hosting it in the eTRIKS Knowledge Management

Portal, or if only lightweight engagement such as consultation or training are required. A particular issue to consider is if eTRIKS will need to work with any personal data held by the project.

In the case that all the project requests of eTRIKS can be managed by a partner that is in both consortia then it is possible that no additional legal agreement will be needed. However, this situation is relatively rare.

Select a suitable template
Selection is determined by the degree of engagement and the nature of the data. For simple engagements a Confidential Disclosure Agreement (CDA) may be sufficient. But if eTRIKS is expected to work on processing the data then a Data Processing Agreement (DPA) will be needed and possibly a Material Transfer Agreement (MTA) if hosting or extensive manipulation of the data is required. All these approaches are covered by an eTRIKS Legal Template.

Agree the scope of the activities
The template provides space to define exactly what activities will fall under the agreement. The data management group within the project and eTRIKS work together with the Account Manager to define the services that eTRIKS will provide. However as the activities are often research based, provision is available to agree modifications to the services between the groups without having to resign the whole agreement.

Propose wording
Having selected the best starting point template it has to be compared with the specific needs of the engagement. For example clauses covering handling of personal data may or may not be needed. Specific activities such as data hosting and training need to be considered

Iterate
It is vanishingly unlikely that agreement will be reached in a single step. Each legal organisation needs to review the implications with their legal experts and propose any changes tat they would like to see. These changes can then be discussed in a teleconference between the legal representatives to reach an understanding that all can agree to.

Agree final wording
The final output is a document that all interested parties have had an opportunity to contribute to. This then forms the agreement between eTRIKS and the project.

Manage signature process
Physically signed documents are required for agreements entered into by partners as part of an IMI project. Copies of the agreement are forwarded to all legal contacts for involved organisations They are required to return 5 signed paper copies to the BioSci office. These are made up into 5 fully signed
It can take many weeks or even months for all parties to sign
For a CDA using the eTRIKS template there is no need for eTRIKS partners to sign the document as it is covered by the project mandate.

Finalise document and archive
After all signatures are received BioSci create 5 full paper copies of the document and send one each to the coordinators and Managing Entities of eTRIKS and the Project. The fifth copy is sent to IMI.

Finally a .pdf copy is made available to all partners for their own records as needed.

Roles

Account Manager
Leads the discussion around what services are needed by the project

eTRIKS legal team
Leads the eTRIKS side of the legal discussions (negotiations) ensures that any issues specifically within the eTRIKS partners (such as Research Contamination) are raised and dealt with in a timely way.
The eTRIKS Legal Team also leads the activities around ensuring the signature process takes place as smoothly as possible.

Project legal team
Leads the Project side of the legal discussions ensuring that any issues within the project partners are addressed.

Project Leader
The ultimate authority from the project on whether the agreement meets the needs of the project

Legal contacts of all involved partners
Tasked with identifying the appropriate person in their organisation (if it is not themselves) who is authorised to sign the legal agreement and ensuring that they understand the details and need for it and sign it promptly.

References
D5.7 – Experience of Harmonisation of Consortium Agreements in Europe

Good Practice for Data Curation

Opportunity

Well curated data is far more valuable as a reusable asset.

In order to explore data in TranSMART it must be loaded into a database accessible by the system. Rarely is the data in the correct format for TranSMART and the process of validating and restructuring the data is called Curation. Curation into tranSMART is a complex and nuanced process that requires experience to drive to a successful conclusion. Many supported projects do not have sufficient depth of knowledge to carry out effective data curation themselves. This knowledge and experience is concentrated in the eTRIKS Curation team.

Challenges

- Lack of well trained curators
- Data sets often have missing information (data and metadata)
- Data owners often do not understand the issues involved in curating a data set
- Many organisations underestimate the resource needed for data curation
- Data curation requires understanding of the data, the receiving system
- Accessing data requires a non disclosure or data processing agreement
- Appropriate standards need to be identified and agreed, the needed standards may not exist
- If the data owner is not available to discuss issues, the curator may need to infer the intended meaning in uncertain situations
- Current curation tools are slow and opaque in their operation
- Reworking poor curation can be very time consuming.

Approach

The process is an iterative one as it is not possible to guarantee that the data set can be curated in a single pass.

In eTRIKS three parallel pillars are needed for efficient data curation activities:

- Project managers oversees the progress and solves issues during the whole process
- Legal team provide support on solving data access constraints and legal agreements
- Technical team with domain expertise IT/data scientist and biology/disease domain experts

Data owners are actively involved throughout the process to ensure the best possible outcome for the project. Data Curation Guidelines have been written to optimise the curation process. Data sets are curated in an isolated, secure instance of tranSMART until release is agreed.

Activities

Engage data owners

It is extremely desirable, where possible, that the data owner is engaged throughout the data curation processes. They best know the data.

If the data owner is not accessible (e.g the data set is from an historical public study) review will have to be done without the Data Owner in which case the curator will have to use their

experience to make decisions on how to manage issues in the data set such as missing or incorrect values.

Put in place appropriate legal agreements
If the data are not public it is essential that the data curator has legally agreed access to the data set. In the case of data sets containing personal data this will include some sort of data processing agreement. Access is agreed and put into place with the help of the eTRIKS Legal team who have a selection of template agreements available to speed the processes.

Initial meta data exchange
Meta data of the datasets (type, volume, access right, quality, availability of detailed metadata, requirement of the curation) are needed to estimate the resources needed and plan making. The structure of the data in TranSMART has a strong influence on what kinds of questions can be easily posed to them. Initial review should include the high-level organisation, diversity and size of the data components.
Use the Data Loading Questionnaire (D4.1) to help understand the data.

Implementation of reproducible pipelines
Before legal agreement is reached, technical team can already start initial implementation using collected metadata and dummy data (that has the same format and structure of the real data).

Performing curation after legal issue solved
Once the legal issues are solved, curation can be performed by tuning and optimizing the initial curation pipeline based on the real data.

Documentation

Documentation about the curation activities will provide information on pipeline (source code and user guide), versioning control on both pipeline and curated data, rationale for any alteration of data values during the curation process, as well as the approval process associated with such alterations including the date, reason and personnel having modified, and approved modifications to, data values.

Roles

Project/account manager
Oversees the progress and solves issues during the whole process. Communicates between data processors and data controller as well as different roles in the project

Curator
Trained experts with adequate IT/data management skills to work on curation tasks that includes the development of curation pipeline and drafting of technical documents

Data Owner
The person who has right of decision over use of the data. The data owner may rely on the assistance of more knowledgeable data investigators and domain experts within their project.

Legal team(s)
Groups with legal knowledge who can agree the terms for legally appropriate data access and provide ongoing support to clear any legal issues that arise.

Domain experts

Biologist or clinical experts that have domain knowledge of the biological/disease domain to provide input to curators during the curation process.

Data Investigators

Researchers and scientists within the supported project who wish to explore the data set. In the case of public studies this is anyone with access to the eTRIKS instance of tranSMART.

eTRIKS Deliverables

D4.1   Data loading Questionnaire

D4.3   Data Provenance Process Management

D4.4   Data Curation Guidelines

D4.12   Final report on curation

D4.14   Document describing the refinements and optimizations of the roadmap implementation of eTRIKS curation and analytics support

List of abbreviations

DoW	eTRIKS Description of Work
eTRIKS KMP	eTRIKS Knowledge Management Platform