Coordinated Research Infrastructures Building Enduring Life-science services
- CORBEL -

Deliverable D3.9
Robust data upload procedures for biosample data and molecular profiling data

WP3 – Community-driven cross-infrastructure joint research - Medical

Lead Beneficiary: ECRIN-ERIC
WP leader: Jacques Demotes (ECRIN-ERIC)
Contributing partner(s): Lygature (EATRIS-ERIC), ErasmusMC, NKI (EATRIS-ERIC)

Contractual delivery date: 31 August 2017
Actual delivery date: 6 September 2017

Authors of this deliverable: Jan-Willem Boiten (Lygature), Mariska Bierkens (NKI), Matthias Hansson (ErasmusMC), Stefan Klein (ErasmusMC), Sjaak Peelen (Lygature)

Grant agreement no. 654248
Horizon 2020
H2020-INFRADEV-1-2014
Type of action: RIA

Content

Executive Summary

The core of the CORBEL common IT framework for genomics and imaging studies in biomarker research consists of the tranSMART (data integration and browsing platform), cBioportal (patient-level genomics platform) and XNAT (image archive platform) solutions. Loading existing project data into these platforms requires some level of data curation and transformation. This is particularly true for tranSMART, which is suited to handle a wide range of data types but also requires careful data modeling of those data before loading. The present document describes a detailed set of guidelines how to execute those data curation and modeling steps for each of the supported data types in tranSMART, cBioportal, and XNAT. It is partially based on previous work in the Dutch TraIT platform which was applied to the CORBEL use cases. Some practical lessons learned from those use cases, in particular the IMI APPROACH project, are given as well.

Project objectives

With this deliverable, the project has contributed to the following objective:

- Establish generic data integration services for image-driven and/or genomics-driven translational studies (e.g. biomarker discovery in heterogeneous diseases) embedding cross-RI services.

Background

Biomarker research projects have to deal with a broad diversity of data (clinical outcomes, images, genomics, biosamples, etc.) each requiring different tools and methods. Individual hospitals often lack the number of patients and the spectrum of techniques required for efficient biomarker research, therefore projects are increasingly supported by large international consortia and public-private partnerships. However, even the largest biomarker research programs face difficulties in timely managing their data sets, for clinical as well as preclinical biomarker research, due to the lack of multimodal data infrastructure for standardized storage, management and processing. CORBEL task 3.4 is designed to implement a common IT framework to support data handling and analysis for both clinical and preclinical biomarker research, across the disease areas. Establishing a common IT framework for preclinical and clinical biomarker research, in any disease area, can be achieved by bundling the expertise and services of the biomedical ESFRIs, most notably EATRIS, EuroBioImaging, ELIXIR and BBMRI, providing a complete solution for biomarker development projects. The effectiveness of the selected approaches will be demonstrated using international research projects in this domain.

One of the key challenges to be tackled is the efficient and high-quality loading of data into the selected solutions constituting the IT framework, which is basically a prerequisite for practical usage of any IT framework in this domain. Therefore, a set of procedures are needed to make the data loading effective, reproducible and broadly accessible by any user. The procedures consist of technical solutions (scripts, programs, etc.), as well documentation describing prerequisites and best practices.

The selected IT framework consists of a number of key solutions briefly introduced below:

- tranSMART: the central data integration platform;
- cBioportal: patient-centric and gene-centric viewing of genomics data;
- XNAT: the central clinical imaging archive.

Each of these solutions require dedicated data loading procedures which have been recorded consistently in the present document.

Data integration platform: tranSMART

tranSMART is a data integration, sharing, and analysis platform for clinical and translational research. It allows users to search, view, and analyze 'final' or 'processed' data through a web interface, thereby allowing easy access to explore such data from multiple domains at a study level. It also enables scientists to develop and refine hypotheses by investigating correlations between genetic, phenotypic and clinical data, as well as assessing their analytical results in the context of published literature and other work. tranSMART constitutes the core of the IT infrastructure in the CORBEL clinical use cases for biomarker research (task 3.4).

Central administration of the available biomedical samples is critical, as well-annotated biomedical samples typically distributed over local biobanks managed by individual project partners. For this, we will develop generic protocols for abstraction of the sample data from its original biobank storage location to tranSMART, enabling distributed sample selection and ordering on the basis of clinical and biomarker queries in the central database. For the actual implementation, we will be able to apply methods developed within WP 6.3 (secure/sensitive data) and synergize with and support Task 3.5 (Biobank cohorts).

Data acquisition, quality control and data processing pipelines are an integral part of the technology platforms for biomarker development projects (represented by our two use-cases). The output of these specific data pipelines will be uploaded into tranSMART using standards for molecular data formats (DNA, RNA, protein) and imaging data, established in conjunction with other partners in the international tranSMART community (coordinated by the tranSMART Foundation in which IMI-eTRIKS, TraIT, and the IMI office are represented in the Board of Directors), ELIXIR, BBMRI-NL, EuroBioImaging, BioMedBridges, as well as the current project (WP6).

Cancer genomics portal: cBioPortal

The cBio Cancer Genomics Portal (http://cbioportal.org) is an open source resource for interactive exploration of multidimensional cancer genomics data sets, currently providing access to data from more than 20,000 tumor samples from almost 100 cancer studies. The cBio Cancer Genomics Portal significantly lowers the barriers between complex genomic data and cancer researchers who want rapid, intuitive, and high-quality access to molecular profiles and clinical attributes from large-scale cancer genomics projects and empowers researchers to translate these rich data sets into biological insights and clinical applications. It now also serves oncologists in their use and interpretation of clinical sequencing data from cancer patients and enables precision oncology.

Imaging platform: XNAT

XNAT (https://www.xnat.org/) is an open source imaging informatics platform developed by the Neuroinformatics Research Group at Washington University. XNAT was originally developed in the Buckner Lab at Washington University, now at Harvard University. It facilitates common management, productivity, and quality assurance tasks for imaging and associated data. Thanks to its extensibility, XNAT can be used to support a wide range of imaging-based projects.
XNAT enables data access via a website (manual upload and download), via the DICOM protocol and via an application programming interface (API), which makes it flexible. Furthermore, XNAT stores not only the images, but also image-derived information, such as annotations and processed versions of the images. It is therefore of interest for the more advanced, technically oriented researchers, and for large studies which require automated image analysis.

Detailed report on the deliverable

CORBEL could lean on previous work done in the Dutch national project TraIT. The existing guidelines have been extended and augmented for additional data types as observed in CORBEL, and based on lessons-learned from the CORBEL use cases (in particular those of IMI APPROACH). For question regarding the practical usage of these guidelines as well for requesting (Excel) templates, the TraIT servicedesk can be approached by any CORBEL user (http://www.traitplaform.org/service-desk).

tranSMART data import processes

When a data owner is ready to hand over the processed data for import into tranSMART it is required that these data adhere to certain specified formats. In this document these formats are explained for each data type currently supported in tranSMART. When adhering to these formats, import will be smooth, integration with different available data types will be possible, and a range of functionality through the tranSMART interface and connected tools will be possible. The conventions below are both a result of the constraints of the platform, as well as best practice established within the TraIT project to allow for consistent and valuable data upload and exchange. These best practices have been adopted in CORBEL and lessons learned have been derived.

In this document the following is specified for every data type:

- The **exchange formats**: the specification of how the data are to be supplied in order to be machine readable; e.g. Variant Call Format, Tab Separated Value format, Comma Separated Value format, or Excel format - with headers as specified by this document.
- The **Minimum Information Guidelines** (MIG): for the data itself certain MIG fields need to be supplied by the data owner in order for tranSMART to correctly interpret the specific type of data. Though not mandatory, some MIGs for the accompanying metadata are also provided, e.g. for tracing back the raw data in public databases, or for version information of data items (questionnaires, TNM staging, how were Z-scores calculated, etc.).
- The **vocabularies**: the terminologies and exact content allowed within data fields required for correct interpretation of the data. For modelling the data correctly into tranSMART, further processing of the available data is generally required. The upload procedure is

preferably supported with relevant validation rules, i.e. what values are acceptable for an item such as Gender; if not this value then an error is raised.

tranSMART differentiates between data types as listed below. Low dimensional data includes clinical data as well as non-high-throughput molecular profiling data, but may also encompass imaging parameters and biobank data. In principle, these are simple data, for which there is one numerical or categorical observation for a concept for each subject (e.g. subject gender (Female/Male), or called mutation status (whether mutated or wild type) for a specific gene of interest of interest). High-dimensional data are usually derived from high-throughput molecular profiling experiments and encompasses a large number of concept observations for a subject (e.g. copy number status per probe, gene or region level derived from microarray data, or normalized read counts per gene with NGS data).

General remarks

Data curation as performed by tranSMART data curation experts will only involve transformation and slight adjustment of the data in such a way that the data can be loaded into tranSMART using existing ETL-pipelines (Extract, Transform, Load). The *content* of the provided data files is not altered during this process. Typically, annotating data with gene names, gene symbols or changing the actual content of the data is the responsibility of the data owner, which should be done offline before loading into tranSMART. After the loading process has been finished, the loaded data will have to be checked by the tranSMART data curation expert first; next, feedback will be given to the data owner to ensure correct upload of their data. It is still possible to do modifications of the data after initial upload (and subsequently redo the upload into tranSMART), until the data have been optimally modelled and entered into tranSMART.

For each provided datafile an additional description needs to be supplied by the data owner on:

- The data in the file. What is the data type? What level do the data report on? What was the experimental platform used to obtain measurements, version details.
- File name conventions. Especially in case of files with many abbreviations, there must be another file providing more information on the files so that the data curation expert knows what sort of data to expect for correct interpretation.
- Possible metadata. Data owners can supply extra information for an item that will pop up with a right-mouse click. These metadata may provide more extended information of data items, but may also contain extra information on the study or data itself. URLs may also be uploaded, e.g. for tracing back raw data, pre-processed data or images which may be available in another tool.

Additionally, a data tree structure is required describing the relations between different concepts. When a data owner is ready to model their data in tranSMART, an Excel template will be provided so that the data owner can specify for a certain number of data levels to which items the data corresponds, highlighting the files in which the data are present as well as the particular locations within this file. After filling in this Excel template, the data curation expert will then perform further

modelling of the study in tranSMART with the available information and consult the data owner for feedback.

Please be careful when opening and saving data using Microsoft Excel or other spreadsheet programs as these programs might change your source data. Excel tries to guess what kind of data is displayed and in the process of doing so might change the actual value of the data when saved to disk (e.g. trailing zeros are removed, or some fields might be interpreted as a date changing the structure of the data).

It is not possible to upload other types of data without at least the bare minimum of clinical data (a subject identifier is the bare minimum, but it is recommended to provide all clinical characteristics of interest at the beginning so modelling can take place). With the minimum clinical data available, it will be possible to upload other low- and high-dimensional data. For all data types the subject identifiers used in the clinical data need to be present in a "subject-sample mapping" file (this is referred to in the high dimensional section; see below), so that all data will be linked to the correct subject.

A data curation report should be generated describing the transformation steps, assumptions, decisions made with regards to the ETL process, iterations and possible improvements to the data structure tree.

The default ETL pipeline used is the transmart-batch pipeline. Extended documentation on data formats accepted for input and the transformation steps (e.g. calculating Z-score) during upload can be found here: https://github.com/thehyve/transmart-batch/tree/master/docs.

How to hand over 'low-dimensional' data

Low dimensional or clinical data describe the subject characteristics. The minimal data to be supplied are the subject identifiers that are used in the study (to which all data will be mapped). Other clinical data (e.g. demographics, outcomes) and non high-throughput molecular profiling biomarkers may also be provided for mapping to the corresponding subjects.
Clinical & Non high-throughput molecular profiling data

- **Exchange formats** for data, metadata and vocabularies; Tab Separated Value, Comma Separated Value, Excel

- **Minimum Information Guidelines** (MIG);
    - *data*: each **row** needs to contain a unique subject or sample identifier. The **columns** thereafter may contain one concept each (e.g. Column 1= Subject ID; Column 2= Age, Column 3= Gender, Column 4= Height). The data can be numeric (integer, decimal values), text or a date.

    - *metadata*: for each item in tranSMART it is possible to provide metadata. See the Excel mapping template and the table below for examples.

- *Note: the more extensive the metadata describing the data items, the better the data can be evaluated in context.*

Optionally the metadata can be provided in separate datafiles instead of in the clinical data mapping template. Such a file should have at least three columns: (1) the name and location of the data item to which the metadata should be added, (2) a title for the metadata (e.g. Info, Summary, Citation rule), (3) the actual text that composes the metadata. A 4th optional column may also be present as a rank (integer value) indicating which line should be displayed first.

- **Vocabularies**: a codebook containing terminologies and value codes for the items to be uploaded into tranSMART. For example, Gender may be coded in the datafile with 1 and 2; the codebook then explains: 1=Male, 2=Female. This codebook should also contain for each item the variable type (numeric, text, date etc. "example: Gender= text") and optionally validation rules (indicating values or the ranges a variable can have).

All the supplied information will end up in the tranSMART data tree structure, whereby the data owner has indicated the location of the respective data items in the delivered data using the provided Excel template (later in this document called the 'column mapping template'). It is recommended to try to complete the mapping of the data items in tranSMART in a consistent manner; similar to other studies for consistent data representation and easy traceability of study data later on. This 'modelling' is one of the most time consuming steps for the data owner, but only has to be performed once. Additional data may be added more easily to already present data.

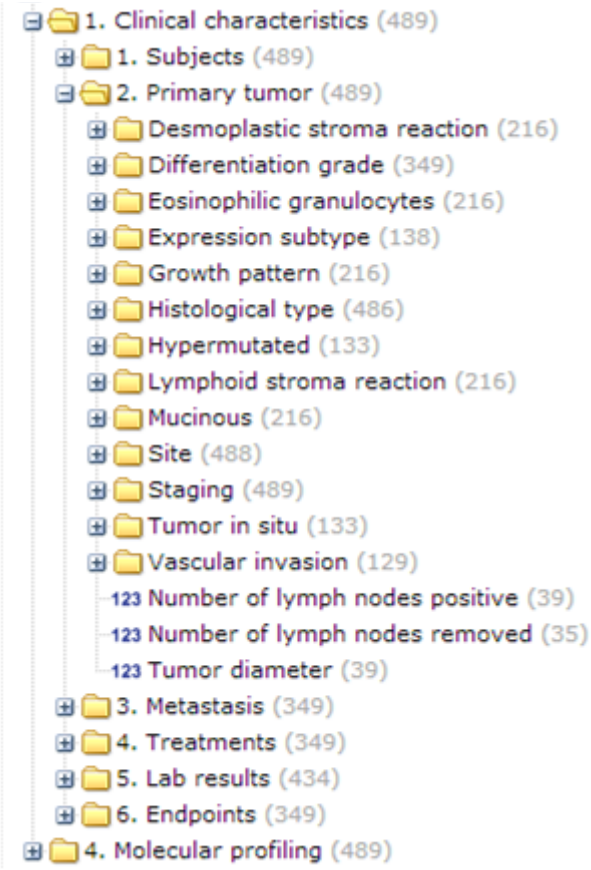
Example data file with just a few concepts (shown as actual values, but coded data may also be supplied according to guidelines above):

SUBJ_ID	Organism	Tissue type	Disease	Ethnicity	Age	Gender
SubjectID_001	*Homo sapiens*	Colon	Colorectal adenocarcinoma	Caucasian	72	Female
SubjectID_002	*Homo sapiens*	Prostate	Prostatic carcinoma	Caucasian	59	Male

tranSMART tree structure: Excel template/'column mapping template'

One of the uses of tranSMART is to explore your data. To be able to do this the data are shaped into a hierarchical structure, what is referred to as the tranSMART data tree. The figure to the right shows a part of a study as it appears in tranSMART.

Shown are items from the clinical data files as they are mapped to the tranSMART data tree using the 'column mapping template'.



1. Tree structure

To correctly fill in the template, please provide the filename of the clinical data in the **Filename** column, the column number of the column you would like to map in **Column number** and each level in the following columns (**Level 1 to 6**). There is room for comments in the **Comments** column and the metadata you would like to be added to the tree can be indicated in the **Metadata** column.

Example *Tree structure* template:

Filename	Column number (starting from 1)	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Comments	Metadata
		Studyname	1. Clinical characteristics	1. Subjects	Gender				
					Age				Info: Age at .....
				2. Primary tumor	Differentiation grade				
					DNA microsatellite instability	PCR xx Promega test xx	MSI, MSS		
					Histological type	Adenocarcinoma, Mucinous carcinoma, Signet-ring cell carci			
					Location tumor	Cecum, Colon ascendens, Flexura hepatica, Colon transversi			
					Staging	T stage			Info: .... edition TNM staging
						N stage			Info: .... edition TNM staging
						M stage			Info: .... edition TNM staging
					Stage grouping				
				3. Metastasis					
				4. Treatments					
				5. Lab results					
				6. Endpoints	Overall survival	Overall survival: event	Alive, Dead, No information		
						Overall survival after primary resection		or after resection liver metastases	Info: In days or months
					Disease-free survival	Disease-free survival: event	Relapse, No relapse, No information		
						Disease-free survival after primary resection		or after resection liver metastases	Info: In days or months
				7. Questionnaires					

2. Value substitution

The *Value substitution* sheet can be used to remap the values that are displayed in the tranSMART data tree. For example, the data file has the gender encoded as 'M', 'F' and '-1', when these should actually be shown as 'Male', 'Female' and 'No information'. By filling in this sheet the values in the datafile will be replaced during the loading of the data.

Example:

Filename	Column number	From value	To value
datafile.txt	3	M	Male
datafile.txt	3	F	Female
datafile.txt	3	-1	No information

How to hand-over 'high-dimensional' data

For all high-dimensional data (omics data types such as microarray, proteomics, NGS) there are at least two files needed for upload:

(1) the platform annotation file. This file provides information on the structure of the data file (what was measured? what items are present in the data?)
   (a) note: for the different high-dimensional data types, see the MIG section for the different data types for examples on what should be minimally supplied by the data owner and examples on optional fields

(2) the ***data file*** with the actual measurements. In the data file the items that were measured are listed, now with reported measurements per subject or sample.
   (a) The data are provided in 'wide format' with script header names as shown in the example files.

High-dimensional data nodes, or high-throughput molecular profiling data equal to one data node in tranSMART, similar to an R data object or an NGS MAF file. This node contains a multitude of measurements for numerous subjects; for example, $\log^2$ ratios for 44,000 probes of an mRNA microarray for 30 subjects. The data of each subject in this node adhere to the same layout (i.e. as for the example above, each subject will have possible values for the 44,000 probes, though some null-values may be present). This means that for example regions need to be binned using one set of bins (primarily for sequencing and region-level data) that are the same for all the samples that are being grouped.

In the following sections, mandatory and optional data information fields are described per high-dimensional data type. It is recommended to provide as much relevant information as possible to get the maximum out of the analysis in tranSMART.

RNA (gene/miRNA) expression data

RNA expression measured using Microarray platforms

- The **exchange format**: Tab Separated Value

- The **Minimum Information Guidelines** (MIG):
    - *data*: a platform annotation file providing information on the structure of the data file and the ***data file*** with the ***actual measurements***

        - Platform annotation file, should have at least **1 column** with:
            - PROBE_ID/GENE: Unique identifier dependent on the data type. Probe names of the used platform in case of probe level data, e.g. A_23_P217507, otherwise a gene identifier
            - GENE_SYMBOL: HGNC symbol, e.g. ZBED1
            - GENE_ID: preferably Entrez gene ID, e.g. 9189

        - ***Data file***, should have at least **2 columns**.
            - First column: **PROBE_ID/GENE** as listed in the platform annotation
            - Second column and onward: measurements - **each column contains measurements for one sample ID**. The values of the measurements should be clarified through the metadata (see below)

    - *metadata*: describes for the delivered files how the data was obtained as well as the location of the raw data. Describe the:
        - **Location of the (raw) data** obtained from each hybridization (may be a reference to a deposit in a central database, or a reference describing the location of e.g. a hard disk in possession of ...). For example, if deposited in the Gene Expression Omnibus (GEO) database, the Series (GSE) identifier.
            - If the processed data is also available in a central repository, a reference to this location can be provided as well.
        - **Platform** used to obtain the data (e.g. 44K Agilent mRNA expression array and GPL6480, or specify otherwise if a custom platform was used.
        - **Data preprocessing steps**: normalization tools, steps and processing settings, and cut-off determination
        - **Genomic build** used for processing and reporting measurements (e.g. GRCh37).
        - **Value representation** of the measurements (e.g. measurement column headers indicating raw intensity scores, normalized log2 values or significance scores after application of a cut-off).
        - Possible **citation rules** if available.

- The **vocabularies**: a codebook containing terminologies and value codes for the items to be uploaded to tranSMART. Describe what the columns headers in the files represent, their position in the file and the type of data they contain. Also, provide subject-sample mapping, explaining which data columns correspond to which subject.

Example mRNA expression files:

Platform annotation file:

PROBE_ID	GENE_SYMBOL	GENE_ID
A_23_P100001	FAM174B	400451
A_23_P100011	AP3S2	10239
A_23_P217507	ZBED1	9189

Data file:

PROBE_ID	**Sample_ID1**	Sample_ID2	….	Sample_IDN
A_23_P100001	-2.92	0.09	….	-3.21
A_23_P100011	0.12	-0.38	….	-0.13
A_23_P217507	-0.48	-1.04	….	-1.3

miRNA expression measured using Microarray platforms

– The **exchange format**: Tab Separated Value

– The **Minimum Information Guidelines** (MIG):
  - *data*: a platform annotation file providing information on the structure of the data file and the ***data file*** with the ***actual measurements***

    - Platform annotation file, should have at least **4 columns** with:
      - **PLATFORM_ID,** e.g. for 15K Agilent miRNA expression microarray, "GPL20906"
      - **PROBE_ID** indicating which probe was measured (e.g.PROBE_ID 19 equals probe A_25_P00010807 of the platform in this example)
      - **miRNA_ID:** microRNA identifier, e.g. NR_030754.1, hsa-mir-622 or MIR622
      - **ORGANISM**, e.g. *Homo sapiens*
      - GENOMIC COORDINATES (e.g. chr13:90883436- 90883531)
      - CYTOBAND (e.g. hs|13q31.3)

    - ***Data file*** should have at least **2 columns**.
      - First column: **PROBE_ID** defined in the platform file
      - Second column and onward: measurements - **each column contains measurements for one sample ID**. The values of the measurements should be clarified through the metadata (see below)

- *metadata*: for the provided files, extra information is requested that describes how the data was obtained and the location of the raw data. Describe the:
  - **Location of the (raw) data** obtained from each hybridization (may be a reference to a deposit in a central database, or a reference describing the location of e.g. a hard disk in possession of...). For example, if deposited in the Gene Expression Omnibus (GEO) database, the Series (GSE) identifier.
    - If the processed data is also available in a central repository, a reference to this location can be provided as well.
  - **Platform** used to obtain data (e.g. 15K Agilent miRNA expression microarray and GPL20906, or specify otherwise if a custom platform was used)
  - **Data preprocessing steps**: normalization tools, steps and processing settings, cut-off determination
  - **Genomic build** used for processing and reporting measurements (e.g. GRCh37).
  - **Value representation** of the measurements (e.g. measurement column headers indicating raw intensity scores, or normalized log2 values, or significance scores after application of a cut-off).
  - Possible **citation rules** if available.

– The **vocabularies**: a codebook containing terminologies and value codes for the items to be uploaded to tranSMART. Describe what the columns headers in the files represent, their position in the file and the type of data they contain. Also, provide subject-sample mapping, explaining which data columns correspond to which subject.

Example miRNA expression files:

Platform annotation file:

PROBE_ID	**miRNA_ID**	**PLATFORM_ID**	**ORGANISM**
A_25_P00010807	hsa-mir-622	GPL20906	*Homo Sapiens*
PROBE_ID 2	hsa-mir-21	GPL20906	*Homo Sapiens*
PROBE_ID 3	hsa-mir-34a	GPL20906	*Homo Sapiens*

Data file:

REFERENCE_ID	**Sample1**	….	SampleN
A_25_P00010807	-0.84038	….	-0.22553
PROBE_ID 2	0	….	2.42
PROBE_ID 3	18.4212	….	94.8219

RNA expression measured using Next Generation Sequencing platforms (mRNA-Seq) read counts

– The **exchange format**: Tab Separated Value

– The **Minimum Information Guidelines** (MIG):
  - *data*: a platform annotation file providing information on the structure of the data file and the ***data file*** with the ***actual measurements***

    - Platform annotation file, should have at least **4 columns** with:
      - **REGION_NAME:** name of the measured item, for read counts the name can be the same as the GENE_SYMBOL
      - **CHROMOSOME:** chromosome number, indicated by an integer ranging from 1 to 22, for sex chromosomes "X", and/or "Y", for mitochondrial DNA "M"
      - **START_BP**: base pair start position of a region
      - **END_BP**: base pair end position of a region
      - CYTOBAND, e.g. hs|20p13
      - GENE_SYMBOL: HGNC gene symbol
      - GENE_ID: Entrez gene ID

    - ***Data file***, should have at least **2 columns**. For each sample, it is mandatory to provide either the read counts or the normalized read counts (or both).
      - First column: **REGION_NAME**
      - Second column and onward can be used to provide the read counts per sample. The header in this is strict and should be called **SAMPLE_ID.readcount** in your files
      - Normalized read counts can be added by introducing a column named **SAMPLE_ID.normalizedreadcount**
      - Z-scores can be provided for each sample in a column named **SAMPLE_ID.zscore**

  - *metadata*: for the provided files are requested describing how the data was obtained and the location of the raw data. Describe the:
    - **Location of the (raw) data** obtained from each sequencing experiment (may be a reference to a deposit in a central database, or a reference describing the location of e.g. a hard disk in possession of...). For example, if deposited in the Gene Expression Omnibus (GEO) database, the Series (GSE) identifier, if deposited to the European Genome-phenome Archive (EGA), the Dataset Accession number (EGAD).
      - If the processed data is also available in a central deposit, a reference to this location can be provided as well.
    - **Platform** used to obtain data (e.g. Illumina HiSeq2000 and GPL11154, or specify otherwise if a custom platform was used)

- **Data preprocessing steps:** normalization tools, steps and processing settings, cut-off determination, reference genome and build used for read-mapping. It is possible to share a Galaxy history as well in which the processing and results are stored.
- **Genomic build** used for processing and reporting measurements (e.g. GRCh37).
- **Value representation** of the measurements (e.g. measurement column headers indicate Featurecounts, Reads Per Kilobase Million [RPKM], Fragments Per Kilobase Million [FPKM] or Transcripts Per Kilobase Million [TPKM], Transcript length, or significance scores after application of a cut-off).
- Possible **citation rules** if available.

– The **vocabularies**: a codebook containing terminologies and value codes for the items to be uploaded to tranSMART. Describe what the column headers in the files represent, their position in the file, and the type of data they contain. Also, provide subject-sample mapping, explaining which data columns correspond to which subject.

Example mRNA-Seq gene expression files:

Platform annotation file:

REGION_NAME	**CHROMOSOME**	**START_BP**	**END_BP**	CYTOBAND	GENE_ SYMBOL	GENE_ID
FAM174B	15	93162685.	93198889	15q26.1	FAM174B	400451
AP3S2	15	90437617	90437617	15q26.1	AP3S2	10239
SV2B	15	91795050	91835782	15q26.1	SV2B	9899

Data file:

REGION_NAME	sample1.readcount	sample1.normalizedreadcount	sample1.zscore	sampleN.readcount	sampleN.normalizedreadcount	sampleN.zscore
FAM174B	0	0	-0.45	...	...	
AP3S2	573	9.16	0.56	...	...	
SV2B	9	3.17	-0.07	...	...	

Genomic and transcriptomic variants

Small nucleotide variants measured using Next Generation Sequencing platforms

The description below provides information on how to hand over genomic variant data (detected with DNAseq) as well as transcriptomic variants (detected with RNA-Seq).

- The **exchange format**: Variant Call Format (VCF), currently version 4.2. *Important note: All samples for a concept need to be in the same VCF file. Multiple VCF files can be merged with vcf-merge from the VCFtools.*

- The **Minimum Information Guidelines** (MIG):
    - A VCF file should contain at least **4 columns** with:
        - **CHROM**: number of the chromosome, e.g. 20
        - **POS**: genomic coordinate/position, e.g. 14370
        - **REF**: reference allele
        - **ALT**: alternative allele
        - ID: SNP id, e.g. rs6054257
        - QUAL: quality score, e.g. 29
        - FILTER: data filter specifications, e.g. PASS, or q10
            - These columns need to be filled per variant.
        - GT format field
            - Needs to be present for each sample.
    - Recommended is to add:
        - Entrez Gene IDs (in the INFO field - GID)
        - HGNC gene symbols (INFO field - GS)
            - Note: the current version of tranSMART only supports 1 GID and GS per chromosomal location. For example, files and extended documentation on the VCF format please see the VCF version 4.2 guidelines here: https://samtools.github.io/hts-specs/VCFv4.2.pdf.

    - *metadata* for the provided files are requested describing how the data was obtained and the location of the raw data. Describe the:
        - **Location of the (raw) data** obtained from each sequencing experiment (may be a reference to a deposit in a central database, or a reference describing the location of e.g. a hard disk in possession of...). For example, if deposited in the Gene Expression Omnibus (GEO) database, the Series (GSE) identifier, if deposited to the European Genome-phenome Archive (EGA), the Dataset Accession number (EGAD).
            - If the processed data is also available in a central repository, a reference to this location can be provided as well.
        - **Platform** used to obtain data (e.g. Illumina HiSeq2000 and GPL11154, or specify otherwise if a custom platform was used)
        - **Data preprocessing steps**: normalization tools, steps and processing settings, cut-off determination, reference genome and build used for read-mapping. It is possible to share a history from the Galaxy tool as well in which the processing and results are stored.
        - **Genomic build** used for processing and reporting measurements (e.g. GRCh37).
        - **Value representation** of the measurements (e.g. measurement column headers indicate small nucleotide variants or polymorphisms (SNVs, SNPs), or significance scores after application of a cut-off).

- Possible **citation rules** if available.

– The **vocabularies**: a codebook containing terminologies and value codes for the items to be uploaded to tranSMART. This codebook should be included in the VCF file as the header part and describe a.o. The kind of data stored in the FILTER, FORMAT and INFO fields. Subject sample mapping required.

– ***Example VCF file:***

```
##fileformat=VCFv4.0
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
```

#CHROM	**POS**	ID	**REF**	**ALT**	QUAL	FILTER	INFO	**FORMAT**	Sample1
20	14370	rs6054257	C	T	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP	0|0:48:1
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP	0|0:49:3

Protein expression data

Protein expression measured using mass-spectrometry
(also: Immunoassay Rule-based medicine (RBM) – analyte concentrations)

– The **exchange format**: Tab Separated Value

– The **Minimum Information Guidelines** (MIG):
  - *data*: a platform annotation file providing information on the structure of the data file and the ***data file*** with the ***actual measurements***

    - Platform annotation file, should have at least **2 columns** with:
      - **ID_REF:** reference ID given to peptides or proteins in the experiment, e.g. 1767, not a standard, but e.g. a MaxQuant.pgID derived from one experiment and different in another experiment.

- **UNIPROT_ID:** a Uniprot-SwissProt Accession Code, e.g. P04637
- CHROMOSOME: chromosome number, indicated by an integer ranging from 1 to 22, or for sex chromosomes "X", and/or "Y", for mitochondrial DNA "M"
- START_BP: base pair start position of a region
- END_BP: base pair end position of a region

▪ ***Data file*** should contain at least **2 columns**.
- First column: **ID_REF** (note this is *not* a standard ID)
- Second column onward: measurements - **each column contains measurements for one sample**. The values of the measurements should be clarified through the metadata (see below).

o *metadata* for the provided files are requested describing how the data was obtained and the location of the raw data. Describe the:
▪ **Location of the (raw) data** obtained from the original experiment (may be a reference to a deposit in a central database, or a reference describing the location of e.g. a hard disk in possession of...). For example, if deposited in the PRoteomics IDEntifications (PRIDE) database, the Project identifier (PXD).
- If the processed data are also available in a central repository, a reference to this location can be provided as well.

▪ **Platform** used to obtain the data; specify the exact platform used in the experiments and the machines used to generate the data (e.g. liquid chromatography-mass spectrometry [LC-MS] or LC-MS/MS).
▪ **Data preprocessing steps**: normalization tools, steps and processing settings.
▪ **Value representation** of the measurements (e.g. measurement column headers indicate spectral counts or label free quantification [LFQ] values).
▪ Possible **citation rules** if available.

– The **vocabularies:** a codebook containing terminologies and value codes for the items to be uploaded to tranSMART. Describe what the columns headers in the files represent, their position in the file, and the type of data they contain. Also, provide subject-sample mapping, explaining which data columns correspond to which subject.

Example mass-spectrometry protein expression files:

Platform annotation file:

ID_REF	**UNIPROT_ID**	CHROMOSOME	START BP	END BP
1767	P07900	1	1051	16146
2789	P68104	5	5846	562250
1742	P06733	X	451	12625

Data file:

ID_REF	**SAMPLE_1**	….	SAMPLE_N
1767	2.7925e+10	….	2.7111e+10
2789	3.2338e+10	….	2.3628e+10
1742	2.5819e+10	….	3.2304e+10

DNA copy number data

Copy Number Alterations measured using Microarray platforms

Because of the analyses in tranSMART and the processed state of the data it is requested that data owners provide region (bin) level or gene level data for their copy number data, as opposed to probe-level data.

– The **exchange formats**: Tab Separated Value

– The **Minimum Information Guidelines** (MIG):
  - *data*: a platform annotation file providing information on the structure of the data file and the ***data file*** with the ***actual measurements***

    - Platform annotation file, should have at least **4 columns** with:
      - **REGION_NAME/GENE**: e.g. chr20:1000001-4590001; however, the name varies per experimental file, could be a numeric regional ID, the combined name of items "chromosome:start_bp-end_bp", or a gene identifier.
      - **CHROMOSOME**: chromosome number, indicated by an integer ranging from 1 to 22, for sex chromosomes "X", and/or "Y", for mitochondrial DNA "M"
      - **START_BP:** base pair start position of a region
      - **END_BP**: base pair end position of a region
      - NUMBER_OF_PROBES: the number of probes present in a region
      - CYTOBAND: e.g. hs|20p13
      - ***only for gene level data!: GENE_SYMBOL (HGNC gene symbol)
      - ***only for gene level data!: GENE_ID (Entrez gene ID)

    - ***Data file***, should contain at least **2 columns** for gene level data and at least **3 columns** for region level data
      - First column: **REGION_NAME/GENE**, as specified in the platform file
      - Second column onward: **measurements.** Multiple types of measurements are available per sample:

- **Sample.call**: contains the copy number variation call

These calls should be provided as:

- -2 = homozygous loss
- -1 = loss
- 0 = normal
- 1 = gain
- 2 = amplification

- ***mandatory only for region level data! **Sample.ratio:** contains the normalized log2 ratio of sample versus reference

Additional columns in the file may be:

- *Sample.segmented: normalized log2 ratio of segmented copy number data*
- *Sample.probhomloss: probability homozygous loss, for a region segment or a gene*
- *Sample.probloss: probability loss, for a region segment or a gene*
- *Sample.probnorm: probability normal, for a region segment or a gene*
- *Sample.probgain: probability gain, for a region segment or a gene*
- *Sample.probamp: probability amplification, for a region segment or a gene*

- *metadata* for the provided files are requested describing how the data was obtained and the location of the raw data. Describe the:
  - **Location of the (raw) data** obtained from each hybridization (may be a reference to a deposit in a central database, or a reference describing the location of e.g. a hard disk in possession of...). For example, if deposited in the Gene Expression Omnibus (GEO) database, the Series (GSE) identifier.
    - If the processed data is also available in a central repository, a reference to this location can be provided as well.
  - **Platform** used to obtain data (e.g. 180K Agilent SurePrint G3 CGH microarray and GPL8687 or specify otherwise if a custom platform was used)
  - **Data preprocessing steps**: normalization tools, steps and processing settings, cut-off determination.
  - **Genomic build** used for processing and reporting measurements (e.g. GRCh37).
  - **Value representation** of the measurements (e.g. measurement column headers indicate raw intensity scores, or normalized log2 values, or scores after application of a cut-off [homozygous deletion, loss, normal, gain, amplification]).
- Possible **citation rules** if available.

– The **vocabularies**: a codebook containing terminologies and value codes for the items to be uploaded to tranSMART. Describe what the columns headers in the files represent, their position

in the file and the type of data they contain. Also, provide subject-sample mapping, explaining which data columns correspond to which subject.

Example copy number alteration files:

Platform annotation file:

REGION_NAME	**CHROMOSOME**	**START_BP**	**END_BP**	NUM_ PROBES	CYTOBAND	*GENE_ SYMBOL****	*GENE_ID****
1:11869-14409	1	11869	14409	42	1p36.33		
1:14363-29570	1	14363	29570	253	1p36.33		
20:1000001-4590001	20	100000	4590001	59833	20p13		

***Note: Gene_symbol and Gene_ID are optional and do not have to be provided when providing region level data.

Data file:

REGION_ NAME	**x.call**	**x.ratio**	x.segm ented	x.probh omloss	x.problos s	x.prob norm	x.pro bgain	x.prob amp	xN.chip	…	xN.prob amp
x1	0	0.009	0.103	0	0	0.951	0.049	0	-0.292	…	0
x2	0	-0.011	0.103	0	0	0.951	0.049	0	-0.047	…	0
20:1000 001-4590001	0	0.081	0.103	0	0	0.951	0.049	0	-0.452	…	0

***Note: ratio is optional when providing gene level data.

Copy Number Alterations measured using Next Generation Sequencing platforms (shallow quantitative DNAseq: qDNAseq)

Because of the analyses in tranSMART and the processed state of the data it is requested that data owners provide region (bin) level or gene level data for their copy number data.

- The **exchange formats**: Tab Separated Value

- The **Minimum Information Guidelines** (MIG):
  - *data*: a platform annotation file providing information on the structure of the data file and the ***data file*** with the ***actual measurements***

    - Platform annotation file, should have at least **5 columns** with:
      - **REGION_NAME/GENE**: e.g. chr20:1000001-4590001; however, the name varies per experimental file, could be a numeric regional ID, the

combined name of items “chromosome:start_bp-end_bp”, or a gene identifier.

- **CHROMOSOME**: chromosome number, indicated by an integer ranging from 1 to 22, for sex chromosomes “X”, and/or “Y”, for mitochondrial DNA “M”
- **START_BP:** base pair start position of a region.
- **END_BP**: base pair end position of a region.
- CYTOBAND: e.g. hs|20p13
- ***only for gene level data!: GENE_SYMBOL (HGNC gene symbol)
- ***only for gene level data!: GENE_ID (Entrez gene ID)

- ***Data file***, should contain at least **2 columns** for gene level data and at least **3 columns** for region level data
  - First column: **REGION_NAME/GENE**, as specified in the platform file
  - Second column onward: measurements. Multiple types of measurements are available per sample:
    - **Sample.call**: contains the copy number variation call

    These calls should be provided as:
    - -2 = homozygous loss
    - -1 = loss
    - 0 = normal
    - 1 = gain
    - 2 = amplification
  - ***mandatory only for region level data! **Sample.ratio:** contains the normalized log2 ratio of sample versus reference

Additional columns in the file may be:

- *Sample.segmented: normalized log2 ratio of segmented copy number data*
- *Sample.probhomloss: probability homozygous loss, for a region segment or a gene.*
- *Sample.probloss: probability loss, for a region segment or a gene.*
- *Sample.probnorm: probability normal, for a region segment or a gene.*
- *Sample.probgain: probability gain, for a region segment or a gene.*
- *Sample.probamp: probability amplification, for a region segment or a gene.*

- *metadata* for the provided files are requested describing how the data was obtained and the location of the raw data. Describe the:
  - **Location of the (raw) data** obtained from each sequencing experiment (may be a reference to a deposit in a central database, or a reference describing the location of e.g. a hard disk in possession of...). For example, if deposited in the Gene Expression Omnibus (GEO) database, the Series (GSE) identifier, if

deposited to the European Genome-phenome Archive (EGA), the Dataset Accession number (EGAD).
    - If the processed data is also available in a central repository, a reference to this location can be provided as well.
  - **Platform** used to obtain data (e.g. Illumina HiSeq2000 and GPL11154, or specify otherwise if a custom platform was used)
  - **Data preprocessing steps**: normalization tools, steps and processing settings, cut-off determination, reference genome and build used for read-mapping. It is possible to share a Galaxy history as well in which the processing and results are stored.
  - **Genomic build** used for processing and reporting measurements (e.g. GRCh37).
  - **Value representation** of the measurements (e.g. measurement column headers indicate raw intensity scores, or normalized log2 values, or scores after application of a cut-off [homozygous deletion, loss, normal, gain, amplification]).
  - Possible **citation rules** if available.

– The **vocabularies**: a codebook containing terminologies and value codes for the items to be uploaded to tranSMART. Describe what the columns headers in the files represent, their position in the file and the type of data they contain. Also, provide subject-sample mapping, explaining which data columns correspond to which subject.

Examples of DNASeq copy number aberration files can be found in the microarray copy number description.

Metabolomics

– The **exchange formats**: Tab Separated Value

– The **Minimum Information Guidelines** (MIG):
  - *data*: a platform annotation file providing information on the structure of the data file and the ***data file*** with the ***actual measurements***

    - Platform annotation file, should have at least **4 columns** with:
      - **BIOCHEMICAL:** e.g. xylitol
      - **SUPER_PATHWAY:** e.g. carbohydrate
      - **SUB_PATHWAY:** e.g. nucleotide sugars; pentose metabolism
      - **HMDB_ID**: the Human Metabolome Database identifier, e.g. HMDB00568.

    - ***Data file***, should have at least **2 columns**.
      - First column: **BIOCHEMICAL**
      - **Second** column and onward: measurements - **each column contains measurements for one sample ID**. The values of the measurements should be clarified through the metadata (see below).

- *metadata* for the provided files are requested describing how the data were obtained and the location of the raw data. Describe the:
  - **Location of the (raw) data** obtained from each experiment (may be a reference to a deposit in a central database, or a reference describing the location of e.g. a hard disk in possession of...). For example, if deposited in MetaboLights, the study (MTBLS) identifier.
    - If the processed data is also available in a central repository, a reference to this location can be provided as well.
  - **Platform** used to obtain the data; specify the exact platform used in the experiments and the machines used to generate the data (e.g. liquid chromatography-mass spectrometry [LC-MS] or LC-MS/MS).
  - **Data preprocessing steps**: normalization tools, steps and processing settings, cut-off determination
  - **Value representation** of the measurements (e.g. measurement column headers indicate raw intensity scores, or significance scores after application of a cut-off).
  - Possible **citation rules** if available.

– The **vocabularies:** a codebook containing terminologies and value codes for the items to be uploaded to tranSMART. Describe what the columns headers in the files represent, their position in the file and the type of data they contain. Also, provide subject-sample mapping, explaining which data columns correspond to which subject.

Example metabolomics files:

Platform annotation file:

BIOCHEMICAL	SUPER_PATHWAY	SUB_PATHWAY	HMDB_ID
mevalonic acid	Carboxylic acid	Mevalonic acid pathway	HMDB00227
5-isopentenyl pyrophosphoric acid	Phosphoric acid	Cholesterol biosynthesis	HMDB01347
xylitol	Carbohydrate	Nucleotide sugars; pentose metabolism	HMDB00568

Data file:

BIOCHEMICAL	Sample1	….	SampleN
mevalonic acid	7179.5	….	13962.1
5-isopentenyl pyrophosphoric acid	39351.4	….	30034
xylitol	205043.3	….	695346.6

XNAT/tranSMART: Quantitative Imaging Biomarkers

In addition to the procedures for low- and high-dimensional data described above, we have developed optimized solutions for the upload of quantitative imaging biomarkers (QIBs). QIBs are the result of computational analysis of medical imaging data such as computed tomography (CT), magnetic resonance imaging (MRI) and positron emission tomography (PET). Examples are simply morphometric measurements (stored in a QIB session) such as the volume of the hippocampus which is an MRI-derived biomarker for Alzheimer's disease (Zandifar, 2017), the volume of the knee cartilage which is an MRI-derived biomarker for knee osteoarthritis (Dam, 2015) (see Figure 1), or the stenosis grade in the carotid artery as measured on CT angiography (Hameeteman, 2011), but also more advanced computational measures like the texture and shape features commonly used in Radiomics analysis (Aerts, 2014). A popular solution for storage, management, and sharing of (de-identified) medical imaging data for research purposes is the eXtensible Neuroimaging Archive Toolkit (XNAT) (Marcus, 2007). In the BioMedBridges project an XNAT service for sharing imaging data for research was initiated, and XNAT was adopted by the TraIT program as a national imaging archive for multi-center QIB research studies (Klein, 2015). Therefore, in CORBEL, we have developed an optimized upload procedure for importing QIBs into tranSMART that is fully interoperable with XNAT.

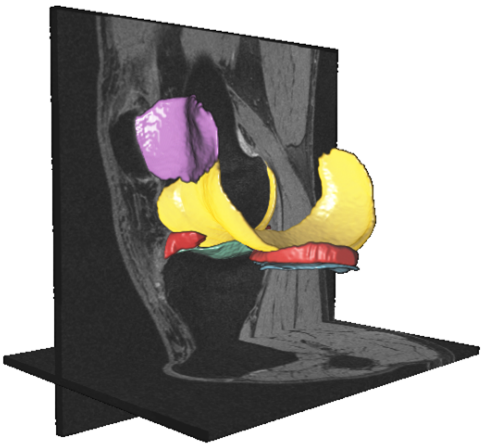


Figure 1: Example of automatic image analysis leading to quantitative imaging biomarkers (QIBs). On the background, axial and sagittal cross-sections of a three-dimensional MRI scan of the knee are shown. The colorful 3D renderings visualize the contours of several anatomical structures in the knee, including the femoral cartilage (yellow). The volume of cartilage serves as a QIB for knee osteoarthritis.

There is a multitude of use cases that could benefit from this data upload pipeline for QIB data. For development and testing, we used data from a study on the prevention of knee osteoarthritis in overweight females - PROOF (Runhaar, 2015). PROOF is a longitudinal study in 407 women who all underwent MRI of both knees, at three time points (baseline, 30 months, and 78 months). In addition, clinical data (i.e. pain scores) and genetic data were acquired. This rich data collection, with multiple samples per subject, including a longitudinal aspect, made it a perfect test case, highly representative for any future QIB studies, including but not limited to other studies on osteoarthritis such as IMI APPROACH.

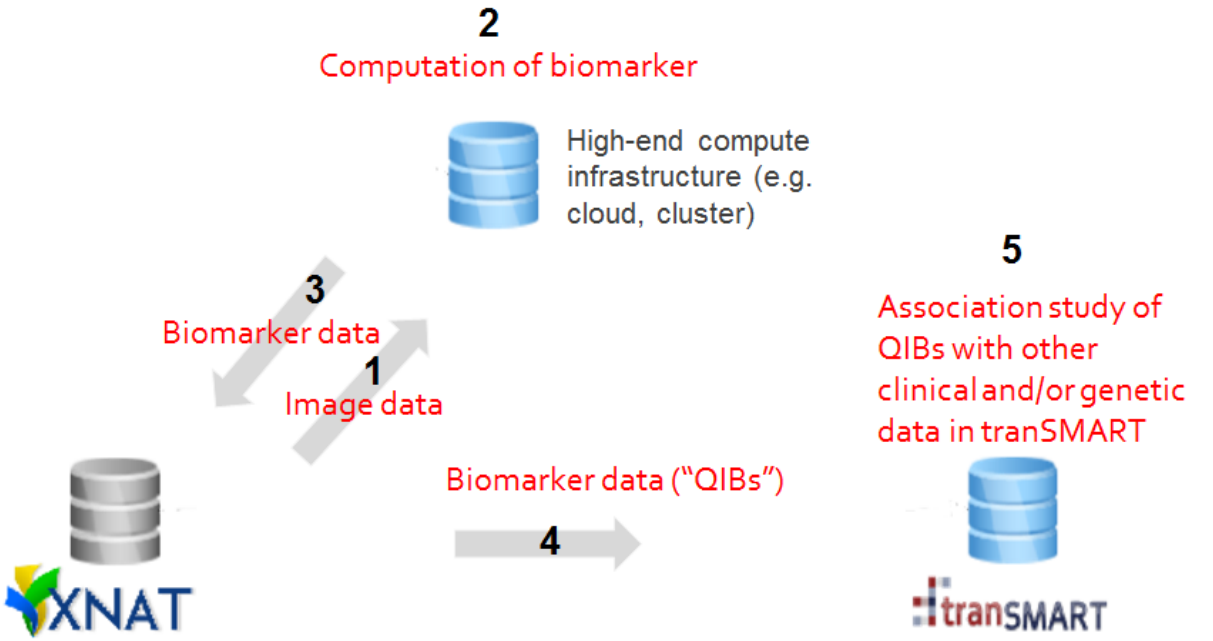


Figure 2: Workflow for uploading quantitative imaging biomarker (QIB) data into tranSMART. It is assumed that de-identified imaging data (usually in DICOM format) is stored in an XNAT research archive. Using automated software for image analysis, image data is downloaded from XNAT (step 1), QIBs are computed (step 2), and pushed back to XNAT and stored in a standardized format for image-derived data (step 3). Subsequently, the QIB data is pulled from XNAT and uploaded into the tranSMART database (step 4). Once all data have been uploaded into tranSMART, association studies between QIBs and other clinical and/or genetic data can be straightforwardly performed.

Figure 2 visualizes the workflow for uploading QIBs into tranSMART. Thanks to the RESTful API of the XNAT web service, all communication with XNAT can be automated. This is a crucial feature, since imaging studies can easily contain hundreds of images, for which manual analysis is not feasible, and automated processing is thus required. To simplify the interface with the XNAT API, a Python toolbox called XnatPy was used, and further developed for our particular purpose (in collaboration with the primary developers) [https://bitbucket.org/bigr_erasmusmc/xnatpy]. Example code making use of XnatPy is shown in Figure 3.

```
# Example code for copying scanner model and manufacturer from a source xnat to a destination
xnat

# import os and xnatpy packages
import os
import xnat
# initialize dictionary where extracted info is kept
experiment_scanner_model = {}
experiment_scanner_manufacturer = {}
# start and end number of subjects
start = 1
end = 408

# open connection to source xnat
with xnat.connect('http://bigr-rad-xnat.erasmusmc.nl') as session_rad:
   project = session_rad.projects['Proof_Study']
```

```
    for nr in range(start, end):
        # subjects are named PROOF001, PROOF002, PROOF003, etc.
        subjectname = 'PROOF{:03d}'.format(nr)
        subject = project.subjects[subjectname]
        print('Processing subject: {}'.format(subject.label))
        # retrieve MRI experiments
        for experiment in subject.experiments.values():
            # get scanner model and manufacturer and put in corresponding dictionary
            experiment_scanner_model[experiment.label] = experiment.scanner.model
            experiment_scanner_manufacturer[experiment.label] = experiment.scanner.manufacturer

# open connection to destination xnat
with xnat.connect('https://xnat-acc.bmia.nl') as session_acc:
    project_acc = session_acc.projects['qibproof']

    for nr in range(start, end):
        name = 'PROOF{:03d}'.format(nr)
        # get subject in destination xnat
        subject = project_acc.subjects[name]
        # replace scanner model and manufacturer in destination experiments with those in source
        # experiments, which have been stored in the dictionaries experiment_scanner_model and
        # experiment_scanner_manufacturer
        for experiment in subject.experiments.values():
            experiment.scanner.model = experiment_scanner_model[experiment.label]
            experiment.scanner.manufacturer = experiment_scanner_manufacturer[experiment.label]
```

Figure 3: Example code snippet for setting properties of images in XNAT directly using the XnatPy API.

Image processing in step 2 of the workflow (see Figure 2) could be done with any suitable software, but for the considered use case, we made use of a modular image processing workflow environment in Python, called FASTR [http://fastr.readthedocs.io], and implemented an automated method for knee cartilage segmentation from high-resolution 3D MRI (Hansson, 2016).

For storage of the resulting QIB data in XNAT, we developed a standardized QIB data type. Initial foundations for this datatype were laid in the BioMedBridges project. In CORBEL, we further refined its design, fixed several bugs, and incorporated feedback from end-users and tranSMART data science experts. After extensive testing, this data type will be scheduled for installation onto the TraIT XNAT production server in the near-future. The QIB data type aims to store all relevant fields which uniquely identifies how and when the derived data was created. A QIB session is connected to its corresponding MR/CT/PET session (i.e., the original image from which it is derived) by a unique identifier, a so-called accession identifier. The QIB session can be uploaded and downloaded as an XML file, which contains the name of tool used to generate the data, as well as its version, start and end time of processing, the paper from which the tool was derived, a short description of the tool, and the units of the biomarker values. In several fields, links to elements from ontologies can be used, which are dynamically expanded in the XNAT interface using the EBI Ontology Lookup Service. The structure of a QIB XML file is visualized in Figure 4. Besides uploading complete XML files, a simplified API for setting and retrieving elements of a QIB session has also been developed, as part of the XnatPy toolbox mentioned above. Source code of the QIB datatype can be found here: https://bitbucket.org/bigr_erasmusmc/qibsession-datatype.

QIBSession

subjectID: *ACCBMIAXNAT_S03471*

analysisTool: *MultiAtlasAppSegm.py*
analysisToolVersion: *0.1*
analysisToolOntologyName: *ToolOntology*
analysisToolOntologyIRI: *ToolIRI*
processingStartDateTime: *2017-03-16T06:14:32*
processingEndDateTime: *2017-03-16T07:30:14*
processingUserName: *mhansson*
processingSiteName: *Erasmus MC*
paperURL: *https://www.ncbi.nlm.nih.gov/pubmed/21937346*
paperTitle: *Automated brain structure segmentation based on atlas registration and appearance models*
paperNotes: *Method is a variant of version described in paper (no MRF model for spatial coherence)*
reviewStatus: *Not reviewed*
reviewer:
description: *MultiAtlas Appearance Model Segmentation with Volume Calculation*
biomarkerCategory: *Cartilage*
biomarker:
ontologyIRI:*NA*
ontologyName: *Uberon*
value: *1920005.3*
unit: *mm^3*
variableDescription: *Femoral Cartilage volume*
biomarker:
ontologyIRI:*NA*
ontologyName: *Uberon*
value: *8758.3*
unit: *mm^3*
variableDescription: *entire scan volume*
baseSession: *(= reference to original image from which this biomarker has been derived)*
baseSession accessionIdentifier: *PROOF001_MRI_L_T0*

Figure 4: Contents of an XML file describing a QIB session. XML syntax has been omitted for clarity.

The script for importing QIB data from XNAT to tranSMART (step 4 in Figure 2) exploits the REST API of XNAT to download the QIB data, and uses the Transmart-Batch import pipeline [https://github.com/thehyve/transmart-batch] for final upload into the tranSMART database. Thanks to the efforts put into standardization of the storage format for QIBs in XNAT, this import script is expected to be reusable in many other QIB research projects. The script has been publicized on GitHub: https://github.com/thehyve/xnat-QIB-TranSMART-import. A screenshot of the appearance of the QIB data in tranSMART is shown in Figure 5.

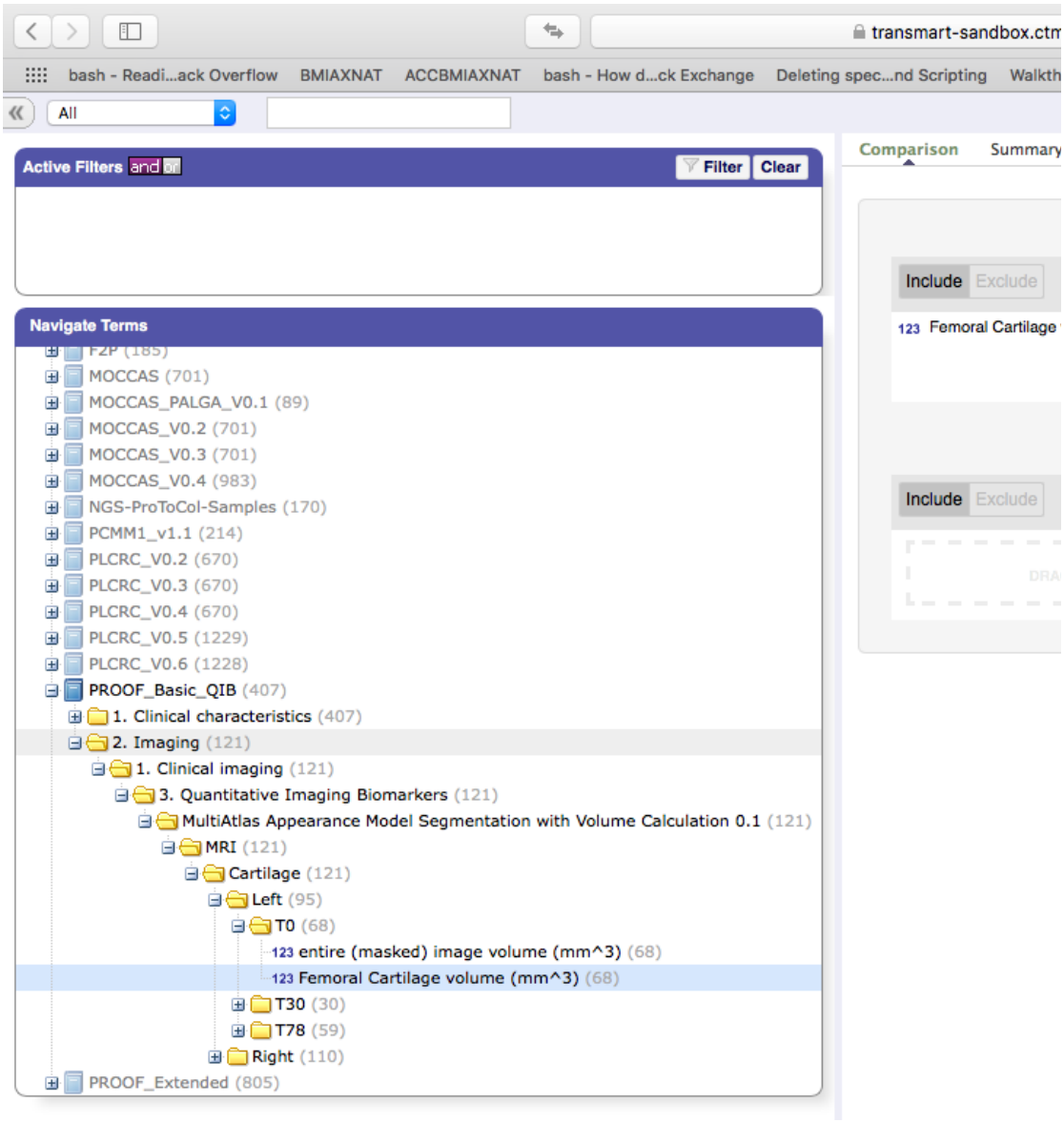


Figure 5: Screenshot showing the appearance of QIB data in tranSMART.

Data integration platform cBioPortal

cBioportal constitutes a second data integration platform, with complementary view and query possibilities of the data compared to tranSMART, which will also play an important role within the CORBEL central IT framework, in particular for the Cancer Core Europe use case. The cBioPortal for Cancer Genomics provides a Web-based resource for researchers to explore, visualize, analyze, and share multidimensional cancer genomics data sets. Users are capable of viewing gene-level data across genes, samples and data types, without requiring bioinformatics expertise.

The preferred route for uploading data to cBioPortal is by first uploading and modelling the data in tranSMART. Next, this curated data will be exported and converted to a cBioPortal suitable format using the tm2cbio script [https://github.com/j-hudecek/tm2cbio]. Some user input is still required, e.g. for identifying specific clinical attributes (OS_STATUS, OS_MONTHS, etc.). Also, data stored in VCF files, e.g. small nucleotide variant files (see above), needs to be converted to MAF format using

vcf2maf [https://github.com/mskcc/vcf2maf]. For copy number alteration data stored in a region or bin format, 'segmented data' in cBioPortal, a liftover of chromosomal coordinates is needed if the coordinates were stored using a genomic build other than hg19.

Though many data types can be stored in tranSMART, not all data types are as of yet accommodated for view and query. In this case, tranSMART may serve as a secure and central platform for storage and distribution of available data. It may also occur that a certain data type is not yet suitable for upload to tranSMART without modifying the existing data format extensively. In this case, it is recommended to curate the original data-owner's files according to the cBioPortal guidelines and import the data directly to cBioPortal.

Sample-level data

For the current tranSMART version (16.2 as of Q3 2017), sample-level data is poorly supported and extensive modelling is needed to store the data in such a manner that questions can be asked of the data. This will be solved in the next release of tranSMART (17.1, estimated to be released in Q3 2018), but in the meanwhile cBioPortal excels at storing both sample and longitudinal data. Thus, if a study has these particular data types, it is possible to upload data straight to cBioPortal. Depending on the specific analysis needs of the user, importing the data to both tranSMART and cBioPortal may still be recommended in this situation, but curation of data files should then occur in parallel, starting from the data-owner obtained files. More detail on cBioPortal supported data types and the required data formats can be found here:
https://cbioportal.readthedocs.io/en/latest/File-Formats.html.

Lessons Learned

APPROACH

The IMI APPROACH project aims to gain a better understanding of disease stratification and acceptance of a guideline to classify osteoarthritis (OA) patients. This will provide clear phenotypes-directed protocols for disease modifying OA drug trials enabling the targeting of subgroups with OA that have uniform disease characteristics, thereby increasing the chances of success. In order to achieve such stratification approach, the APPROACH consortium has the following objectives:

- Implement and establish a new, integrated and comprehensive database platform of existing data from partners that will be extended with newly collected longitudinal data, incorporating novel high quality biomarkers.
- Define subsets of (phenotypically) different patients in the existing cohorts and refine these in the successive new longitudinal extension cohort and subsequently identify the "right patient" to treat for each subset/phenotype via innovative stratification techniques.
- Optimize, introduce and validate the next generation imaging methodologies, human motion analysis and biochemical assays to enable more efficient and reliable diagnoses and treatment of OA patients, refining stratification of phenotypes.
- Identify mechanistic targets for patient subsets, create prediction models and establish guidelines for development of a disease modifying OA drug.

In the APPROACH project, retrospective cohort data from multiple cohorts had to be uploaded to tranSMART. This complicated integration effort gave rise to multiple challenges which had to be overcome. The lessons learned from this process have been incorporated in the present document.
While processing the individual data files for upload into tranSMART, it became apparent that the quality of the metadata of the different cohort data sets varies significantly. Some cohorts have a proper codebook, whereas others only have a minimal description of the parameters used. It also was observed that even in some cases where sufficient metadata were provided, that these data were scattered over different files. These files were geared towards human readability, but were not suited for automated extraction of the required meta data. Furthermore, there was hardly any standardization between the different osteoarthritis cohorts with regards to the naming and description of the parameters used. As a result, the relevant metadata had to be extracted mostly manually, which is not only a time-consuming process, but also rather error-prone.
So, in order to prevent this for future studies, it is recommended that cohorts can export the metadata at any time to a simple and interoperable format like a comma separated text file, Excel file, JSON, or XML. Such format can then easily be imported into tranSMART and would require minimal to none pre-processing. Another aspect requiring more attention in future projects is to make more use of standard (in this case osteoarthritis) ontologies and vocabularies. This will make it much more easy to combine data from different (osteoarthritis) cohorts and would require much less harmonization efforts. Lastly from a usability perspective it is important that sufficient metadata are available in tranSMART, i.e. sufficient data to allow novice users to understand the exact meaning of the parameters being browsed/analyzed.

Next steps

These guidelines are not a static document, but will need to be maintained incorporating new data types, adapting to new software releases, and including lessons learned from running projects. The CORBEL project team intends to keep the document up-to-date in close collaboration with the TraIT data team and other collaboration partners. The guidelines will be further validated in practice, in particular in the context of the Cancer Core Europe use case, which is about to begin. That will probably lead to further refinement of these guidelines.
Data loading for sample data is still in its infancy as also highlighted in the cBioportal section. This is mainly due to lack of sample handling capabilities in tranSMART in the current version (16.2). Once these capabilities become available (current expectation: H1 2018) we will refine and elaborate the data loading procedures accordingly.

References

- Zandifar A et al (2017). A comparison of accurate automatic hippocampal segmentation methods. Neuroimage. 155:383-393.
- Dam EB et al (2015). Automatic segmentation of high- and low-field knee MRIs using knee image quantification with data from the osteoarthritis initiative. J Med Imaging. 2(2):024001
- Hameeteman K et al (2011). Evaluation framework for carotid bifurcation lumen segmentation and stenosis grading. Med Image Anal. 15(4):477-88.

- Aerts H et al (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature Communications 5, Article number: 4006.
- Marcus DS et al (2007). The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data. Neuroinformatics. 5(1):11-34.
- Klein S et al (2015). XNAT imaging platform for BioMedBridges and CTMM TraIT. Journal of Clinical Bioinformatics. 5(Suppl 1):S18.
- Runhaar J et al (2015). Prevention of Knee Osteoarthritis in Overweight Females: The First Preventive Randomized Controlled Trial in Osteoarthritis, The American Journal of Medicine, Volume 128, Issue 8, 888 - 895.e4.
- Hansson NM et al. Evaluation of two multi-atlas cartilage segmentation models for knee MRI: data from the Osteoarthritis Initiative, 9th International Workshop on Osteoarthritis Imaging, 2016.
- eTRIKS Standards Starter Package: DOI: 10.5281/Zenodo:50825
- tranSMART batch documentation:
- https://github.com/thehyve/transmart-batch/tree/master/docs

Abbreviations

API	Application Programming Interface
APPROACH	Applied Public-Private Research enabling Osteoarthritis Clinical Headway
CT	Computed Tomography
CORBEL	Coordinated Research Infrastructures Building Enduring Life-science services
DICOM	Digital Imaging and Communications in Medicine
eTRIKS	European Translational Information & Knowledge Management Services
ETL	Extract - Transform - Load
GEO	Gene Expression Omnibus
HGNC	HUGO Gene Nomenclature Committee
IMI	Innovative Medicine Initiative
JSON	JavaScript Object Notation
MAF	Mutation Annotation Format
MIG	Minimum Information Guidelines
MRI	Magnetic Resonance Imaging
NGS	Next Generation Sequencing
OA	Osteoarthritis
PET	Positron Emission Tomography
QIB	Quantitative Imaging biomarkers
REST	Representational state transfer
SNP	Single-nucleotide polymorphism
TraIT	Translational Research IT
VCF	Variant Call Format
XML	eXtensible Markup Language
XNAT	Extensible Neuroimaging Archive Toolkit

Delivery and schedule

The delivery is delayed: No