347. Import from Biomaterial Bank

Relevance & Approach

Large collections of biosamples e.g. from tumors or "remainder" material from routine clinical laboratories play an increasing role in translational research. The intuitive query capabilities of i2b2 make it an ideal platform for querying combined data from clinical or study records, biosamples, and analysis data gained from biomaterial. Importing biosample data into i2b2, however, has so far required individually coded ETL pathways.

Unlike the CDISC ODM format for clinical trial data, there currently exists no established standard format for biomaterial data. Therefore, the IDRT biomaterial extractor takes the approach of a "generic concept hierarchy" that contains data elements typically used for sample collection, and combines it with a "system-specific driver" that can be adapted to the actual source system to be connected.

The data elements of the generic biomaterial concept hierarchy are:

- specimen basic information: sample ID, material type
- storage hierarchy and sample location
- SPREC (Standard PREanalytical Code) describing sample preparation and storage conditions (see Betsou et al. 2010)
- individually defined metadata elements

In order to link all data elements not just to a patient or encounter, but to a specific sample, the full biosample concept hierarchy is implemented with modifier codes (see also: Modifier Support in IDRT)

In the IDRT project, a driver was developed to import biosample data from the STARLIMS Biorepository® by Abbott Informatics (tested with version 10.5 and 10.7).

Implementation

In this section, the specific implementation for the STARLIMS® driver is described. The driver was implemented on the Talend Open Studio platform in order to integrate with the other components of the IDRT toolkit. The driver creates CSV files for the full ontology and fact data generated from the source system, which is then imported using the standard IDRT CSV extractor. A driver for a different biosample management system only needs to implement the extraction and preparation of ontology and fact CSV files.

Extraction

All required raw data tables are copied from the source system into a staging area. Read-access to the database of the production system or a database copy is required. The following tables are transferred:

Context	Table	Description
Sample core data	INVENTORY	Biosample inventory objects (e.g. samples, aliquots, containers), including relevant core attributes (e.g. sample ID, material type)
	INVENTORY_TRANSACTIONS	Actions taken with samples (e.g. splitting into aliquots, moving) including links to parent samples of aliquots
	MATERIALS	material types
	RASPROJECT_INVENTORY	links between samples and projects
Sample metadata	METADATA	metadata content for samples
	METADATA_TEMPLATE_FIELDS	field definitions for flexible sample metadata
Storage hierarchy	DEPARTMENTS	top level of storage hierarchy
	BUILDINGS	buildings inside of departments
	ROOMS	rooms inside of buildings
	LOCATIONS	recursive storage structure inside of a room (e.g. a freezer subdivided into slots, racks and rack positions)

Transformation

Sample core data elements are transformed using standard TOS components into the required ONT and DATA csv files.

The SPREC concept hierarchy is copied from a predefined CSV file included in the IDRT distribution.

The storage hierarchy requires further processing for the "lower" levels to determine the parent/child relations and hierarchy level within the recursive part.

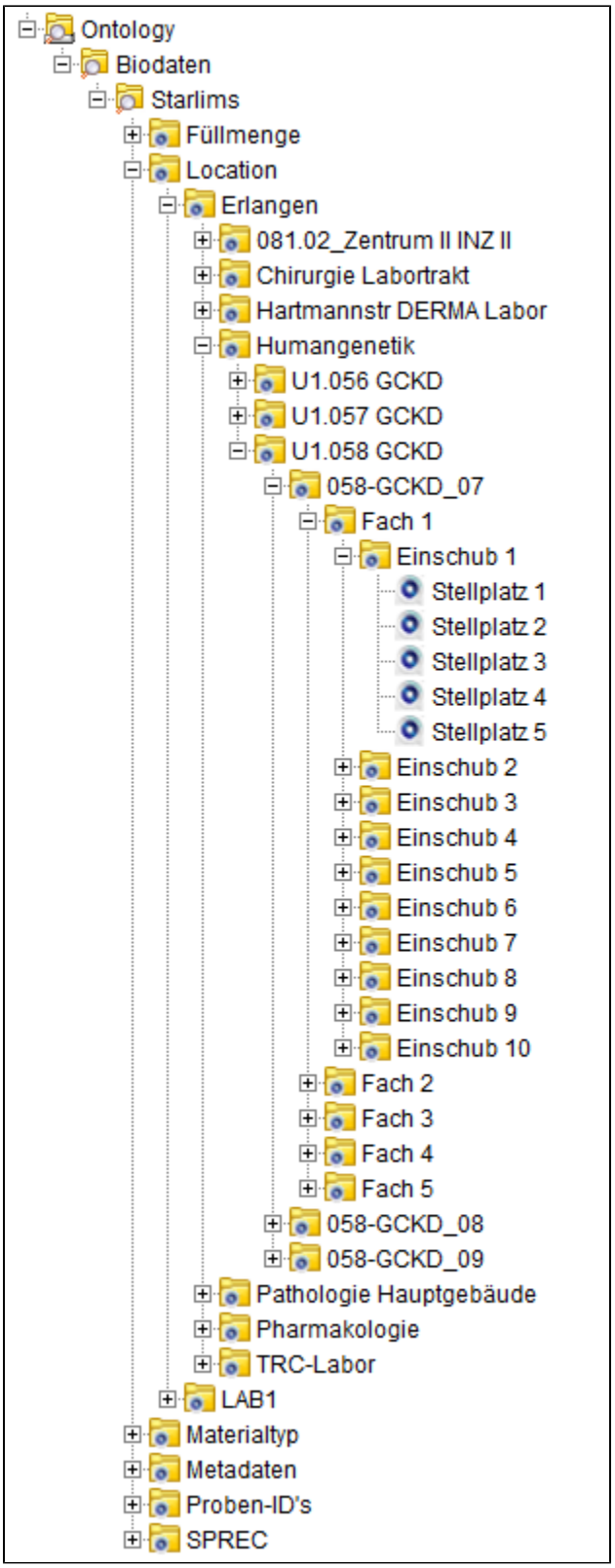
STARLIMS® provides flexible metadata templates which can be designed to fit individual project needs (e.g. to collect additional information about sample quality or processing steps not covered by SPREC). Metadata templates can be defined separately for each project and material type and are versioned. Samples within one project thus may contain metadata from different templates or template versions. In addition to TOS standard components, a small dedicated Java program is needed to generate a consistent concept hierarchy of templates, versions, fields and values from the raw source data.

To enable queries over both specimen and (clinical) phenotype data, patient and encounter IDs need to be linked to the samples. Depending on data protection requirements, the biospecimen database may contain only sample IDs, with a separate ID management database required to provide the link between samples and patients. Inserting patient and encounter IDs into the import dataset thus requires an additional, project-specific transformation step which is not included in the IDRT distribution.

Loading

The ONT and DATA csv files generated in the Transformation step are loaded using the standard CSV extractor module of the IDRT toolkit.

An example of concept hierarchy after loading biosample data with the IDRT biomaterial extractor is shown in the following figure:


Ontology
Biodaten
Starlims
Füllmenge
Location
Erlangen
081.02_Zentrum II INZ II
Chirurgie Labortrakt
Hartmannstr DERMA Labor
Humangenetik
U1.056 GCKD
U1.057 GCKD
U1.058 GCKD
058-GCKD_07
Fach 1
Einschub 1
Stellplatz 1
Stellplatz 2
Stellplatz 3
Stellplatz 4
Stellplatz 5
Einschub 2
Einschub 3
Einschub 4
Einschub 5
Einschub 6
Einschub 7
Einschub 8
Einschub 9
Einschub 10
Fach 2
Fach 3
Fach 4
Fach 5
058-GCKD_08
058-GCKD_09
Pathologie Hauptgebäude
Pharmakologie
TRC-Labor
LAB1
Materialtyp
Metadaten
Proben-ID's
SPREC