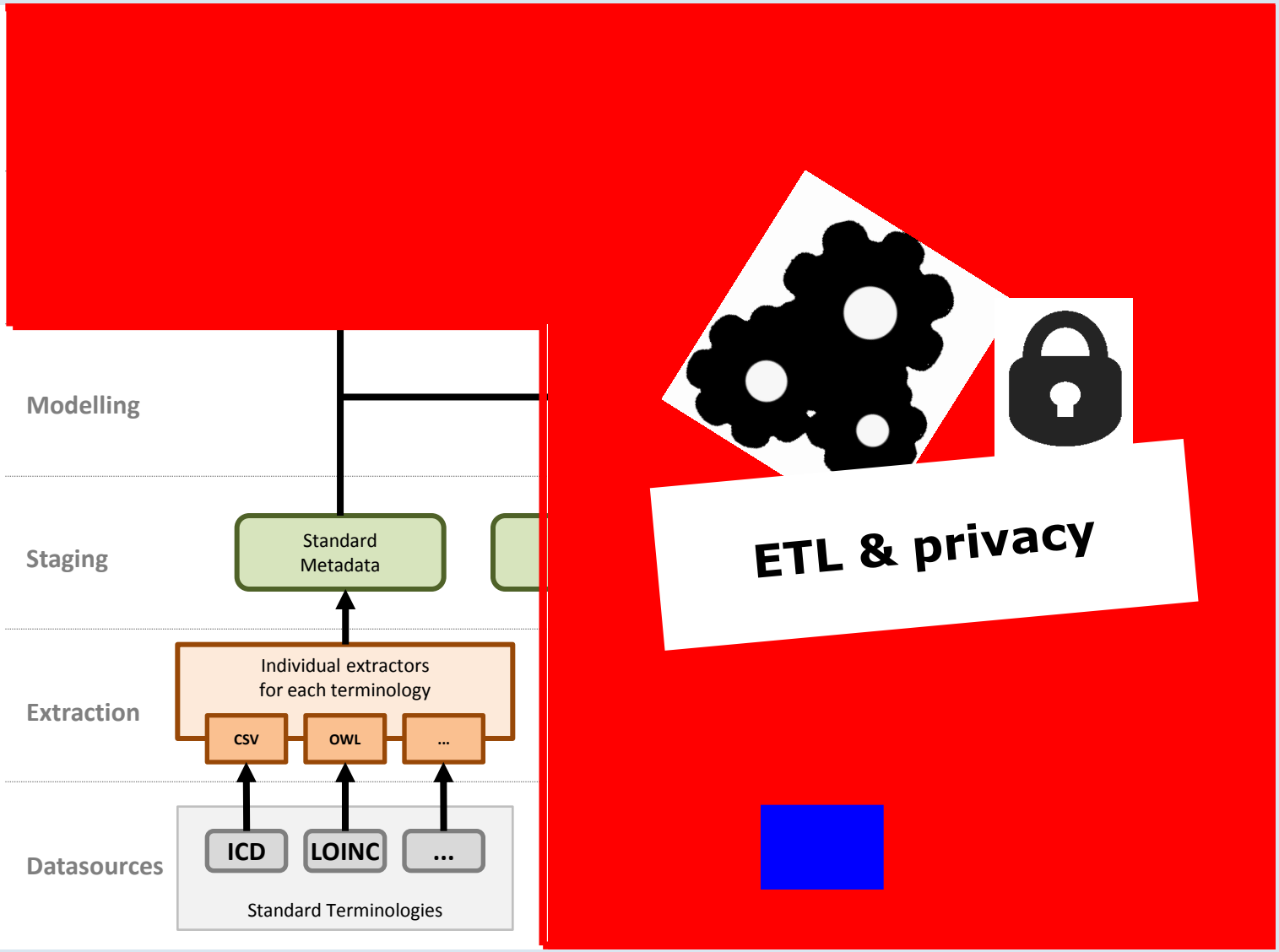
First European i2b2 Academic User Meeting

IDRT: Unlocking Research Data Sources with ETL for use in a Structured Research Database

The IDRT Team (in alphabetical order):

Christian Bauer (presenter), Benjamin Baum, Jan Christoph, Igor Engel, Thomas Ganslandt, Matthias Löbe, Sebastian Mate, Daniel Plog, Hans-Ulrich Prokosch, Matthias Quade, Ulrich Sax, Sebastian Stäubert, Lars Voitel, Alfred Winter

IDRT Architecture

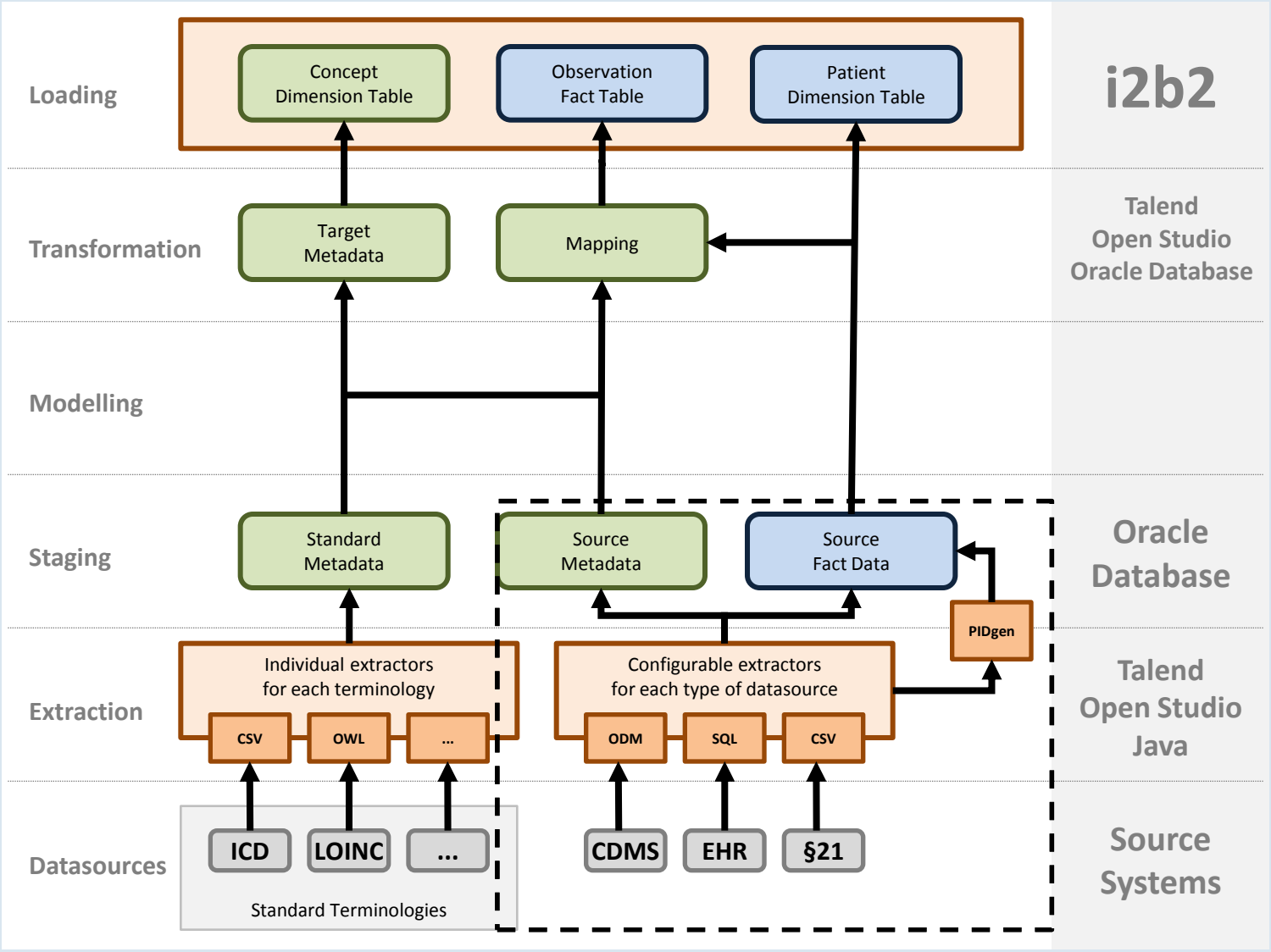


IDRT - Goal

- create tools for a simple and easy import of medical data into the i2b2 database

- challenges
  - How do we get the data into the database?
  - How do we get i2b2 ontologies for the data?

ETL

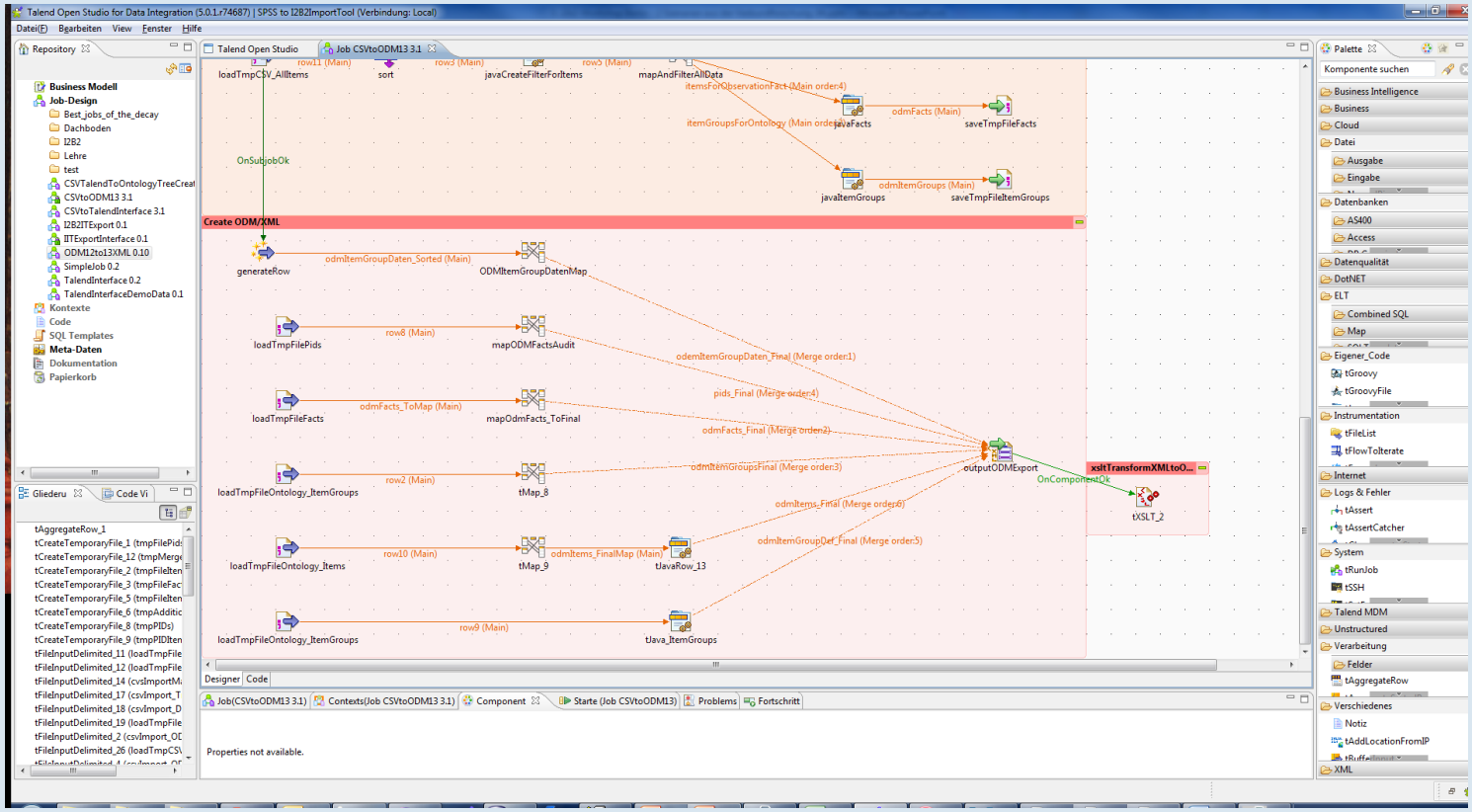

Loading
Transformation
Modelling
Staging
Extraction
Datasources
Concept Dimension Table
Observation Fact Table
Patient Dimension Table
Target Metadata
Mapping
Standard Metadata
Source Metadata
Source Fact Data
PIDgen
Individual extractors for each terminology
CSV
OWL
...
Configurable extractors for each type of datasource
ODM
SQL
CSV
ICD
LOINC
...
Standard Terminologies
CDMS
EHR
§21
i2b2
Talend Open Studio Oracle Database
Oracle Database
Talend Open Studio Java
Source Systems

ETL

- Which data formats do we need?
  - CSV
  - SQL
  - CDISC-ODM
- How can we import the data into the i2b2 database?
  - create generic ETL jobs for the data formats
- How can we get i2b2 ontologies for the patient data?
  - use configuration files to get some user input
- How will we create the ETL?
  - use Talend Open Studio

ETL / Talend Open Studio

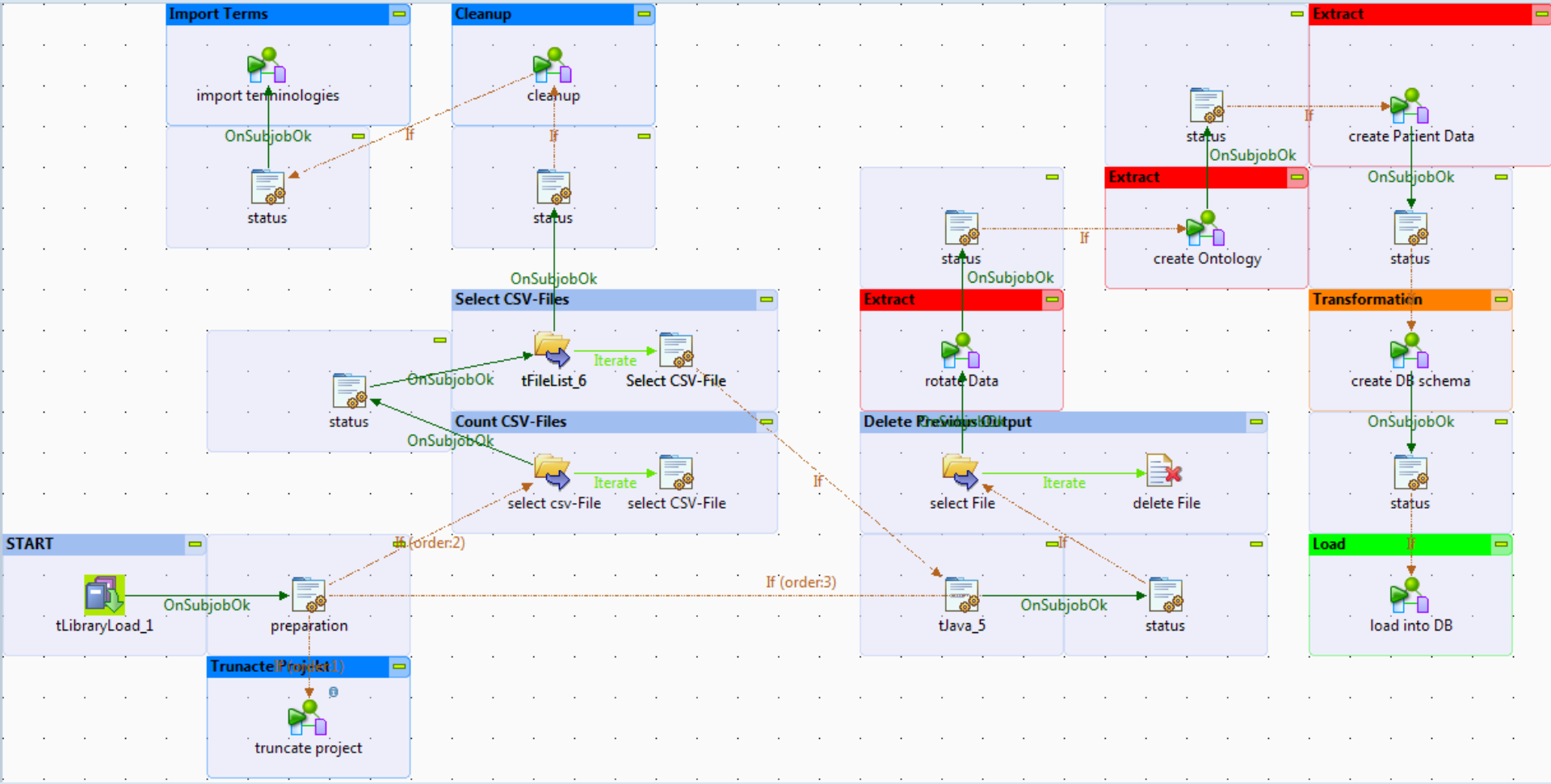
- open source data integration program
  - used for the creation of ETL (extract – transform – load)
  - graphical code generator (Java)



CSV/SQL-ETL

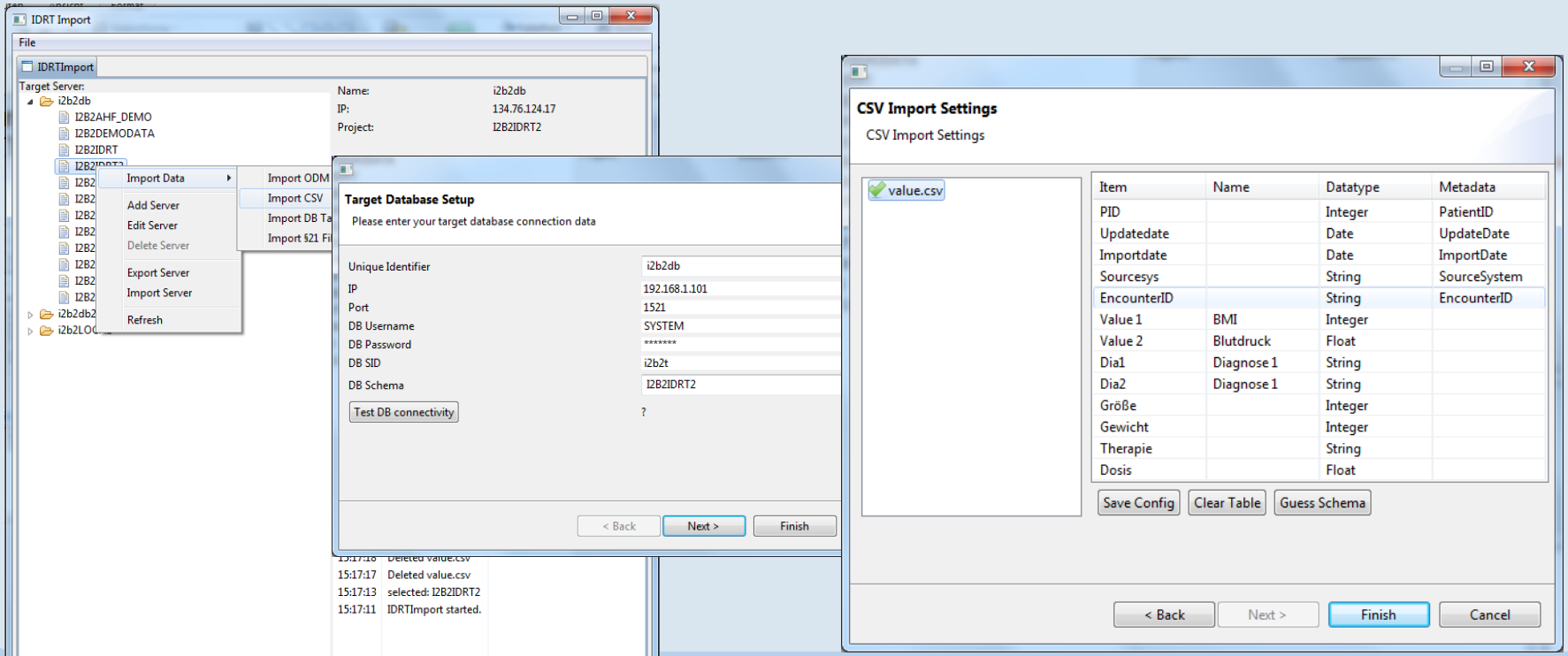
- user creates configuration file
  - nice names, data types, patient ids
- job loads files/database patient data
- job creates i2b2 ontology and patients based on configuration and patient data
- job loads i2b2 specific data and patient data into the i2b2 database

CSV/SQL-ETL

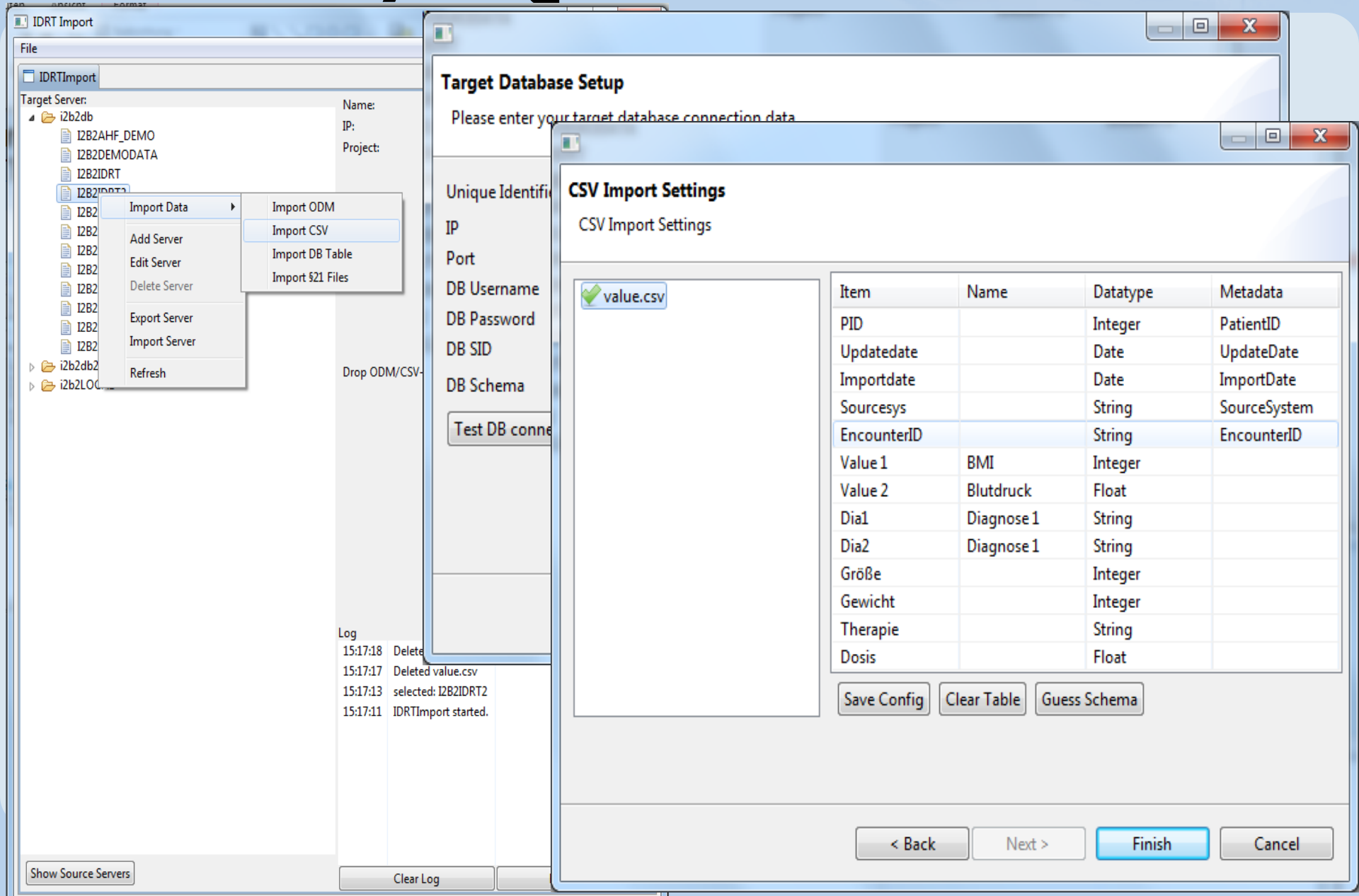

Import Terms
import terminologies
OnSubjobOk
status
If
Cleanup
cleanup
If
status
OnSubjobOk
Select CSV-Files
tFileList_6
Iterate
Select CSV-File
status
OnSubjobOk
Count CSV-Files
OnSubjobOk
select csv-File
Iterate
select CSV-File
If
START
tLibraryLoad_1
OnSubjobOk
preparation
If (order:2)
If (order:3)
Truncate Project
truncate project
status
OnSubjobOk
Extract
rotate Data
Delete Previous Output
select File
Iterate
delete File
If
tJava_5
OnSubjobOk
status
If
Extract
create Ontology
status
OnSubjobOk
Extract
create Patient Data
OnSubjobOk
status
Transformation
create DB schema
OnSubjobOk
status
Load
load into DB

CSV/SQL-ETL

- ETL jobs can be run inside Talend Open Studio, as java code or in a gui
- gui has automatic creation of configurations
- gui has easy editing and automatic saving of configurations
- gui has server browser

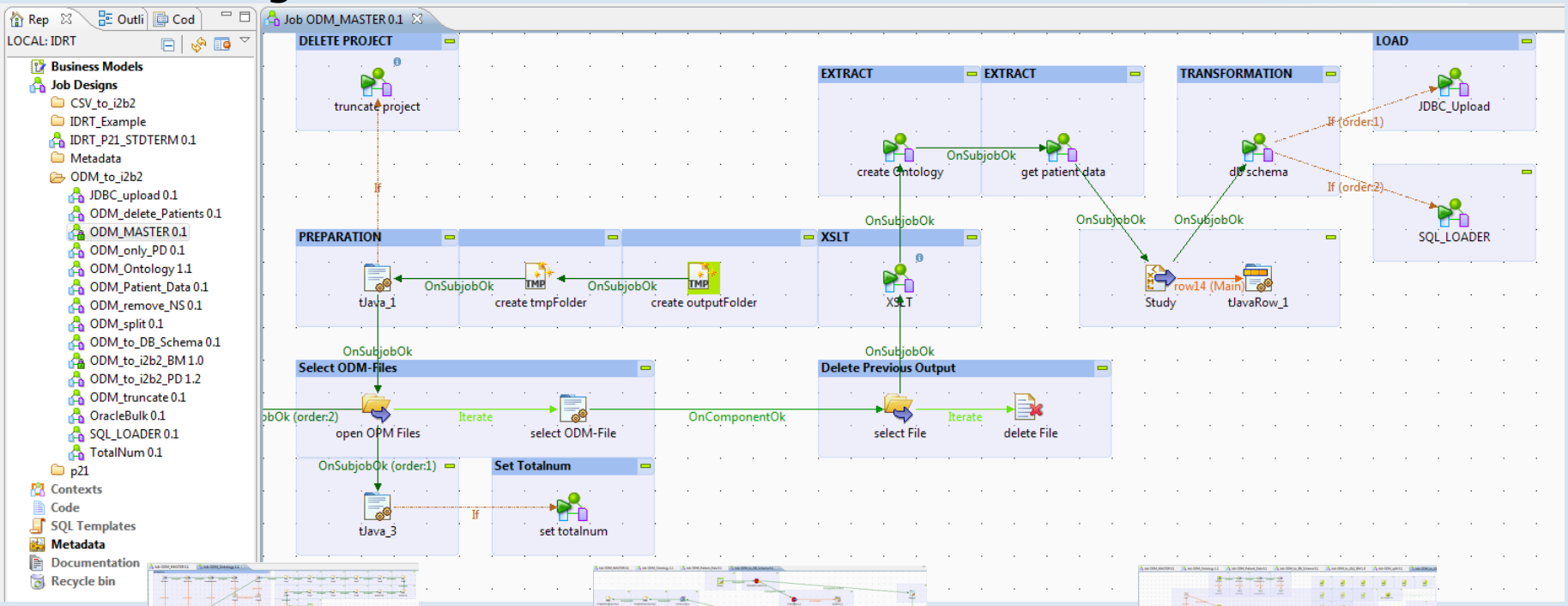


CSV/SQL-ETL


IDRT Import
File
IDRTImport
Target Server:
i2b2db
I2B2AHF_DEMO
I2B2DEMODATA
I2B2IDRT
Import Data
Add Server
Edit Server
Delete Server
Export Server
Import Server
Refresh
Import ODM
Import CSV
Import DB Table
Import §21 Files
Name:
IP:
Project:
Drop ODM/CSV-
Log
15:17:18 Delete
15:17:17 Deleted value.csv
15:17:13 selected: I2B2IDRT2
15:17:11 IDRTImport started.
Show Source Servers
Clear Log
Target Database Setup
Please enter your target database connection data
Unique Identifi
IP
Port
DB Username
DB Password
DB SID
DB Schema
Test DB conne
CSV Import Settings
CSV Import Settings
value.csv
Item
Name
Datatype
Metadata
PID
Integer
PatientID
Updatedate
Date
UpdateDate
Importdate
Date
ImportDate
Sourcesys
String
SourceSystem
EncounterID
String
EncounterID
Value 1
BMI
Integer
Value 2
Blutdruck
Float
Dia1
Diagnose 1
String
Dia2
Diagnose 1
String
Größe
Integer
Gewicht
Integer
Therapie
String
Dosis
Float
Save Config
Clear Table
Guess Schema
< Back
Next >
Finish
Cancel

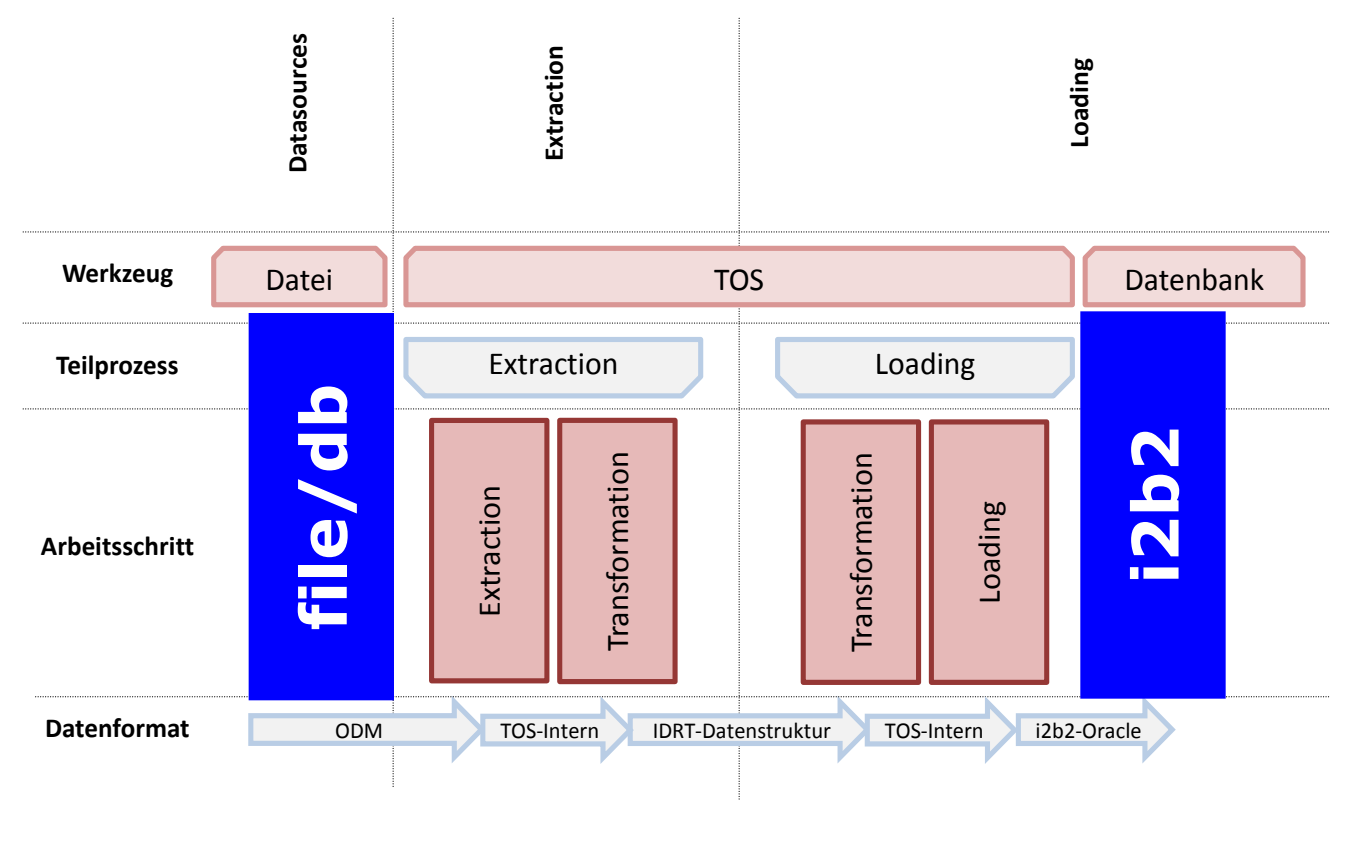
ODM-ETL

- CDISC xml standard
- represents a paper based trial study (study -> events -> forms -> item groups -> items)
- no configuration needed!



ETL

- all ELT jobs have two sections
- writing to i2b2 database is the same job for all the IDRT imports

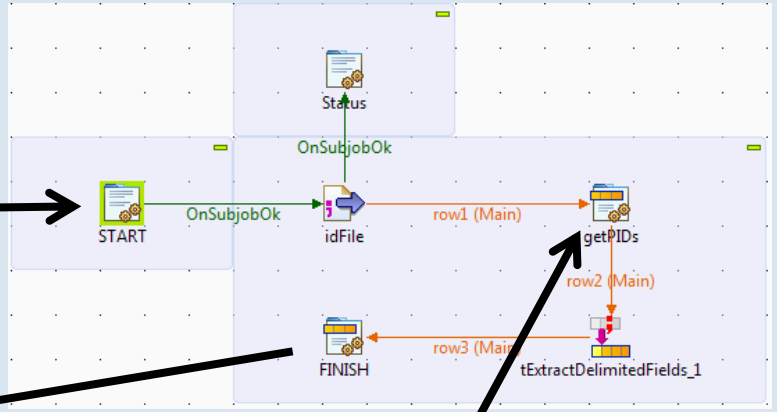


Security: integration of a patient pseudonymization service

patient data with personal information

OWN_ PID	VORNAME	NACHNAME	GESCH LECHT	GEBTAG	MELDE DATUM	Menge	Einheit
0	Martina	Müller	f	1911-01-01	2010-11-21	5	ml
1	Martina	Schmidt	f	1911-11-21	2010-07-20	8	dl
2	Albert	Berg	m	1910-02-02	2011-08-10	10	dl
3	Albert	Einstein	m	1915-05-23	2008-05-16	14	dl

Talend Open Studio sub job



patient data with psyeudoym

PID	Attribute	Value
ZLTGHE3D	Menge	5
ZLTGHE3D	Einheit	ml
AIE8U6GH	Menge	8
AIE8U6GH	Einheit	dl
JEHAOP73	Menge	10
JEHAOP73	Einheit	dl

answer from the patend identifaction

Success:ZLTGHE3D:Es wurde ein passender Fall gefunden.

Discussion & Outlook

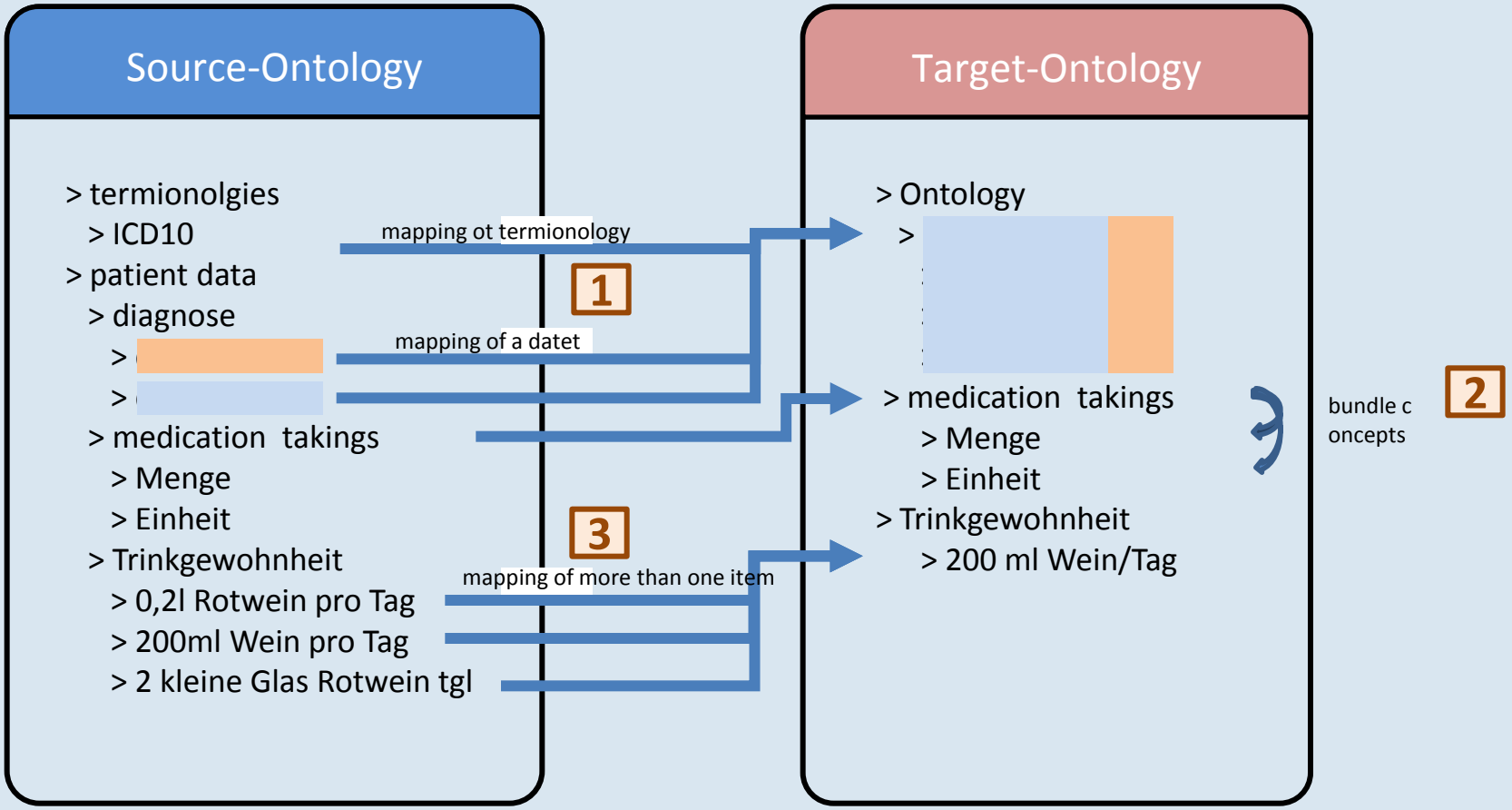
IDRT ETL provides an easy to use package for importing patient data into i2b2. But …

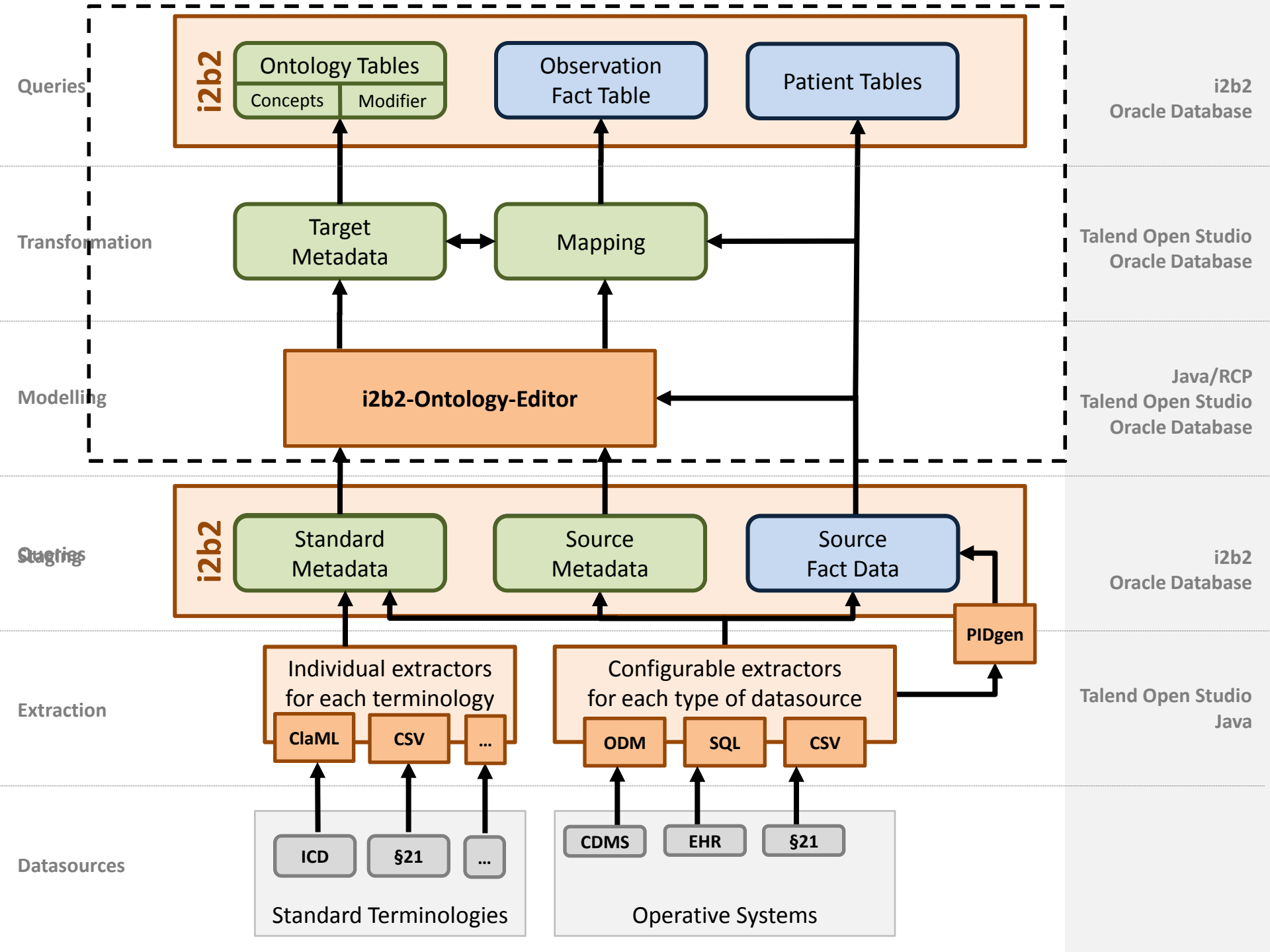
- CSV/SQL i2b2 ontologies are unstructured and often not nice to look at
- no complexe i2b2 ontologies
- we need sub data elements (1 patient -> n biomaterial specimen)

IDRT 2 solutions:

- expand the ETL to incorporate sub data elements
- create a editor to manipulate and create ontologies

IDRT2 - Editor




Queries
Transformation
Modelling
Staging
Extraction
Datasources
i2b2
Ontology Tables
Concepts
Modifier
Observation Fact Table
Patient Tables
Target Metadata
Mapping
i2b2-Ontology-Editor
i2b2
Standard Metadata
Source Metadata
Source Fact Data
PIDgen
Individual extractors for each terminology
ClaML
CSV
...
Configurable extractors for each type of datasource
ODM
SQL
CSV
ICD
§21
...
Standard Terminologies
CDMS
EHR
§21
Operative Systems
i2b2 Oracle Database
Talend Open Studio Oracle Database
Java/RCP Talend Open Studio Oracle Database
i2b2 Oracle Database
Talend Open Studio Java

Discussion & Outlook

IDRT ETL provides an easy to use package for importing patient data into i2b2. But …

- CSV/SQL i2b2 ontologies are unstructured and often not nice to look at
- no complexe i2b2 ontologies
- we need sub data elements (1 patient -> n biomaterial specimen)

IDRT 2 solutions:

- expand the ETL to incorporate sub data elements
- create a editor to manipulate and create ontologies