Data Analysis and ETL Process

Document author: arho.virkki@tyks.fi Contributors: anna.hammais@tyks.fi, juhana.valo@medbit.fi, katja.kanerva-leppanen@medbit.fi

Data Enrichment and Analysis Steps

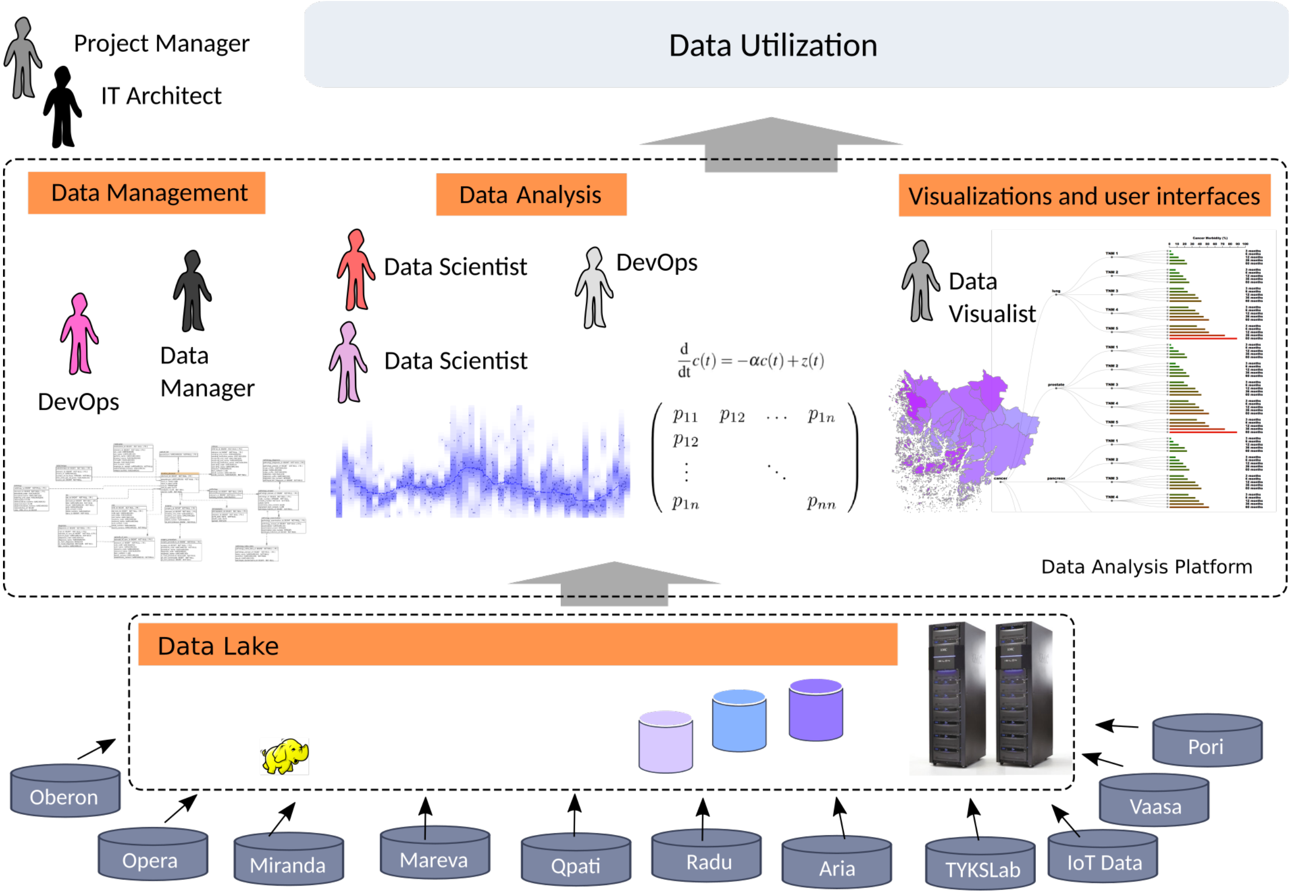

Figure [PDF, SVG]. Process Overview.

Data Flow

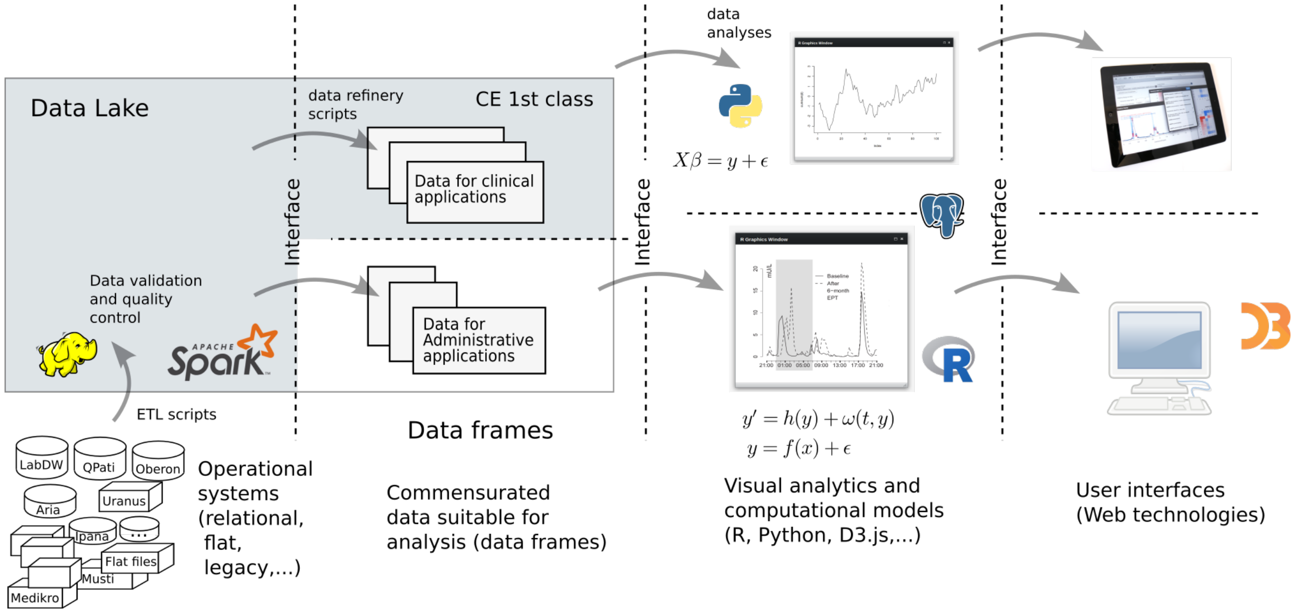
Data Lake
data refinery scripts
CE 1st class
Data for clinical applications
Interface
Data validation and quality control
Data for Administrative applications
ETL scripts
Data frames
LabDW
QPati
Oberon
Uranus
Aria
Ipana
...
Flat files
Musti
Medikro
Operational systems (relational, flat, legacy,...)
Commensurated data suitable for analysis (data frames)
data analyses
$X\beta = y + \epsilon$
Interface
$y' = h(y) + \omega(t, y)$
$y = f(x) + \epsilon$
Visual analytics and computational models (R, Python, D3.js,...)
Interface
User interfaces (Web technologies)

Figure [PDF, SVG]. Data Analysis Workflow.

Research Process Steps

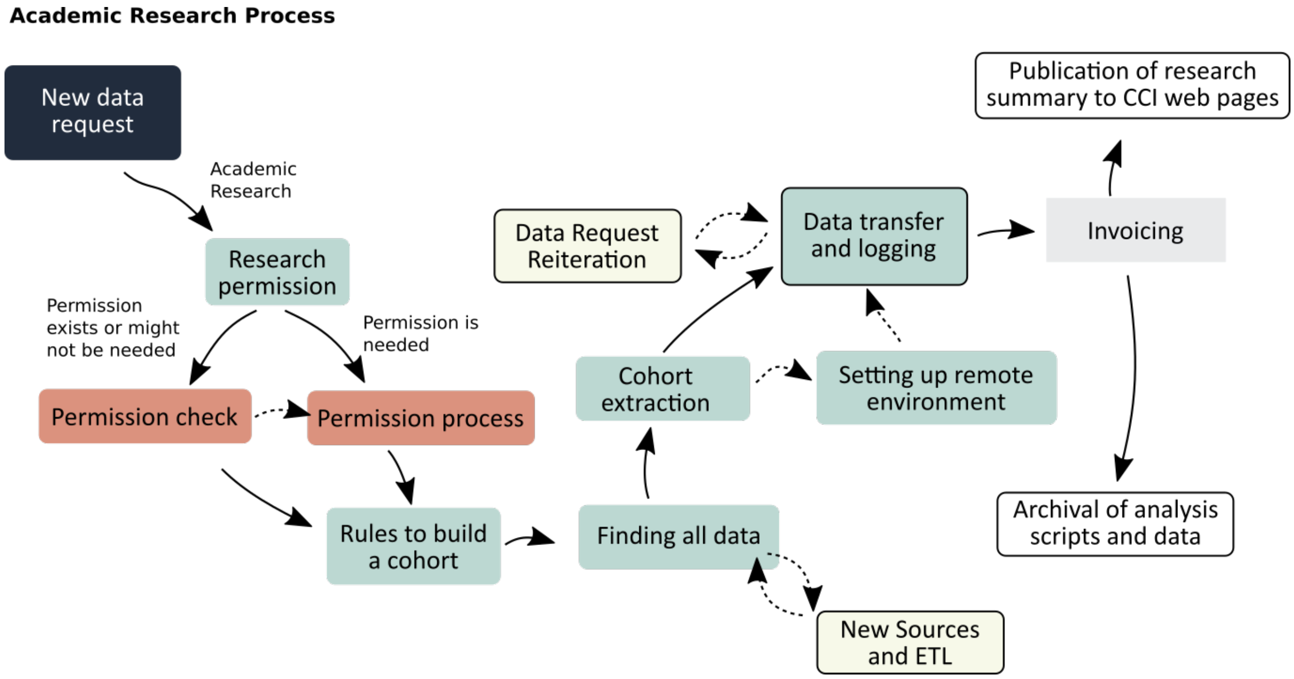

Figure [PDF, SVG]. Academic research process.

Contract Research Process

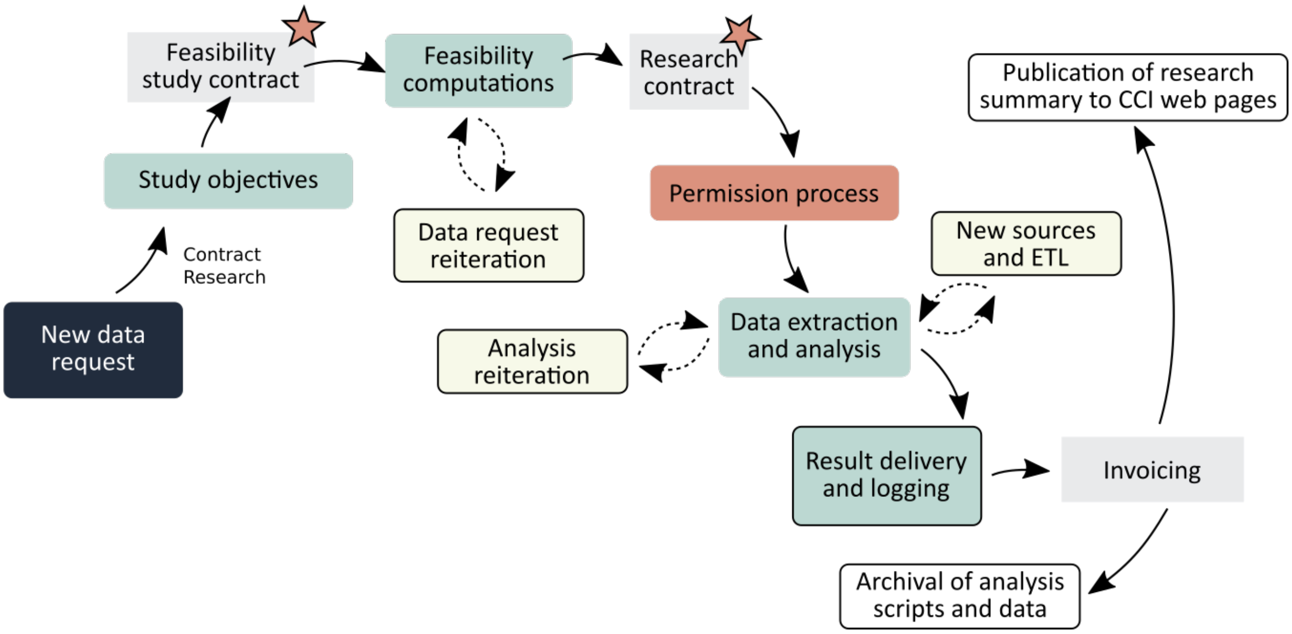

Figure [PDF, SVG]. Contract research process.

New Data Sources and Data Harmonization

New Data Source Process

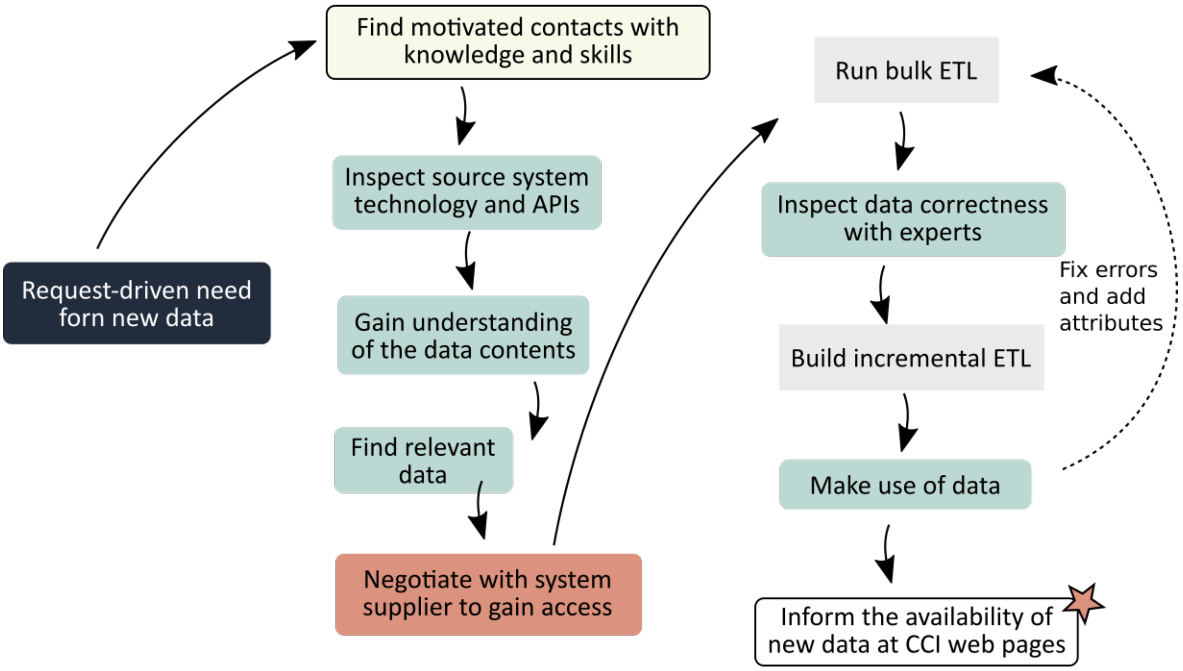
Find motivated contacts with knowledge and skills
Run bulk ETL
Inspect source system technology and APIs
Inspect data correctness with experts
Request-driven need forn new data
Fix errors and add attributes
Gain understanding of the data contents
Build incremental ETL
Find relevant data
Make use of data
Negotiate with system supplier to gain access
Inform the availability of new data at CCI web pages

Figure [PDF, SVG]. New data source process.**

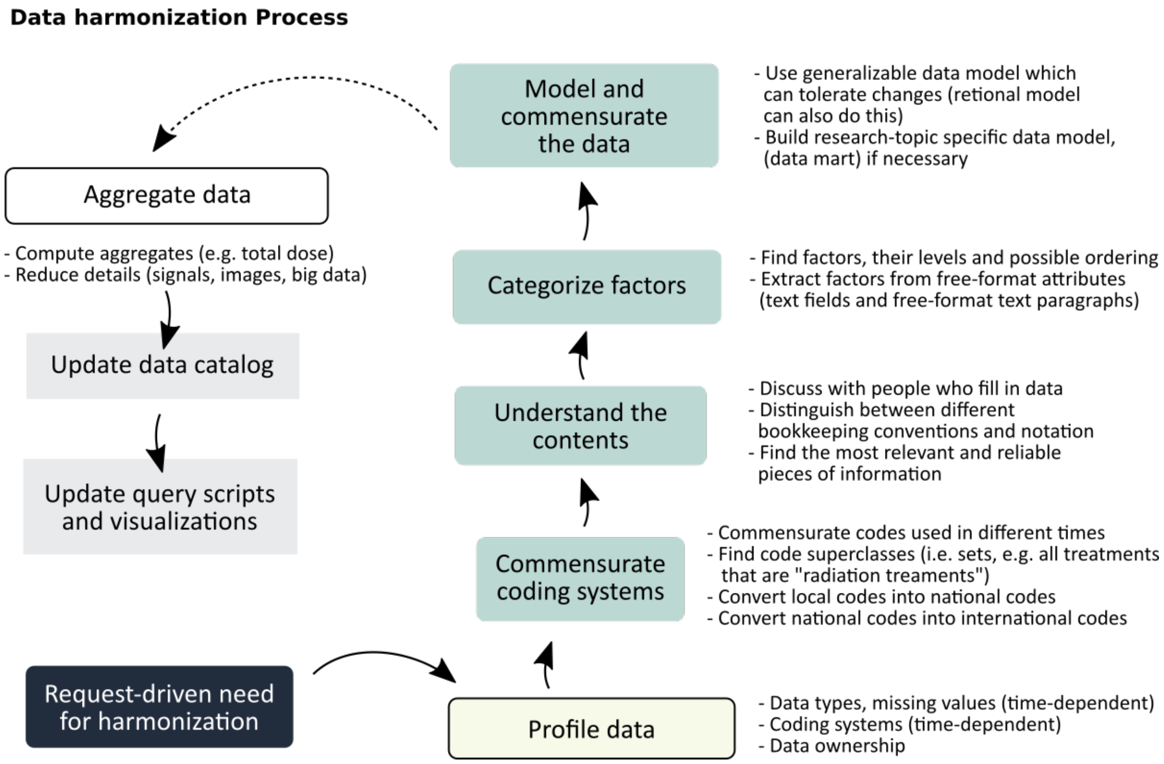
Data harmonization Process
Model and commensurate the data
- Use generalizable data model which can tolerate changes (retional model can also do this)
- Build research-topic specific data model, (data mart) if necessary
Aggregate data
- Compute aggregates (e.g. total dose)
- Reduce details (signals, images, big data)
Categorize factors
- Find factors, their levels and possible ordering
- Extract factors from free-format attributes (text fields and free-format text paragraphs)
Update data catalog
Understand the contents
- Discuss with people who fill in data
- Distinguish between different bookkeeping conventions and notation
- Find the most relevant and reliable pieces of information
Update query scripts and visualizations
Commensurate coding systems
- Commensurate codes used in different times
- Find code superclasses (i.e. sets, e.g. all treatments that are "radiation treaments")
- Convert local codes into national codes
- Convert national codes into international codes
Request-driven need for harmonization
Profile data
- Data types, missing values (time-dependent)
- Coding systems (time-dependent)
- Data ownership

Figure [PDF, SVG]. Data harmonization process.

ETL Steps

1. Data extraction from source
2. File upload
3. Format conversion
4. Type conversion
5. Data integration
6. Semantic unification

ETL Script Repository

The ETL scripts are saved to the Git-repository `ktp@ktpgit:/opt/git/ETL.git`. Two working copies are used in production:

```
ktphadoop.vsshp.net:/var/lib/hive/ETL/
ktpanalytics.vsshp.net:/opt/ktp/ETL/
```

The first is for pre-processing the flat text files at the Hadoop machine, and the second is used with Pentaho Kettle in the data integration phase. The automated etl scripts are run by the *ktp* user. For details, see *crontab -e* as *ktp* at *ktpanalytics*.

Details

Hadoop Environment
PostgreSQL Setup