Data Integration Architecture of DigiMed Bayern

Florian Kohlmayer (Bitcare GmbH)
• Integration of different data types:
  • structured clinical data, genetic data, images, narratives,…

• Integration from different domains:
  • research, clinical care

• Integration from different sources:
  • hospital information system (HIS), laboratory information system (LIS), biobank information management system (BIMS), picture archiving and communication system (PACS), …

• Integration of data from different "clinical pictures"

• Integration of cross-sectional and longitudinal data
Goal: Improve Research and Care

• Improve:
  • Data quality
  • Collaboration
  • Traceability
  • Reproducibility
  • Data security
  • Data protection
  • Transparency
  • Reusability
Risk: Highly Sensitive Data

- Special category of personal data (Art. 9 GDPR: genetic data, data concerning health)

- Many attributes, big collectives, different data sources, ubiquitous data collections (e.g. “quantified self“)

Conflict: **Research ↔ Privacy**
Design Principles / Corner Stones

- Cooperate for concepts, architecture and tools with DIFUTURE and other projects
- Use case driven
- Agile/iterative and pay-as-you go approach
- Security and privacy by design
- Minimize number of interfaces
- Data lake approach
- Multiple extract-transform-load (ETL) steps, separating syntactic and semantic harmonization
- Common data block format
- Only pseudonymized data will be loaded into the data lake
- Separation of components with strict access permissions
Architecture: Overview

- Ancillary system
- Hospital information system
- Ancillary system
- Picture management system
- Research system
- OMICS management system

Connectors:
- Pseudonymization, record linkage, consent management, provenance tracking

Data lake (common data format)
- Data types: structured, OMICS (links and/or raw data), images (links and/or raw data),...

Pseudonymization, risk management

Cohort selection
- Data mart 1
- Data mart 2
- Analysis tools

Analysis platform

Clinical DWH

syntactic harmonization

semantic harmonization
Data Lake – Data Model:
Data Block Example (pre-pseudo)

```
"blockVersion": 1,
"blockContent": "DATA",
"identity": {
  "subjectID": "12345",
  "otherIDs": [
    { "type": "CASE_ID",
      "value": "55566778"
    }
  ],
  "directlyIdentifyingInfos": [
    { "key": "Lastname",
      "value": "Tester",
      "datatype": "STRING"
    },
    { "key": "Firstname",
      "value": "Test",
      "datatype": "STRING"
    }
  ],
  "pseudonymized": false,
  "sourceSystem": "HK_DB"
},
"importTimestamp": "2019-05-06T12:50:27.718",
"contentType": {
  "type": "LAB_DATA",
  "version": 1
},
"beginTimestamp": "2019-04-08T11:32:00",
"endTimestamp": "2019-04-09T00:00:00",
"consent": ...
"provenance": ...
```

"values": [
  [ { "key": "Param",
      "value": "Kalium",
      "datatype": "STRING"
    },
    { "key": "Value",
      "value": "5.2",
      "datatype": "DOUBLE"
    },
    { "key": "Unit",
      "value": "mmol/L",
      "datatype": "STRING"
    }
  ],
  [ { "key": "Param",
      "value": "INR",
      "datatype": "STRING"
    },
    { "key": "Value",
      "value": "1.0",
      "datatype": "DOUBLE"
    }
  ]
]
Software Solutions Used (Examples)

- **Collection**
  - DIS

- **Workflow pipelines / OMICS**
  - Galaxy
  - OpenBIS

- **Integration**
  - DIFUTURE und DigiMed Softwarelösungen

- **Analysis**
  - tranSMART
  - I2b2
  - Re-dash
  - Clinical Knowledge Graph
Prototype: Segments

Hospital-intranet

- HK-DB, EPU-DB (Oracle)
- DIS + add. research data (MySQL, CSV)
- Medico data (Ingres)
- QIMS data (Advantage)

Trustcenter-intranet

- Record linkage
- Pseudonymization

Internet

- Data lake (RESTHeart, MongoDB)
- tranSMART-Exporter
- tranSMART-Instance

Research-intranet

- Data scientist
- Internal researcher

DMZ

- External researcher

Portal, data marts, analysis platform
Prototype: Key-Facts

- Integration of:
  - Lab results, medication, risk factors
  - Catheter, lesions, medication, follow-up
  - Electrophysiological examination
  - Research data, genotypic metadata

- Data lake dimension:
  - Five collections (MEDICO, HK_DB, EPU_DB, DIGIMED_CSV, GWAS_CSV)
  - ~19 mio. datablocks
    - MEDICO: ~ 6.9 Mio
    - HK_DB: ~12 Mio
    - EPU_DB: ~ 285.000
    - DIGIMED_CSV: ~2.000
    - GWAS_CSV: ~9.000

- Export into tranSMART
  - Four “studies”, integrating data e.g. from HK_DB and Medico
  - Total of ~9 Mio. facts (patients: ~170.000)
Proof-of-Concept: Screenshot “Datalake Browser”

Raw medication data block (truncated)
Proof-of-Concept: Screenshot tranSMART

Summary statistics: diabetes vs. non-diabetes
One size fits all?

- Different process steps
  - Collection
  - Preprocessing
  - Integration
  - Analysis

- Every process step has different requirements
  - Data security and privacy
  - Storage of data and processing (OLAP vs OLTP)
  - Functionalities
  - Workflows

→ Heterogeneous software solutions, but generic/common architecture and security and privacy concept