FOSDEM 2020, Brussels

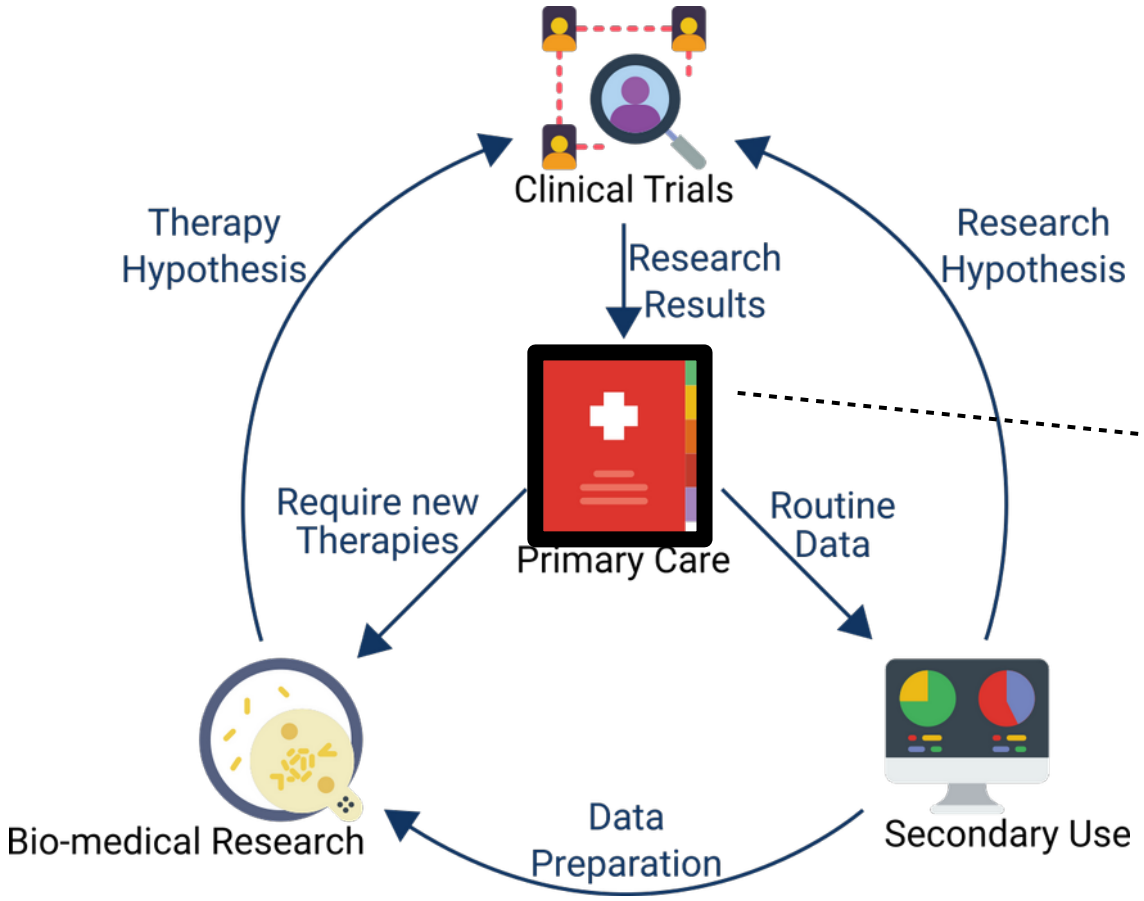
Spotlight on Free Software Building Blocks for a Secure Health Data Infrastructure

Marcel Parciak, Markus Suhr, Tibor Kesztyüs, Dagmar Krefting

University Medical Center Göttingen
Department of Medical Informatics
Germany

http://mi.umg.eu

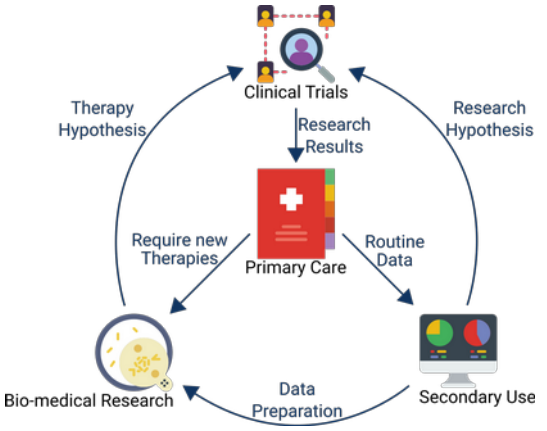
Flows of Information in Medicine



The dreaded
Black Box
of (mostly) proprietary
software

Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

Fields in Medical Research

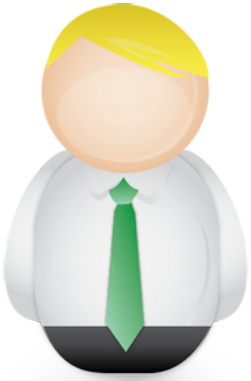


- Each field usually operates on its own
  - Each one consists of own IT-solutions and / or IT-infrastructure
  - Different laws, guidelines and organizational constraints apply

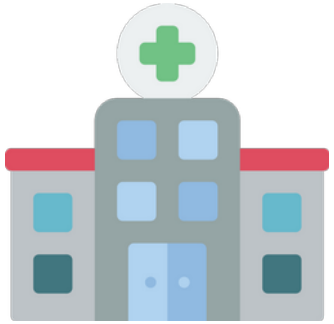
- As a result, each field operates different software solutions forming a heterogeneous IT-landscape

Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

Meet Bob...



Bob suffers from chronic heart insuffiency

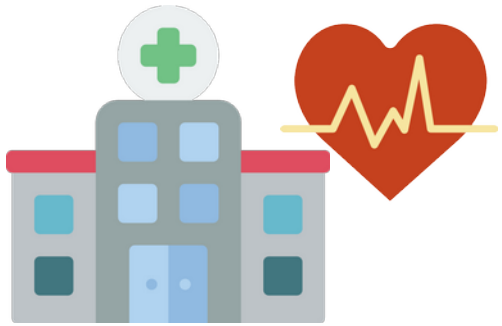


The development of Bob‘s condition is
monitored through yearly check-ups at the Hospital

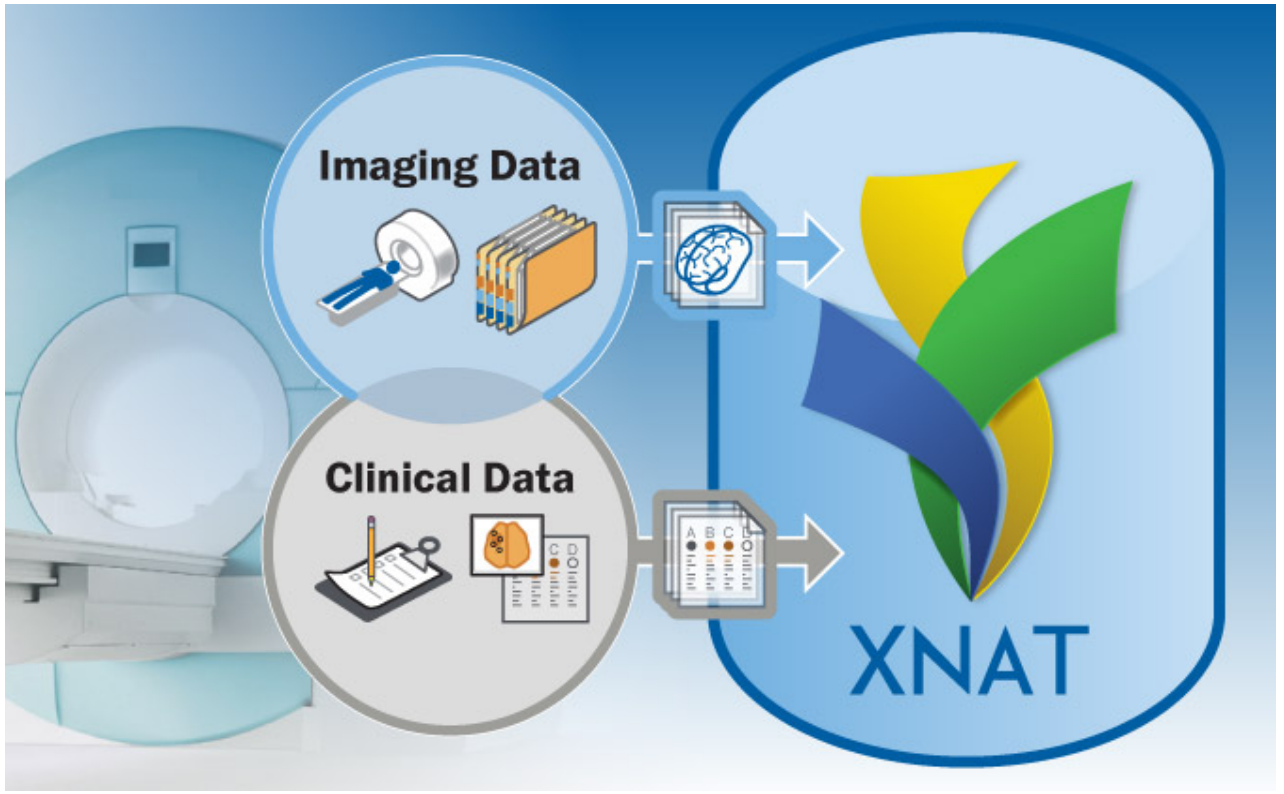
Echocardiography, Vital parameters,
Medication are collected

...and stored in specialised
Clinical information systems

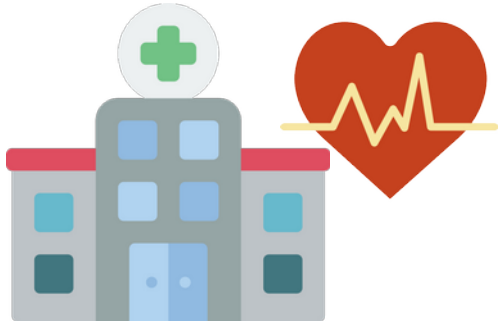
Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de



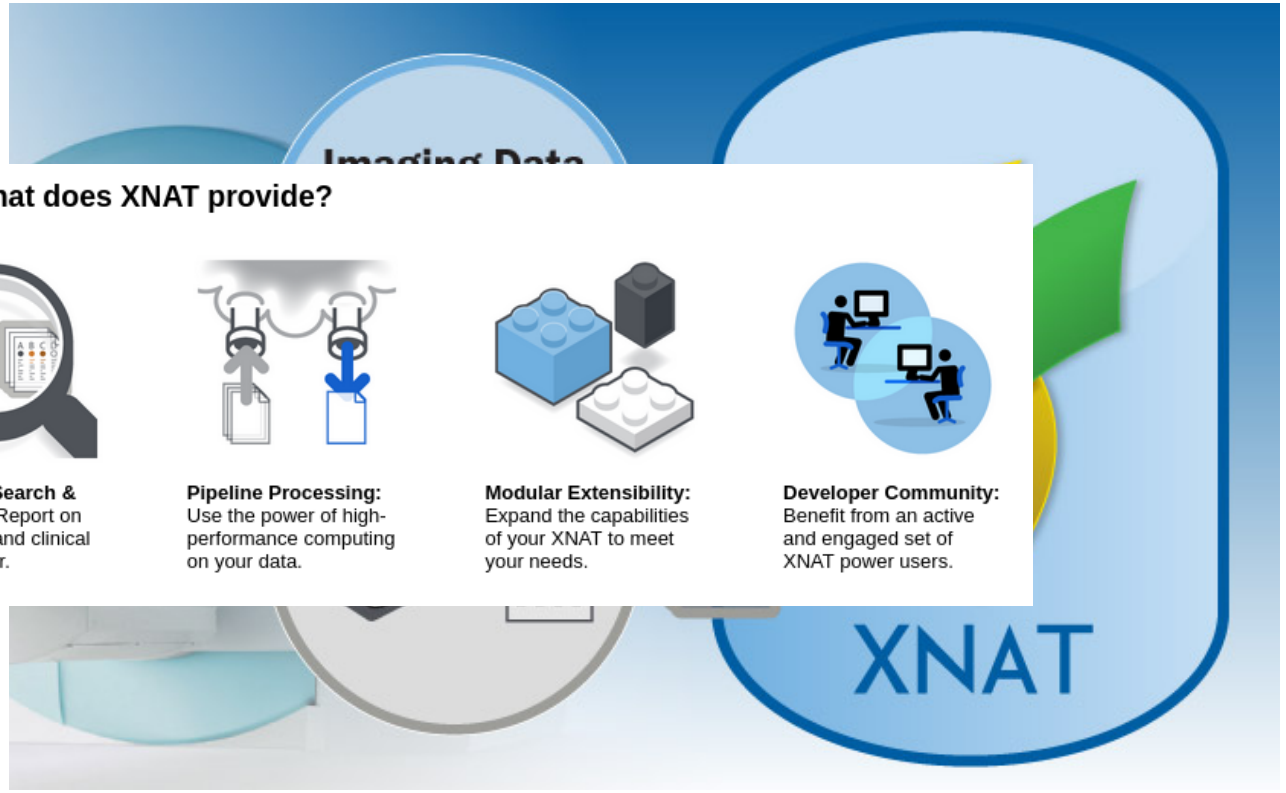
The Hospital uses the open source software **XNAT** to store echocardiography data (images, videos) and measured vital parameters



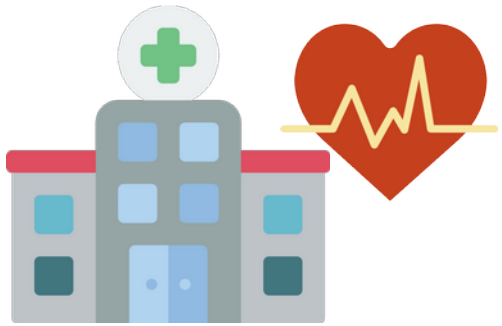
Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de



The Hospital uses the open source software **XNAT** to store echocardiography data (images, videos) and measured vital parameters



Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

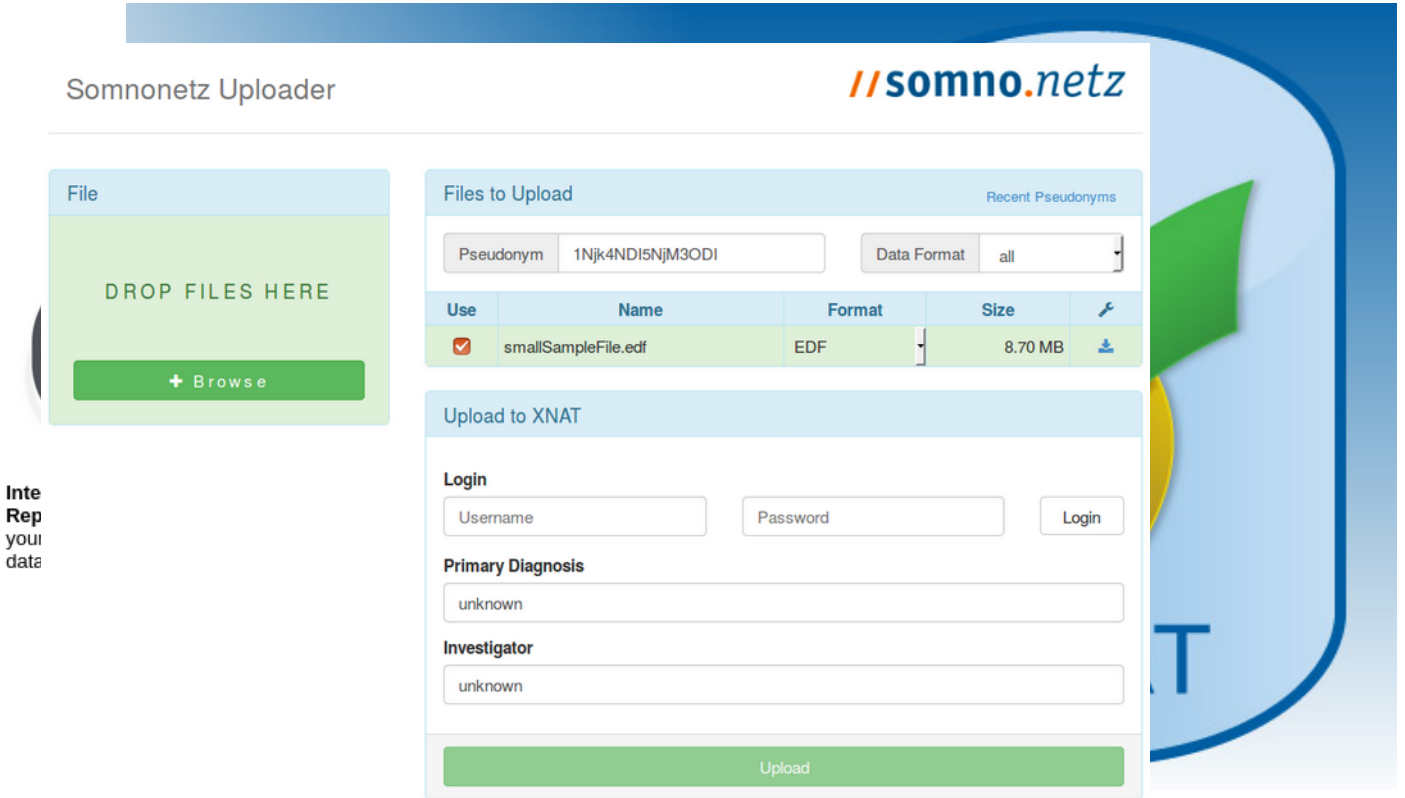


The Hospital uses the open source software **XNAT** to store echocardiography data (images, videos) and measured vital parameters

Full DICOM Integration and Anonymization:
Get image data in, and keep PHI out.

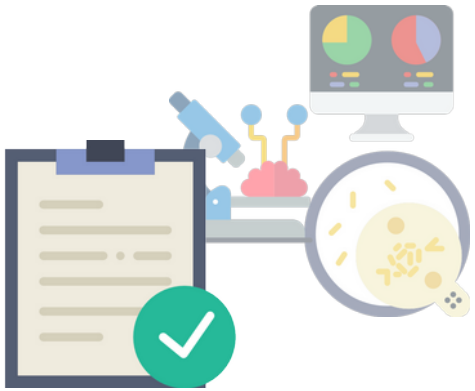
Secure Access & Permission Control:
You decide who does what with your data.

Inte
Rep
your
data



Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

Bob has given **consent** that his routine medical data may be used for **research purposes**



Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

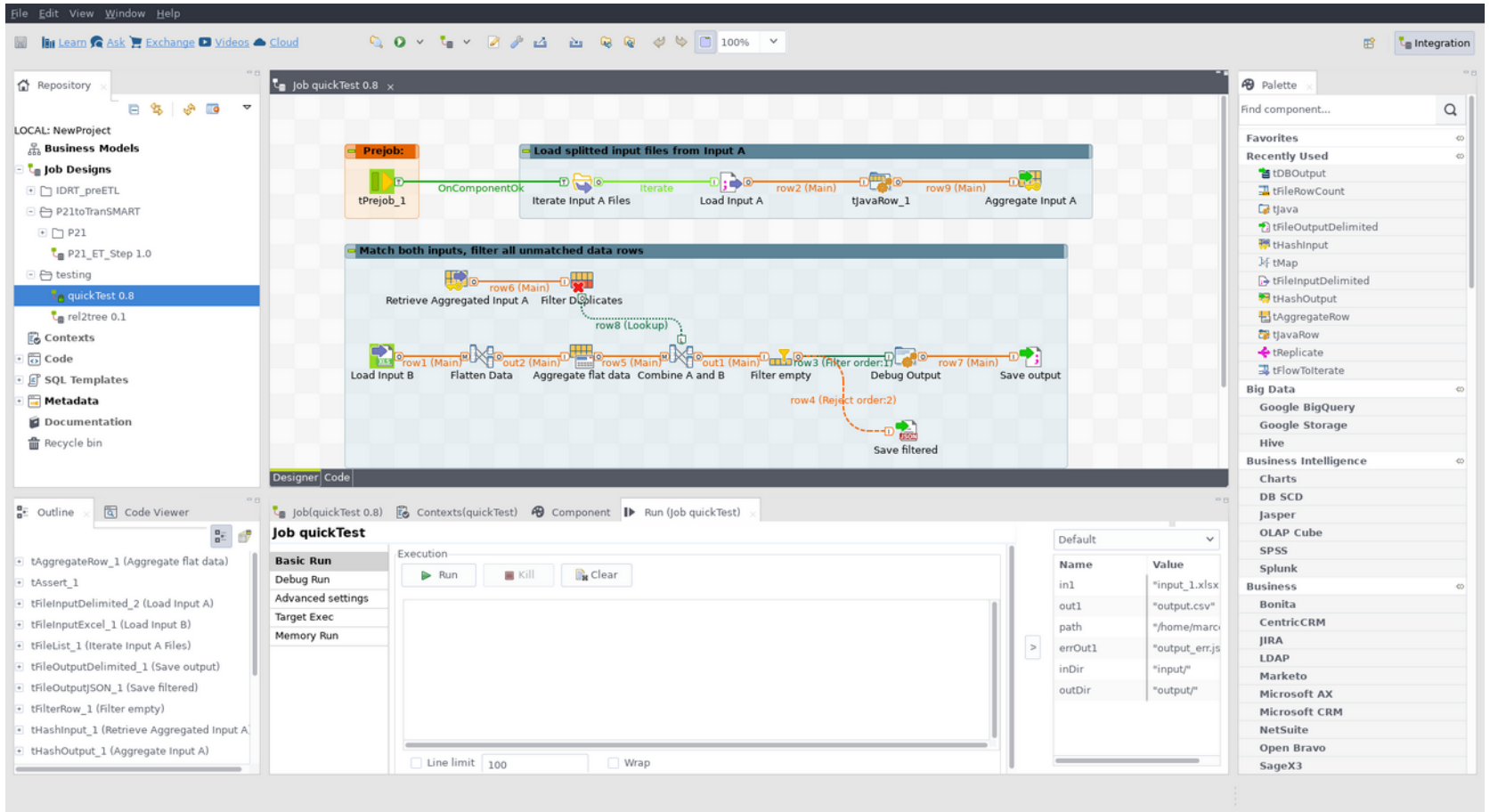
Which leads us to... Alice!

Alice is a health data engineer at the hospitals‘ medical data integration centre

Alice creates data integration pipelines that
extract data from clinical systems,
mask identifiers of patients,
transform extracted data and
load them into a research data repository

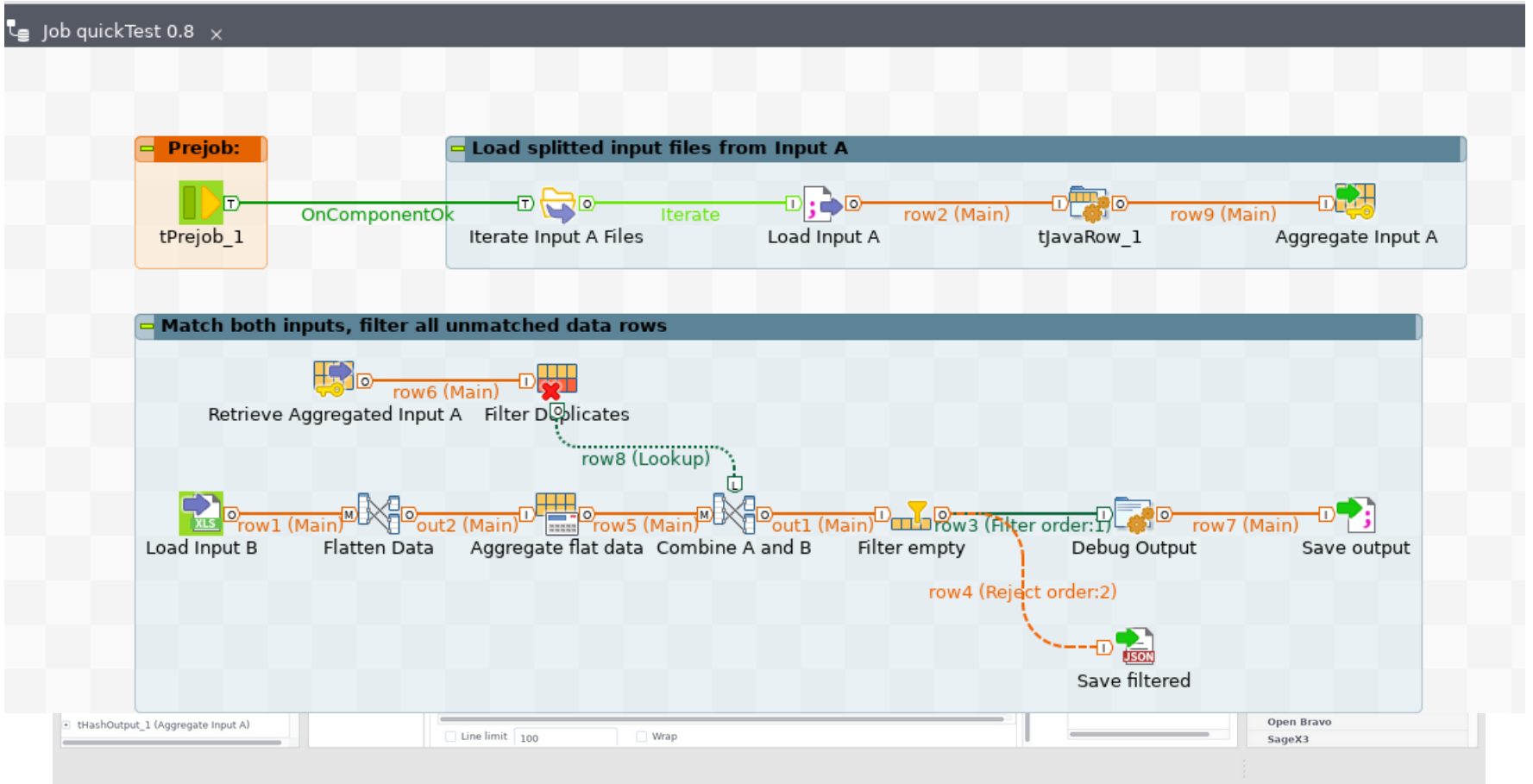
Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

Alice uses **Talend Open Studio for Data Integration**
to create data integration pipelines



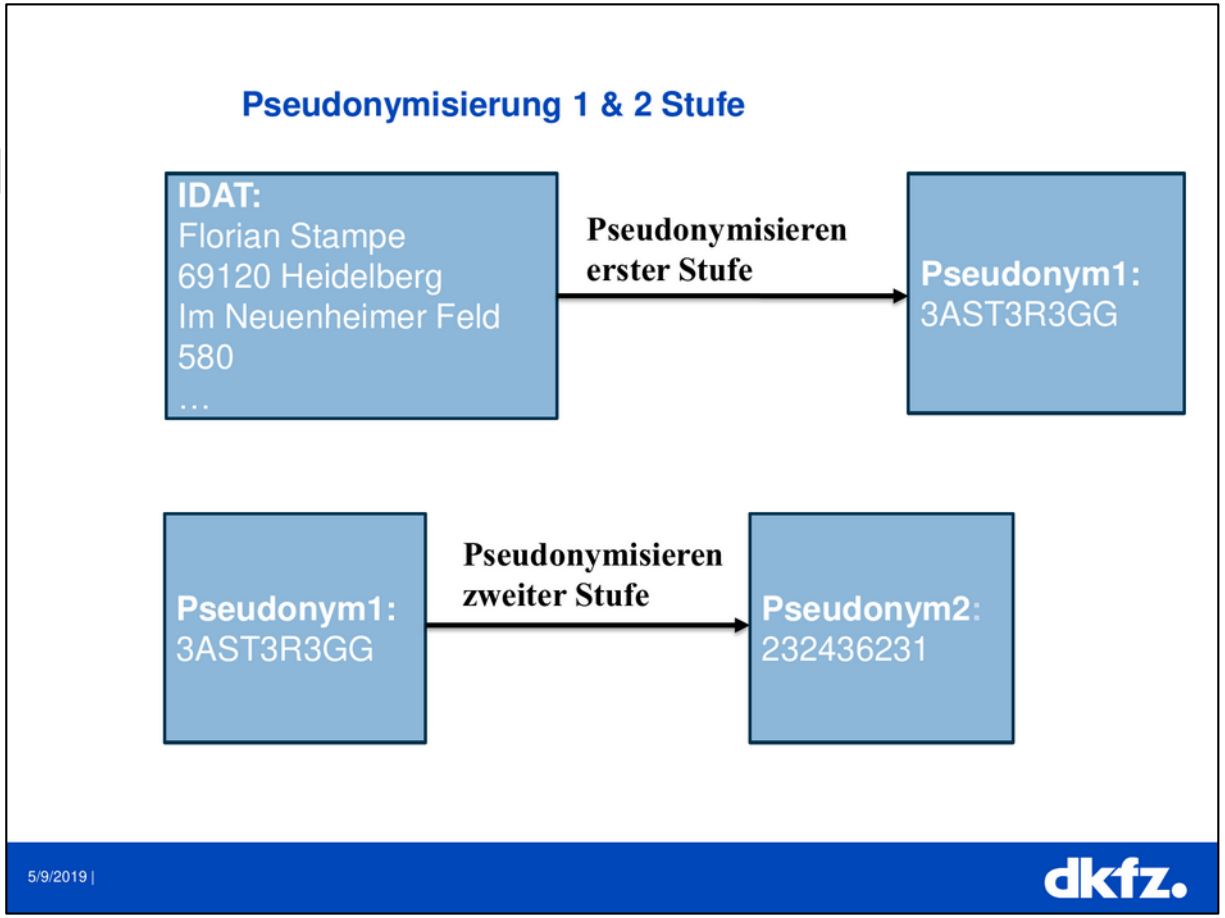
Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

Alice uses **Talend Open Studio for Data Integration**
to create data integration pipelines



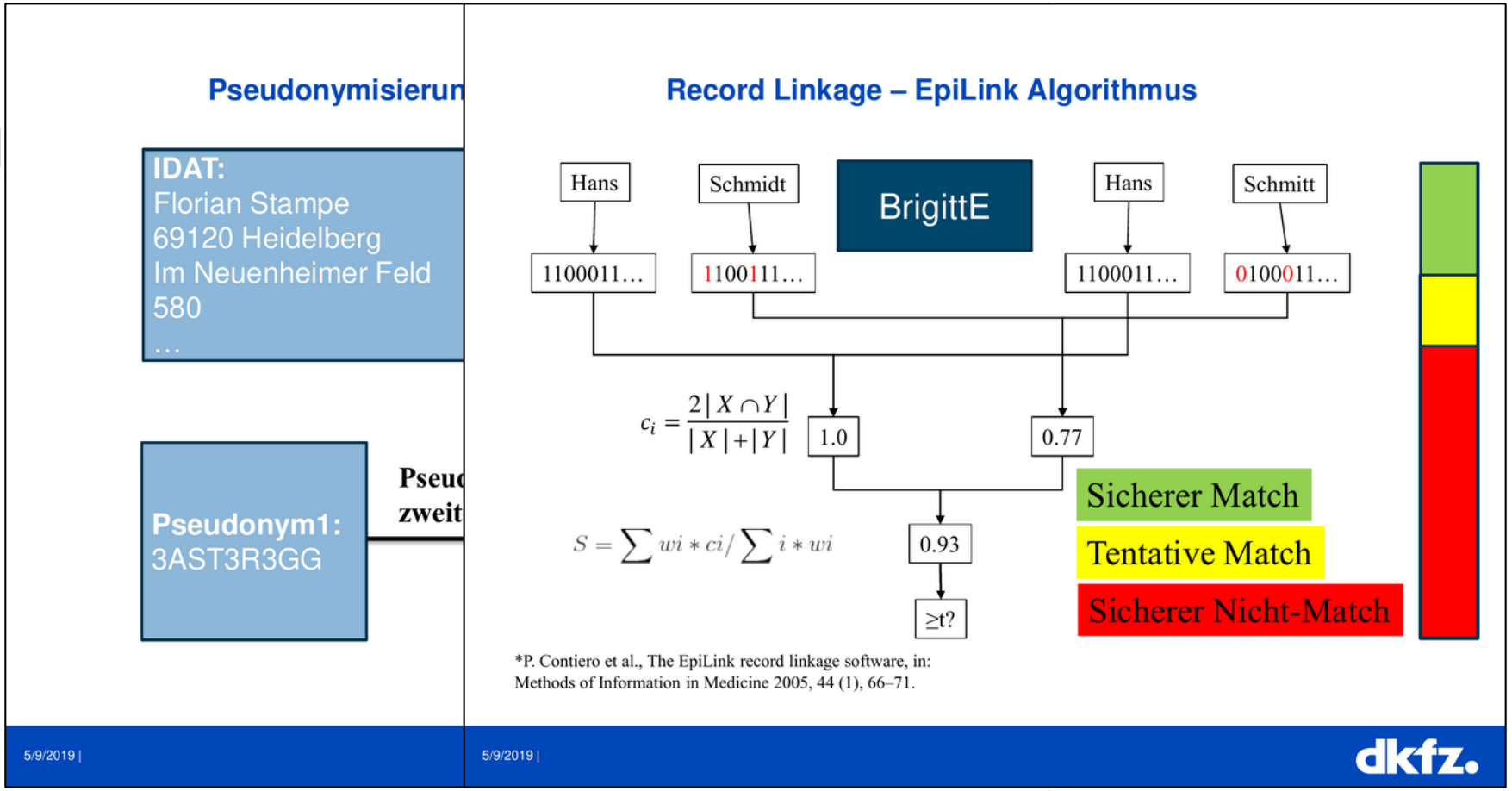
Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

The **Mainzelliste** is used to mask identifiying attributes in medical data



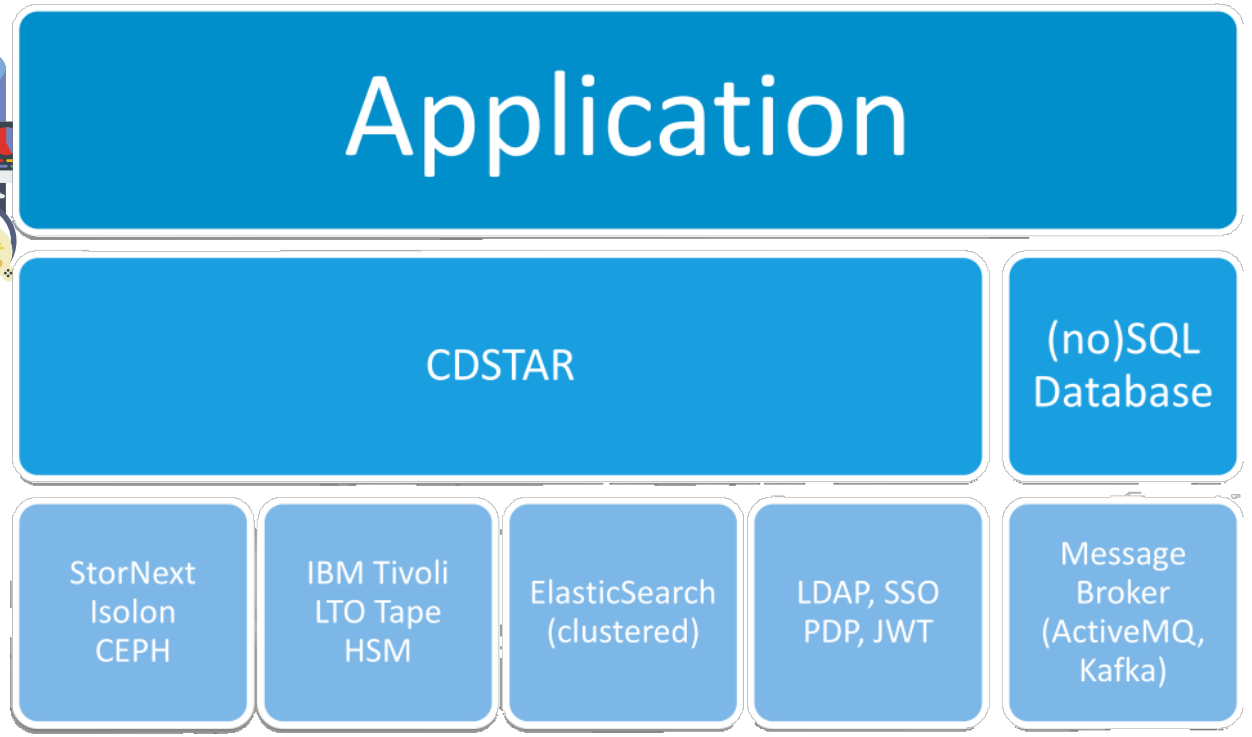
Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

The **Mainzelliste** is used to mask identifiying attributes in medical data



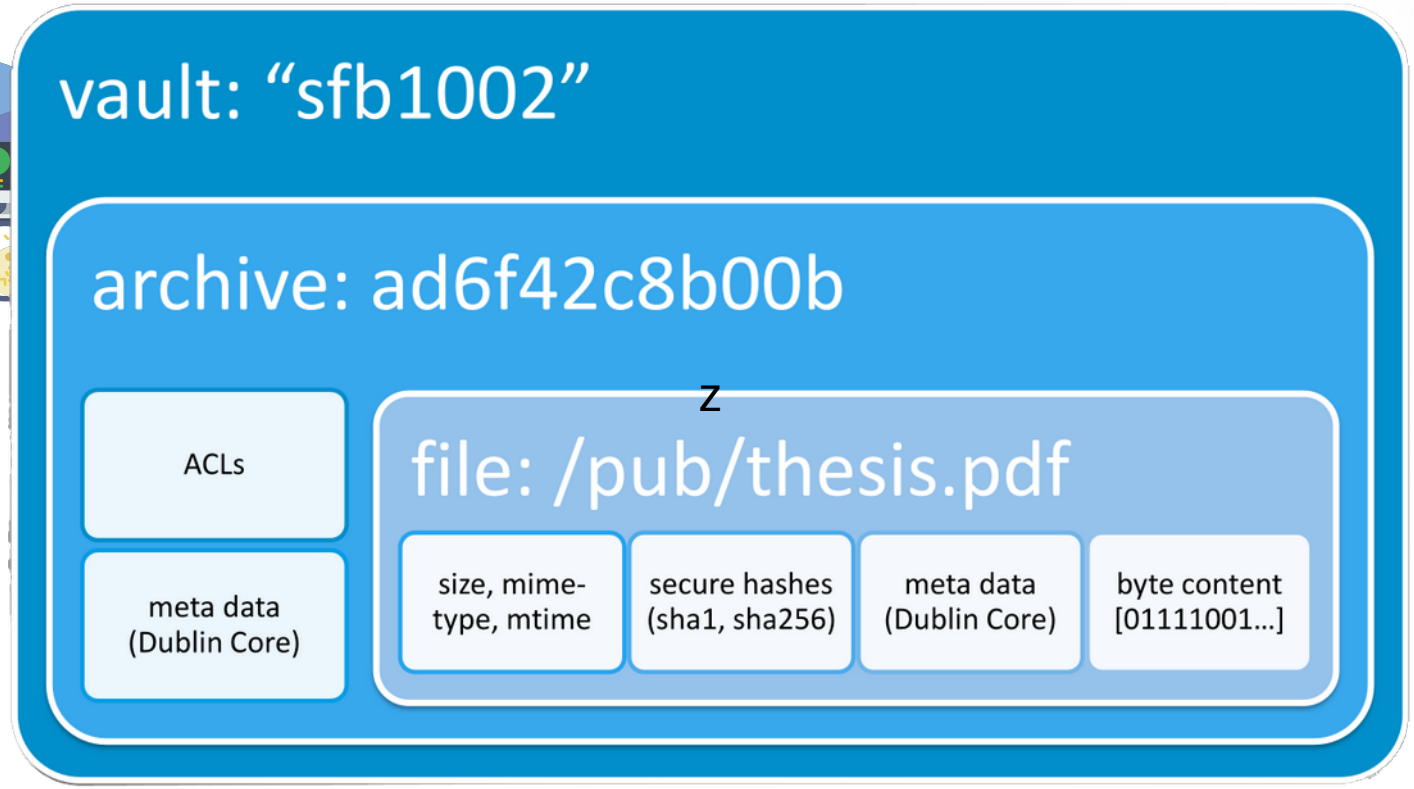
Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

Each data artefact created along the data integration pipeline is stored persistently using **CDSTAR**



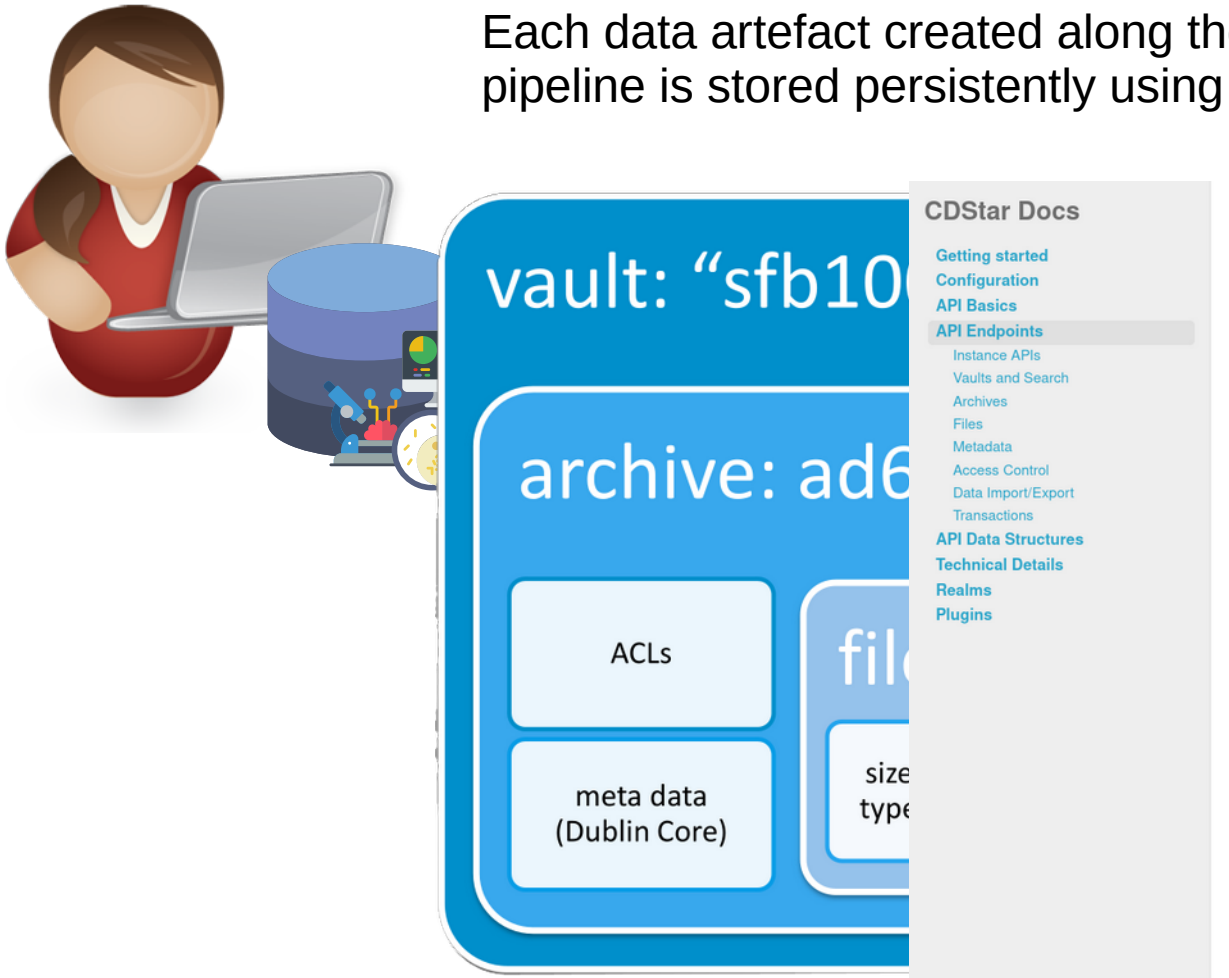
Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

Each data artefact created along the data integration pipeline is stored persistently using **CDSTAR**



Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

Each data artefact created along the data integration pipeline is stored persistently using **CDSTAR**



CDStar Docs

- Getting started
- Configuration
- API Basics
- API Endpoints
  - Instance APIs
  - Vaults and Search
  - Archives
  - Files
  - Metadata
  - Access Control
  - Data Import/Export
  - Transactions
- API Data Structures
- Technical Details
- Realms
- Plugins

API Endpoints

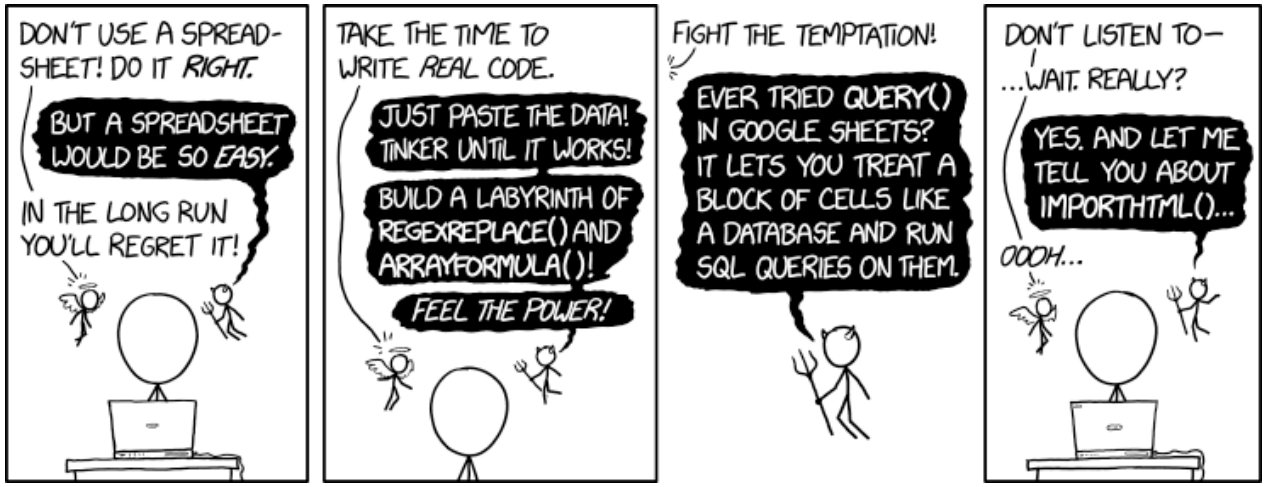
This chapter lists and describes all web service endpoints defined by the standard CDStar HTTP API. Requests are routed to the appropriate endpoint based on their HTTP method, content type and URI path. Some endpoints also require certain query parameters to be present. Path parameters (variable parts of the URL path) are marked with curly brackets.

Table 9. HTTP Endpoints: Overview

Title	Method	URI Path	Consumes
Instance APIs			
Service Info	GET	/v3/	
Service Health	GET	/v3/_health	
Vaults and Search			
List Vaults	GET	/v3/	
Get Vault Info	GET	/v3/{vault}	
Search in Vault	GET	/v3/{vault}?q	
List all Archives in a Vault	GET	/v3/{vault}?scroll	
Archives			
Create Archive	POST	/v3/{vault}/	multipart/form-data application/x-www-form-urlencoded
Get Archive Info	GET	/v3/{vault}/{archive}	
Export Archive	GET	/v3/{vault}/{archive}?export	
Update Archive	POST	/v3/{vault}/{archive}	multipart/form-data application/x-

Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

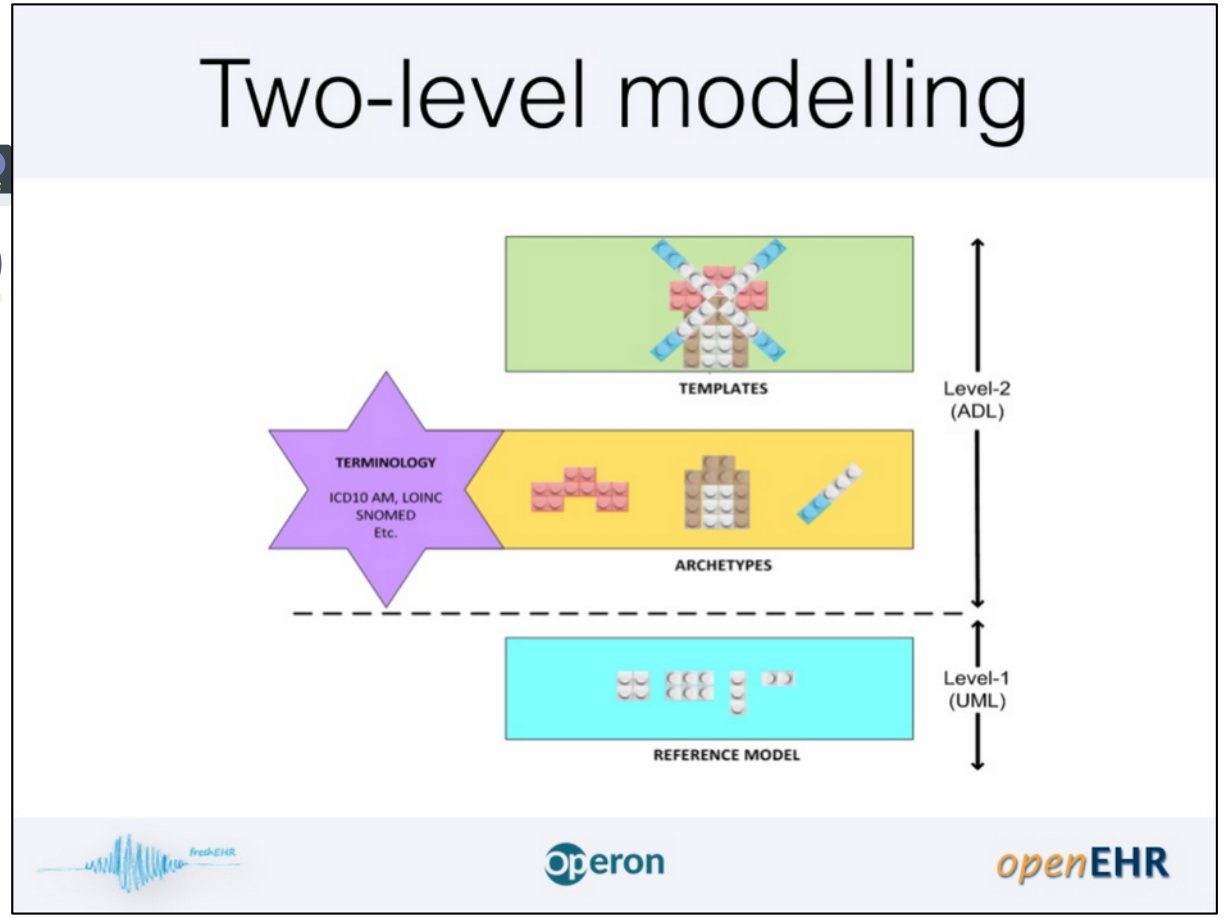
(Speakers awkwardly switch microphone while audience is distracted by a funny cartoon)



https://xkcd.com/2180/

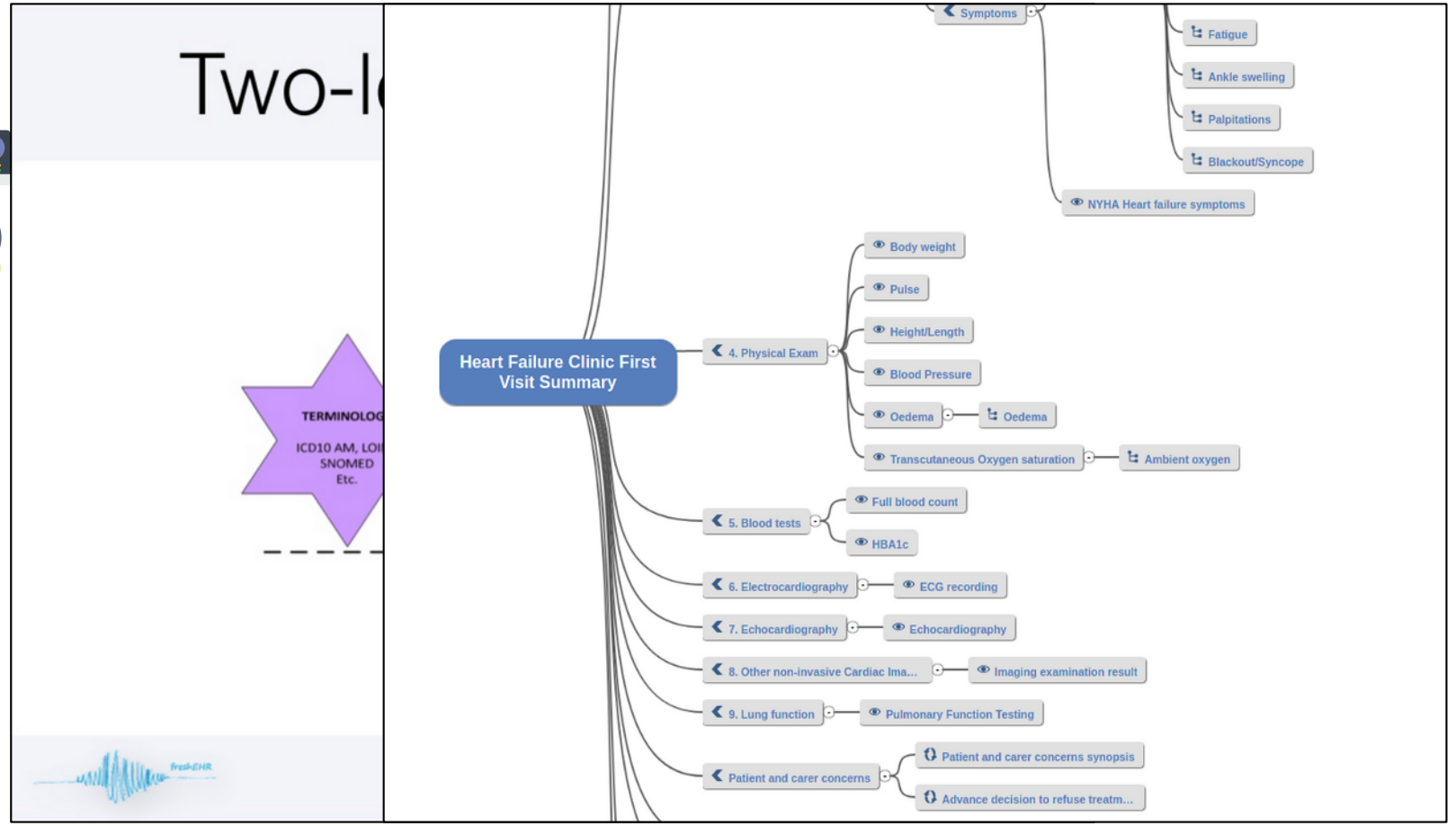
Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

Output from all the data integration pipelines is formatted according to a semantic data model specified using **openEHR**.

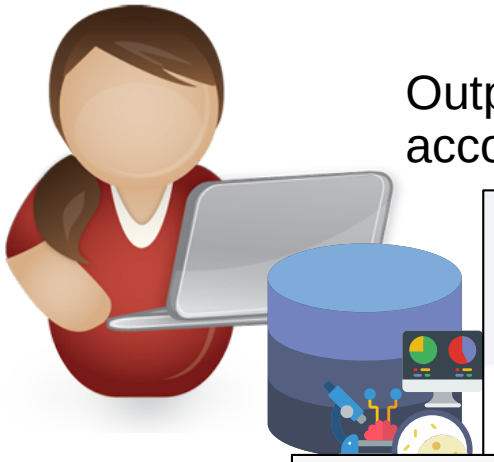


Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

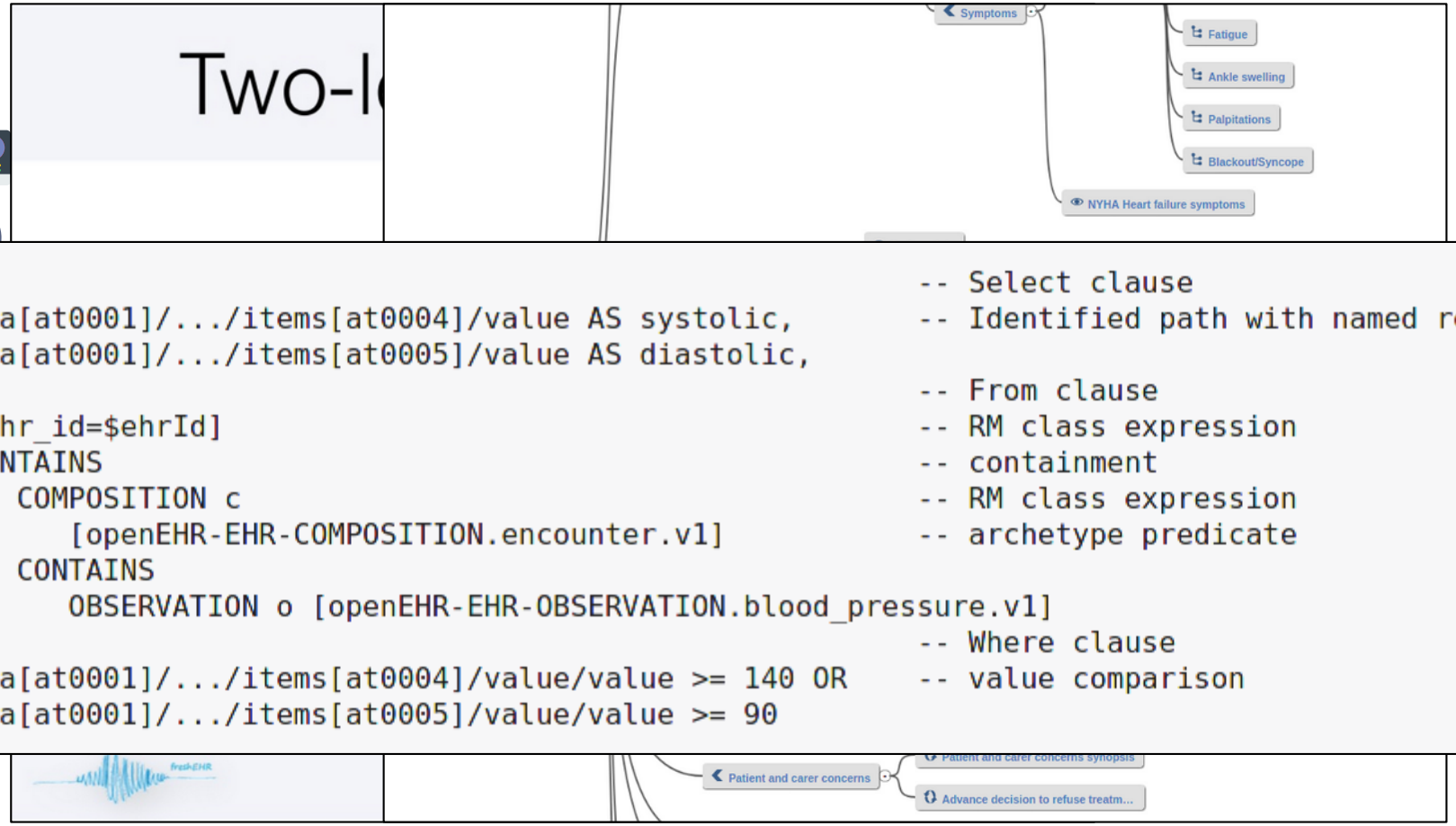
Output from all the data integration pipelines is formatted according to a semantic data model specified using **openEHR**.



Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de



Output from all the data integration pipelines is formatted according to a semantic data model specified using **openEHR**.




```
SELECT                                                    -- Select clause
   o/data[at0001]/.../items[at0004]/value AS systolic,    -- Identified path with named result
   o/data[at0001]/.../items[at0005]/value AS diastolic,
FROM                                                      -- From clause
   EHR[ehr_id=$ehrId]                                     -- RM class expression
      CONTAINS                                            -- containment
         COMPOSITION c                                    -- RM class expression
            [openEHR-EHR-COMPOSITION.encounter.v1]        -- archetype predicate
         CONTAINS
            OBSERVATION o [openEHR-EHR-OBSERVATION.blood_pressure.v1]
WHERE                                                     -- Where clause
   o/data[at0001]/.../items[at0004]/value/value >= 140 OR -- value comparison
   o/data[at0001]/.../items[at0005]/value/value >= 90
```


Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

This is medical researcher Carmen!

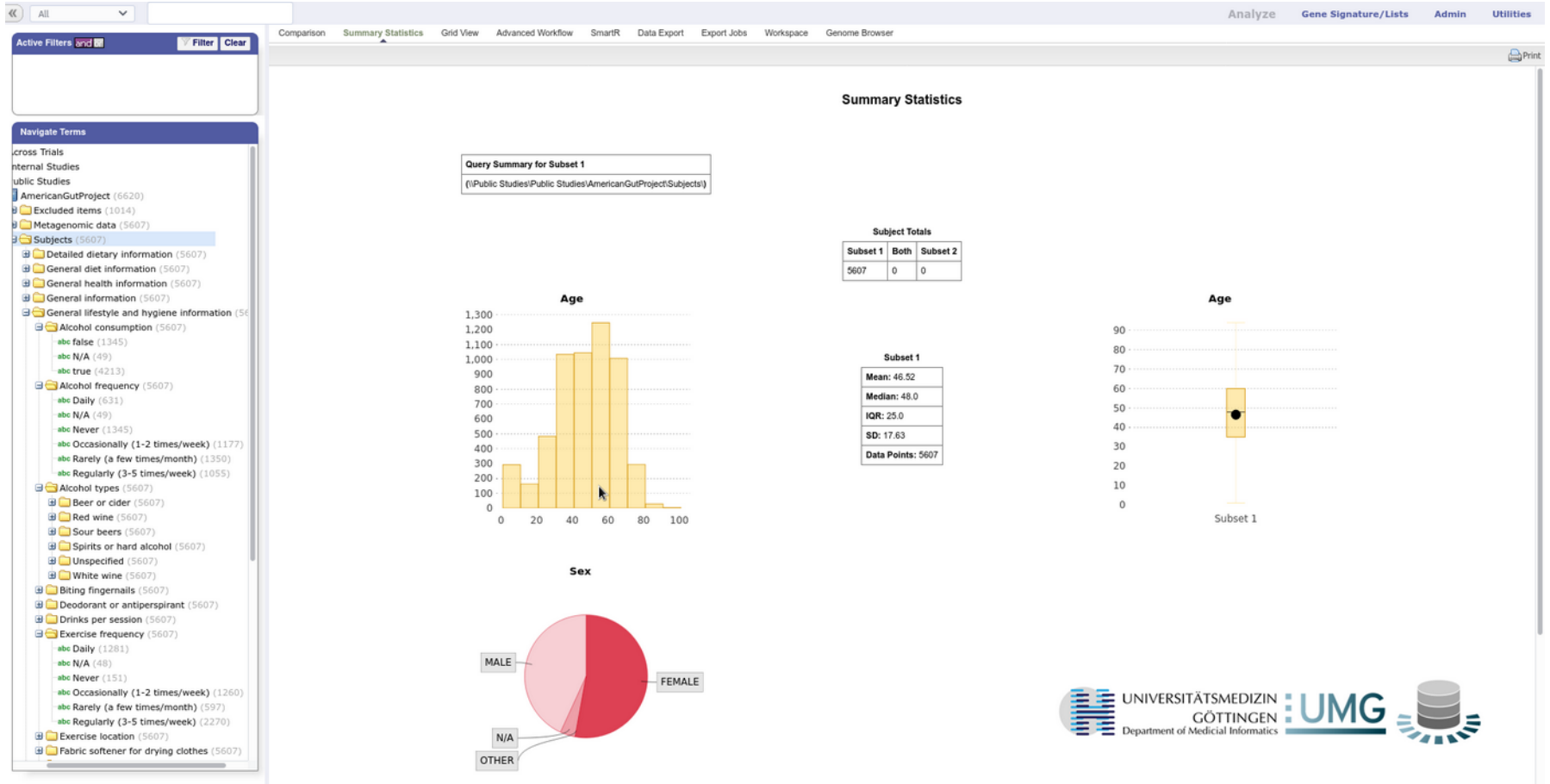
Carmen currently runs a research project on chronic Heart insufficiency

(like Bob‘s...)

Based on openEHR,
Carmen can specify search queries to identify patient cohorts that share certain medical conditions

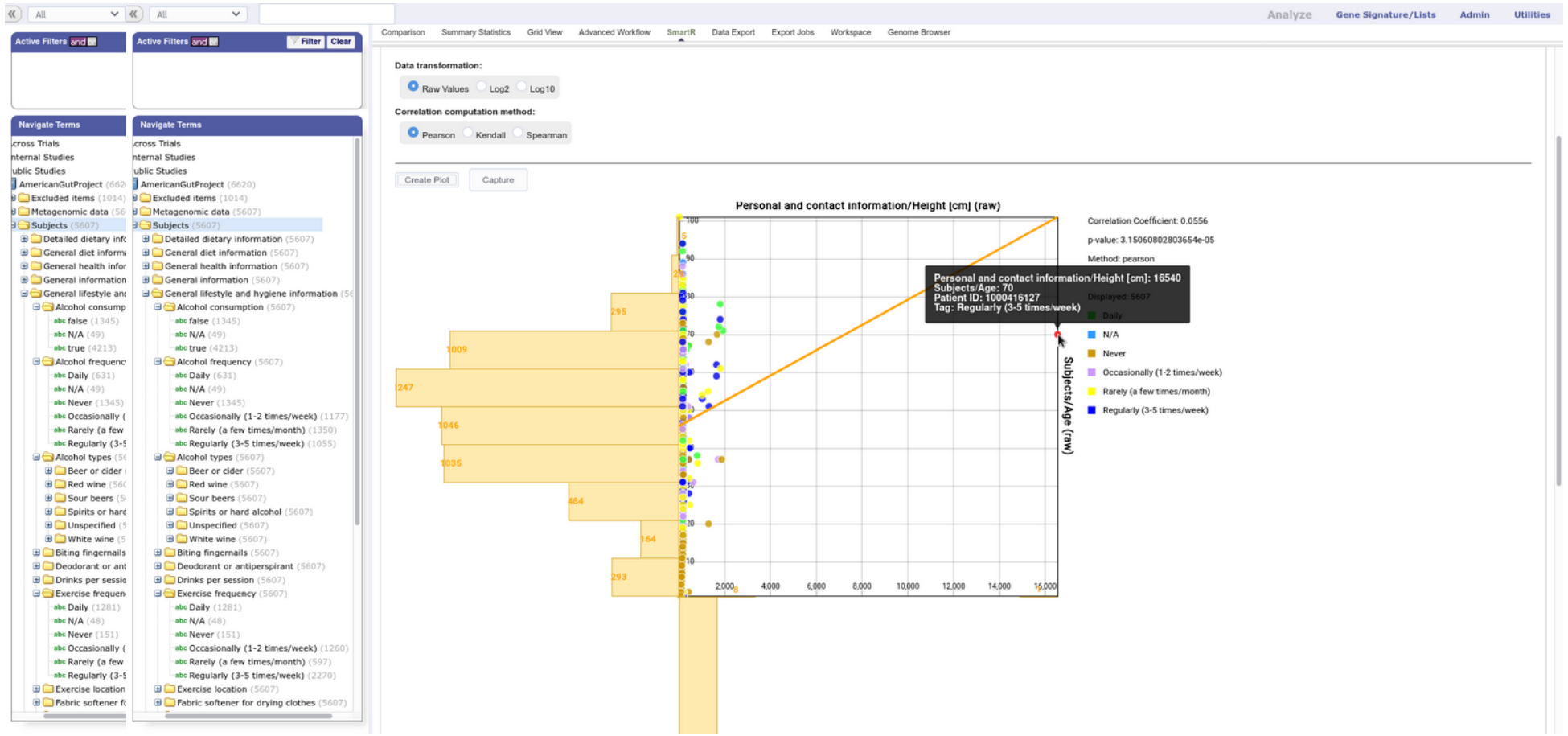
Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

Carmen retrieves a dataset, specifies a heart insufficiency research hypothesis and uses **i2b2 tranSMART** to perform simple analytics.



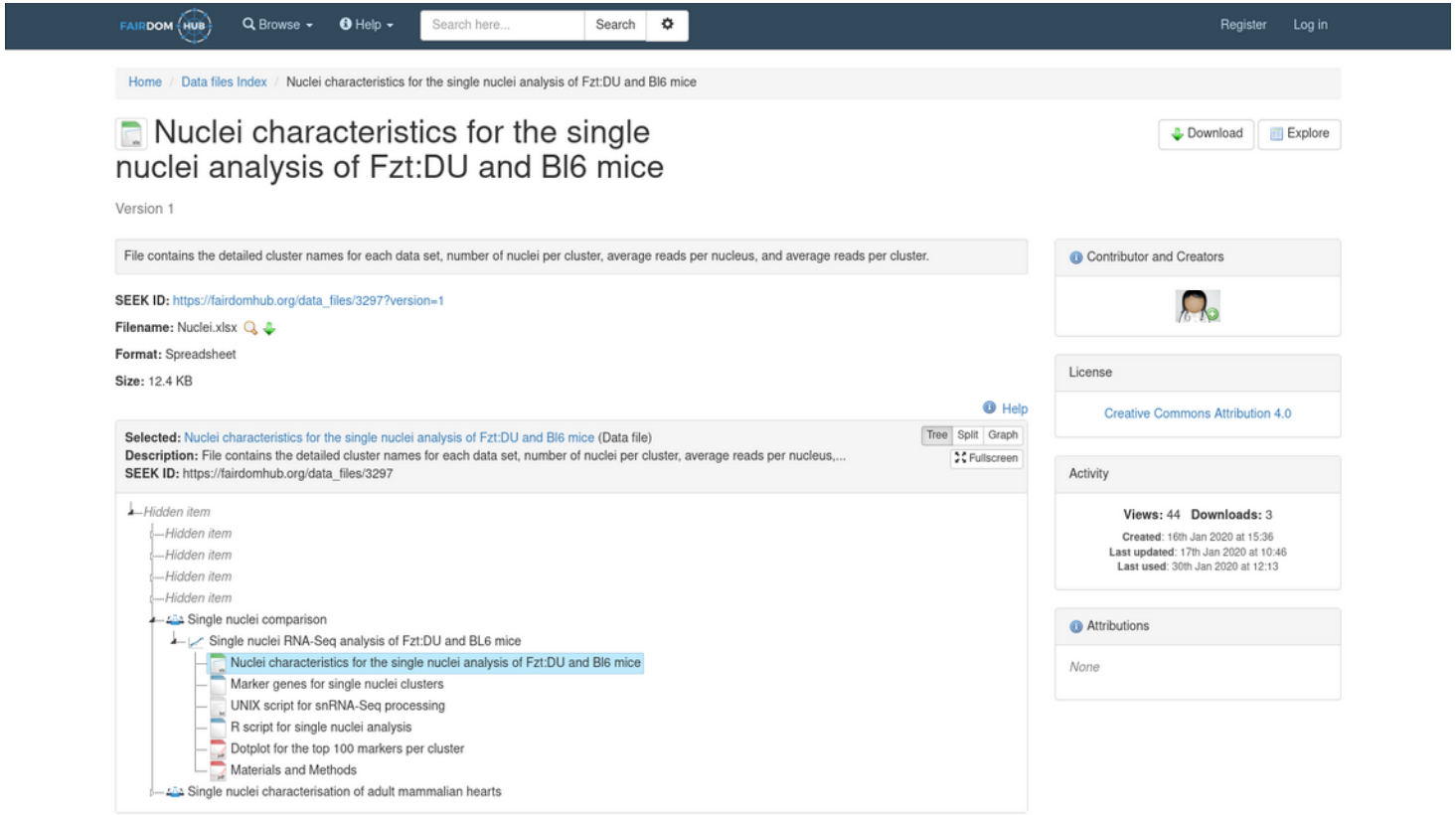
Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

Carmen retrieves a dataset, specifies a heart insufficiency research hypothesis and uses **i2b2 tranSMART** to perform simple analytics.



Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

Carmen publishes her research results in a scientific open access journal.
The original datasets and the experiment setup are documented at an **FAIRdom/SEEK** instance operated at the hospital.



Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

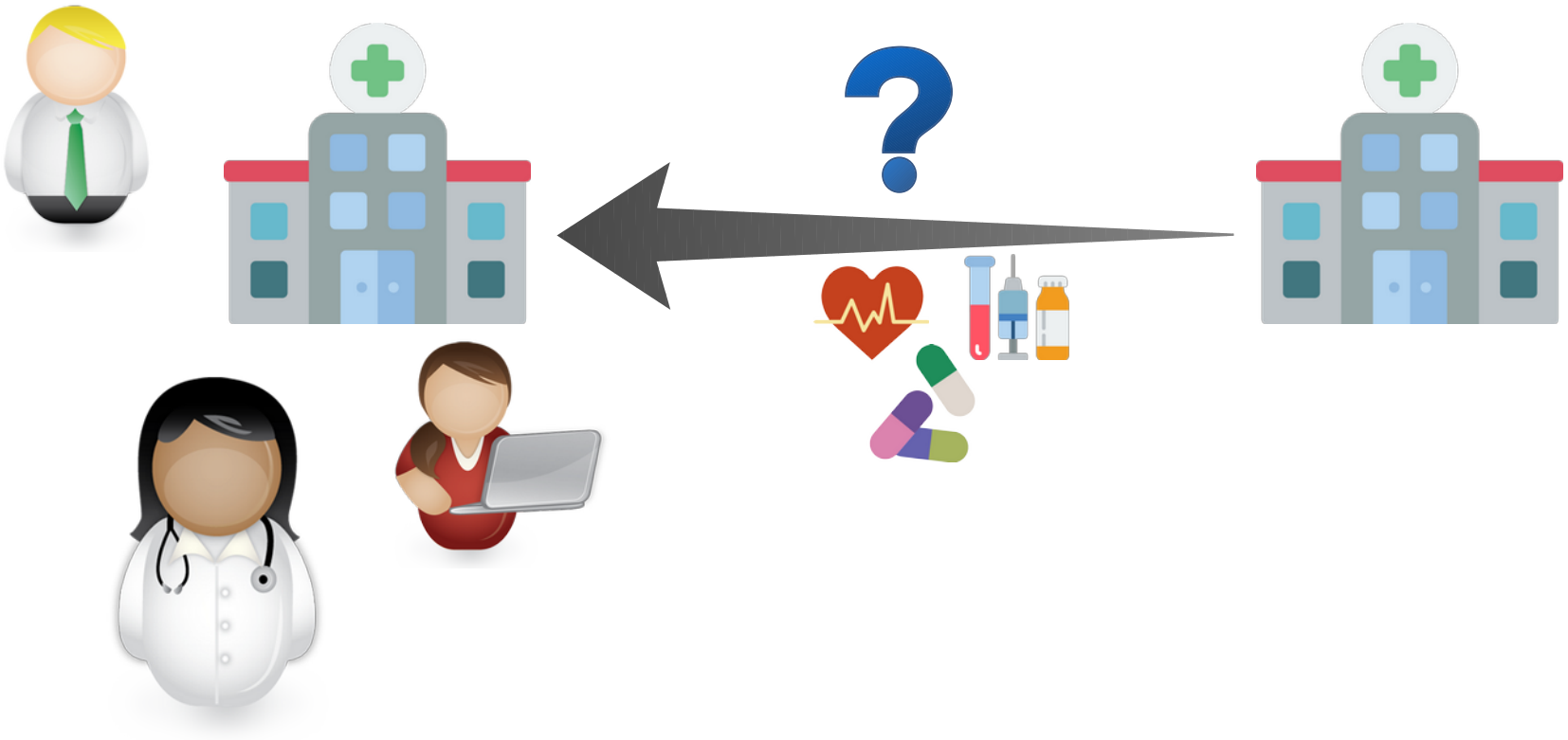
Thanks to all the mentioned software tools

And people involved in data-driven medical research

Treatment for Bob‘s condition can be improved,
allowing him and fellow patients to lead longer and happier lifes

Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

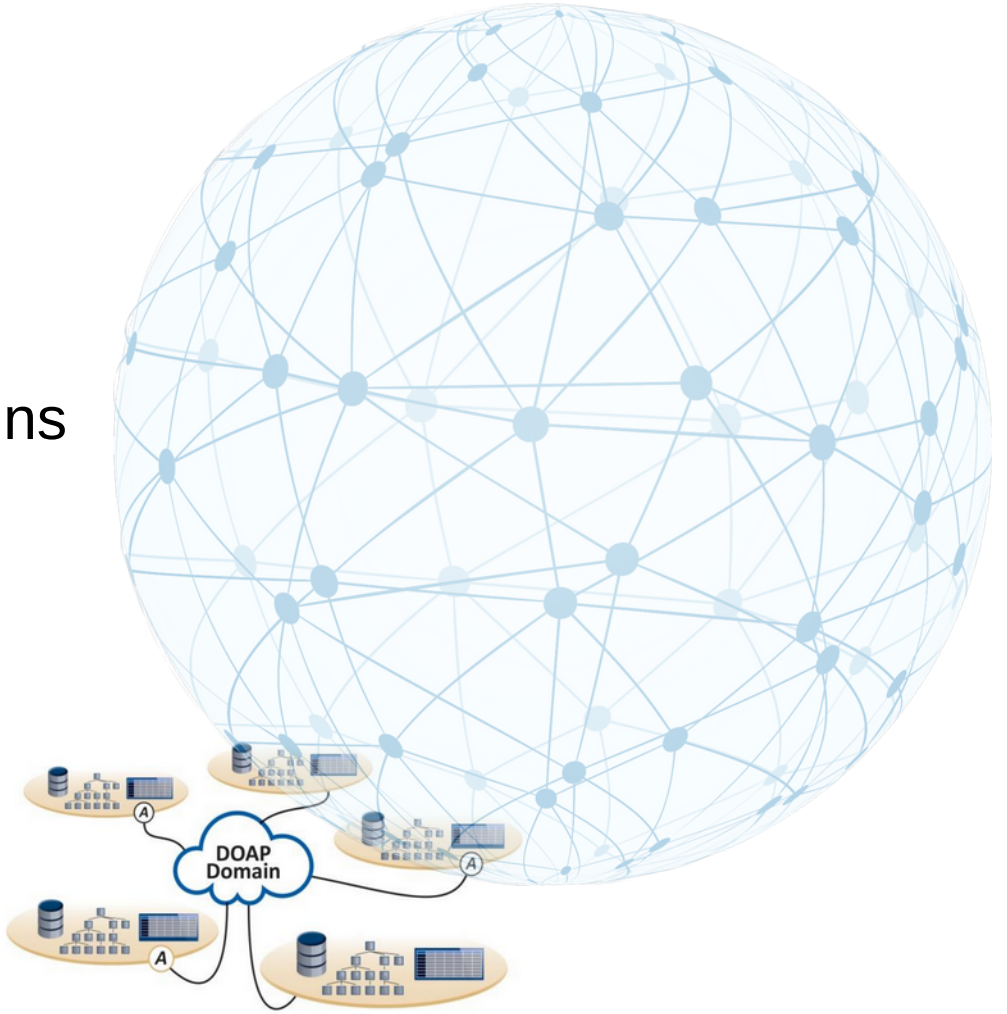
What about data sharing with other hospitals?



Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

Global Research Data Infrastructure

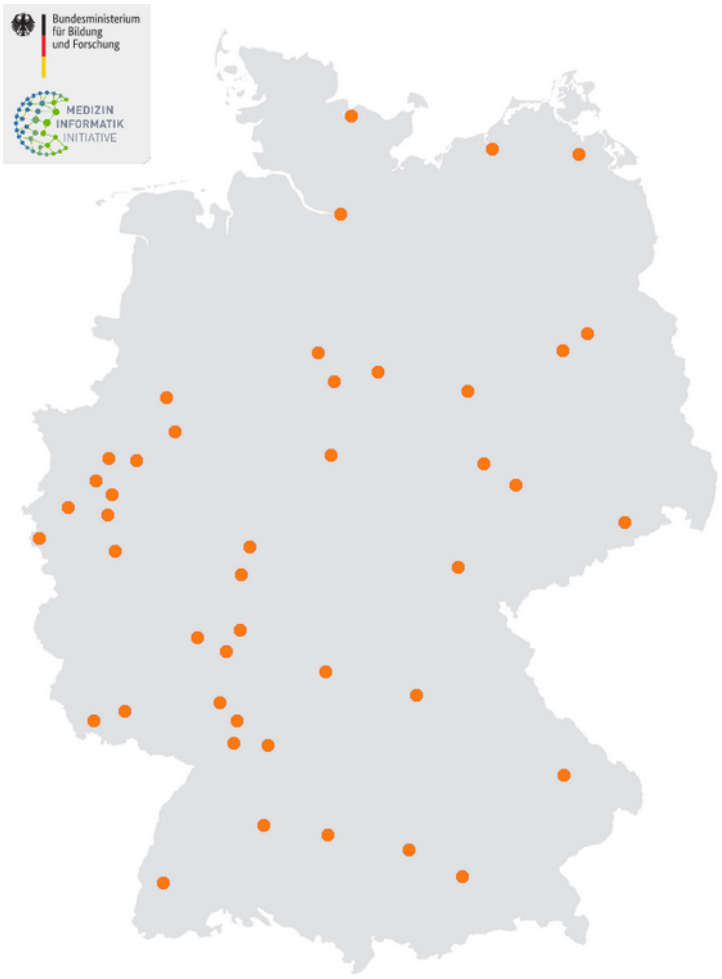
- A new infrastructure is emerging: „Internet of Data Objects“
- Works on the creation of linked data stores and applications are spread across many research domains
- Multiple developments in MI:
  - National scale (Germany): Medical Informatics Initiative
  - International scale: OHDSI, EHDEN
- Cross-domain developments:
  - Global scale: RDA, W3C Data on the Web



DOAP: digital object access protocol, figure from Wittenburg & Strawn, 2018

Secure Health Data Infrastructure

- Sensitive medical data is re-purposed for research application, which bears both: a **high value** for research and **potential for misuse**
- To benefit from linked medical data and improve healthcare with data-driven insights, we need to:
  - Create a **secure** IT-infrastructure
  - Enable **accountable** and **transparent** dataflows
  - **Empower** the patient



Sites of the Medical Informatics Initiative in Germany.

Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

Do we need a
political campaign for
free and decentralized
software
in the healthcare domain?

Public Money

Public Code

publiccode.eu

Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

MOTD

- Global data infrastructures for medical research are being built
- Decentralized and free technologies lead to secure IT-infrastructures
- Typically, medical information systems are not FOSS territory
- FOSS tools are available and frequently used in Medical Informatics research
- Voice your opinion on free software in healthcare!

Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

References

Wittenburg P, Strawn G. Common Patterns in Revolutionary Infrastructures and Data 2018. https://doi.org/10.23728/b2share.4e8ac36c0dd343da81fd9e83e72805a0

Image sources:
Lego bricks: homero chaper, stockvault.net
Spotlights: Designed by upklyak / Freepik
XNAT image: https://xnat.org and Dagmar Krefting
Mainzelliste images: Created by Florian Stampe and Galina Tremper
CDSTAR schema: Created by Marcel Hellkamp, cdstar.gwdg.de
openEHR images: Created by Ian McNicoll, https://ckm.openehr.org and https://specifications.openehr.org
FAIRDOM/Seek screenshot: https://fairdomhub.org/data_files/3297
Network globe: Designed by macrovector_official / Freepik
Map of Germany: https://medizininformatik-karte.de/
Wooden signpost: Designed by Freepik
Public Money, Public Code: CC-BY-SA 4.0 Free Software Foundation Europe
Bob, Alice, Carmen and some arrows: taken from the LibreOffice Gallery (yes, we‘re that cheap)
The rest of the icons: made by Smashicons from www.flaticon.com

Acknowledgements

GWDG
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

DZHK
DEUTSCHES ZENTRUM FÜR HERZ-KREISLAUFFORSCHUNG E.V.

HiGHmed
Medical Informatics

SPONSORED BY THE
Federal Ministry of Education and Research

Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

Interested in medical informatics?

Contacts us, we hire

mi.umg.eu

(website will relaunch very soon, don‘t judge us based on current design)

http://mi.umg.eu

– end of live presentation –

Additional slides

Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

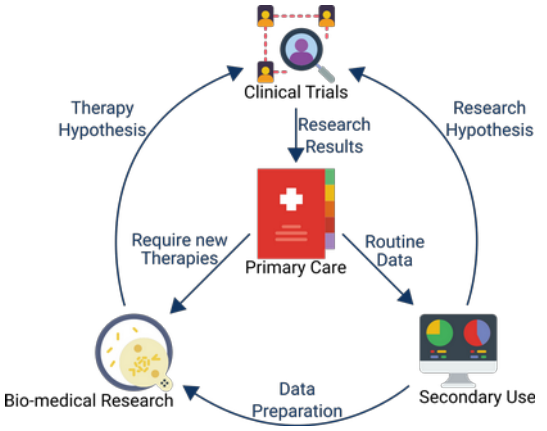
Tool Summaries

Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

Software Name

Name and short description

Med Inf research field indicator



This is a quick summary of what this tool does

- **aim**: a few more details on what should be achieved...
- **usage**: a few more details on how this tool can be used...

Sofware description template

- **developed since**: 2011
- **bus factor**: nr of core developers (community size)

Software license

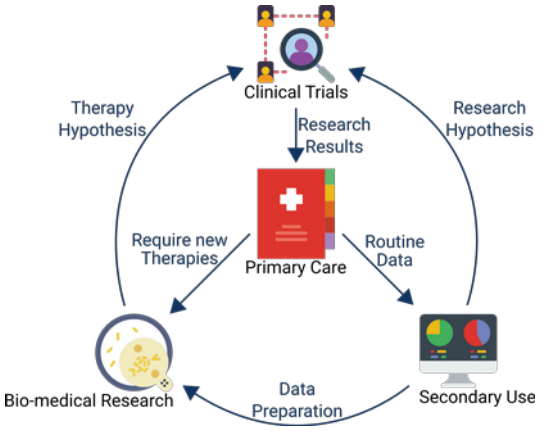
SPDX short identifier

Link(s)

https://weblink.to

Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

XNAT (Extensible Neuroimaging Archive Toolkit)



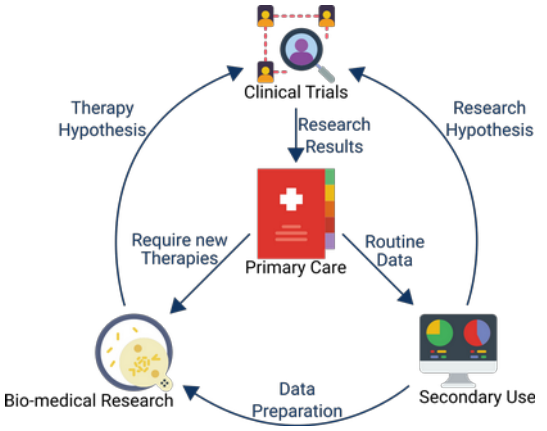
The leading open source medical imaging platform.

- aim: „*XNAT's core functions manage importing, archiving, processing and securely distributing imaging and related study data*“
- usage: Upload, manage, share (medical) images; integrated processing pipelines, analysis scripts, etc.

XNAT Software License (~MIT)

https://xnat.org
https://bitbucket.org/xnatdev/

Talend Open Studio for Data Integration



An enterprise-grade tool to model, create and run extract, transform and load processes.

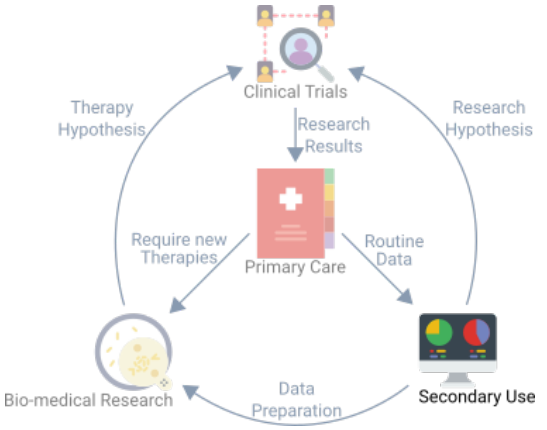
- aim: creation of ETL-processes to integrate heterogeneous data into desired formats and schemas
- usage: extract data from heterogeneous sources, transform them into common data formats and load them into research tools for Secondary Use
- developed since: 2005
- bus factor: Talend SA, 1,200+ employees

Apache-2.0

https://talend.com

Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

Mainzelliste



An application that allows to mask patient data and store a patient-to-mask mapping.

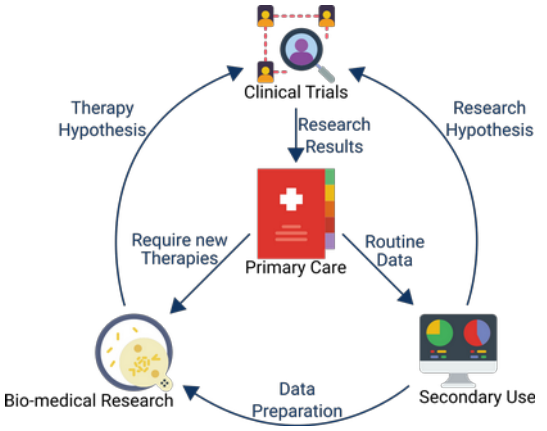
- aim: separate identifying from medical data but securely keep a link between both as well as match similar patients based on identifiying data
- usage: satisfy legal constraints (e.g. GDPR) when using data from primary care in research
- developed since: 2013
- bus factor: 3 maintainers + 15 contributors

AGPL-3.0

http://mainzelliste.de

Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

CDSTAR



Package-oriented storage and data archive middleware.

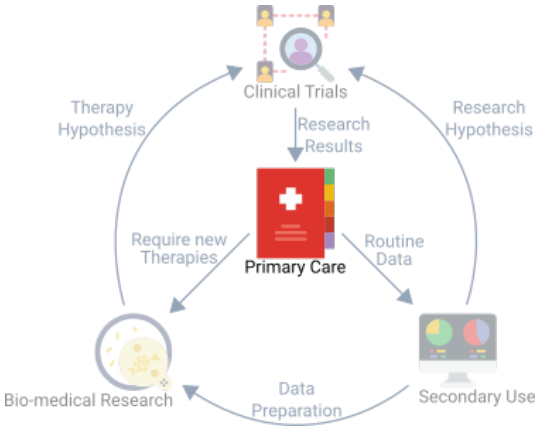
- aim: store multiple files as a package with metadata via webservice communication
- usage: utilize as storage back-end similar to object storage services; transparently handles tiered storage configuration
- developed since: 2016
- bus factor: 1 developer + GWDG professional support

Apache-2.0

https://gitlab.gwdg.de/cdstar

Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

openEHR



A technology to model, capture, store and query electronic health records.

- aim: define means to store medical data in re-usable and vendor-independent data models
- usage: storage of patient data in primary care (electronic health record)
- developed since: 2000
- bus factor: openEHR foundation with several industrial and academic partners

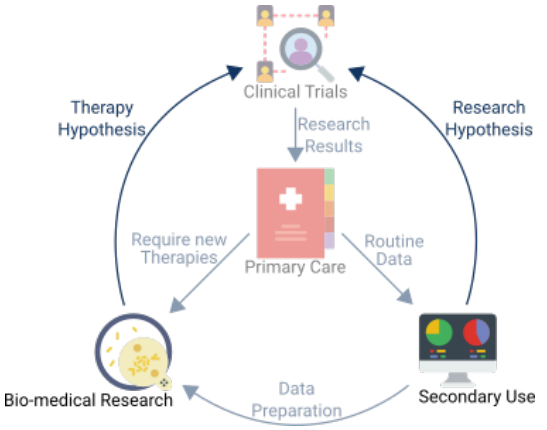
Apache-2.0

specifications: CC licenses

https://openehr.org
https://discourse.openehr.org

Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

i2b2 tranSMART



A modular, web-based data warehouse solution for clinical data for exploration and analysis of medical datasets.

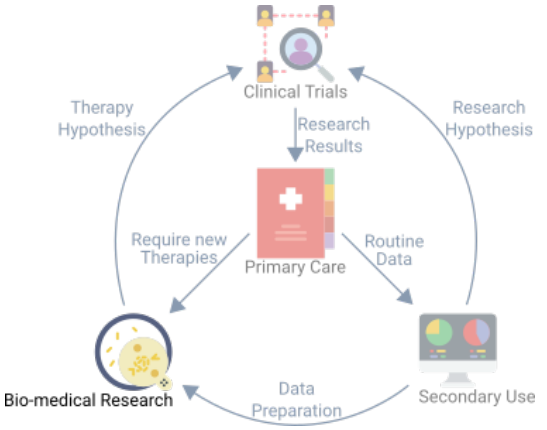
- aim: provide a ontology-driven storage and data analytics platform for clinical datasets
- usage: exploration of patient cohorts and simple analytics in Secondary Use
- developed since: 2004
- bus factor: i2b2 tranSMART foundation with four sustaining sponsors

i2b2: MPL-2.0
tranSMART: GPL-3.0

https://i2b2transmart.net
https://github.com/i2b2-tranSMART

Marcel Parciak, Markus Suhr {marcel.parciak|markus.suhr}@med.uni-goettingen.de

FAIRdom/SEEK



An „*open source web platform for sharing scientific research assets, processes and outcomes*“.

- aim: facilitate reproducible documentation of biomedical experiments by transparently enforcing ISA ontology.
- usage: document projects, investigations, experiments, samples, files, scripts, publications and their relation; available as open platform at https://fairdomhub.org or for self-hosting
- developed since: 2011
- bus factor: 4 active maintainers, FAIRDOM Association with 11 funding partners

BSD-3-Clause

https://seek4science.org
https://github.com/seek4science