



ETL-Prozesse und FHIR im MIRACUM-Projekt

**Dr. Detlef Kraska, Universitätsklinikum Erlangen,
Medizinischen IK-Zentrum (MIK)**

09.09.2019

GMDS Workshop 39

Rahmenbedingungen für die Architektur der Datenintegrationszentren in MIRACUM

- Unterschiedliche organisatorische Voraussetzungen an den Standorten
- Unterschiedliche datenschutzrechtliche Vorgaben
- Verteilte Entwicklung, getrieben durch Competence Center
- Einsatz von Open Source
- Integration mit Fokus auf ETL-Strecken
- Ziel-Repositories i2b2/tranSMART, OMOP (OHDSI) und XNAT

2018 war es noch nicht möglich, die Architektur der MIRACUM Datenintegrationszentren komplett bis Ende 2021 zu definieren

- Entscheidung, mittels einer Phasenplanung immer nur die nächsten 6 bis 9 Monate feiner zu planen
- die Deliverables 2018 zielten vor allem darauf ab, die Arbeiten der MI-I Konzeptphase an alle Standorten produktiv zu nehmen
- 2019 werden die Komponenten ausgeliefert, die die Standorte brauchen, um den Kerndatensatz und die Daten für die ersten Use Cases erschließen zu können inkl. der Integration eines Pseudonymisierungsdienstes

2018: §21-Daten in 2 Repositories

Auslieferung von 2 vmWare-Containern

1. ETL-Job auf Basis Talend zum Einlesen der §21-Daten in i2b2, dort Harmonisierung und Pseudonymisierung
2. ELT-Job auf Basis Talend zum Einlesen der §21-Daten in OMOP, dort Harmonisierung und Pseudonymisierung

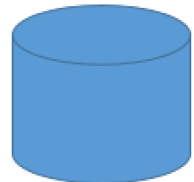
-> Entwicklung an nur wenigen Standorten

-> Schnelles Ausrollen möglich

- Auslieferung von Docker-Containern statt kompletten vmWares
- Ein zentrales Gitlab als Versionsverwaltungssystem
- Einsatz von Continuous Integration (CI) zur Sicherstellung der Softwarequalität
- Umstieg von Talend auf Pentaho als ETL-Tool
- Eine zentrale Docker Registry für die freigegebenen Container
- Modularisierung der Architektur

Etablierung von Sicherheitszonen

Netzsegment Klinik



Klin. Warehouse
oder klin. System

Netzsegment H2R (HospitalToResearch)

Netzsegment Forschung



i2b2



OMOP

Netzsegment Treuhandstelle



Treuhandstelle

MITGLIED DER



GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

FKZ 01ZZ1801A

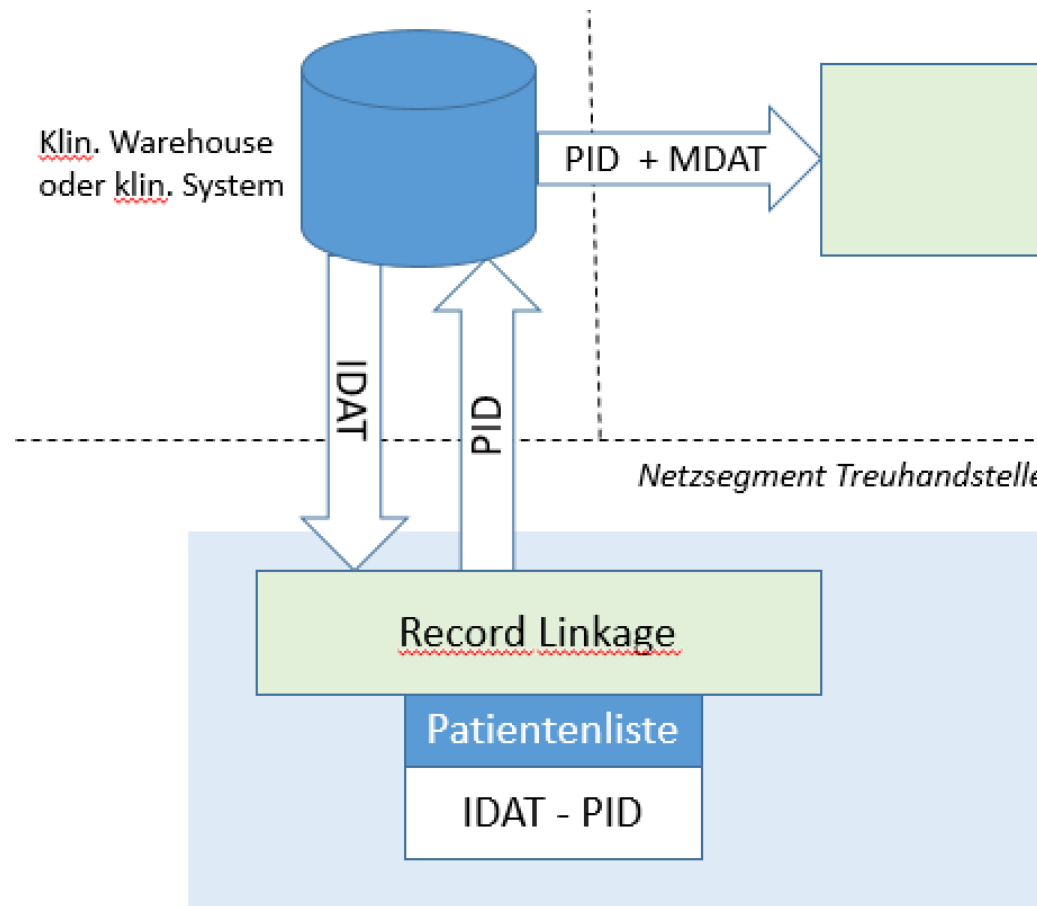


FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

Universitätsklinikum
Erlangen



Optionen für die Anbindung der Treuhandstelle



MITGLIED DER



GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

FKZ 01ZZ1801A

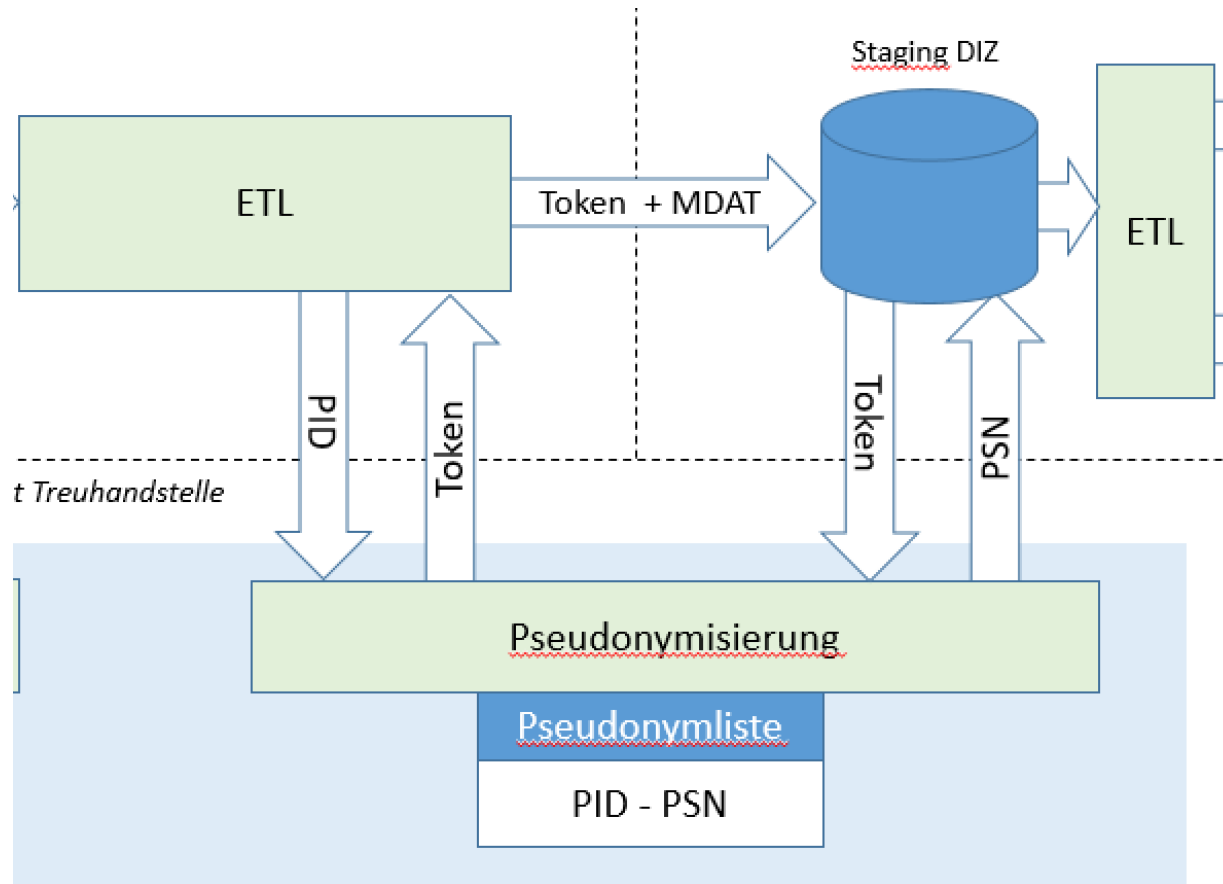


FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

Universitätsklinikum
Erlangen



Optionen für die Anbindung der Treuhandstelle



MITGLIED DER



GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung
FKZ 01ZZ1801A

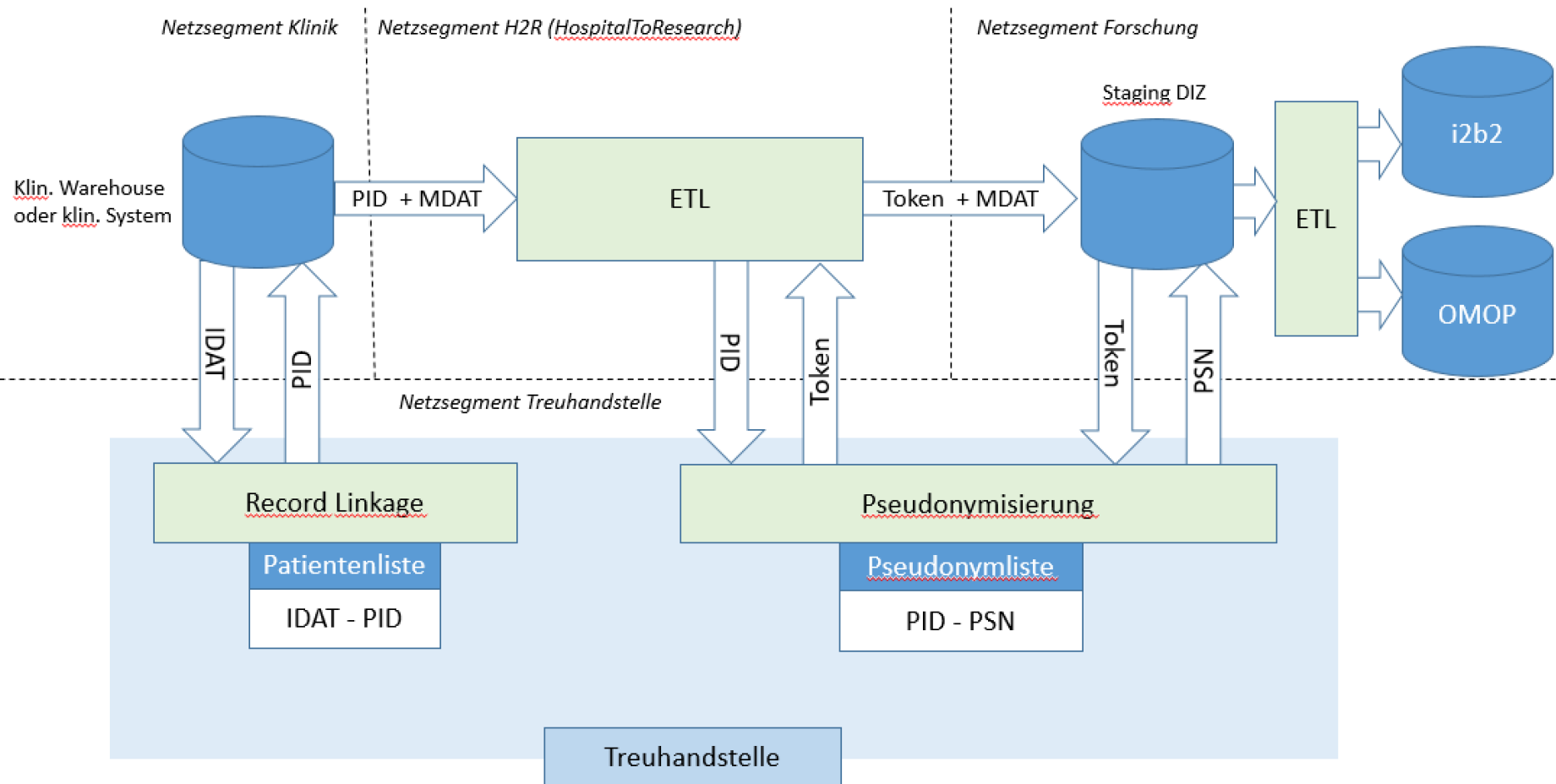


FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

Universitätsklinikum
Erlangen



Optionen für die Anbindung der Treuhandstelle



MITGLIED DER



GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung
FKZ 01ZZ1801A



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

Universitätsklinikum
Erlangen



Die Integration von Quellsystemen an 10 Standorten und das Füllen von mind. 4 Ziel-Repositories macht den Einsatz eines zentralen Zwischenformats sinnvoll / notwendig

- Aus Sicht vom ETL-Design wäre eine zentrale Datenbank am geeignetsten mit Staging-Bereich und Tabellen für das Zwischenformat
- Die Standorte füllen aus ihren Quellsystemen das Zwischenformat, die Competence Center für die Repositories leiten daraus die Daten ab
- Aber: wer definiert das Zwischenformat, woran orientiert man sich?

Die NSG AG Interop hat inzwischen den Einsatz von FHIR für die Definition des Kerndatensatzes empfohlen

- In den nächsten Monaten entstehen Implementierungsleitfäden für administrative Patientendaten, Labor und Medikation
- Wenn wir FHIR als unser Zwischenformat nehmen, bekommen wir damit die Dokumentation „geschenkt“
- Aber: wie sieht die technische Implementierung aus?

- ETL funktioniert auf Datenbanken gut und kann dann auch große Datenmengen in vertretbarer Zeit verarbeiten
- Es gibt an den Standorten Daten, bei denen wir mehrere 100 Millionen Ressourcen vorliegen haben (Labor)
- FHIR war ursprünglich für die Verarbeitung von Daten einzelner Patienten gedacht, das skaliert über die klassischen REST-Schnittstellen nicht gut

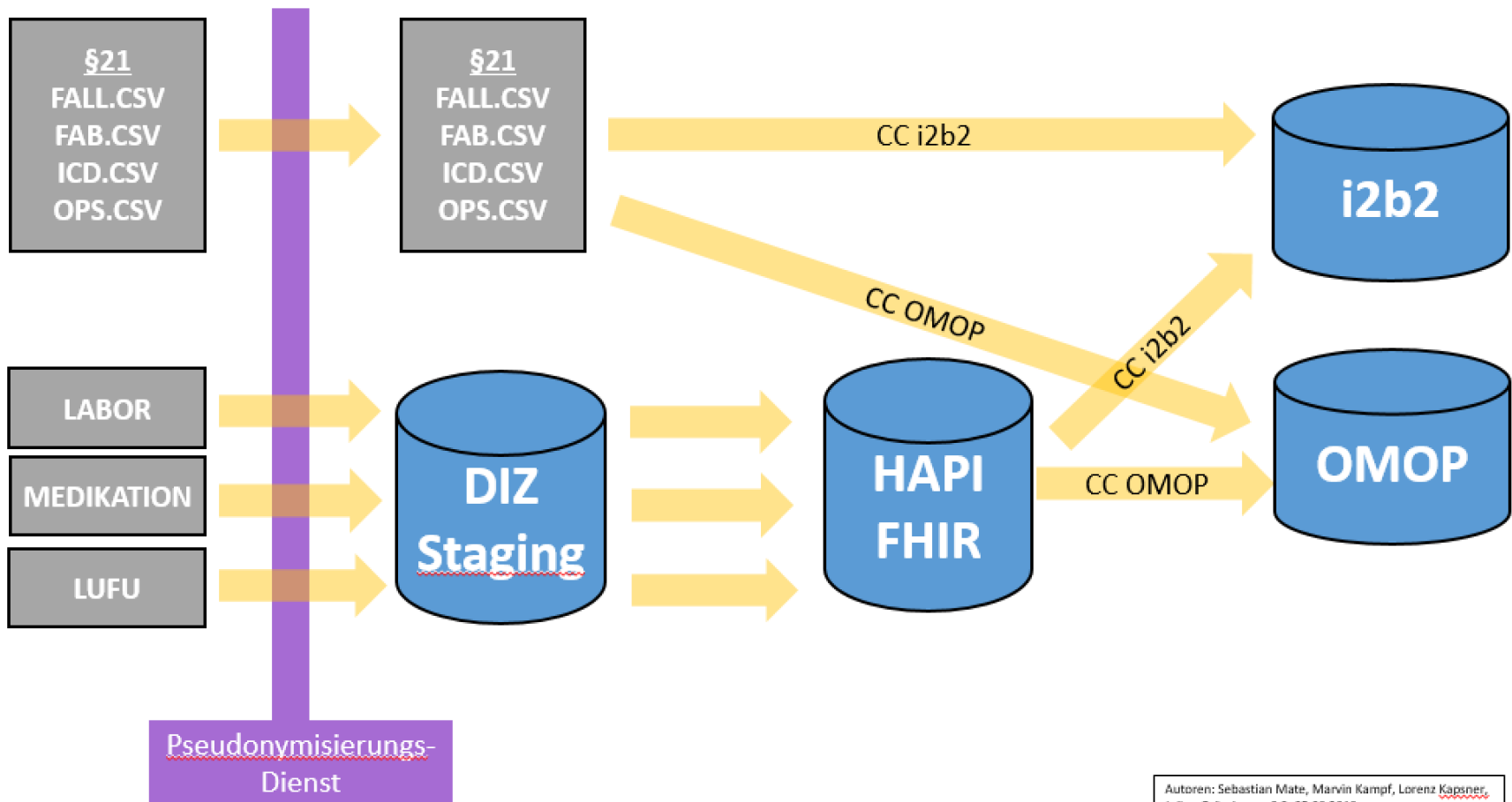
Alternativen

- ETL mit einem Zwischenformat in Form von relationalen Tabellen, die sich an den FHIR-Ressourcen orientieren
- ETL, das FHIR-Ressourcen in JSON in relationalen Tabellen als Zwischenformat ablegt¹
- ETL, das mittels FHIR Bundle oder FHIR Bulk ein FHIR Repository als „Zwischenformat“ nutzt

-> Das alles in den Varianten „komplette Beladung“ oder „Delta-Updates“

¹ Das ist das Prinzip von fhirbase, s. <https://www.health-samurai.io/fhirbase>

Arbeiten am aktuellen Release



Autoren: Sebastian Mate, Marvin Kampf, Lorenz Kapsner,
Julian Gründner, v0.3, 05.09.2019

MITGLIED DER



GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

FKZ 01ZZ1801A



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

Universitätsklinikum
Erlangen



Integriert in die Umgebung ist das MIRACUM DQA-Tool in den Varianten für i2b2 und OMOP, näheres hierzu in der Veranstaltung

VS 13: Datenintegration, Datenmodelle & Datennutzung

MI 11.09.2019, 13:00 – 14:30 Uhr, Raum 189, Emil-Figge-Str. 44

“Moving Towards an EHR Data Quality Framework: The MIRACUM Approach”

Kapsner et. al.

„Erweiterung des MIRACUM Datenqualitäts-Frameworks für das OMOP Datenmodell“

Bathelt et. al.

MITGLIED DER



GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

FKZ 01ZZ1801A



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

Universitätsklinikum
Erlangen



Vielen Dank!



Kontakt: detlef.kraska@uk-erlangen.de



MITGLIED DER



GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

FKZ 01ZZ1801A



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

Universitätsklinikum
Erlangen

