Data Sharing als integraler Teil des Forschungsdatenmanagements

Toralf Kirsten

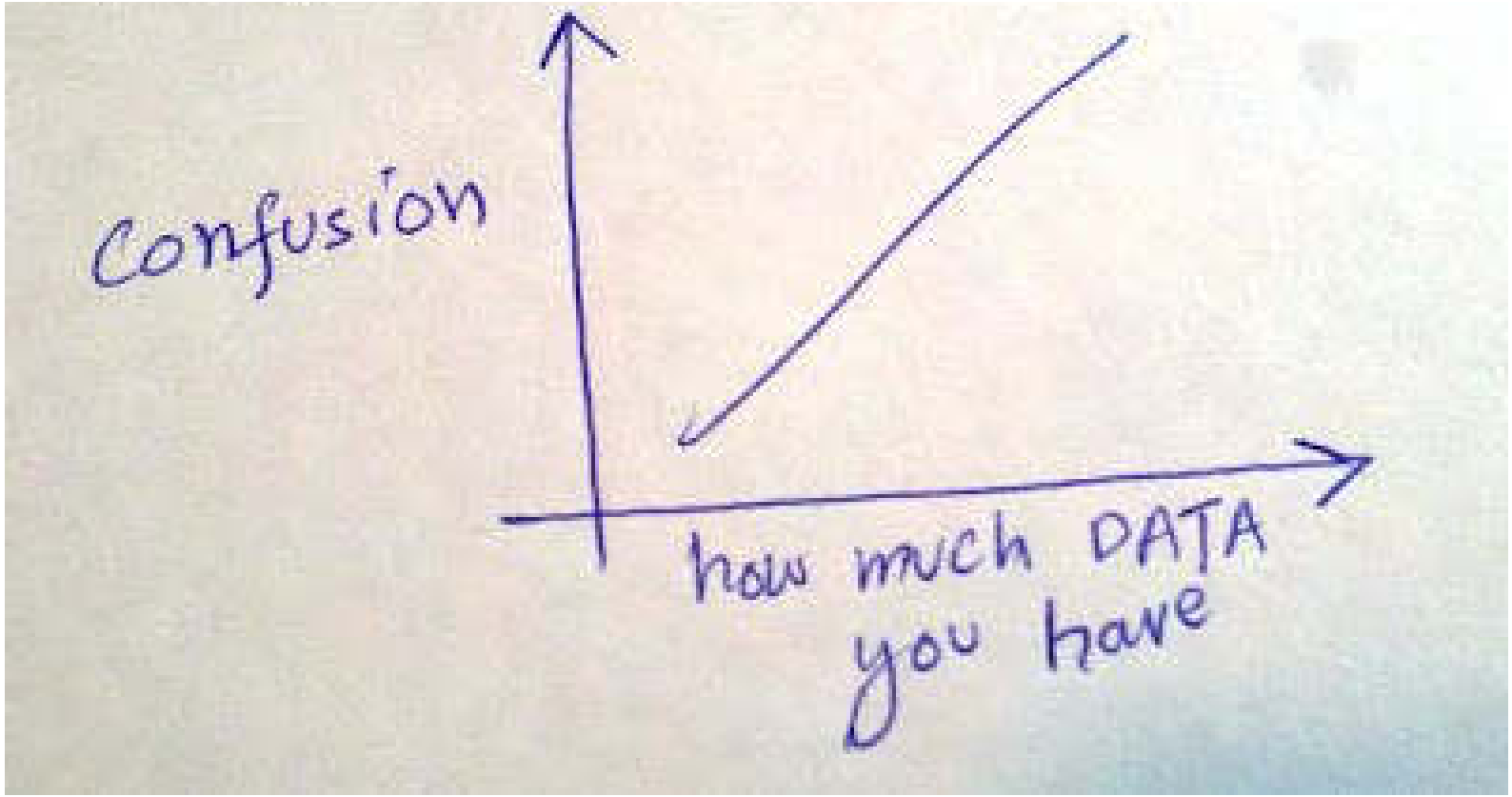
LIFE Leipziger Forschungszentrum für Zivilisationserkrankungen

UNIVERSITÄT LEIPZIG
Medizinische Fakultät

HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

SaxFDM, Dresden, 19.09.2019

Data-Confusion-Principle

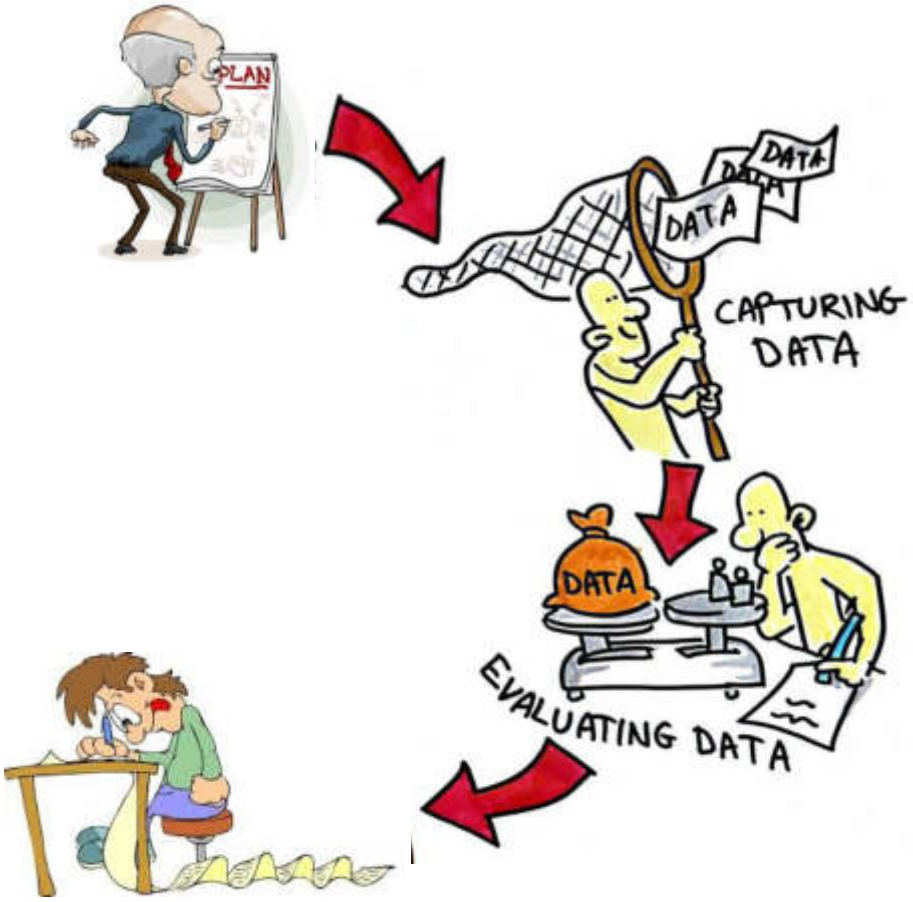


Quelle: Hans Binder, Universität Leipzig

Gründe für die Konfusion

- Datenvolumen
- Heterogenität der Daten(-quellen)
  - Verschiedene Datengranularitäten
  - Unterschiedliche Formate, Strukturen (Schemas) und Semantik
- Keine umfassenden Metadaten
- Unklare Datenherkunft (Data Provenance)
- Keine Dokumentation zu Generierung und Verarbeitung verfügbar
- ...

Traditioneller Weg



- Planung des Experiments
  - Hoch spezialisiert
  - Fokus auf zu untersuchende Fragestellung
- Datenerzeugung (Feldarbeit)
- (Datenkuration)
- Analyse
- Publikation der Ergebnisse

Aber was passiert, wenn

- Jemand wissen möchte
  - Wie sind die Daten erzeugt worden?
  - Was bedeuten die einzelnen Datenelemente (Fragen, Parameter, …)?
  - Wie sind die Daten geprüft / kuratiert worden?
  - Wie sind die Daten analysiert worden (Skript?)?

Aber was passiert, wenn

- Jemand wissen möchte
  - Wie sind die Daten erzeugt worden?
  - Was bedeuten die einzelnen Datenelemente (Fragen, Parameter, …)?
  - Wie sind die Daten geprüft / kuratiert worden?
  - Wie sind die Daten analysiert worden (Skript?)?
- Projekt beendet ist, keine Förderung mehr vorhanden
  - Wo kann jemand Einblick in die Daten gewinnen?
  - Kann ein Doktorand (MA, BA, …) die Daten für weiterführende Auswertungen verwenden?

Ebenen des Data Sharing

- Daten im Verbleib des Datenproduzenten
- Zugriff auf Daten in der Arbeitsgruppe
- Datenzugriff in der Abteilung / Klinikum
- Weitergabe unter Nutzung persönlicher Kontakte
- Daten in zugriffsbeschränkten Repositorien
- Frei verfügbare Daten in Repositorien

Leipzig Health Atlas

DRYAD

GEO Gene Expression Omnibus

Welche Daten?

- Projektbeschreibungen: Studienziel, Ein- und Ausschlusskriterien, Einzugsgebiet, ...
- Metadaten: dedizierte Beschreibung der Daten - Merkmale
- Abgeleitete Daten: Befunde, Scores wie BMI, SOFA Score, ...
- Kurierte Daten: Qualitätsgesicherte Rohdaten

- Rohdaten: Einzeldaten & Dokumente im Zustand der Datenaufnahme

Warum tun wir uns so schwer?

- Datenschutz: Personenbezogene Daten, Information über Unternehmen

Warum tun wir uns so schwer?

- Datenschutz: Personenbezogene Daten, Information über Unternehmen
- Fehlende Infrastrukturen
  - Daten sind nicht auffindbar → **F**indability
  - Daten sind nicht zugreifbar → **A**ccessibility
  - Daten sind heterogen → **I**nteroperability
  - Unklare Regeln zur Weitergabe → **R**euse

Warum tun wir uns so schwer?

- Datenschutz: Patientenbezogene Daten, Information über Unternehmen
- Fehlende Infrastrukturen
  - Daten sind nicht auffindbar → **F**indability
  - Daten sind nicht zugreifbar → **A**ccessibility
  - Daten sind heterogen → **I**nteroperability
  - Unklare Regeln zur Weitergabe → **R**euse

FAIR Prinzipien als Voraussetzung für ein Data Sharing

Quelle: Wilkinson MD, Dumontier M, …, Mons B: The FAIR Guiding Principles for scientific data management and Stewardship. Nature Scientific Data. 3:160018 (2016)

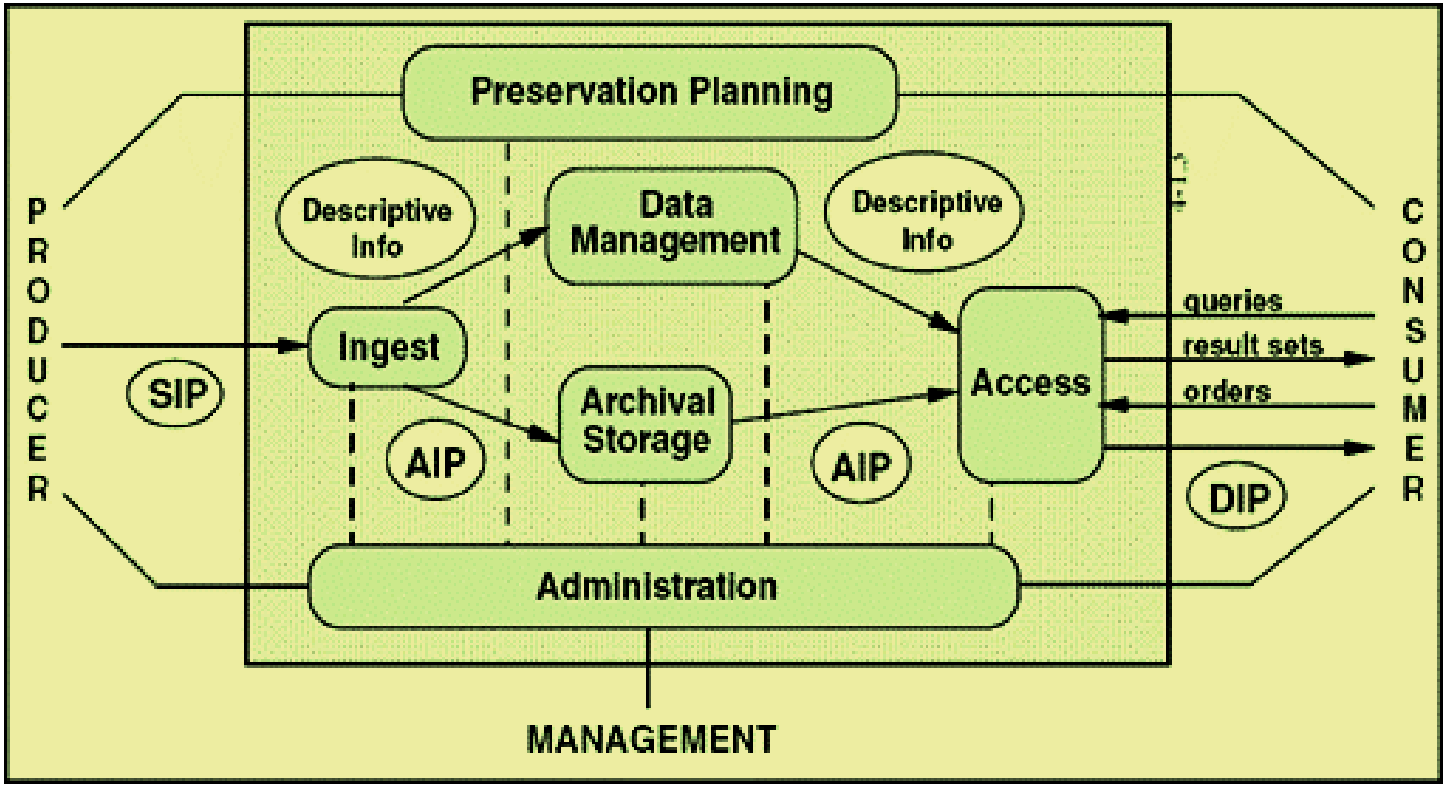
Data Sharing Projekte & Initiativen

- Epidemiologische Studien und Forschungszentren
  - KORA (Augsburg/München)
  - Ship (Greifswald)
  - LIFE (Leipzig)
  - NaKo
- Leipzig Health Atlas (LHA)
- Medizininformatik Initiative (MII)
- Nationale Forschungsdateninfrastrukturen (NFDI)

Voraussetzungen für ein Data Sharing

- Organisatorische Infrastrukturen
  - Aufbauorganisation
  - Ablauforganisation
- Technische Infrastrukturen zur Unterstützung des Data Sharing

Organisatorische Strukturen



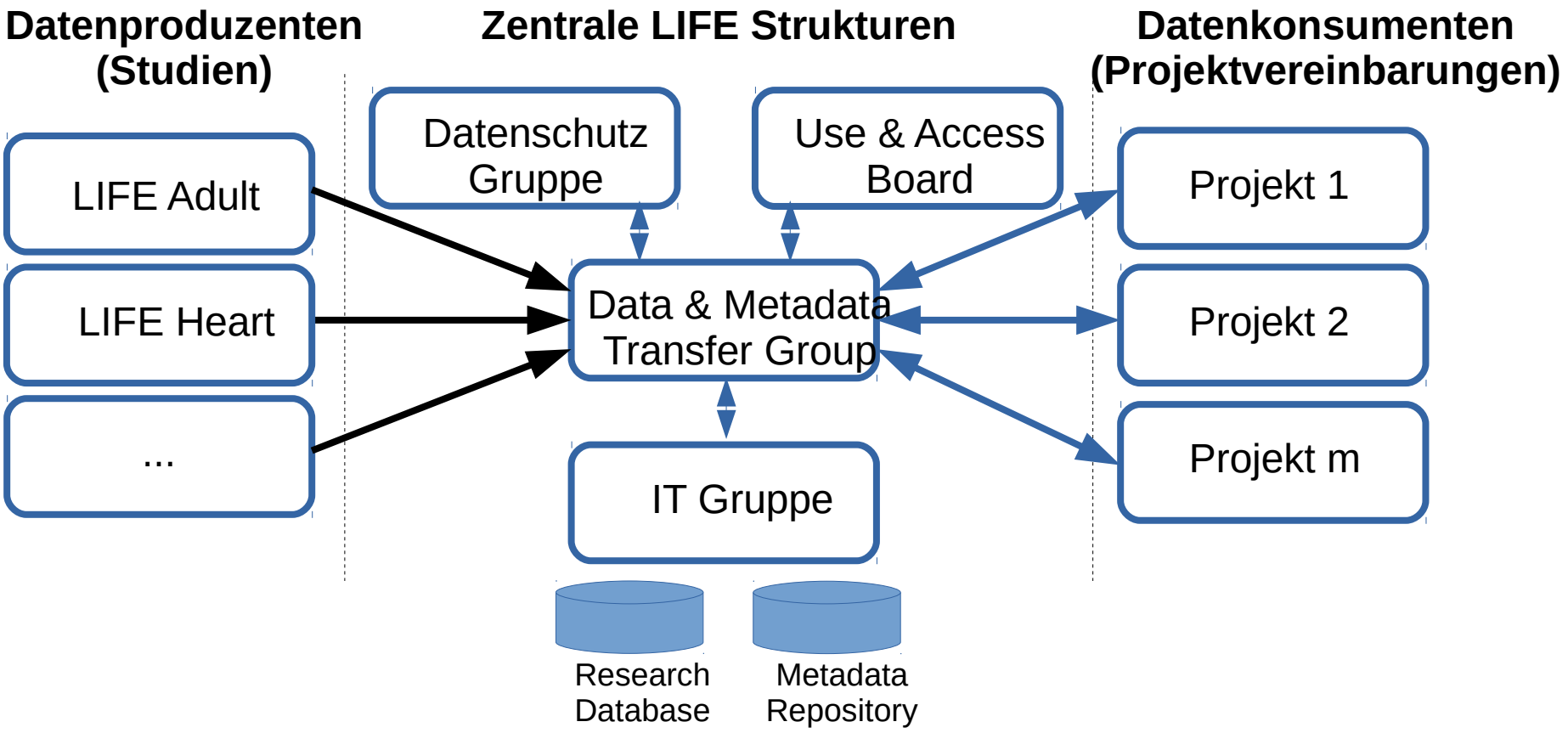
Open **A**rchiving **I**nformation **S**ystem

Referenzmodell
ISO 14721:2012

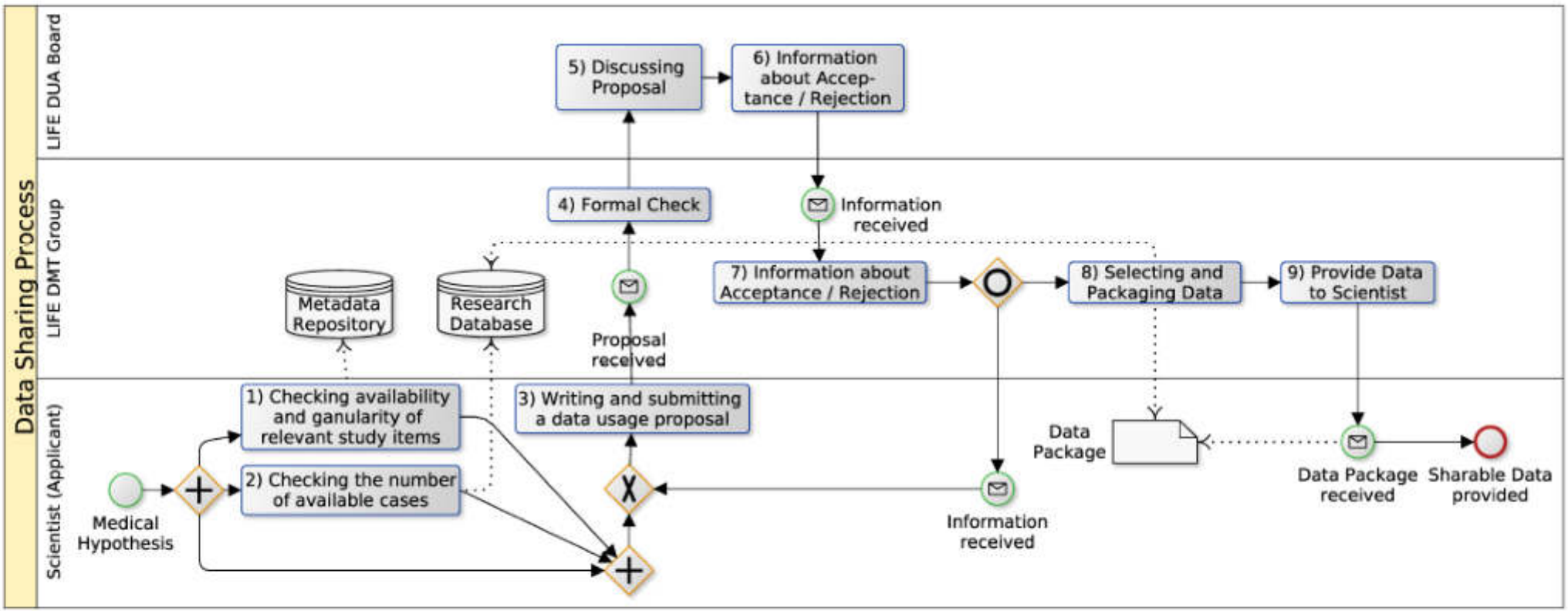
Quelle: CCSDS: Reference Model for an Open Archival Information System. 2012 https://public.ccsds.org

SIP – Submission Information Package
AIP – Archival Information Package
DIP – Dissemination Information Pakage

Organsiationsstruktur am LIFE FZ

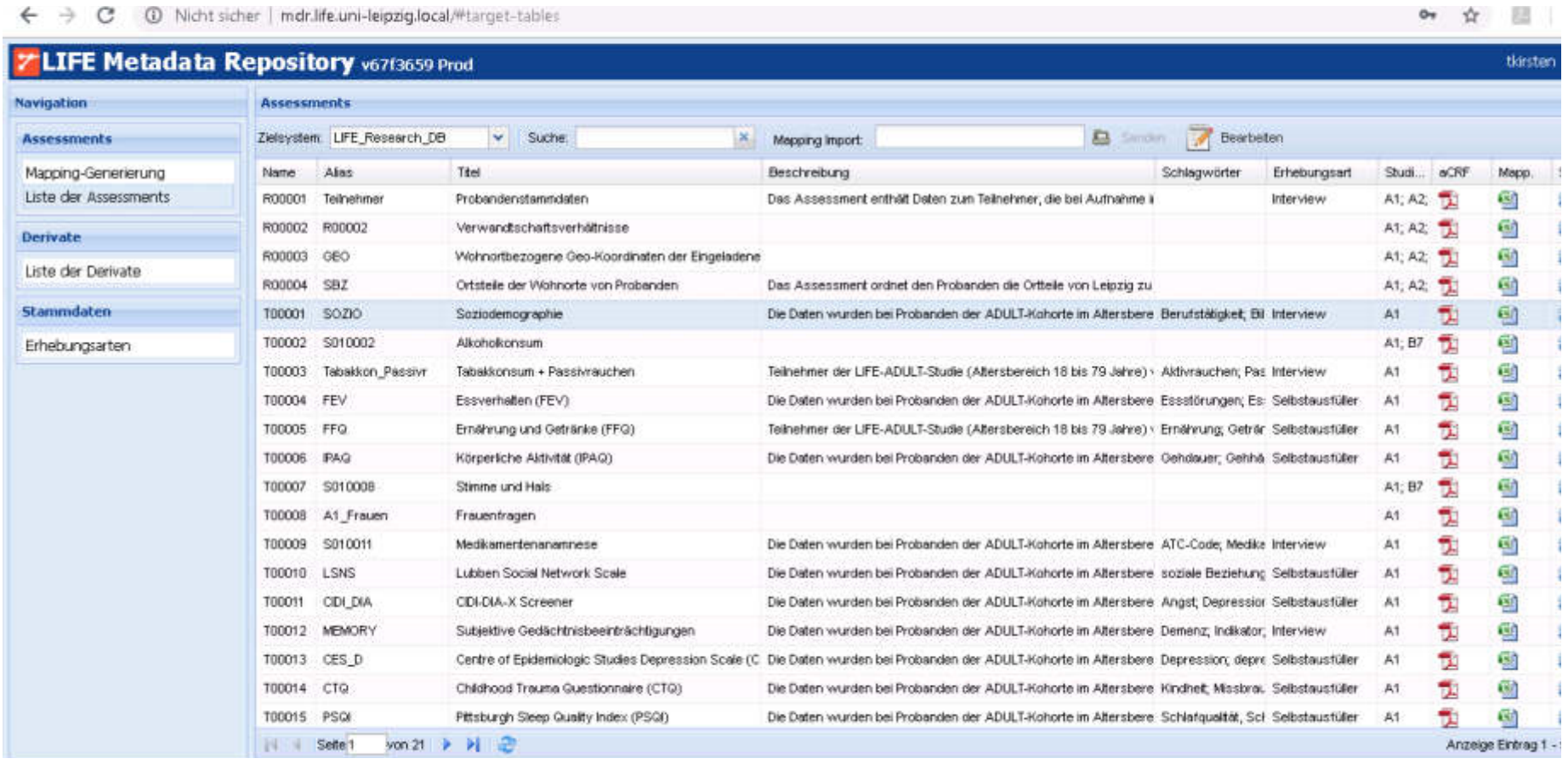


Data Sharing Prozess in LIFE

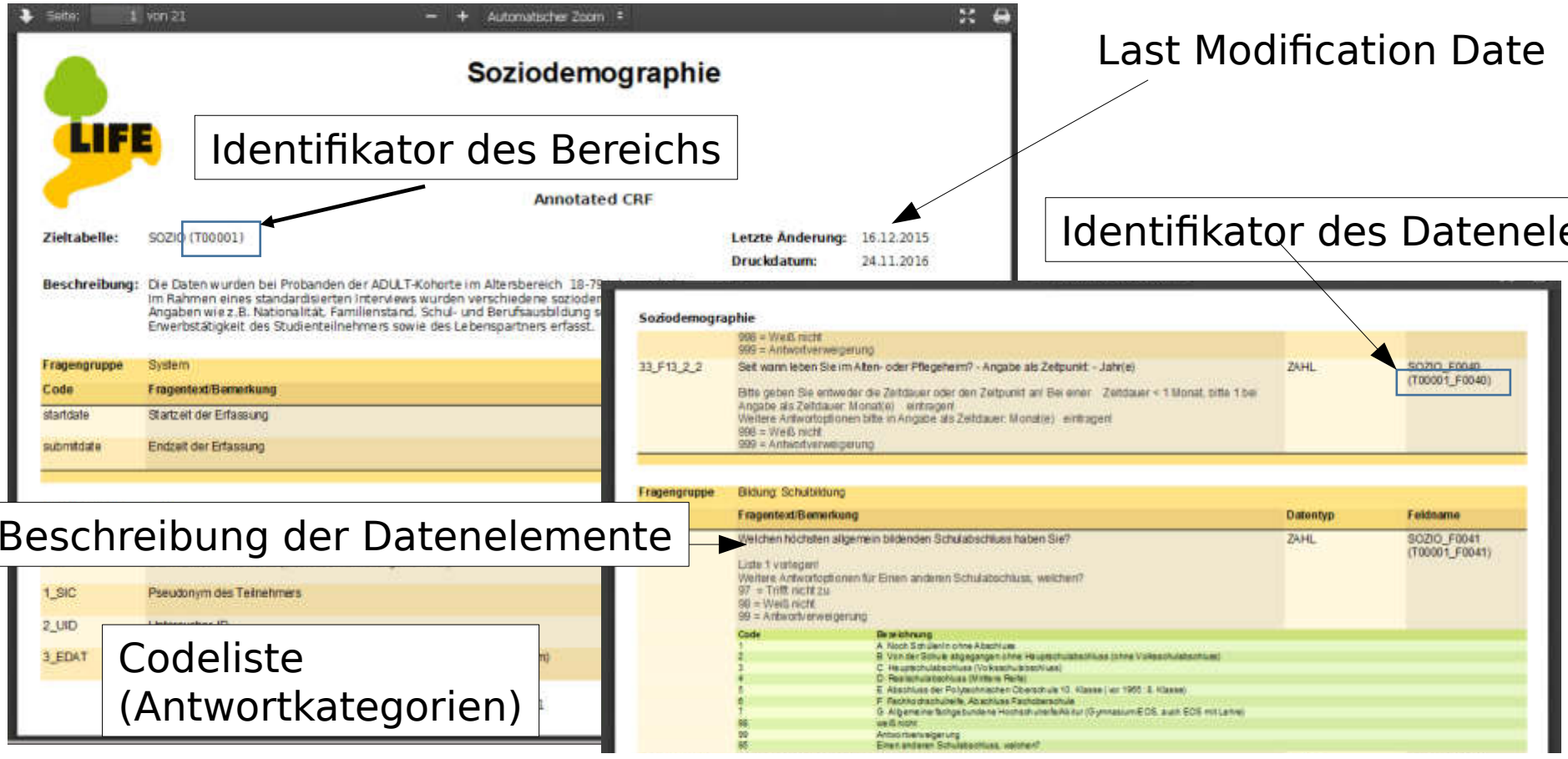


Kirsten, T *et al.* Selecting, Packaging and Granting Access for Sharing Study Data – Experiences and Recent Software Developments from the LIFE Study. Workshop on Digital Processes and Information Systems for Research Data Management. Conference of the German Informatics Society, Lecture Notes of Informatics, 2017

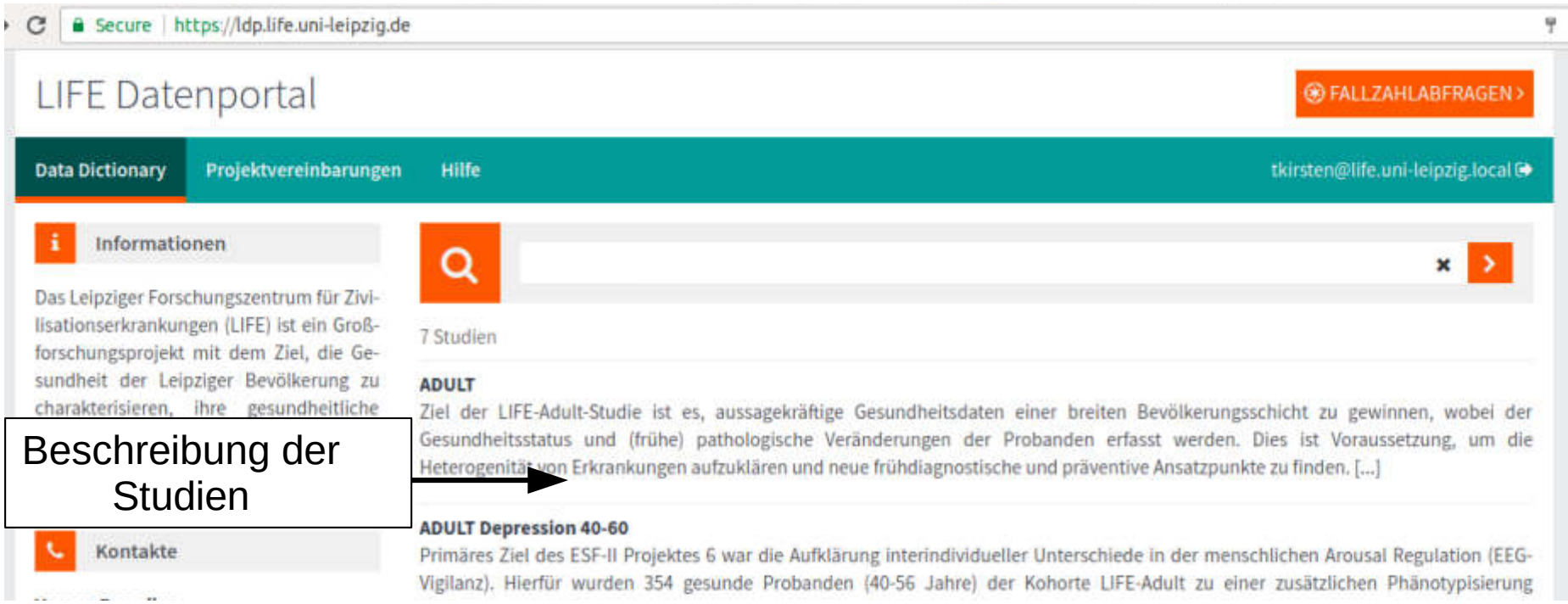
LIFE Metadaten Repository



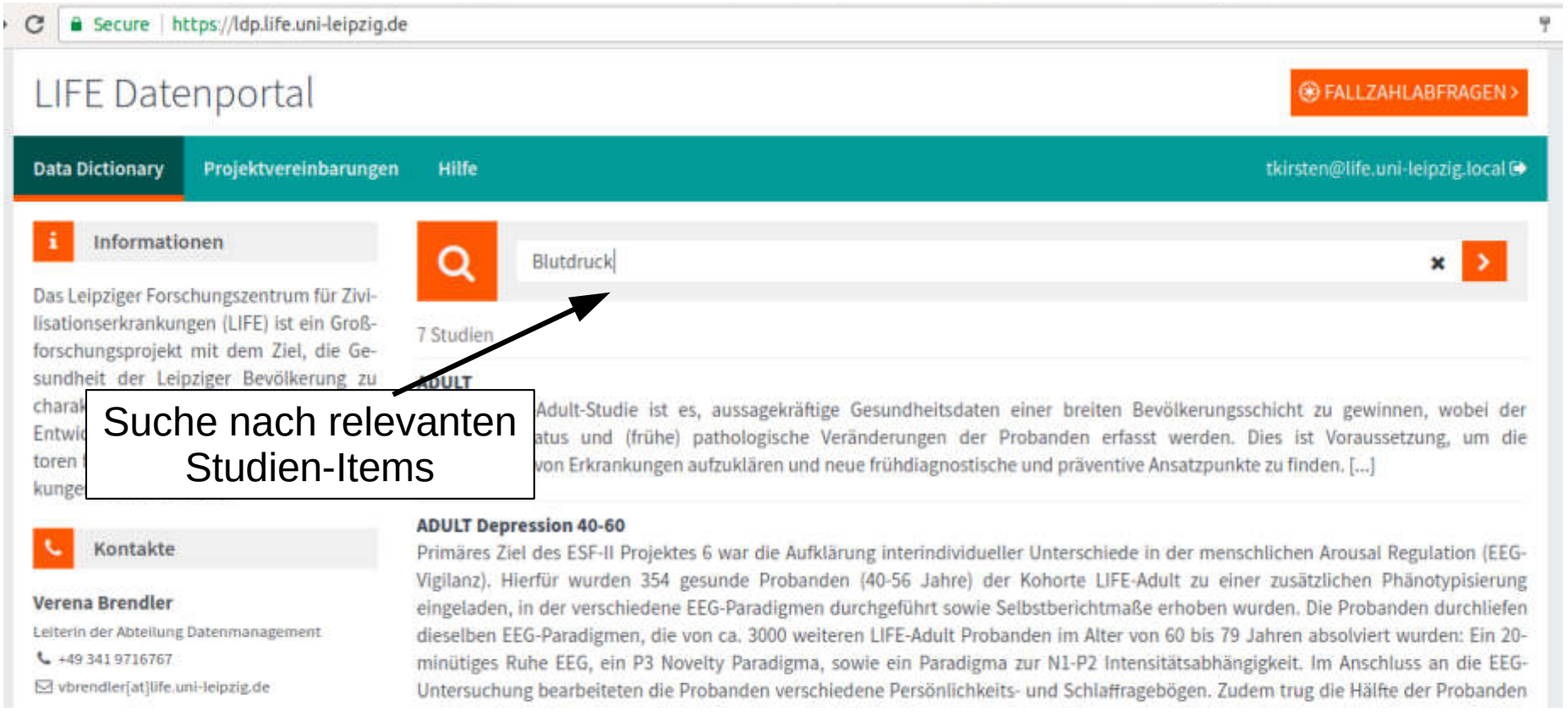
Metadaten – zum Mitnehmen



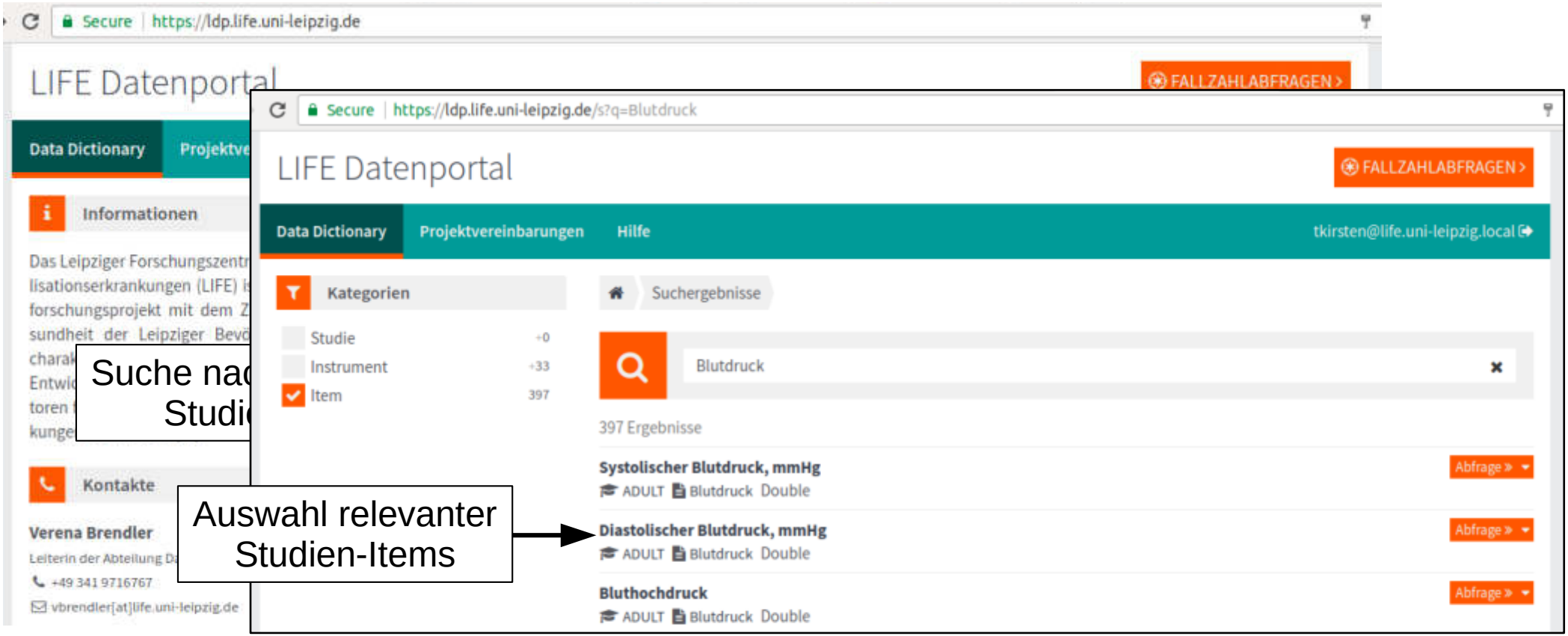
LIFE Datenportal



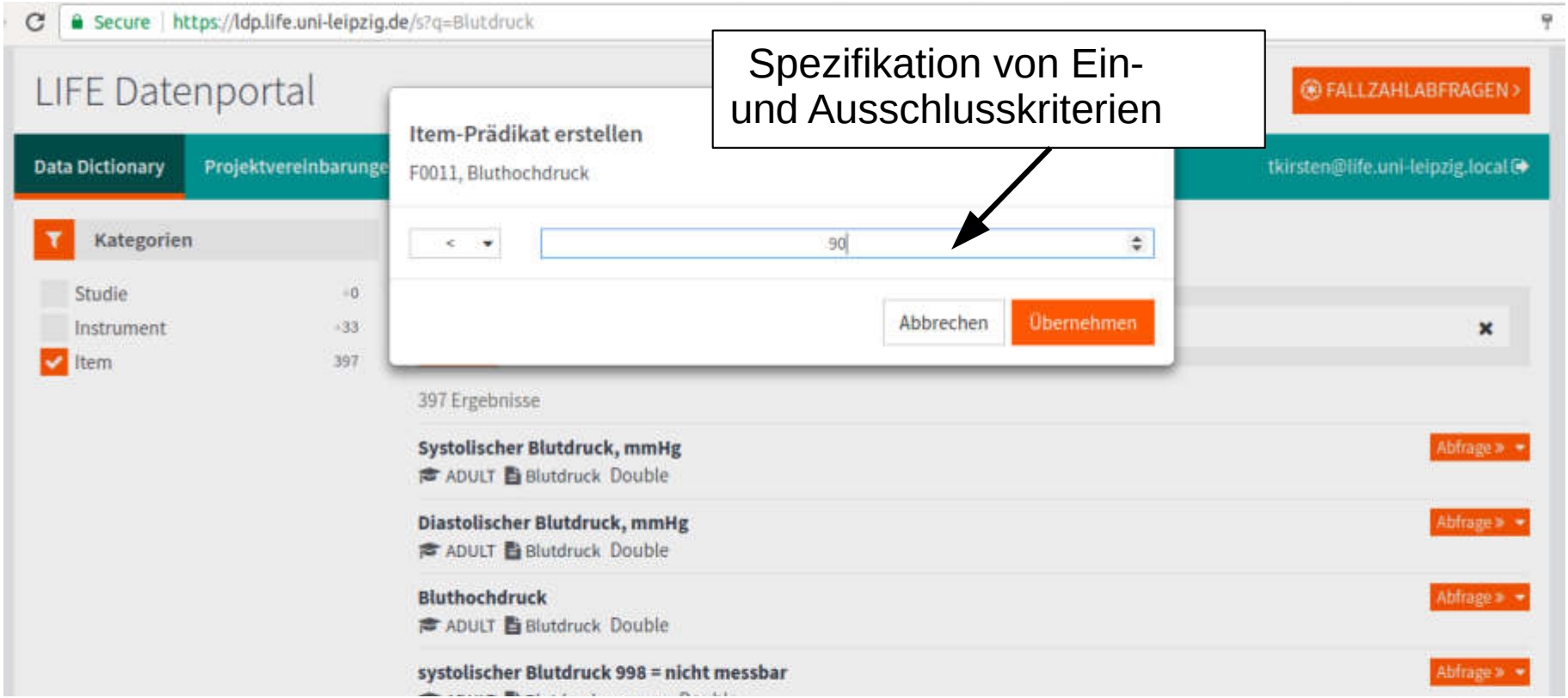
LIFE Datenportal: Metadaten



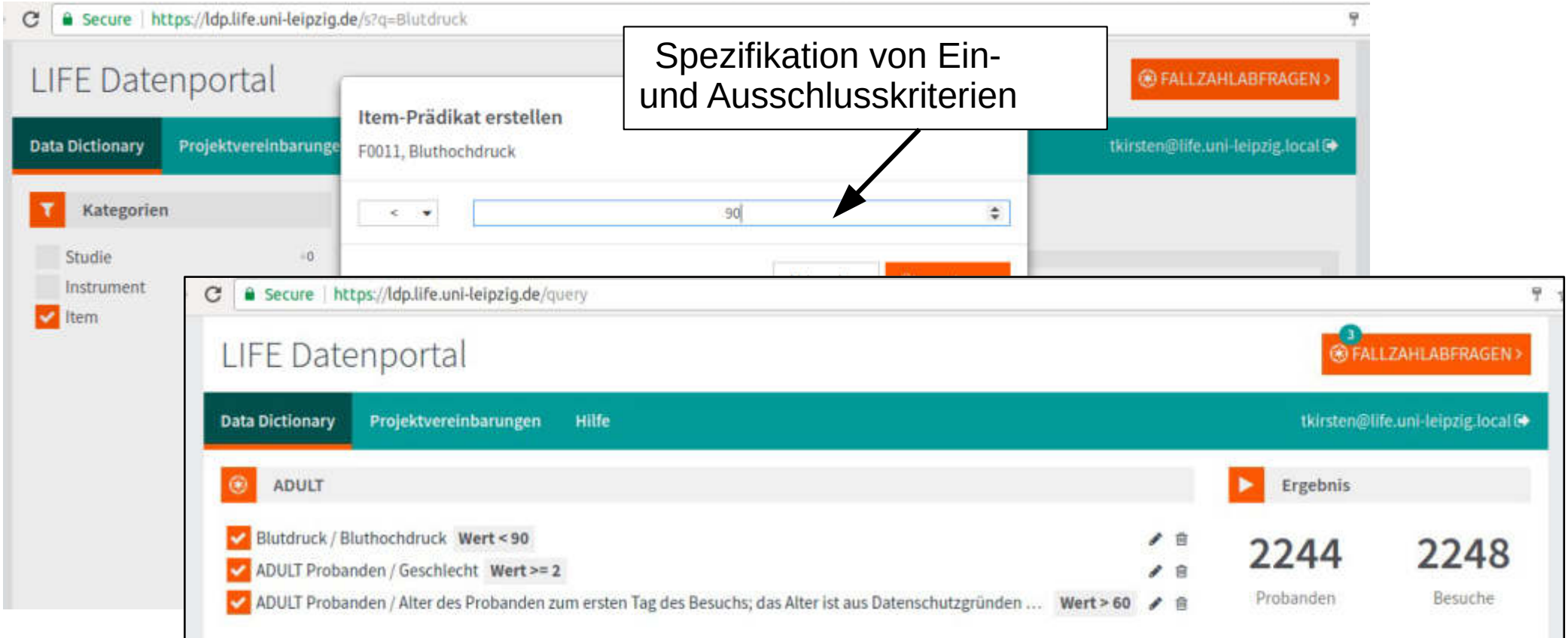
LIFE Datenportal: Metadaten



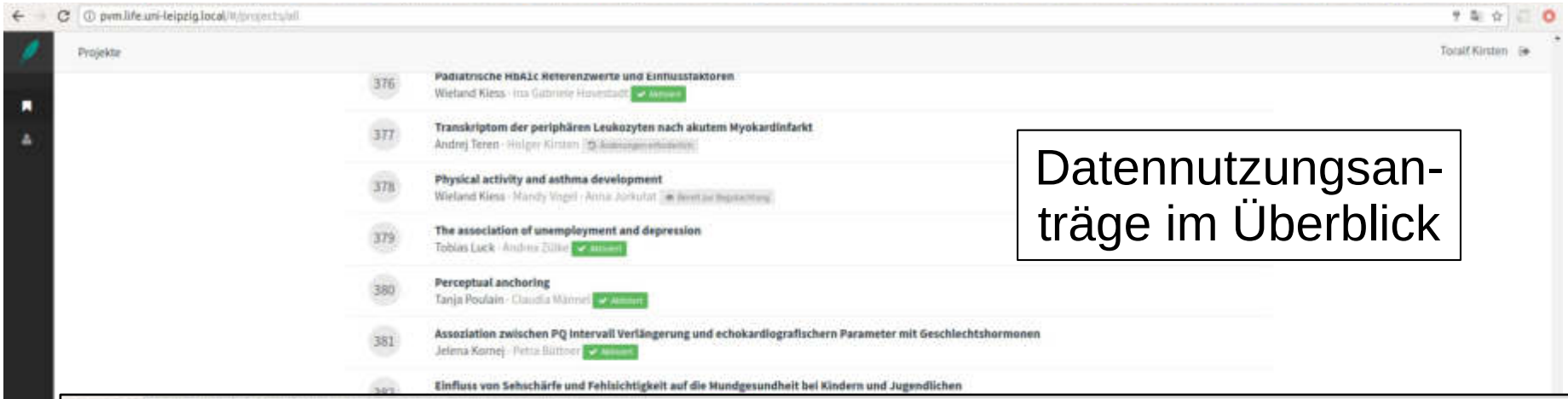
LIFE Datenportal: Fallzahlabfragen



LIFE Datenportal: Fallzahlabfragen



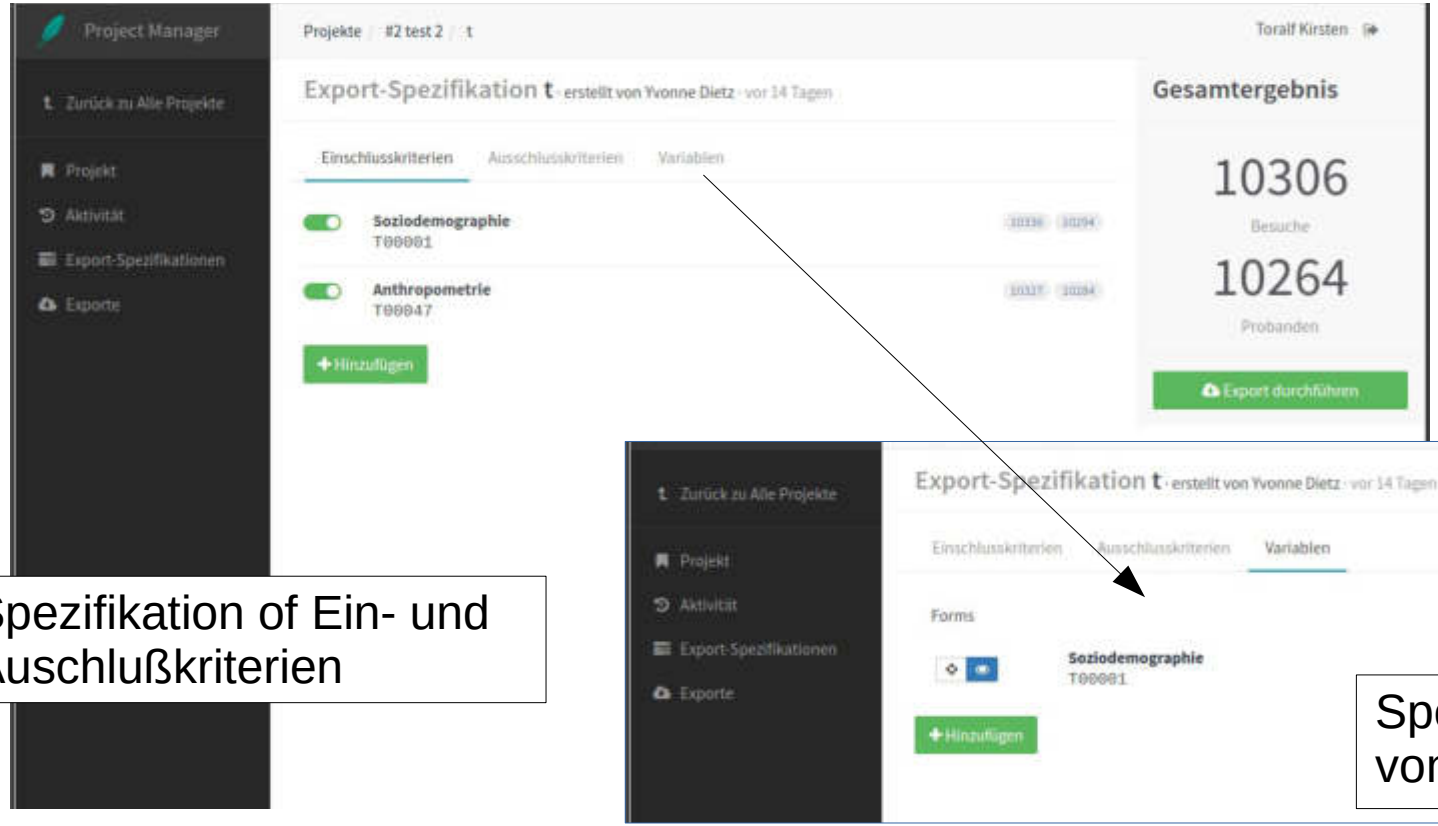
Verwaltung von Nutzungsanträgen



Datennutzungsanträge im Überblick

Details über einen Datennutzungsnantrag mit kompletter Historie

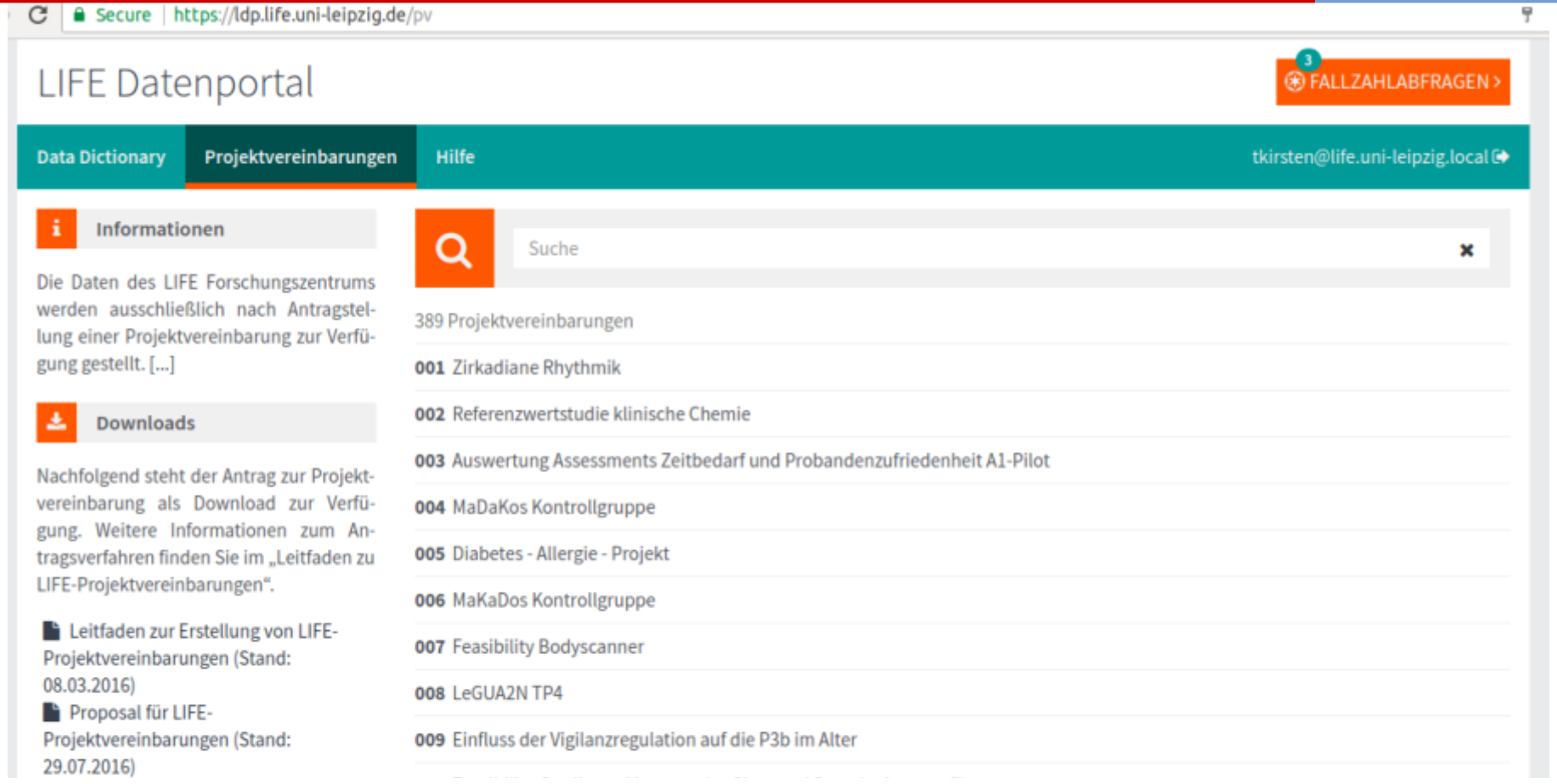
Verwaltung von Nutzungsanträgen (2)



Spezifikation of Ein- und Auschlußkriterien

Spezifikation von Zielgrößen

Verwaltung von Nutzungsanträgen (3)



Data Sharing mit Leipzig Health Atlas

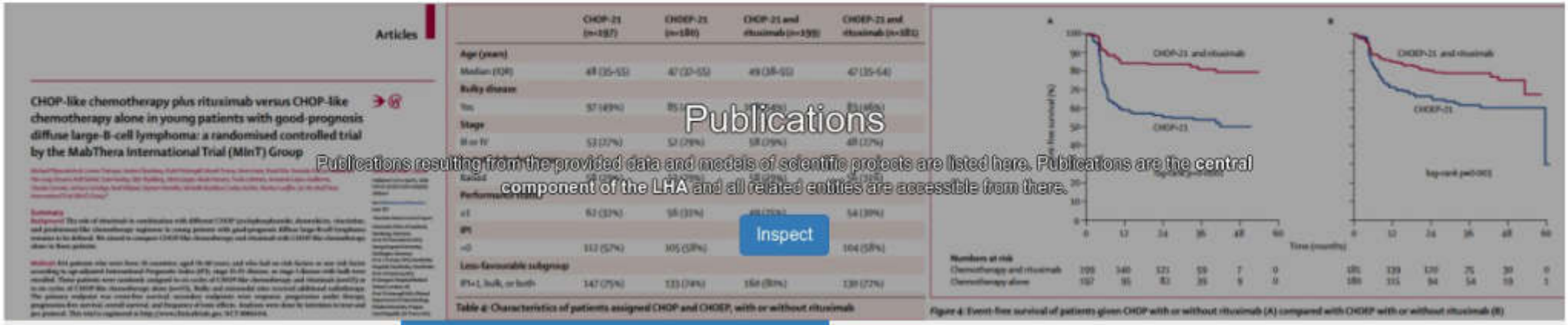
- Zentrale Verwaltungsplattform für
  - Publikations- und Projektdaten
  - Modelle (z.B. Risiko-, Vorhersagemodelle, …)
  - Statistische Analysemethoden (Pipelines)
- Einsatz eines Datenportals zur Einsichtnahme in Merkmale und Fallzahlabfrage
- Demonstratoren

https://www.health-atlas.de/

LHA Plattform

Leipzig Health Atlas

Browse Help Search here... Search Register Log in



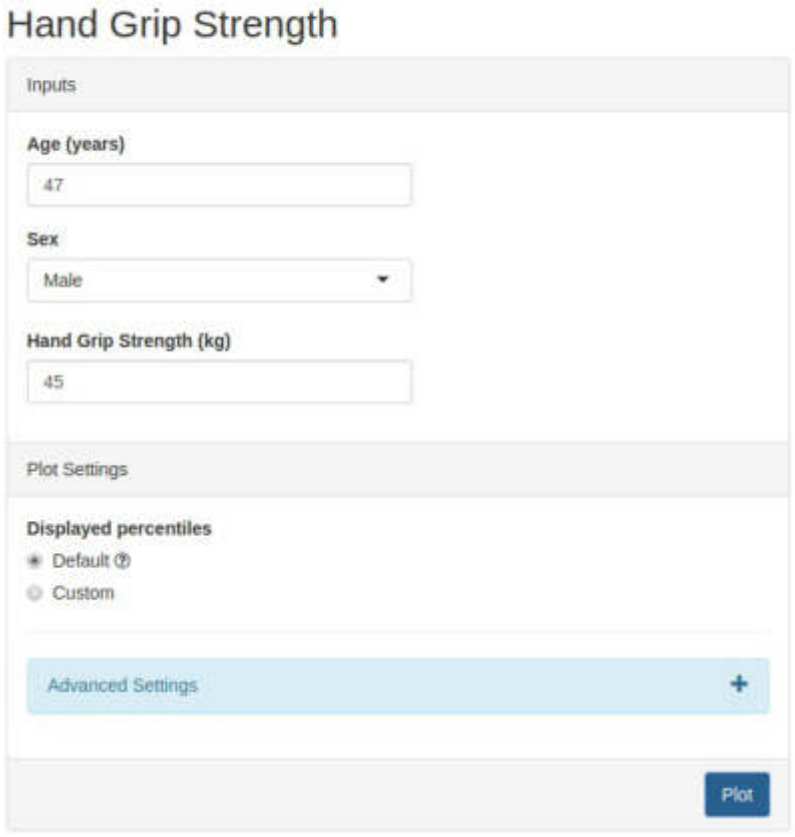
23 Scientific Projects | 265 Publications | 156 Data Files | 24 Models

The Leipzig Health Atlas (LHA) is an alliance of medical ontologists, medical systems biologists and clinical trials groups to design and implement a multi-functional and quality-assured atlas. It provides models, data and metadata on specific use cases from medical research fields in which our team has scientific and clinical expertise. Two basic characteristics are:

1. an interoperable ontology-based semantic platform to share highly annotated data, novel ontologies, usable models and working software tools;
2. an advanced, application-oriented analytic pipeline for a clinical and scientific user community to provide disease-related phenotype classifications, omics based disease sub-classifications, risk predictions and simulation models for diseases and organ functions

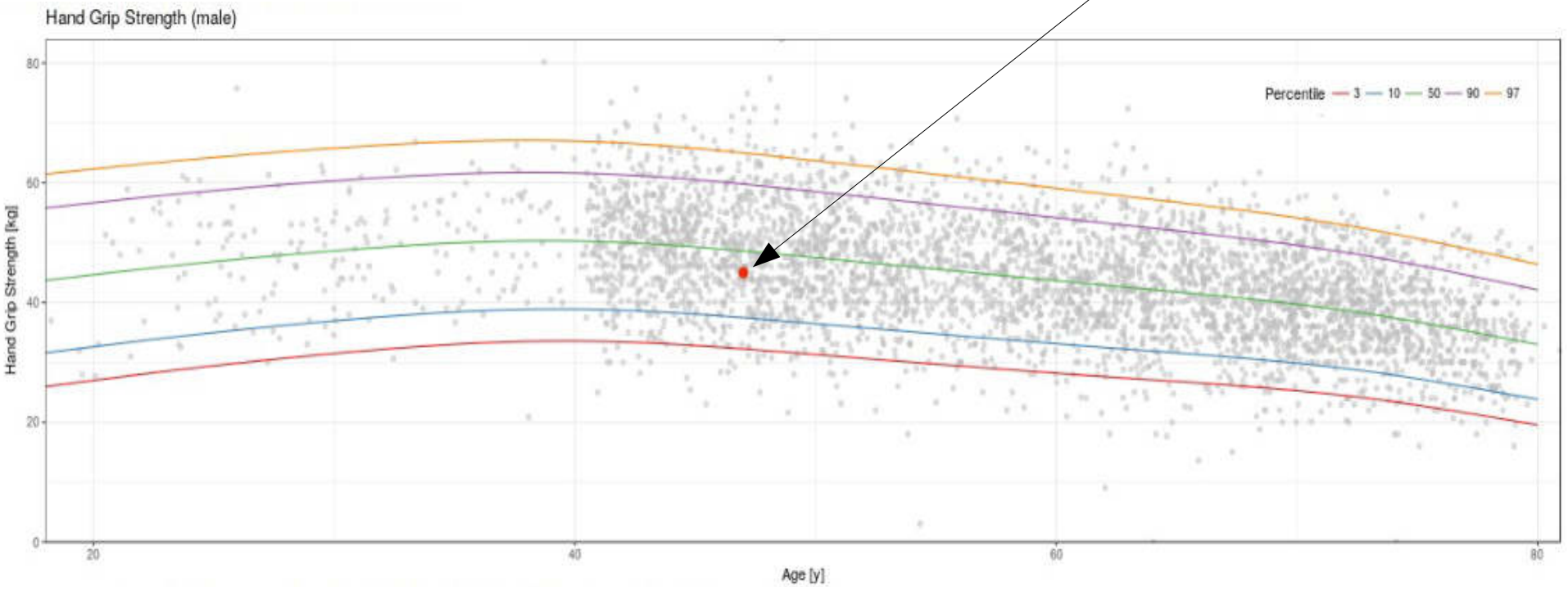
Demonstratoren – Beispiel Normwerte

- Basis: Daten der LIFE Adult Sudie (10.000 Probanden)
- Normwerte oftmals nur für wenige Phänotypen verfügbar
- Web-Applikation zur Anzeige von Normwerten verschiedenster Phänotypen
- Nutzer: Kliniker, Wissenschaftler, Patienten/Probanden



Demonstratoren – Normwerte (II)

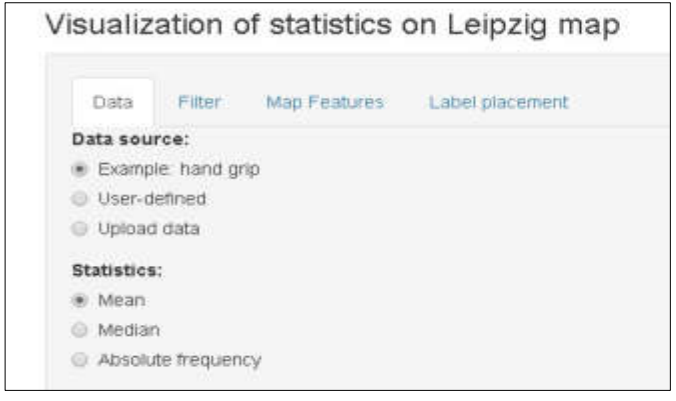
https://apps.health-atlas.de/handgrip

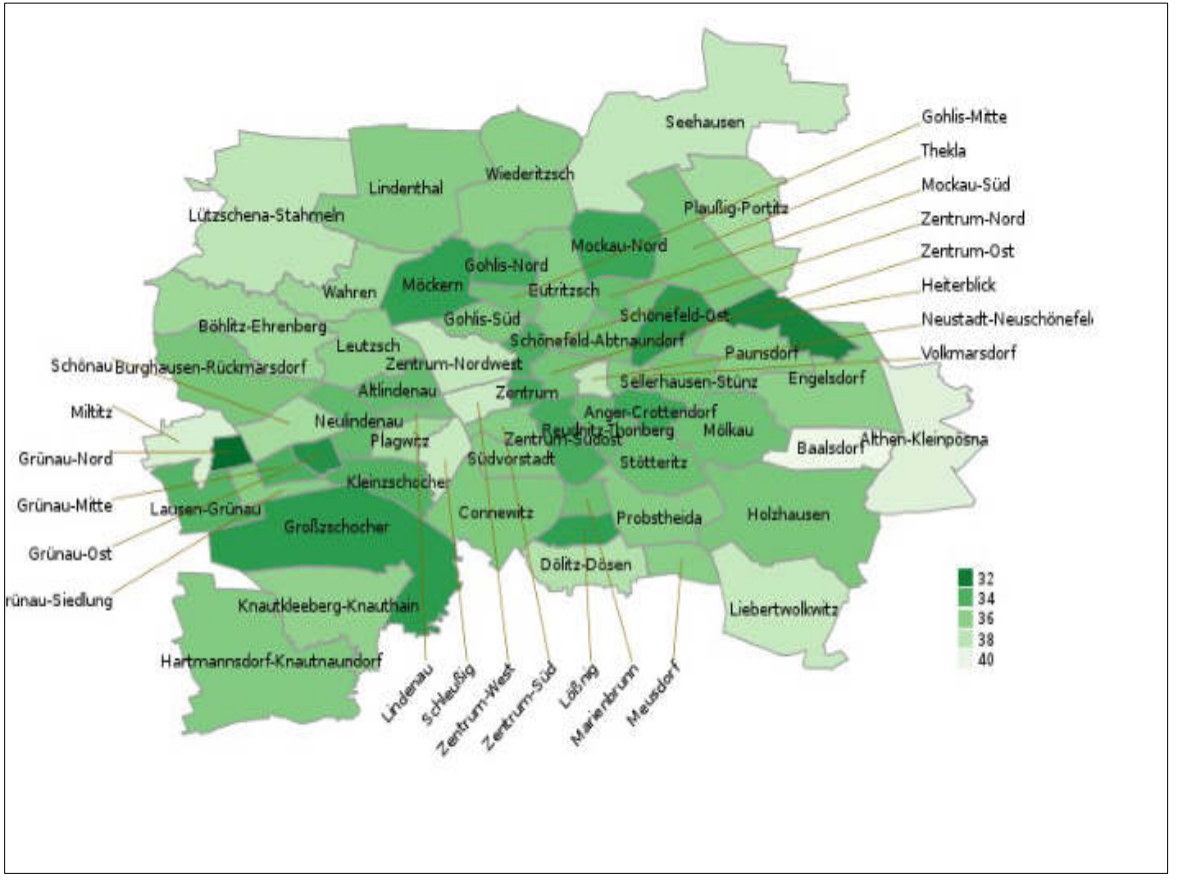


Raum-Zeit-bezogene Analysen

- Basis: LIFE Adult
- Zuordnung der Individual-daten zu Wohnorten

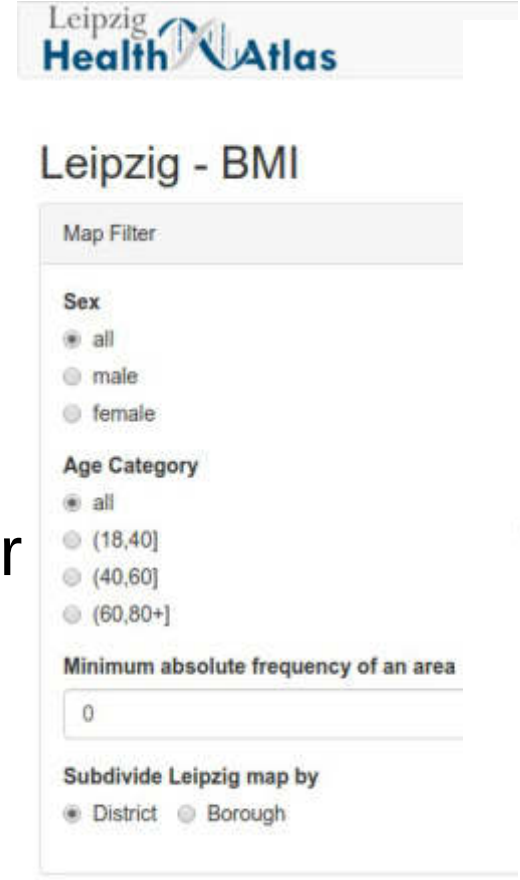
→ Aggregation je Ortsteil

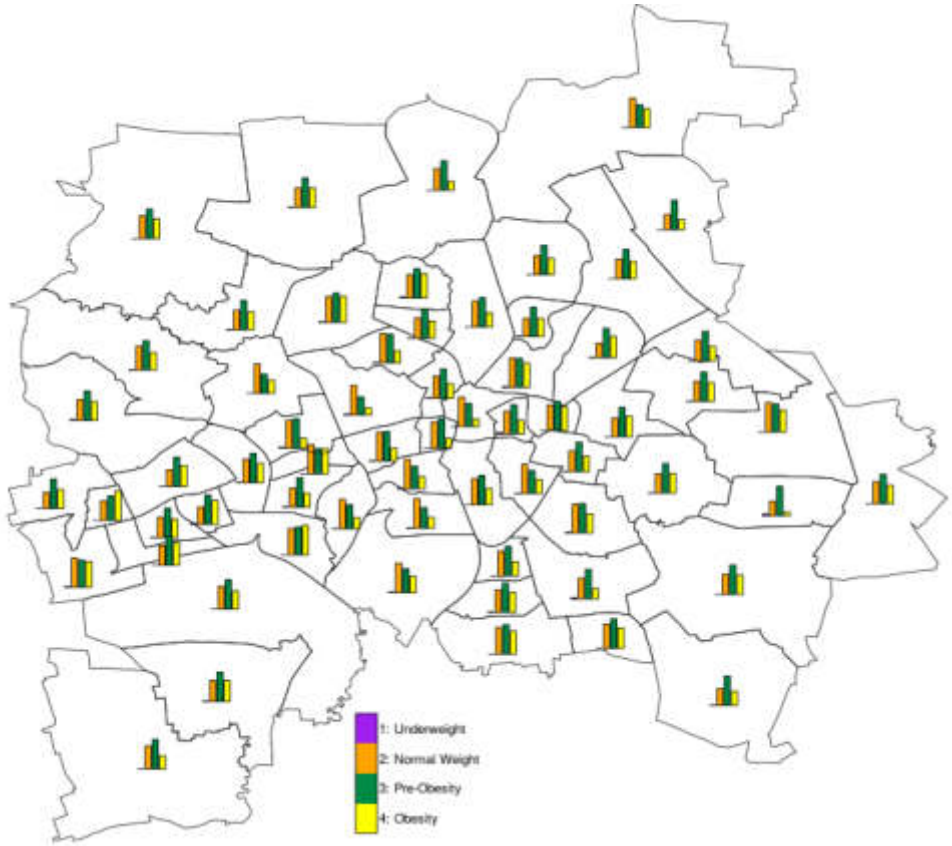




Raum-Zeit-bezogene Analysen (II)

- Basis: LIFE Adult
- Verteilung der Probanden je BMI Kategorie über die Leipziger Ortsteile



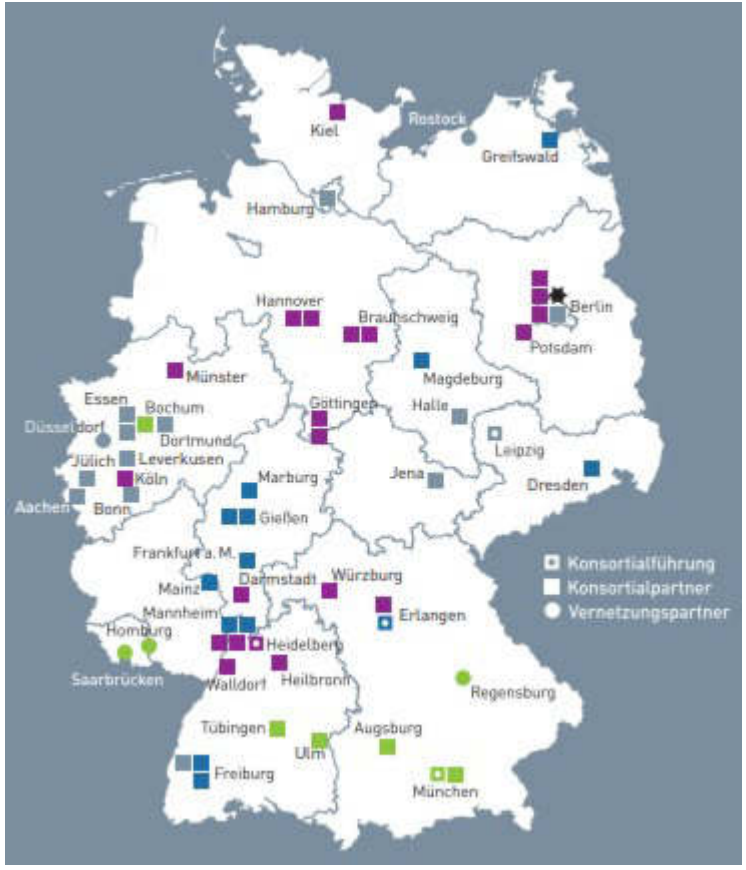


Medizininformatik Initiative

- Ziel: Aufbau von FAIR Data Points (Datenintegrationszentren)
  - Etablierung organisatorischer Einheiten und Prozesse
  - Aufbau von technischen Infrastrukturen
- Universitätsklinika aktuell im Fokus
- SMITH als eines von vier geförderten Konsortien

https://www.smith.care

DIFUTURE
HiGHmed
MIRACUM
SMITH
Koordinationsstelle



Quelle: https://www.medizininformatik-initiative.de

Nutzungsordnung - Reuse

Medizininformatik-Initiative
Begleitstruktur – Koordinationsstelle des Nationalen Steuerungsgremiums

MEDIZIN INFORMATIK INITIATIVE

Übergreifende Muster-Nutzungsordnung

zum

Austausch von Patientendaten, Biomaterialien, Analysemethoden und -routinen im Rahmen der Medizininformatik-Initiative

AG Data Sharing

insbesondere unter Mitwirkung von T. Kirsten, R. Jahns, T. Leddig, C. Schade-Brittinger, A. Scherag, M. Löffler

Version: 1.02

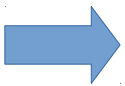
Personal Health Train (PHT)

- Ansatz, um verteilte Daten für gemeinsame Analysen nutzbar zu machen
- Kooperation zwischen Aachen (Decker, Beyan), Tübingen (Kohlbacher), Leipzig (Kirsten, Löbe)
- Zentrale Idee
  - Bringe Analyseroutinen zu Daten (FAIR Data Points)
  - Erzeuge Zwischenergebnisse an FAIR Data Points
  - Aggregiere Zwischenergebnisse

Karim M; ... **Kirsten T** … Beyan O: A Distributed Analytics Platform to Execute FHIR-based Phenotyping Algorithms. In: Proc. of the 11th International Conference Semantic Web Applications and Tools for Life Sciences (SWAT4HCLS 2018) Bd. CEUR Workshop Proceedings 2275, 2018

Forensik im Wandel

- Wir arbeiten in Mittweida aktuell an ...
  - … einem Metadaten Repository für forensische Daten
  - … Crime Atlas, in dem Daten, Metadaten und Algorithmen / Pipelines hinterlegt werden
  - … Infrastrukturkomponenten zur Abbildung des Use&Access für Forensik-Daten
  - ... übergreifenden Infrastruktur, um verteilte Datenbestände unter Wahrung des Datenschutzes zu analysieren



NFDI4Crime

Zusammenfassung

- Data Sharing notwendig für Forschung und Versorgung
  - Kulturwandel in einzelnen Disziplinen
  - Nachhaltigkeit und Schonung von Ressourcen
- Infrastrukturen
  - Organisatorische Einheiten und Prozesse
  - Technische Infrastrukturen
- Verschiedene Data Sharing Projekte und Initiativen
- Erweiterung des Data Sharing für Algorithmen, Modelle