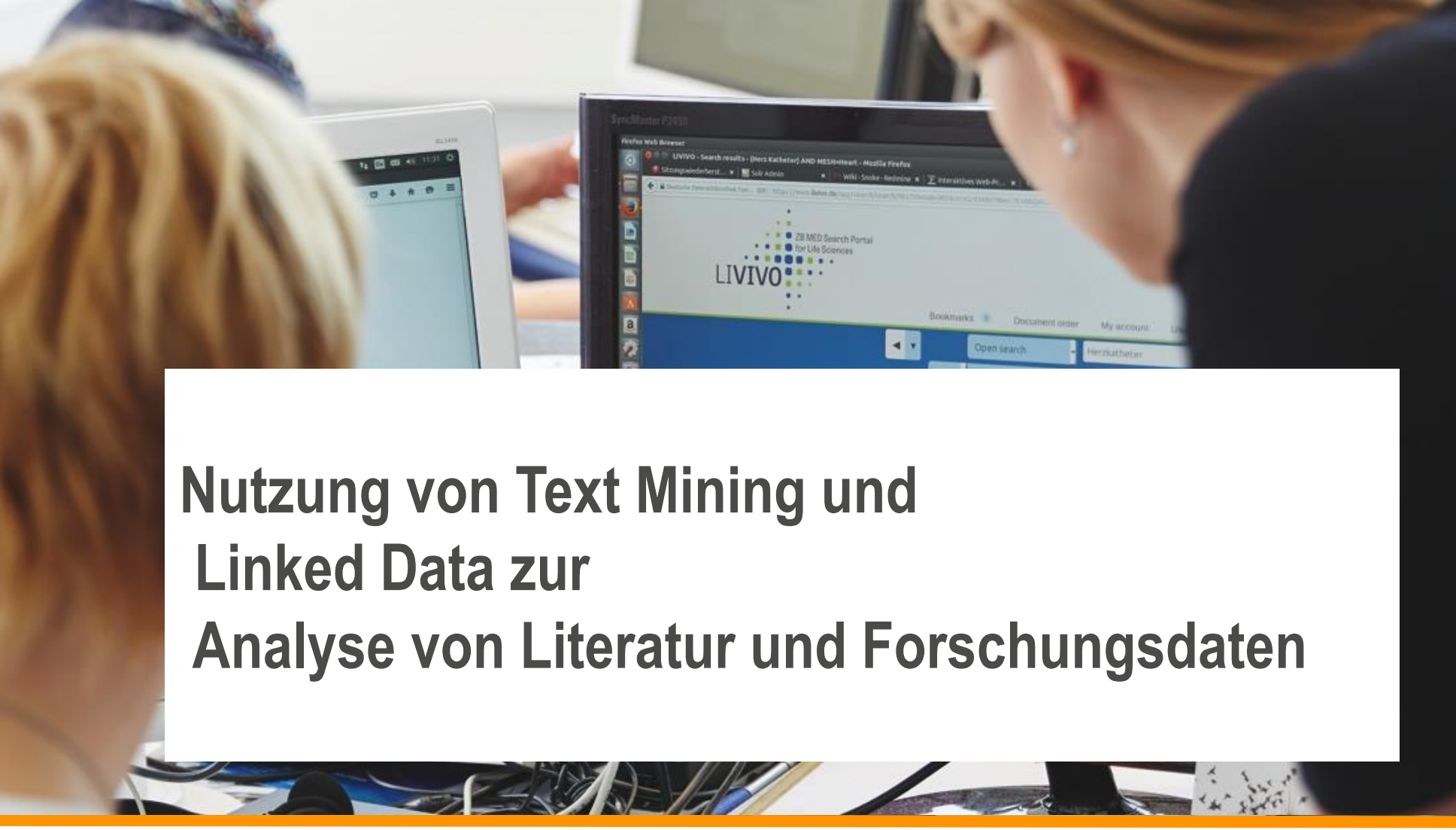


Nutzung von Text Mining und Linked Data zur Analyse von Literatur und Forschungsdaten

ZB MED Informationszentrum Lebenswissenschaften

Prof. Juliane Fluck,

UNIVERSITÄT BONN

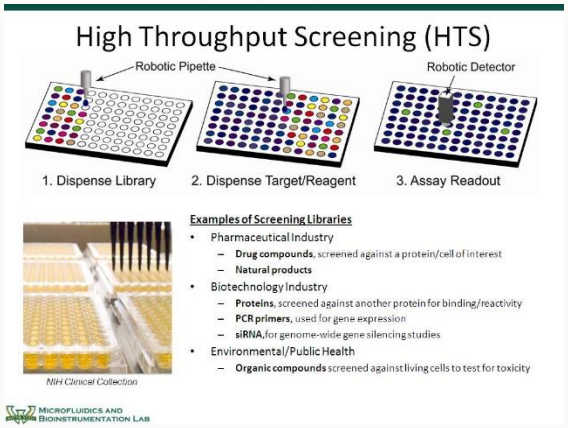
9.10. 2019 DINI Tagung

THESE 1

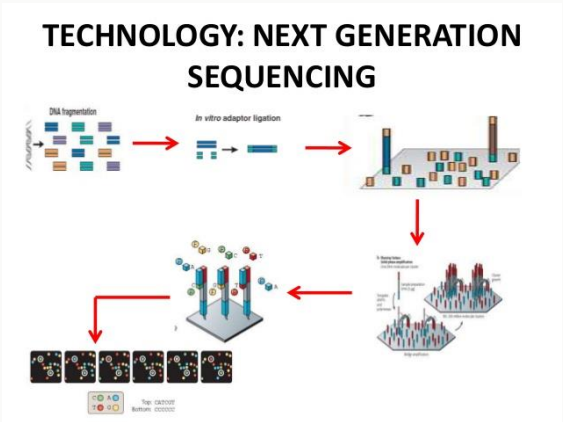
Wachsender Bedarf an elektronisch verfügbarer Semantik

ZUNAHME VON LARGE SCALE EXPERIMENTEN – BEISPIEL BIOMEDIZINISCHE FORSCHUNG

- High TroughputScreening (HTS)
- Next Generation Sequencing



https://nanohub.org/

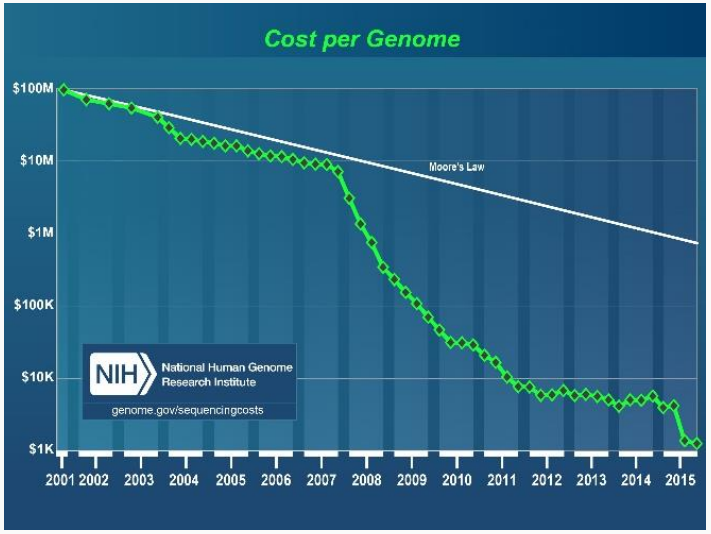


https://www.slideshare.net/lampweb/clinical-applications-of-next-generation-sequencing

ZUNAHME VON LARGE SCALE EXPERIMENTEN – BEISPIEL BIOMEDIZINISCHE FORSCHUNG

und drastische Reduktion der Kosten:

Beispiel Genomsequenzierung



https://www.genome.gov/sequencingcosts/

heute nur noch 1000 USD/Genome

WACHSENDER BEDARF AN ELEKTRONISCH VERFÜGBARER SEMANTIK

Höherer Wissensbedarf zur Interpretation der Daten

Zunahme an Large Scale Methoden

- Man braucht von allen gemessenen Entitäten und deren Interaktionen Information, diese kann keiner ohne elektronisch verfügbarer Information interpretieren

In der Medizin Shift zu personalisierter Medizin

- braucht für jeden Patienten abhängig von seiner persönlicher Ausprägung andere Information

WACHSENDER BEDARF AN ELEKTRONISCH VERFÜGBARER SEMANTIK II

Wachstum unstrukturierter Information

- Literaturwachstum ist viel schneller als Wachstum der Wissens-Datenbanken
  - PubMed zurzeit mehr

  als 29 Millionen Abstracts

  - immer mehr Informationen werden als Supplement abgelegt

WACHSENDER BEDARF AN ELEKTRONISCH VERFÜGBARER SEMANTIK III

Wachstum von Datenpublikationen

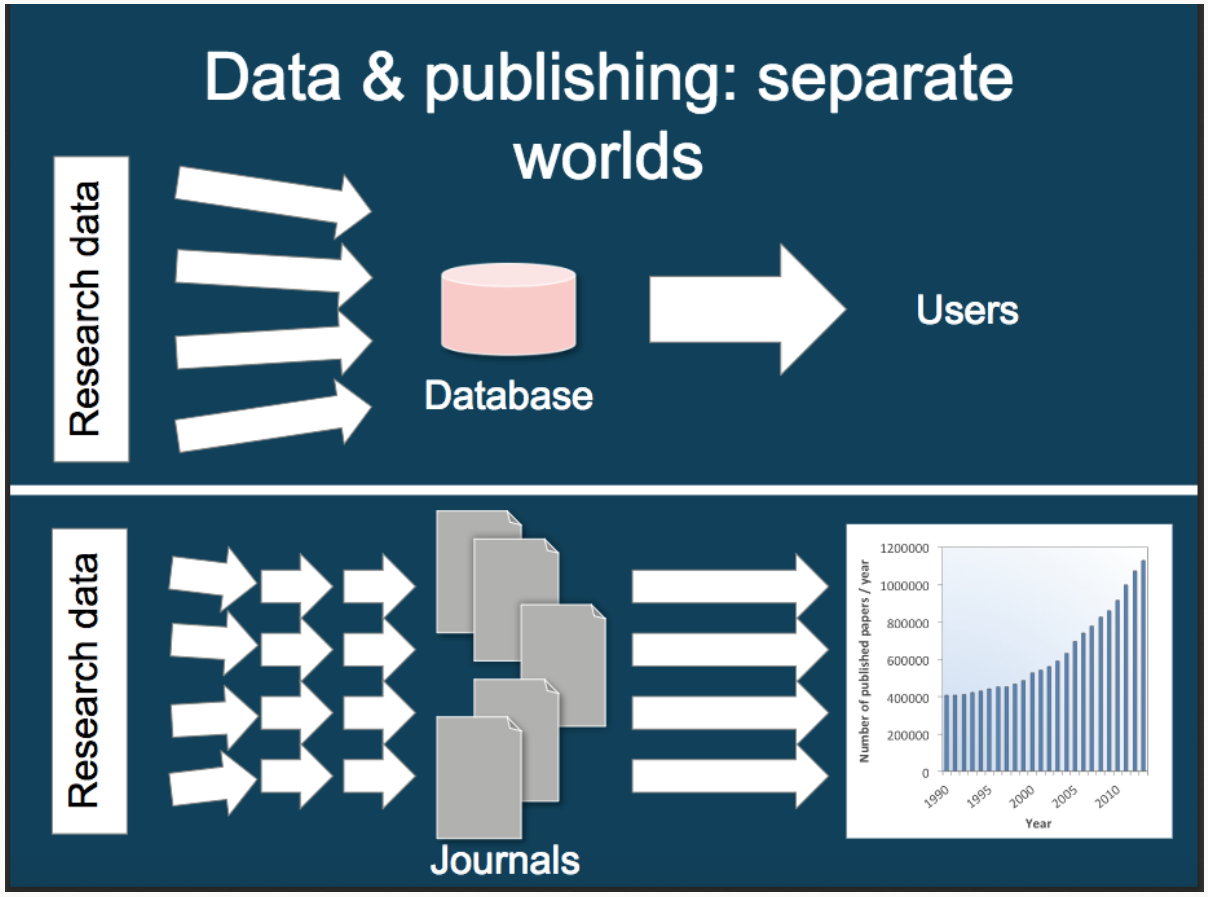
- Wenn Daten abgelegt werden, sind sie nur unvollständig annotiert und selten interoperabel
- Daten Kuration und Integration wird zum Hauptfaktor für Analyse heterogener Daten

Konsequenz auf Ebene der Förderer:

Förderung einer nationalen Forschungsdateninfrastruktur

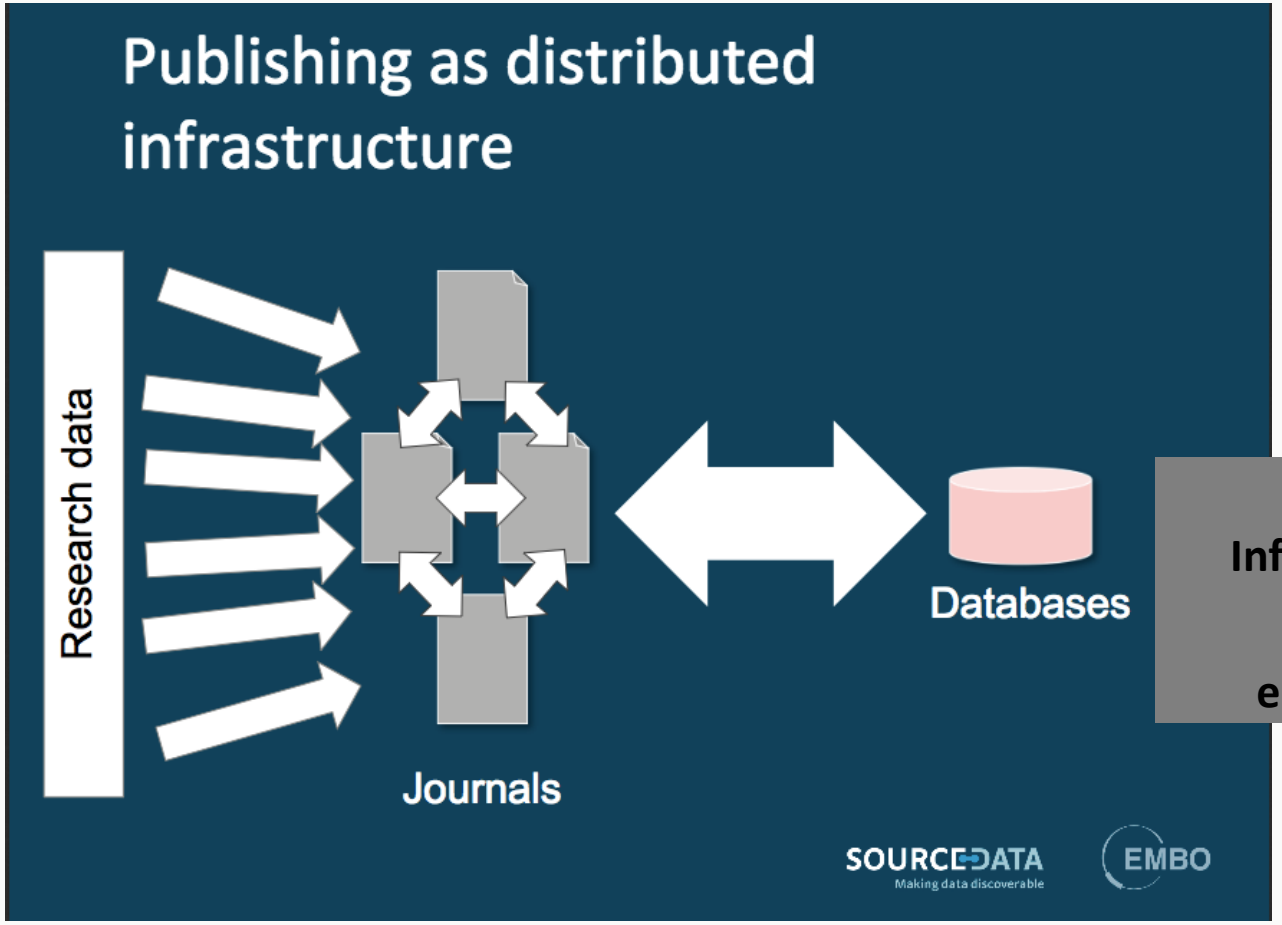
ZB MED ist stark beteiligt, in 2019 Einreichung NFDI4Health (Leitung), NFDI4Agri (Beteiligung)

Traditionales Veröffentlichen:



Dr. Thomas Lemberger (EMBO): SourceData – bridging scientific publishing and open science
https://os.helmholtz.de/fileadmin/user_upload/os.helmholtz.de/Workshops/helmholtz_oswebinar44_lemberger.pdf

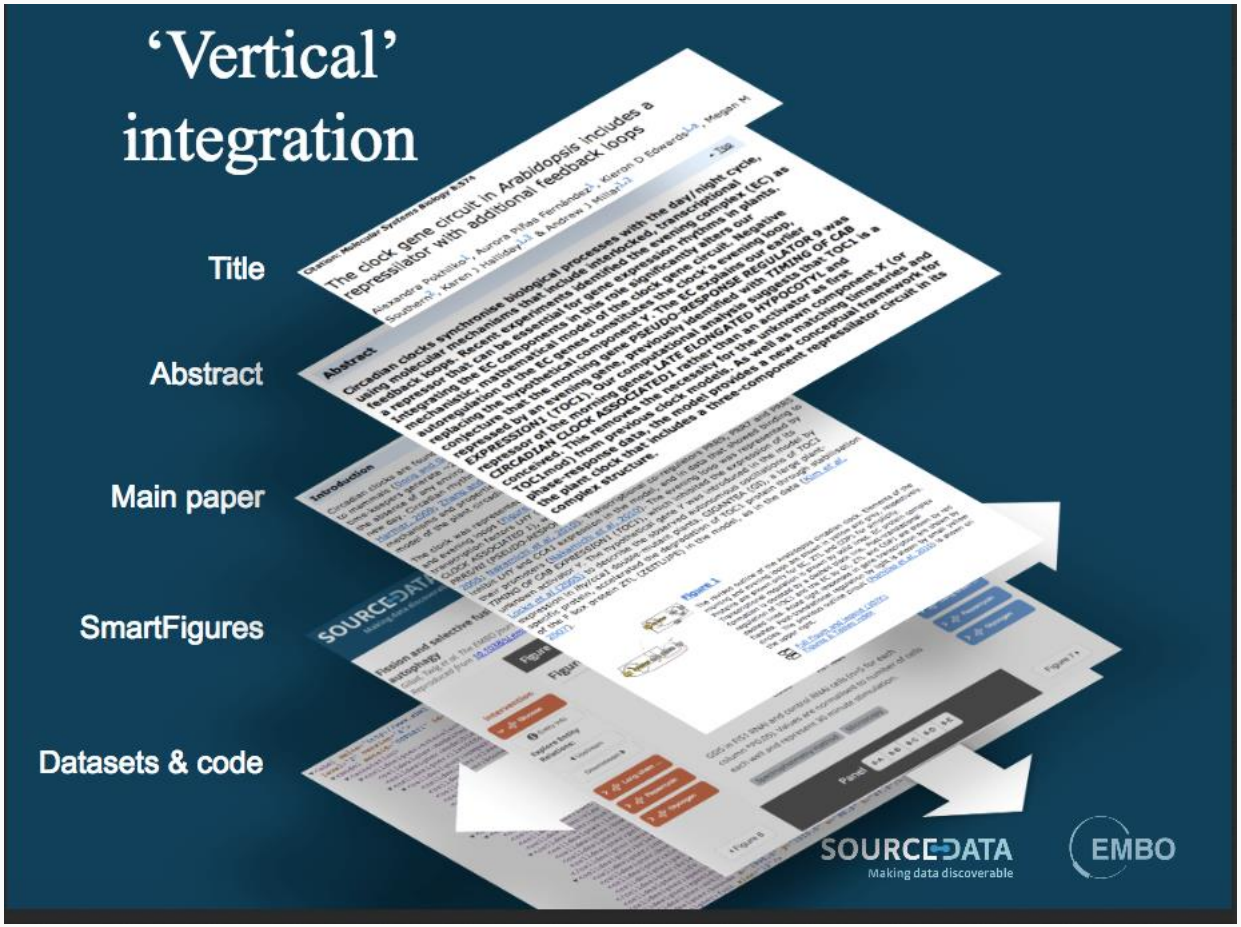
VISION



Ein Großteil der Information/des Wissens ist strukturiert und elektronisch verfügbar

Dr. Thomas Lemberger (EMBO): SourceData – bridging scientific publishing and open science
https://os.helmholtz.de/fileadmin/user_upload/os.helmholtz.de/Workshops/helmholtz_oswebinar44_lemberger.pdf

PROJEKT SOURCEDATA:



Dr. Thomas Lemberger (EMBO): SourceData – bridging scientific publishing and open science
https://os.helmholtz.de/fileadmin/user_upload/os.helmholtz.de/Workshops/helmholtz_oswebinar44_lemberger.pdf

IM VERGLEICH: MOMENTANER PUBLIKATIONSPROZESS

Wissenschaftler reichen Paper in Computer-lesbaren Formaten ein:

- Zumeist Word oder TEX
- mit Struktur für Titel, Abstract, Methoden, Results, Abbildungslegenden
- Tabellen oft auf Ecxel basierend
- Verlage generieren intern oft XML

PUBLIZIERT WIRD – PDF

PDF

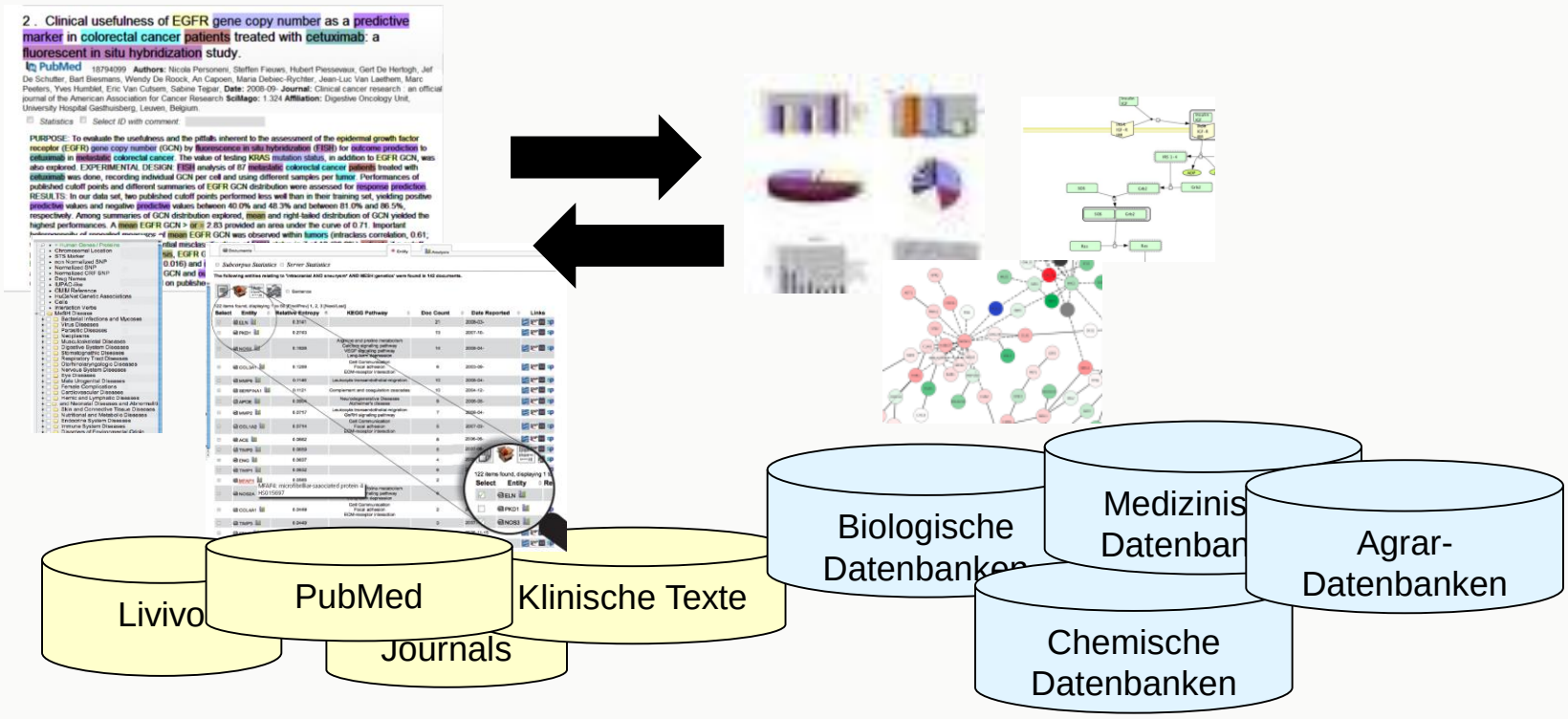
Graphisches Format

- gut für den Leser
- gut zum Versand
- Struktur geht verloren
- aufwendige, immer mit Fehlern behaftete Extraktion
- Massiver Verlust elektronisch verfügbarer Information

Wir brauchen elektronisch lesbare Publikationen,
die außerdem automatisch extrahiert werden dürfen

THESE 2:

KLARER BEDARF, INFORMATION IN LITERATUR MIT INFORMATION IN DATENBANKEN ZU VERBINDEN



INTEROPERABILITÄT ZWISCHEN DATEN UND LITERATUR?

Wichtigster Faktor:
Interoperabilität zwischen Ontologien/Terminologien zwischen Bibliotheken und Daten

Beispiel Lebenswissenschaften:

– MeSH wird in PubMed benutzt, enthält Gen und Proteinnamen!
– Gene/Protein: Referenzen zu Sequenzdatenbanken genutzt

Mapping zwischen den Welten ist notwendig oder automatische Erkennung der gleichen Terminologien/Ontologien

AUTOMATISCHE MÖGLICHKEIT DER INDIZIERUNG

Natural Language Processing (NLP) Methode

Named Entity Recognition (NER) = Erkennung von Terminologieklassen in Texten

Beispiele:

- Personennamen
- E-Mailadressen,
- Maßeinheiten
- Gene
- Krankheitsnamen

AUTOMATISCHE MÖGLICHKEIT DER INDIZIERUNG

Natural Language Processing (NLP) Methode

Named Entity Recognition (NER) = Erkennung von Terminologieklassen in Texten

Beispiele:

- Personennamen: Müller, Meier, Schmidt…

Besonders wertvoll, wenn für die Terminologie noch die Referenz zu einer semantischen Entität gegeben wird

Normalisierung/Mapping: Peter Aloysius Müller

(*25. 10.1955 in Illingen) ist ein ehemaliger deutscher Politiker...

(Quelle Wikipedia)

ERKENNUNG UND NORMALISATION RELEVANTER TERMINOLOGIE

Ist basaler Grundbaustein für

- Informationsretrieval
- Informationsextraction
- Metadatenannotation
- Bau/ Referenz zu Wissensgraphen

ONTOLOGIEN – IN DER BIOLOGIE WEIT VERBREITET

BioPortal

755 Ontologien mit 9,476,739 Klassen

Open Biological and Biomedical Ontology (OBO) Foundry

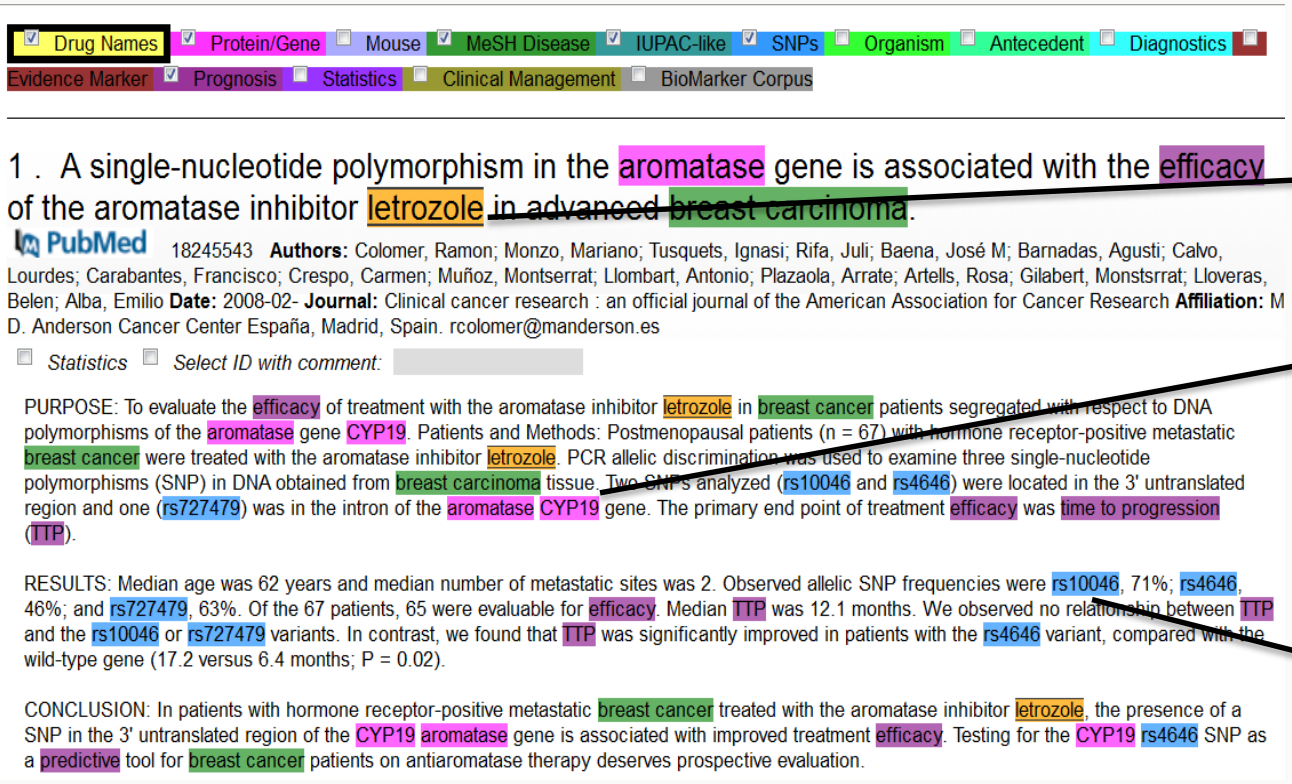
Ontology lookup service

- In der biomedizinischen Domäne existiert eine große Anzahl an Terminologien und Ontologien
- Datenbanken in der Biomedizin nutzen mittlerweile standardisierte Ontologien (Biocurator Community)

ProMiner Normalisierung in SCAIView

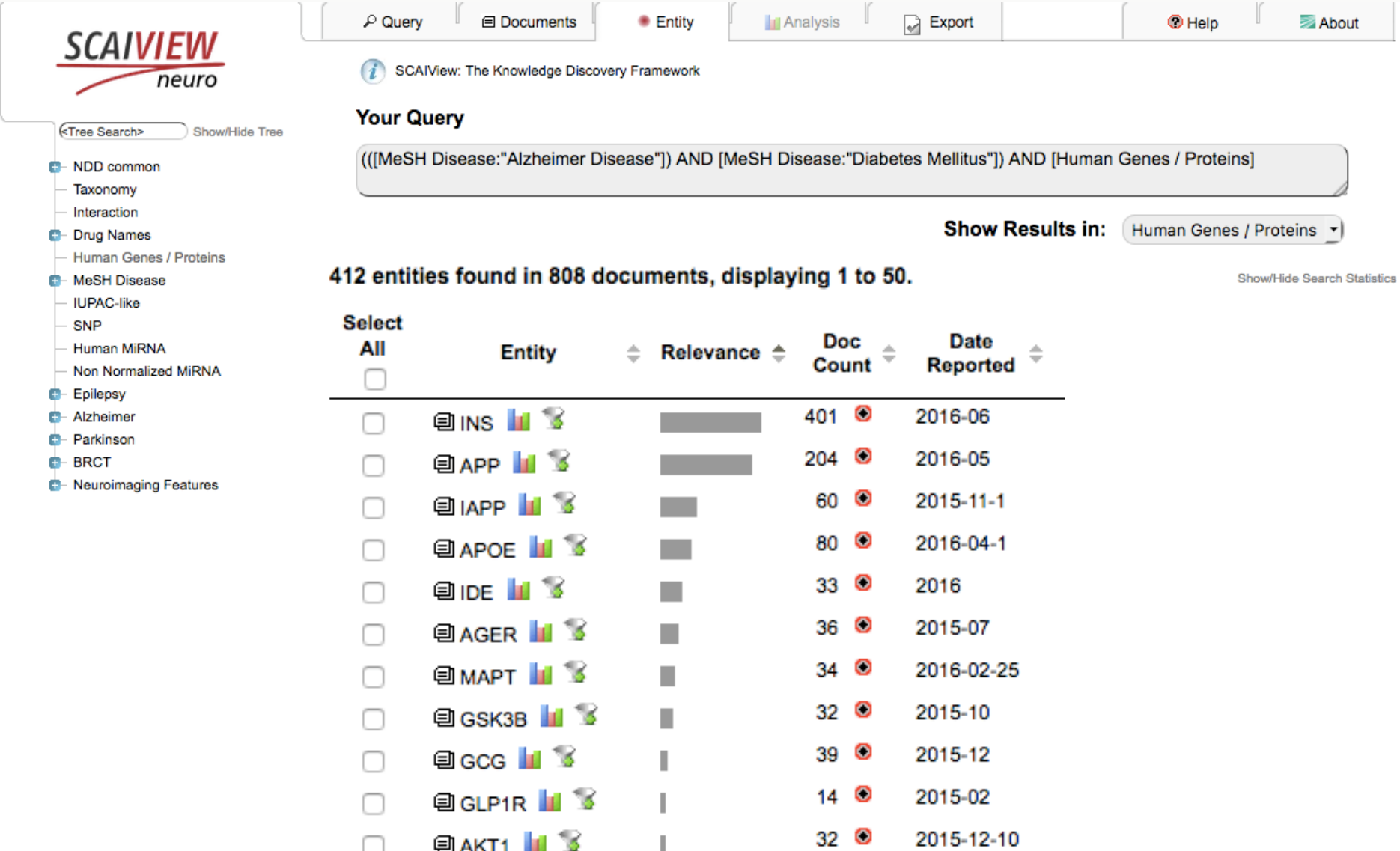
Fraunhofer SCAI

https://www.scai.fraunhofer.de/de/geschaeftsfelder/bioinformatik/produkte.html



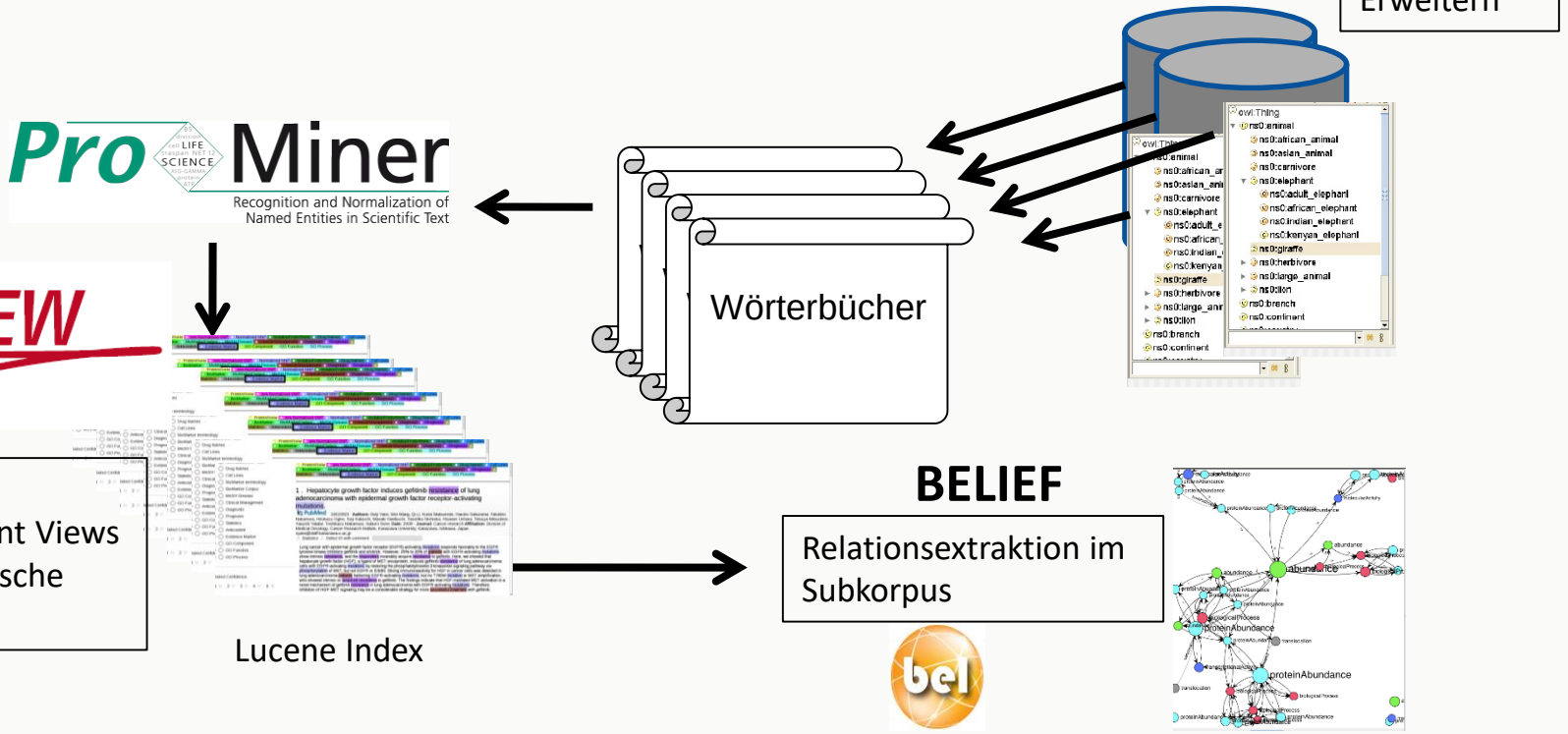
Interoperabilität erlaubt die Verknüpfung von Information und die Bildung von Wissensgraphen

WELCHE GENE SIND BEI DEN KRANKHEITEN ALZHEIMER UND DIABETES RELEVANT?



ZIELGERICHTETE INFORMATIONSEXTRAKTION ZUR ERZEUGUNG VON WISSENSGRAPHEN

Extrahieren
Kuratieren
Erweitern



https://www.scai.fraunhofer.de/de/geschaeftsfelder/bioinformatik/produkte.html

AUFBAU EINER ‚ROUTINE TEXT MINING INFRASTRUKTUR IST KEINE NEBENAUFGABE!

- Workflow-Engines mit einfachen Integrationsmöglichkeiten für neue Anwendungen
- Modularer Aufbau bei komplexen semantischen Suchumgebungen wie SCAIView
- Large Scale Prozessieren mit regelmäßigen Updates von Terminologie/Erkennern/Texten bedeutet eine enorme Herausforderung

➡ **benötigt Einsatz in Compute-Umgebung, Software Infrastruktur und IT- Support**

WIE GUT IST TEXT MINING?

Kommerzielle Lösungen wie z.B. IBM Watson haben viele Hoffnungen geweckt –

Wie bekommen wir ein Einschätzung über die Qualität/Einsetzbarkeit der Lösungen?

- In der akademischen Welt werden Challenges angeboten, um Methoden und Tools zu überprüfen
- Sehr gut Methode, um die Funktionaliät einzuschätzen
- Sollte aus der Entwicklerwelt in die Anwendungswelt überführt werden

‚TAKE HOME MESSAGES‘

- Text und Data mining (Image/Tabellenmining) benötigt aufgearbeitete Formate kein PDF!
- Interoperable Meta-Daten ermöglichen automatische Verknüpfung von Literatur und Daten
- Der Aufbau und Erhalt einer Textminingpipeline erfordert IT Ressourcen (Server, Entwickler und Support)
- Für viele Anwendugnen reichen schon Metadaten/Referenzanalyse oder einfache Termanalysen in entsprechenden Suchumgebungen

Herausforderung: Intiutive Benutzeroberflächen!

Danke für die Aufmerksamkeit!