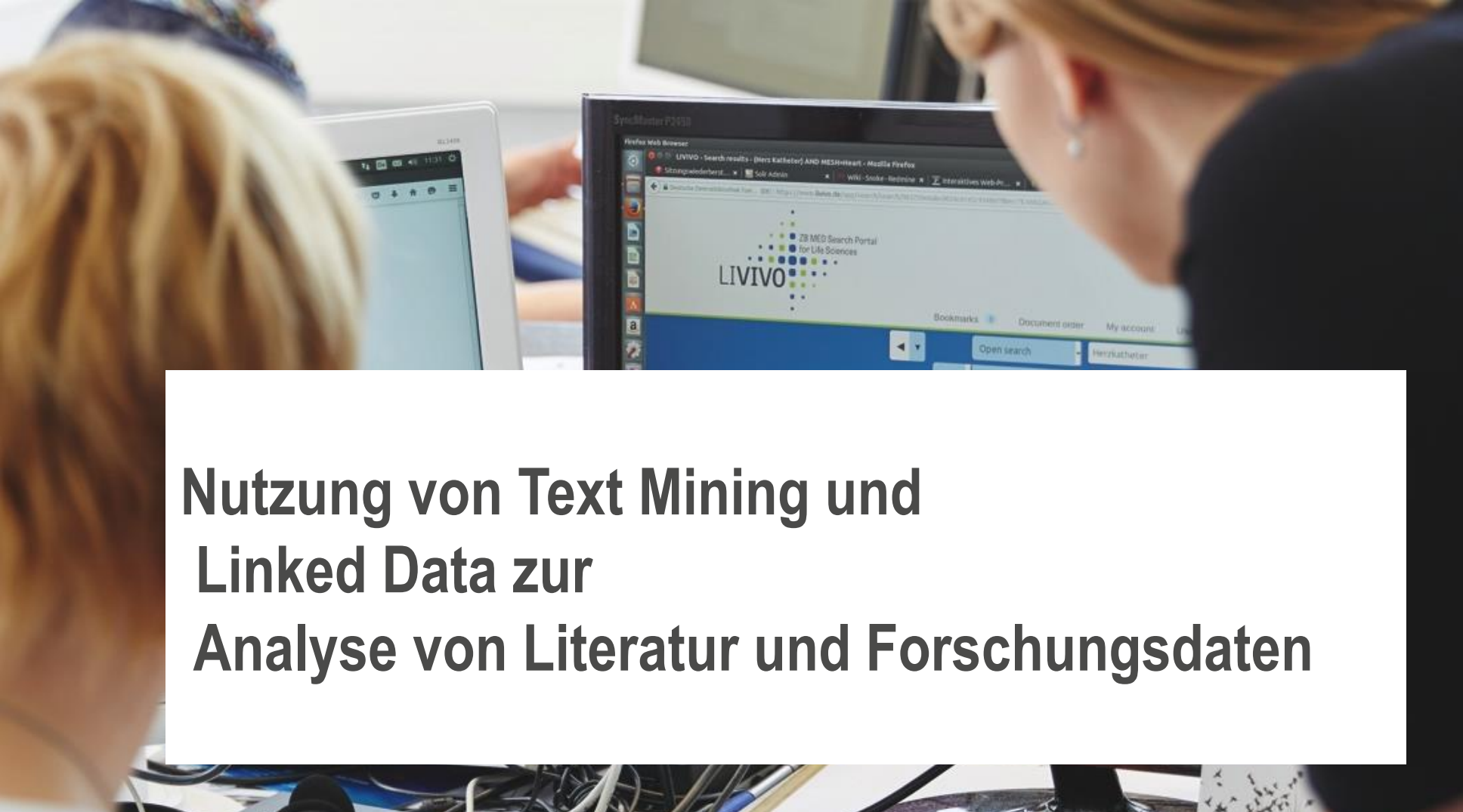




Nutzung von Text Mining und Linked Data zur Analyse von Literatur und Forschungsdaten

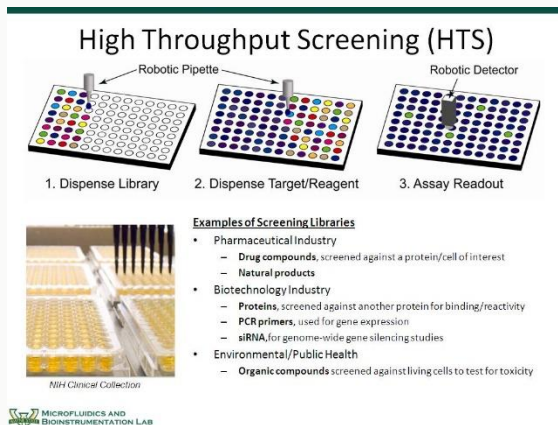
THESE 1

Wachsender Bedarf an elektronisch verfügbarer Semantik

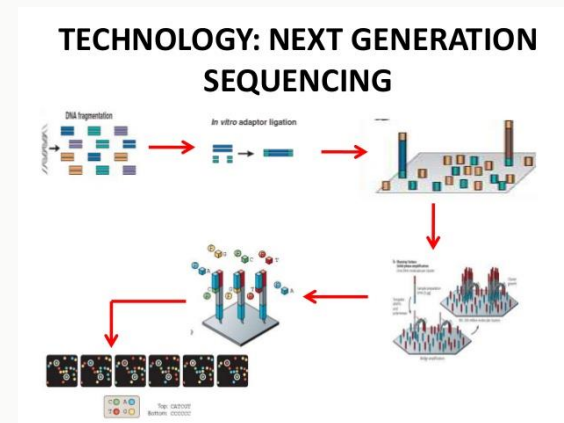
ZUNAHME VON LARGE SCALE EXPERIMENTEN – BEISPIEL BIOMEDIZINISCHE FORSCHUNG

3

- High Throughput Screening (HTS) ➤ Next Generation Sequencing



<https://nanohub.org/>



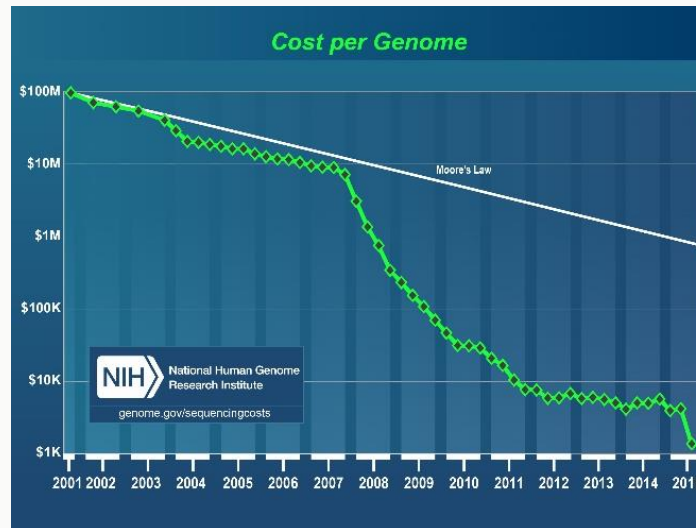
<https://www.slideshare.net/lampweb/>

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

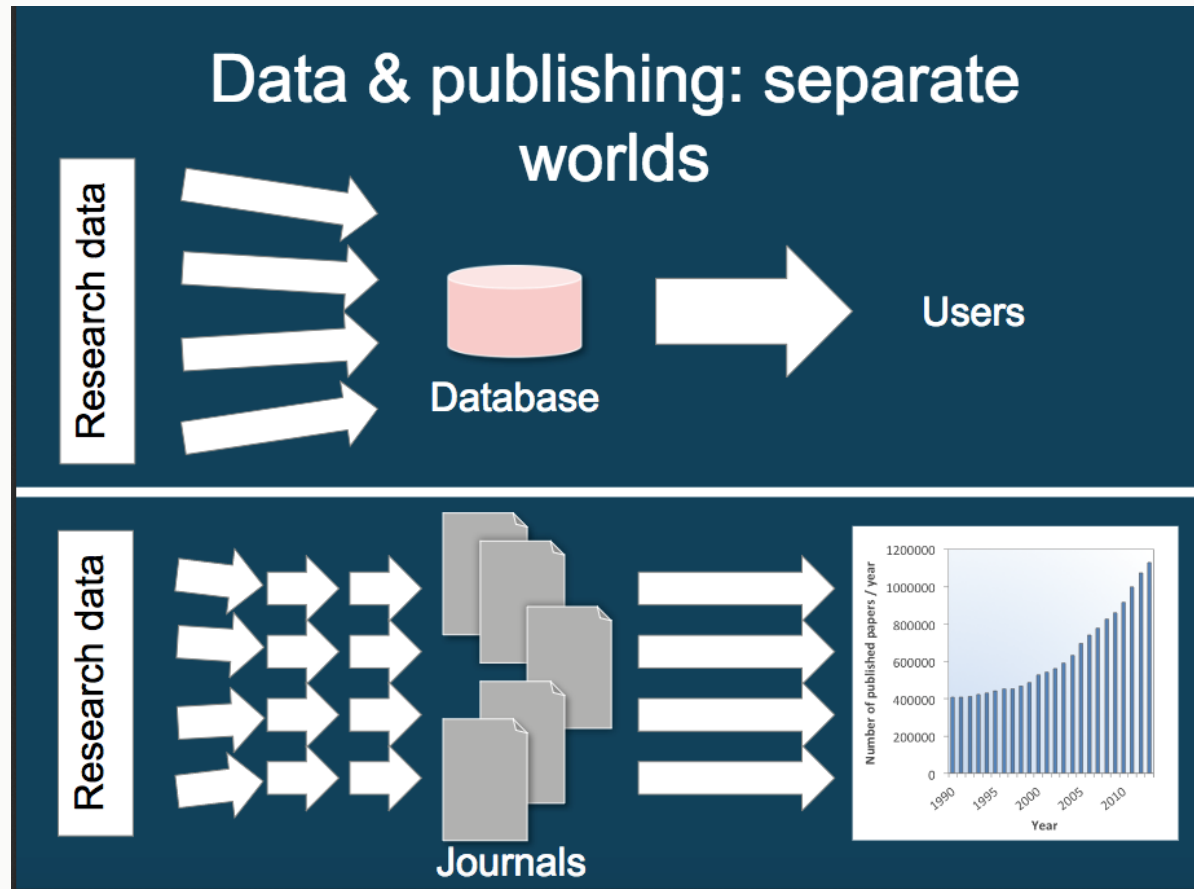
- Wenn Daten abgelegt werden, sind sie nur unvollständig annotiert und selten interoperabel
- Daten Kuration und Integration wird zum Hauptfaktor für Analyse heterogener Daten

Konsequenz auf Ebene der Förderer:

Förderung einer nationalen Forschungsdateninfrastruktur

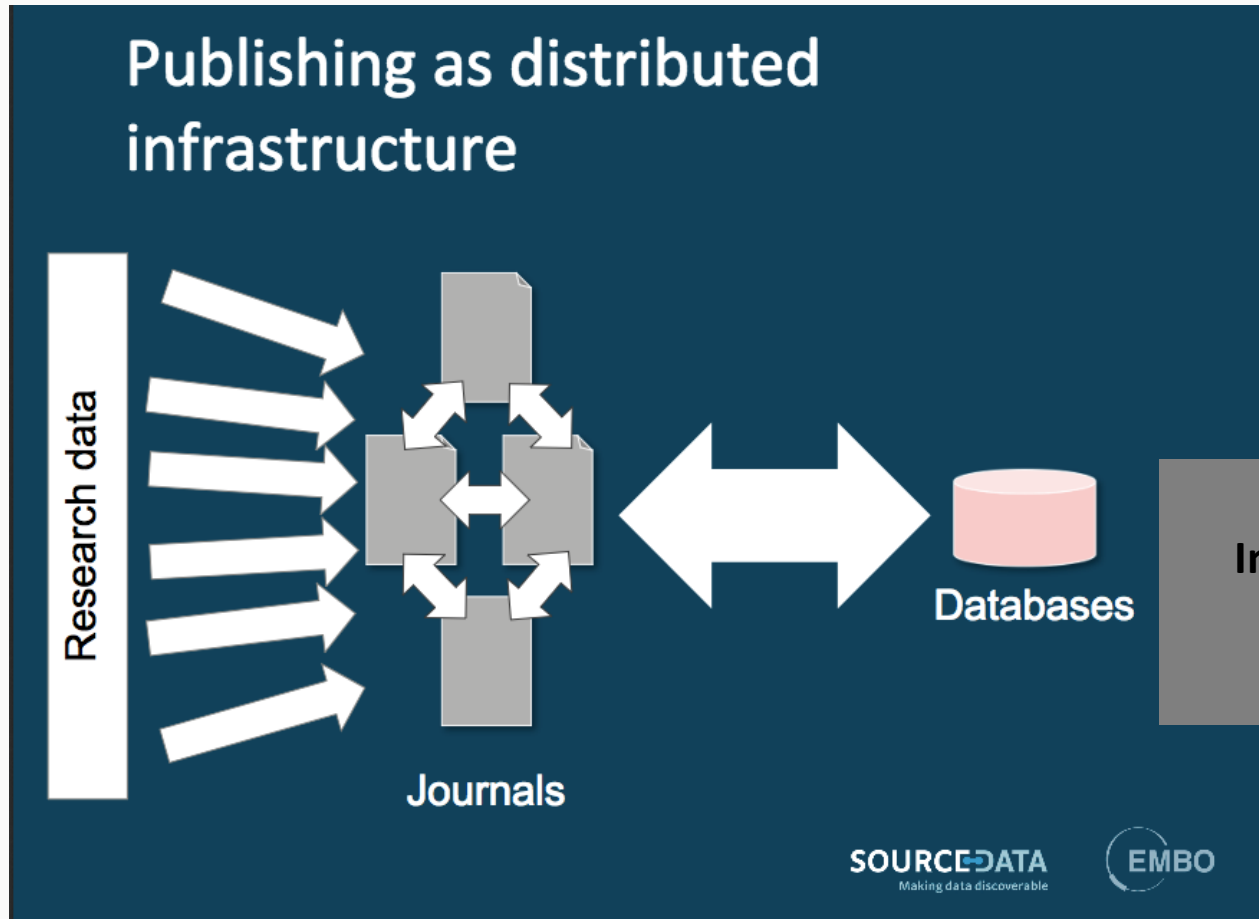
ZB MED ist stark beteiligt, in 2019 Einreichung NFDI4Health (Leitung), NFDI4Agri (Beteiligung)

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

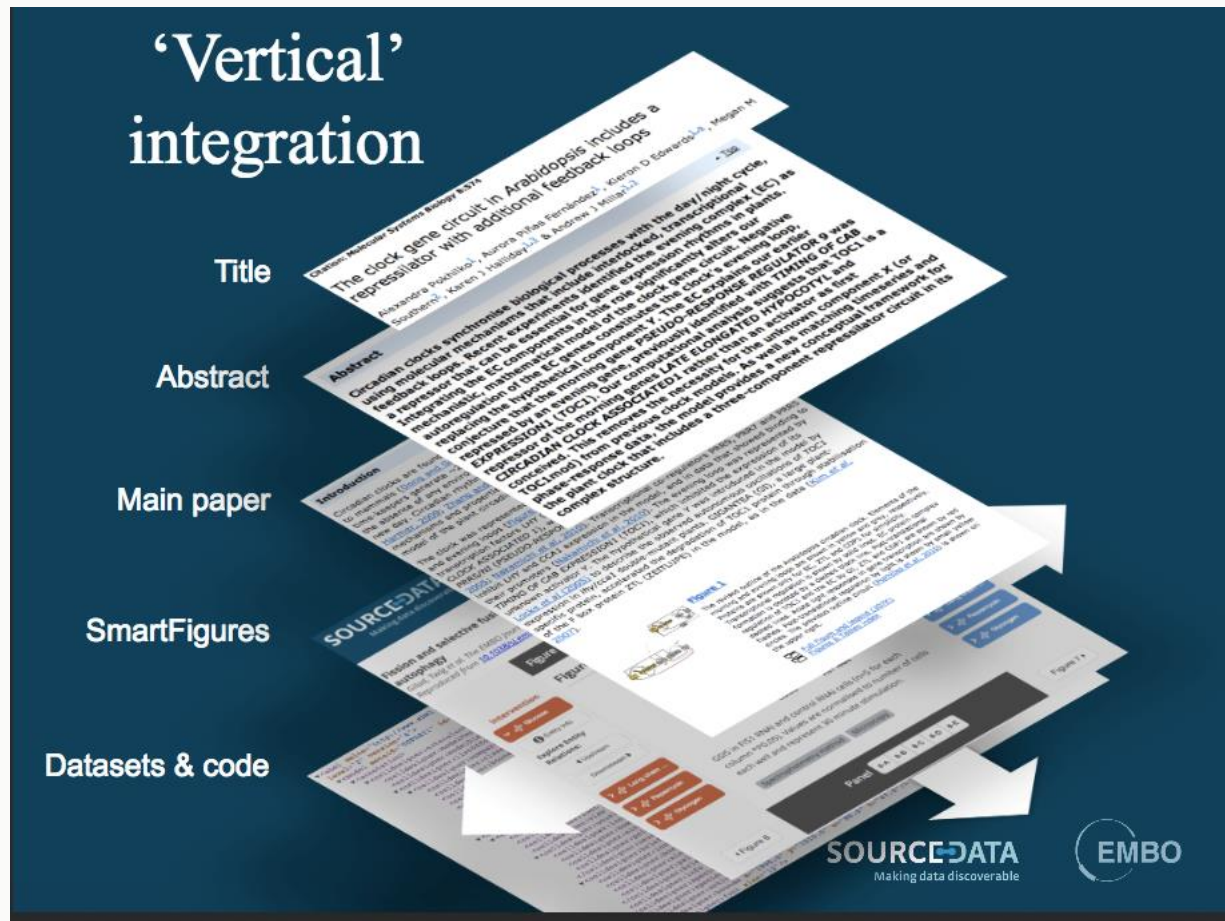


Ein Großteil der Information/des Wissens ist strukturiert und elektronisch verfügbar

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 Excel basierend
- Verlage generieren intern oft XML

PUBLIZIERT WIRD – PDF

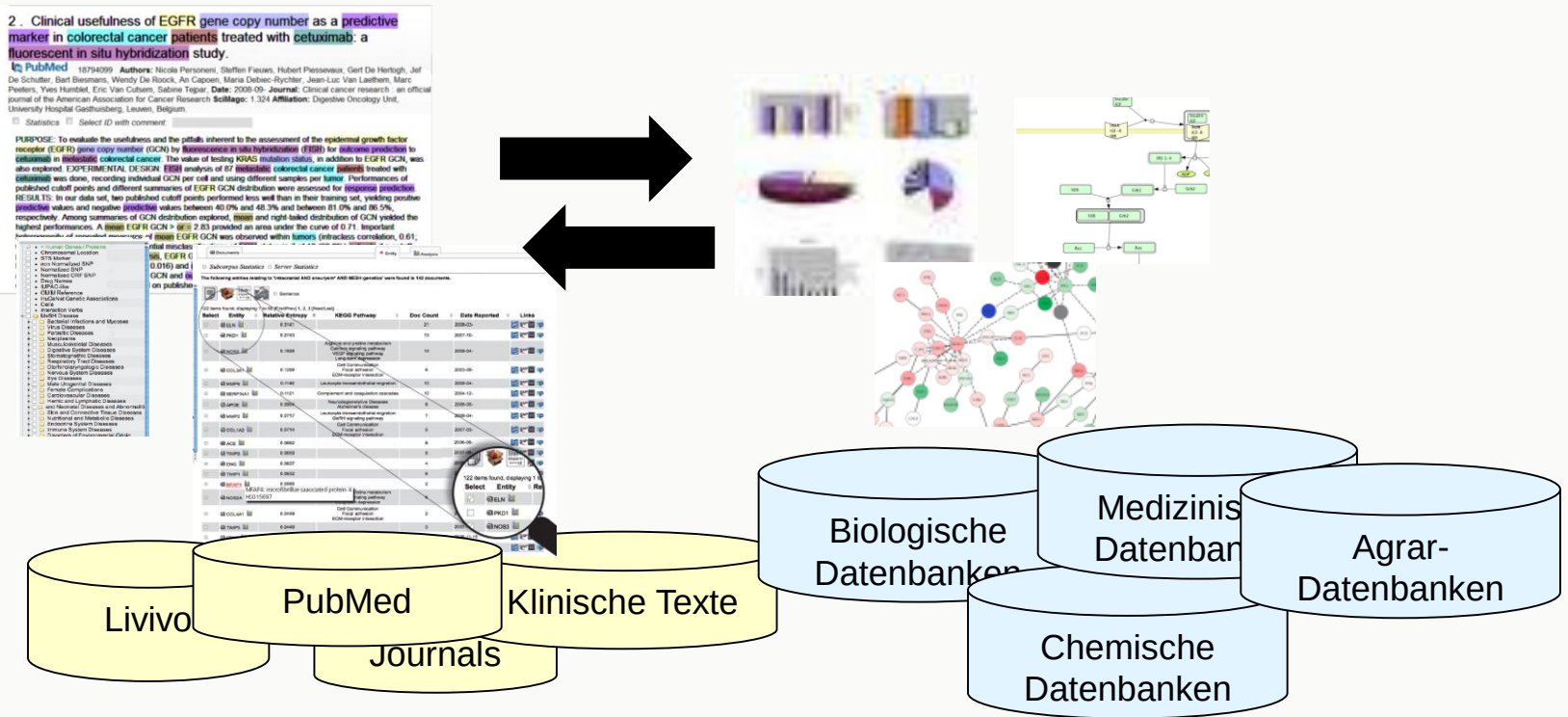
PDF

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**

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



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

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

☒ Drug Names ☒ Protein/Gene ☐ Mouse ☒ MeSH Disease ☒ IUPAC-like ☒ SNPs ☐ Organism ☐ Antecedent ☐ Diagnostics ☐ Evidence Marker ☒ Prognosis ☐ Statistics ☐ Clinical Management ☐ BioMarker Corpus

1. A single-nucleotide polymorphism in the **aromatase** gene is associated with the **efficacy** of the aromatase inhibitor **letrozole** in advanced **breast carcinoma**.

PubMed 18245543 **Authors:** Colomer, Ramon; Monzo, Mariano; Tusquets, Ignasi; Rifa, Juli; Baena, José M; Barnadas, Agusti; Calvo, Lourdes; Carabantes, Francisco; Crespo, Carmen; Muñoz, Montserrat; Lombart, Antonio; Plazaola, Arrate; Artells, Rosa; Gilabert, Montserrat; Lloveras, Belen; Alba, Emilio **Date:** 2008-02- **Journal:** Clinical cancer research : an official journal of the American Association for Cancer Research **Affiliation:** M D. Anderson Cancer Center España, Madrid, Spain. rcolomer@manderson.es

☐ Statistics ☐ Select ID with comment:

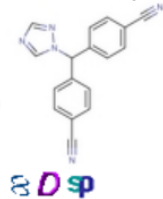
PURPOSE: To evaluate the **efficacy** of treatment with the aromatase inhibitor **letrozole** in **breast cancer** patients segregated with respect to DNA polymorphisms of the **aromatase** gene **CYP19**. **Patients and Methods:** Postmenopausal patients (n = 67) with hormone receptor-positive metastatic **breast cancer** were treated with the aromatase inhibitor **letrozole**. PCR allelic discrimination was used to examine three single-nucleotide polymorphisms (SNP) in DNA obtained from **breast carcinoma** tissue. Two SNPs analyzed (**rs10046** and **rs4646**) were located in the 3' untranslated region and one (**rs727479**) was in the intron of the **aromatase CYP19** gene. The primary end point of treatment **efficacy** was **time to progression** (TTP).

RESULTS: Median age was 62 years and median number of metastatic sites was 2. Observed allelic SNP frequencies were **rs10046**, 71%; **rs4646**, 46%; and **rs727479**, 63%. Of the 67 patients, 65 were evaluable for **efficacy**. Median **TTP** was 12.1 months. We observed no relationship between **TTP** and the **rs10046** or **rs727479** variants. In contrast, we found that **TTP** was significantly improved in patients with the **rs4646** variant, compared with the wild-type gene (17.2 versus 6.4 months; P = 0.02).

CONCLUSION: In patients with hormone receptor-positive metastatic **breast cancer** treated with the aromatase inhibitor **letrozole**, the presence of a SNP in the 3' untranslated region of the **CYP19 aromatase** gene is associated with improved treatment **efficacy**. Testing for the **CYP19 rs4646** SNP as a **predictive** tool for **breast cancer** patients on antiaromatase therapy deserves prospective evaluation.

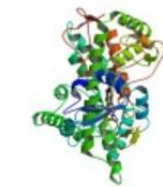
Letrozole:

Letrozole is an aromatase inhibitor used in the treatment of breast cancer.



CYP19A1:

cytochrome P450, family 19, subfamily A, polypeptide 1



Pfam HUGO GO UCSC

rs10046:

Substitution C/T at chromosome 15 location 49290278 CYP19A1 [Illumina Human1M]

SGO SP e! KEGG

Interoperabilität erlaubt die Verknüpfung von Information und die Bildung von Wissensgraphen

WELCHE GENE SIND BEI DEN KRANKHEITEN ALZHEIMER UND DIABETES RELEVANT?

20



Query

Documents

Entity

Analysis

Export

Help

About



SCAIView: The Knowledge Discovery Framework

Your Query

(((MeSH Disease:"Alzheimer Disease")) AND [MeSH Disease:"Diabetes Mellitus"]) AND [Human Genes / Proteins]

Show Results in:

Human Genes / Proteins

412 entities found in 808 documents, displaying 1 to 50.

Show/Hide Search Statistics

Select

All

Entity



Relevance



Doc
Count



Date
Reported



INS

APP

IAPP

APOE

IDE

AGER

MAPT

GSK3B

GCG

GLP1R

AKT1



401

204

60

80

33

36

34

32

39

14

32



2016-06

2016-05

2015-11-1

2016-04-1

2016

2015-07

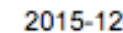
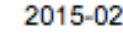
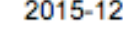
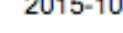
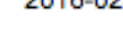
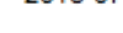
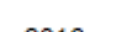
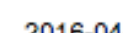
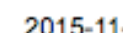
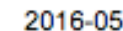
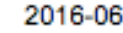
2016-02-25

2015-10

2015-12

2015-02

2015-12-10



<Tree Search>

Show/Hide Tree

- NDD common
- Taxonomy
- Interaction
- Drug Names
- Human Genes / Proteins
- MeSH Disease
- IUPAC-like
- SNP
- Human MiRNA
- Non Normalized MiRNA
- Epilepsy
- Alzheimer
- Parkinson
- BRCT
- Neuroimaging Features



AUFBAU EINER ,ROUTINE TEXT MINING INFRASTRUKTUR IST KEINE NEBENAUFGABE!

22

- Workflow-Engines mit einfachen Integrationsmöglichkeiten für neue Anwendungen
- Modularer Aufbau bei komplexen semantischen Suchumgebungen wie SCAIView
- Large Scale Prozessieren mit regelmäßigen Updates von Terminologie/Erkennen/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 eine Einschätzung über die Qualität/Einsetzbarkeit der Lösungen?

- In der akademischen Welt werden Challenges angeboten, um Methoden und Tools zu überprüfen
- Sehr gute Methode, um die Funktionalität 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 Anwendungen reichen schon Metadaten/Referenzanalyse oder einfache Termanalysen in entsprechenden Suchumgebungen

Herausforderung: Intuitive Benutzeroberflächen!

Danke für die Aufmerksamkeit!