Einführung in die Angewandte Bioinformatik: Datenbanken, Publizieren, ISI, PubMed 30.04.2009

Prof. Dr. Sven Rahmann

Neues Thema

Datenbanken

Daten, Wissen, Datenbanken

Daten (Singular: Datum) :=
alles, das gesammelt, gespeichert und wieder gelesen werden kann.
Daten für sich sind nicht interpretiert, haben keine Bedeutung.

Wissen :=
interpretierte Daten, aus Daten abgeleitete Fakten

Beispiel: 108
Die Ziffernfolge 1-0-8 kann man als die Zahl 108 interpretieren,
diese wiederum als Hausnummer, oder als den 1. August, oder als Geldbetrag, ...

Datenbank (DB) :=
strukturierte Sammlung von Daten

Hinweis zur Notation
Das Zeichen := bedeutet „ist definiert als“, d.h., Sie lesen eine Definition.
Der Doppelpunkt steht auf der Seite des zu definierenden Begriffs.

Inhalt von Datenbanken

Datenbanken enthalten
alles, das gesammelt, gespeichert und wieder gelesen werden kann.
Daten für sich sind nicht interpretiert, haben keine Bedeutung.

Die Datenbank ist nur der „Container“,
der die Daten in strukturierter Form enthält,
so dass sie wiedergefunden werden **können**.

Das ist nutzlos,
solange es keine Möglichkeit gibt, auf die Daten zuzugreifen,
Daten hinzuzufügen, zu löschen, sie zu verändern, etc...

DB / DBMS / DBS

Datenbank-Managementsystem (DBMS) :=
Software, die den Zugriff auf eine Datenbank erlaubt –
bietet überhaupt erst die Möglichkeit, eine Datenbank zu nutzen.
Kann shell-ähnlich sein (man muss Befehle eintippen)
Kann browser-ähnlich sein (graphische Benutzeroberfläche, GUI)
Beispiele: MS-Access, OpenOffice.org Base, mySQL, ...

Datenbanksystem (DBS) :=
DB + DBMS (taucht im Grunde immer zusammen auf,
man kann aber dieselbe DB möglicherweise mit verschiedenen DBMS bearbeiten)

Beispiele für Datenbanken

Allgemeine Beispiele

- Kundendatenbanken (besitzt jede Firma)
- Adressdatenbanken (hat mein email-Programm; auch mein Adressbuch ist eine)
- Filmdatenbanken (z.B. imdb.com)
- Literaturdatenbanken (amazon braucht so etwas für alle Arten von Literatur)

Beispiele, die uns in dieser Vorlesung interessieren

- wissenschaftliche Literaturdatenbanken (z.B. PubMed – später)
- biologische Sequenzdatenbanken
  - für DNA / speziell für RNA / UniProt für Proteinsequenzen
- Proteinstrukturdatenbanken, z.B. PDB
- Datenbanken zu Reaktionsnetzwerken
  - metabolische Netze
  - Protein-Reaktionsnetzwerke / Signaltransduktionsnetzwerke

Textdateien („flat files“) als Datenbanken

Textdateien („flat files“)

Eine Textdatei ist eine Datei, in der die gespeicherte Information direkt, ohne Formatierung oder Meta-Informationen steht.
Beispiel: FASTA-Dateien,
Gegenbeispiel: Word (.doc)-Dateien mit Formatierung und Autoren-Information.

Historische Bedeutung von Textdateien als Datenbanken

Datenbank + DBMS: Infrastruktur nötig; Aufwand
1970er: Erste DNA-, Protein-Sequenzen (wenige!)
einfach und übersichtlich in Textdatei zu speichern
gut austauschbar
„historisch gewachsen“; heute noch immer Verwendung von Textdateien!

Vor- und Nachteile von „flat files“ als Datenbanken

Vor- und Nachteile von Textdateien

+ Man kann sie mit jedem Text-Editor lesen und bearbeiten
  (Shell-Befehle: `cat`, `more`; keine spezielle Software!)
- Suche oft ineffizient
- Datenintegrität und Datensicherheit werden nicht unterstützt

Vor- und Nachteile „echter“ Datenbanken

- Man benötigt ein DBMS zum Zugriff.
  Wer das „richtige“ DBMS nicht hat, kann auf die Informationen nicht zugreifen.
+ Erlaubt effiziente Suchanfragen (z.B. durch Erstellen eines Index)
+ Das DBMS kümmert sich um Integrität und Zugriffsrechte.

Neues Thema

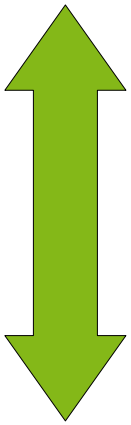
Wissenschaftliches Publizieren und Zitations-Datenbanken

Grundlagen wissenschaftlicher Veröffentlichungen

Wissenschaftler werden aus Steuergeldern bezahlt.
Forschungsergebnisse sollen der Allgemeinheit zur Verfügung stehen.
Es gibt mehrere Möglichkeiten, Ergebnisse zu veröffentlichen.

Arten der Veröffentlichung

- Konferenzbeiträge
- Forschungsartikel in Fachzeitschriften
- Übersichtsartikel in Fachzeitschriften
- Buchkapitel
- studentische Abschlussarbeiten
- Lehrbücher



Neue Resultate
schnelle Veröffentlichung

ältere Resultate
langsame Veröffentlichung

Wer trägt die Kosten wissenschaftlicher Veröffentlichungen?

2 Modelle:

Klassisch:
Der Leser (durch Bestellen der Zeitschrift)

Neuer („open access“):
Der Autor bezahlt den Verlag für die Druckkosten,
häufig gibt es „flat rates“ für eine ganze Universität oder Institute.

Der Verlag, nicht der Forscher, verdient Geld mit Publikationen.

Qualitätssicherung wissenschaftlicher Veröffentlichungen

Grundprinzip

Veröffentlichung von Forschungsergebnissen bedeutet Fortschritt:
Andere machen weiter, wo einem selbst die Ideen ausgegangen sind.
Schlecht, wenn (viel) Unsinn veröffentlicht wird.
Generell sollte gelten: Was gedruckt wird, stimmt.

Qualitätssicherung durch Peer Review

Geschriebene Artikel werden nicht einfach gedruckt.
Andere Fach-Wissenschaftler lesen und kommentieren die Artikel vor Erscheinen.
Wenn alle Fragen und Bedenken ausgeräumt sind, wird der Artikel gedruckt.
„Peer Review“: Kollegen (an anderen Instituten) begutachten die Arbeit.
peer: Gleichrangiger, Gleichgestellter
Problematisch: Kostet viel Zeit und Arbeit, wird nicht bezahlt.
Aber: Ohne peer review bricht guter Wissenschaft das Fundament weg.

Bewertung wissenschaftlicher Veröffentlichungen

Grundidee

Nicht alle Artikel sind gleich wichtig, vor allem auf lange Sicht.
Man zitiert einen Artikel, wenn man sich in seiner Arbeit darauf beruft.
Richtungsweisende Artikel werden häufiger gelesen und häufiger zitiert.

Versuch einer Einteilung

„Write-only“-Artikel:
- werden nicht gelesen
- Verschwendung von Zeit und Geld
- dienen dazu, die eigene Publikationsliste zu verlängern

Kleine Fortschritte, Verbesserungen bestehender Resultate:
- häufigste Art der Publikation

Große Fortschritte, Lösung langer offener Probleme:
- seltener, erzeugen größere Aufmerksamkeit in der Fachwelt

Grundlagen eines neuen Forschungsfeldes:
- sehr sehr selten, Nobelpreisverdächtig

Qualitätsstandards wissenschaftlicher Zeitschriften

Verschiedene Zeitschriften haben verschiedene Qualitätsstandards.

Immer:
- Die Arbeit muss methodisch einwandfrei sein;
- keine offensichtlichen Fehler!

Unterschiede bei:
- Wie groß ist der erzielte Fortschritt?
- Wie aktuell ist das Thema gerade?

Ziel jedes Wissenschaftlers:
- möglichst „hoch“ publizieren („gute“ Zeitschrift)
- eigentlich Unsinn, denn es sollte um den Inhalt des Artikels gehen; Zeitschrift lebt von guten Artikeln, nicht Artikel von guten Zeitschriften (oder doch??)

„Einfluss“ von Zeitschriften

Wie den Einfluss (impact) eines *Artikels* messen?
Übliches Maß: impact factor der *Zeitschrift*!

- Robuster zu berechnen als für jeden Artikel einzeln
- Ungenauer: unbedeutende Artikel können in guten Zeitschriften erscheinen

Zeitschriften achten sehr auf einen hohen impact factor,
und akzeptieren nur Artikel, die diesen vermutlich halten oder übertreffen.

Definition des Impact Factors

$$\frac{\text{Zahl der Zitate im Jahr } x \text{ auf Artikel der Zeitschrift in den Jahren } x\text{-}2 \text{ und } x\text{-}1}{\text{Zahl der Artikel der Zeitschrift in den Jahren } x\text{-}2 \text{ und } x\text{-}1}$$

Kritik

IF misst durchschnittliche Zitierhäufigkeit aller Artikel einer Zeitschrift

- Durchschnitt sagt nichts über individuelle Artikel
- Zitierhäufigkeit sagt etwas über Bekanntheit, nicht Relevanz
- Zeitschrift sagt nichts über individuelle Forscher;
  trotzdem ist kumulativer / durchschnittlicher IF wichtig bei Bewerbungen

Alternative

- Hirsch-Index eines Wissenschaftlers [Übung]

Ermittlung des Impact Factors

Wer macht das mit welchem Aufwand?

Thomson Scientific (früherer Name: Insititue of Scientific Information, ISI)
http://scientific.thomson.com/free/essays/journalcitationreports/impactfactor/
Literaturliste jeder(!) Veröffentlichung durchgehen,
in Bezug zu vergangenen Veröffentlichungen setzen
Manuelle Nachbearbeitung; teuer!

Datenbank dieser Verknüfpungen:

Science Citation Index Entended (SCIE) im ISI Web of Science / of Knowledge
http://www.isiwebofknowledge.com/

Zukunft:

Automatisierte Suchmaschinen, z.B. Google Scholar (http://scholar.google.com)
Grundlegende Probleme dabei: Korrektheit, Vollständigkeit

Neues Thema

Literatur-Datenbanken

Literatursuche

Wichtige Frage, bevor man ein Forschungsprojekt beginnt:

„Was gibt es schon?“

Um dies herauszufinden, benutzt man Literaturdatenbanken.

Früher:
spezielle Review-Zeitschriften mit Zusammenfassungen anderer Artikel

Datenbanken und Systeme zur biomedizinischen Literatur

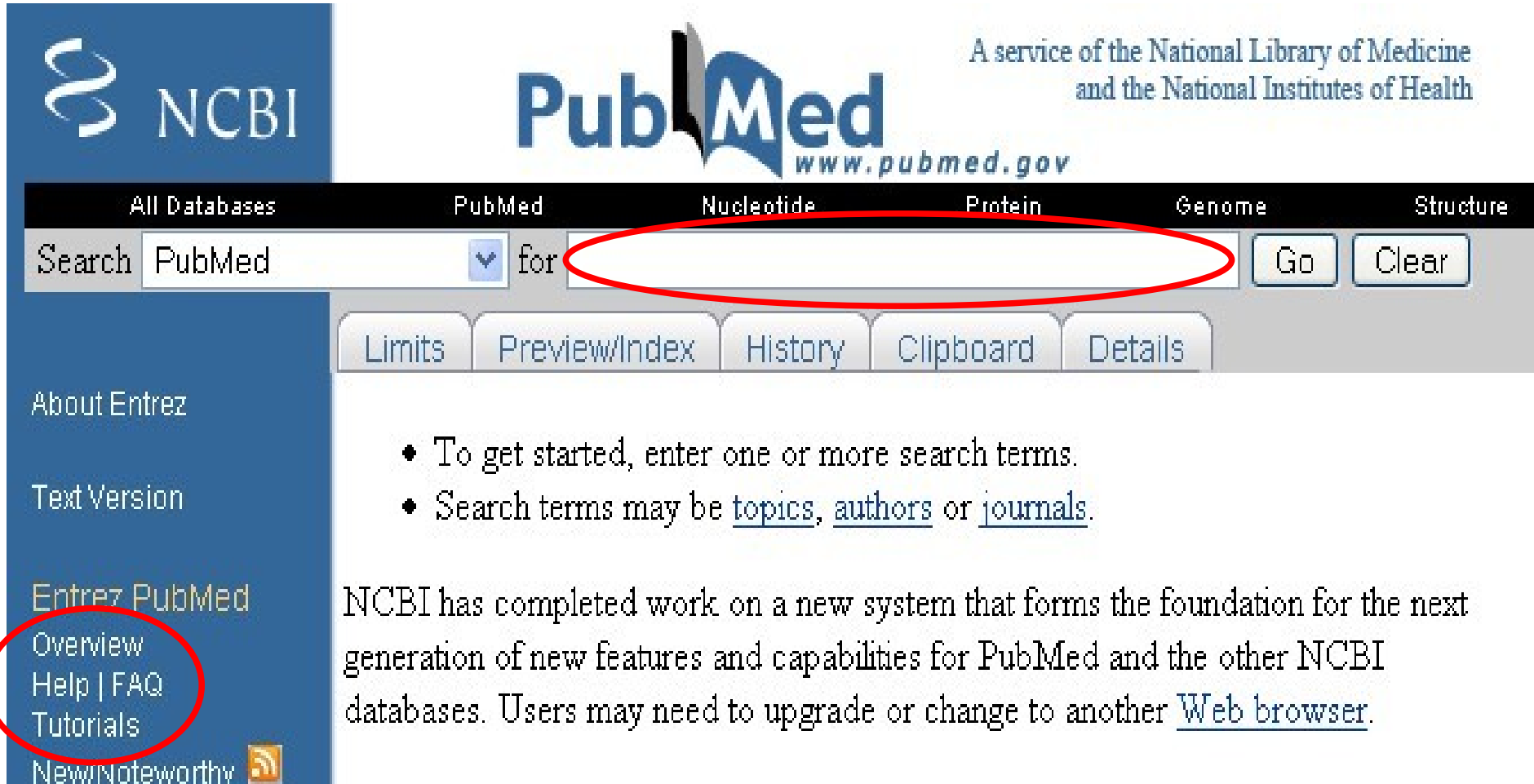
MEDLINE :=
öffentliche Datenbank mit (im weitesten Sinne biomedizinischen) Artikeln, Querverweisen, Zusammenfassungen, ca. 5000 verschiedene Zeitschriften

PubMed :=
frei zugreifbares online-System, das die MEDLINE-Datenbank enthält und komplexe Abfragen erlaubt [damit befassen wir uns jetzt!]

PubMedCentral :=
frei zugängliches digitales Archiv von Artikeln aus den Lebenswissenschaften

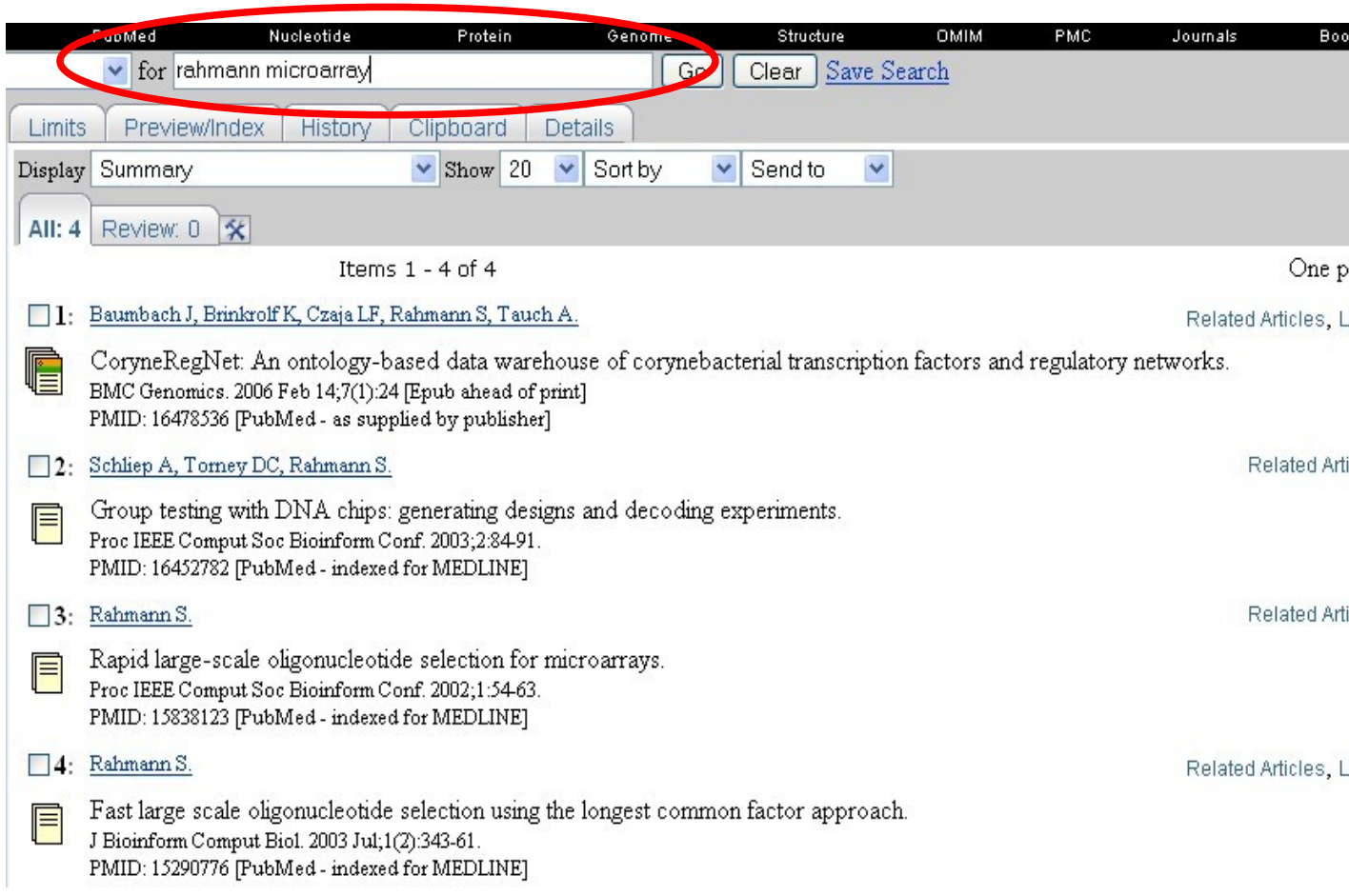
Entrez :=
System, das eine gemeinsame Oberfläche und Abfragesystem für PubMed und andere Datenbanken bietet

Zugriff auf MEDLINE mit PubMed über Entrez unter
http://www.ncbi.nlm.nih.gov/sites/entrez/
(Was sind NCBI, NLM, NIH?)



Einfache PubMed-Suche
http://www.ncbi.nlm.nih.gov/sites/entrez/

Stichwörter in das „for“-Feld eingeben, z.B. Autorennamen, Titel-Teile



PubMed-Suchfunktionen

Ähnlich wie bei Internet-Suchmaschinen:

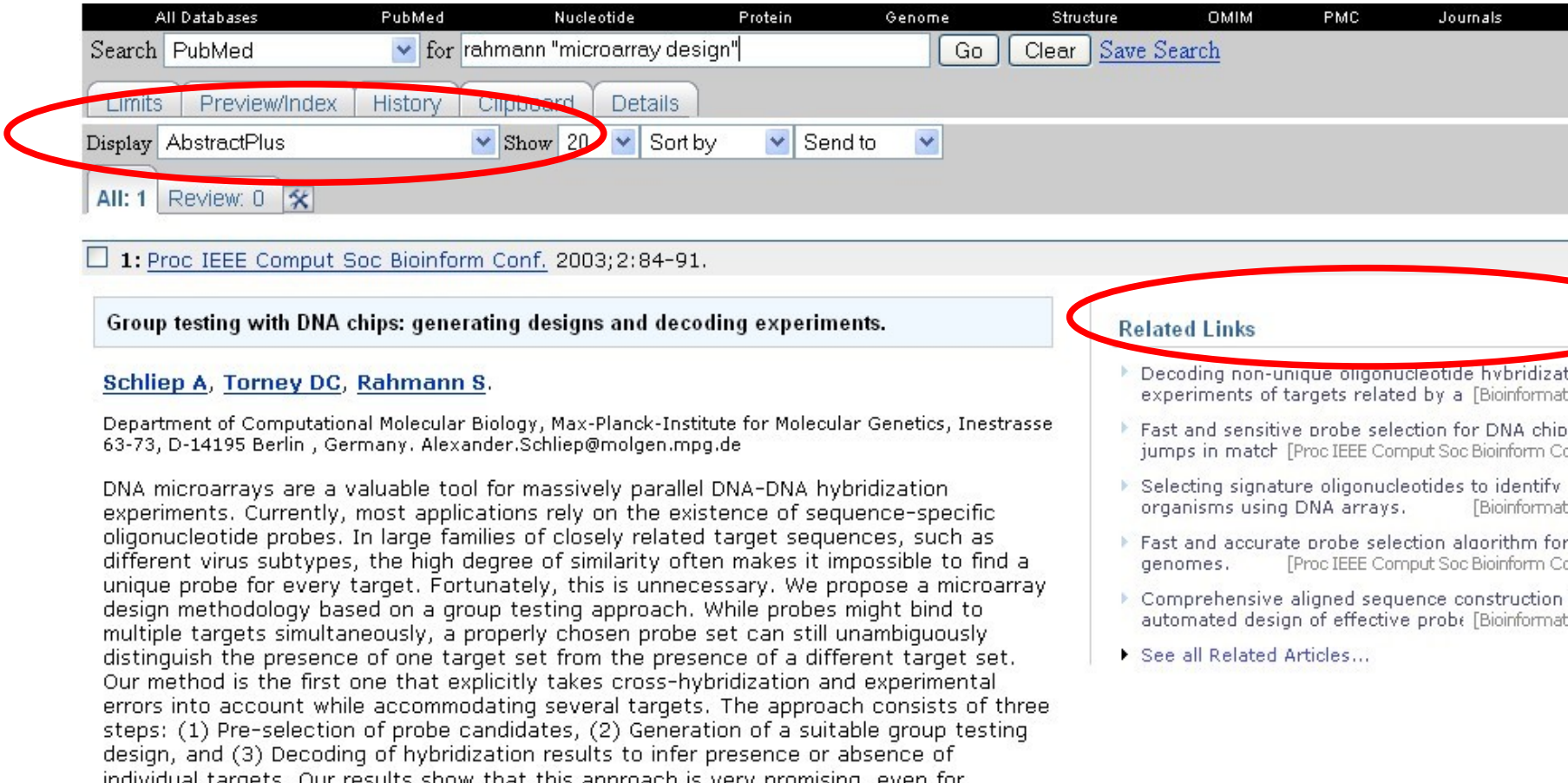
- Alle eingegebenen Wörter müssen vorkommen
- Reihenfolge der Wörter spielt keine Rolle
- Anführungszeichen legen Zusammenhang und Reihenfolge der Wörter fest
- Man kann generell nach Autor, Titel, Wörtern im abstract (Zusammenfassung), Jahreszahlen, ..., suchen.
- Groß - und Kleinschreibung spielt keine Rolle.

Beispiele

- Rahmann Microarray
  es wird eine Liste der passenden Artikel angezeigt
- Rahmann “Microarray Design”
  bei nur einem Treffer werden direkt mehr Details gezeigt

Zugriff auf weiterführende Informationen

Display-Auswahlbox: „Abstract Plus“; sowie „Related Links“ rechts



Feldrestriktionen

Problem

Die Suchanfrage **Down Syndrome** findet so ziemlich alles:
Artikel über das Down-Syndrom, Artikel von Dr. Down über irgendein Syndrom, ...

Lösung

Man legt fest, in welchem Feld (Autor, Titel, ...) man jeweils suchen will.

[PMID]	PubMed ID; eindeutige Nummer, die einem Artikel zugeordnet ist
[TI]	Titel
[AB]	Abstract, Zusammenfassung
[AD]	Adresse des Instituts des publizierenden Autors
[FAU]	Full author name, Voller Name des Autors
[AU]	Autor (Nachname + Initialen)
[SO]	(Abkürzung des Zeitschrift-Namens)

Feldrestriktionen - Beispiele

Rahmann S [AU]
: alle Artikel von allen Leuten, die S. Rahmann heißen

„Down Syndrome“ [TI]
: zusammenhängender Begriff im Titel

down [AU] AND syndrome [TI]
: Artikel von Dr. Down über Syndrome

12345678 [PMID]
: das durch diese Nummer eindeutig identifizierte paper

Tipp bei der Literatursuche:
PMIDs von relevanten Artikeln notieren

Boole'sche Verknüpfungen

Hintergrund des Namens:

George Boole, engl. Logiker, 1815 – 1864

Operatoren

AND (und, Standard, alle Terme müssen vorkommen)
OR (oder, ein Term muss vorkommen)
NOT (nicht, der Term darf nicht im genannten Zusammenhang vorkommen)

Beispiele

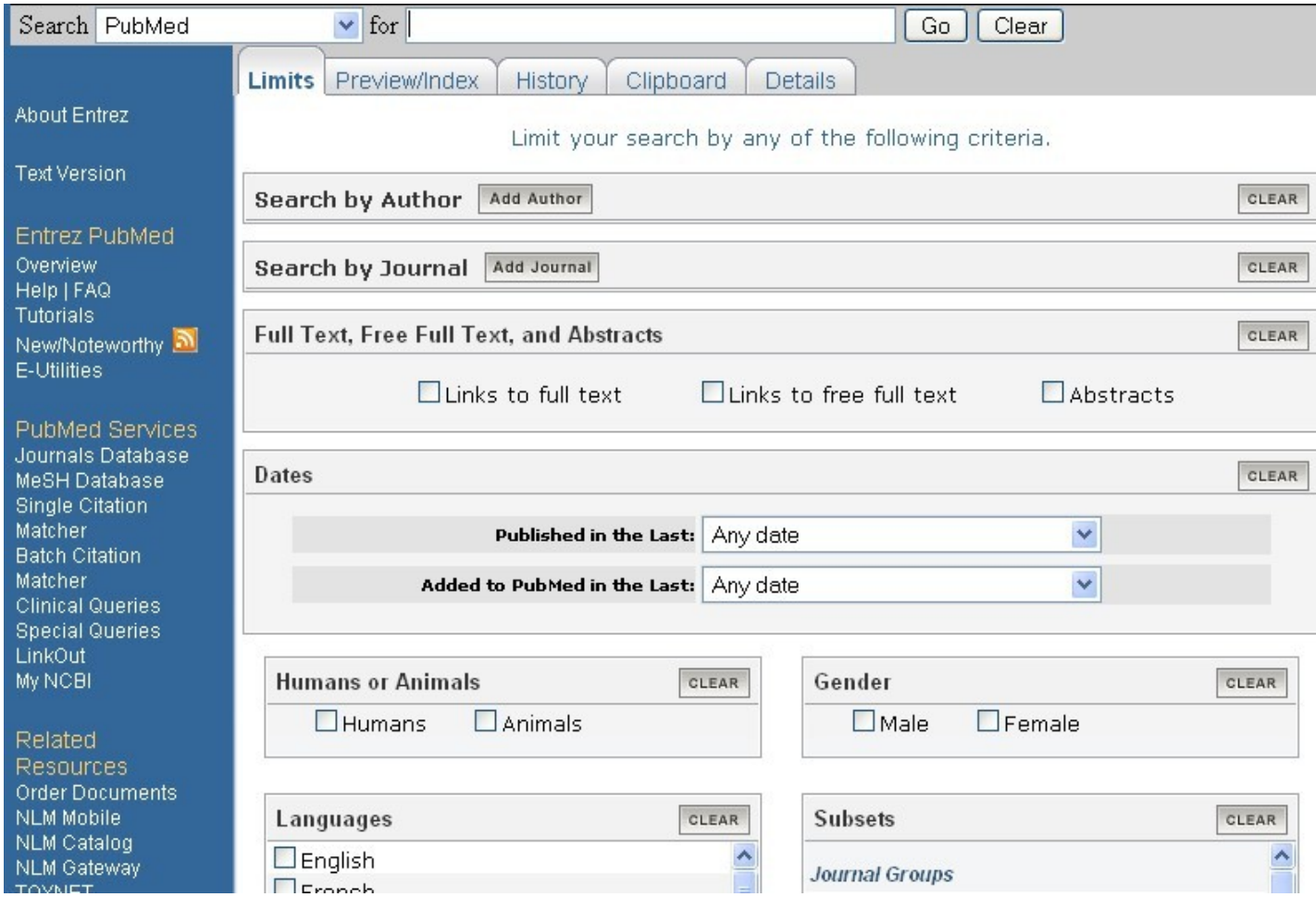
microarray [ti] AND Dortmund [ad]

Finde Microarray-Experten in Dortmund

microarray [ab] NOT Rahmann S [au]

Finde Arbeiten über Microarrays, die nicht von S.Rahmann sind

Limits



Statt die Suchanfrage „per Hand“ einzugeben,
kann man die Einschränkungen bequemer in einer Suchmaske eingeben;
aber: weniger flexibel bei komplexen Anfragen

Tipps beim Suchen

- Anführungszeichen benutzen, wo möglich
- Initialen der Autoren benutzen, wenn bekannt
- PMIDs notieren
- Bei zu vielen Ergebnissen, Suche weiter einschränken
- Bei zu wenig Ergebnissen, Suche erweitern

Problem: Synonyme
Dasselbe Konzept, dieselbe Idee
wird durch verschiedene Wörter oder Wortkombinationen ausgedrückt.
Man müsste alle ausprobieren, um sicher zu sein, alles zu finden!

Lösung
Standardisiertes Vokabular: MeSH-Terme (Medical Subject Headings)

MeSH-Terme

MeSH – Medical Subject Headings :=
standardisiertes und kontrolliertes Vokabular
zur Indizierung von Artikeln in MEDLINE / PubMed

MeSH-Terme erlauben,
auf konsistente Weise Informationen zu Themen zu erhalten,
die sich mit verschiedenen Begriffen beschreiben lassen.

Neues Problem dabei

Woher bekomme ich die richtigen MeSH-Terme zu einem Thema?

Lösung

Durchsuche die MeSH-Datenbank,
die wie PubMed über Entrez zugänglich ist

Beispiel zu MeSH-Termen

Suche nach MeSH-Term, der bioinformatische Arbeiten zur Lösung des Microarray-Design-Problems umfasst.

Dann Verwendung in PubMed Suche mit Feldnamen [MH],

rahmann [AU] AND „oligonucleotide array sequence analysis“ [MH]