Università degli Studi di Trieste

Corso di Laurea Magistrale in
INGEGNERIA CLINICA

NATURAL LANGUAGE PROCESSING: ESEMPIO APPLICATIVO

Corso di Informatica Medica

Docente Sara Renata Francesca MARCEGLIA

Dipartimento di Ingegneria e Architettura

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

DA DOCUMENTO NON STRUTTURATO…

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

EXAMPLE 1 – Medical report at enrollment visit

Piacenza, Nov 12, 2015
Case History. Mr. John Doe complains having difficulty in hearing in both ears since 10 years ago when he had a car accident resulting in head injury. He suffers from diabetes type 2, hypertension and arteriosclerosis. He reports tinnitus in both ears since last 2 years. He was diagnosed with Meniere. No family history of hearing loss.
Mr. Doe denies ear pain, ear fullness and balance problems.

Evaluation.
Otoscopic inspection: normal bilaterally
Immittance: type A bilaterally
Pure tone audiometry: bilateral sensorineural hearing loss at 2000-8000 Hz, moderate at right ear, profound at left ear.
Speech audiometry: max intelligibility 100% at 60 dB right ear and 0% at 90 dB at left.
Word discrimination: 80%
Word recognition: 85%
Sound field speech audiometry in noise: max intelligibility 35%, S/N 5 dB.
CT scan: normal
MRI: normal

Psychometric Results
Impact of tinnitus: THI score 76, grade 4

Conclusions

Based on history and current exams, Mr. Doe has a bilateral sensorineural hearing loss of profound degree for the left ear with Meniere disease and tinnitus. It is recommended to have a cochlear implant at the left ear.

… A DOCUMENTO STRUTTURATO

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Section	Sub-section	Coded concept	Auditory disorders axis (and sub-class)	Risk factor group
Medical history evidences	Presenting complaint(s)	[C1384666] Hearing impairment	Hearing disorder ()	-
	Significant diseases/illnesses/disorders/signs having still present relevance	[C0011860] Diabetes Mellitus, Non-Insulin-dependent	-	General medical conditions
		[C0020538] Hypertensive disease	-	General medical conditions
		[C0003850] Arteriosclerosis	-	Vascular disease
		[C0040264] Tinnitus	Hearing disorder ()	-
		[C0025281] Meniere Disease	Ear disorders (peripheral vertigo)	-
	Traumatic injuries	[C0744612] Head injury trauma	-	General medical conditions
	Family history of HL	None	-	-
Examinations – audiometric	Audiometric thresholds for tones	Date: 2015/11/12. Diagnostic procedure: [C2022139] Pure tone threshold audiogram. Type of HL: [C0018784] Sensorineural Hearing Loss (disorder). Laterality: [C0238767] Bilateral. HL degree right ear: [C0205081] Moderate (severity modifier). HL degree left ear: [C0439808] Profound. Range of frequencies (Hz): 2000/8000.	Hearing disorder (hearing loss)	-

NATURAL LANGUAGE PROCESSING

- È possibile effettuare analisi su testi non strutturati?
- È possibile estrarre informazione anche se il testo non è codificato/standardizzato?

↓

TECNICHE DI NATURAL LANGUAGE PROCESSING

Esempio: http://linguistic-annotation-tool.italianlp.it

I PASSI PRINCIPALI DEL NLP

SENTENCE TOKENIZER

- Identificazione dei confini delle frasi (token)
- Può considerare parole (congiunzioni), spazi, o segni di punteggiatura

PART OF SPEECH (POS) TAGGER

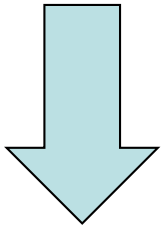
- Definizione della categoria grammaticale (nome, verbo, etc)

PARSER

- Identificazione delle relazioni grammaticali tra le parti identificate

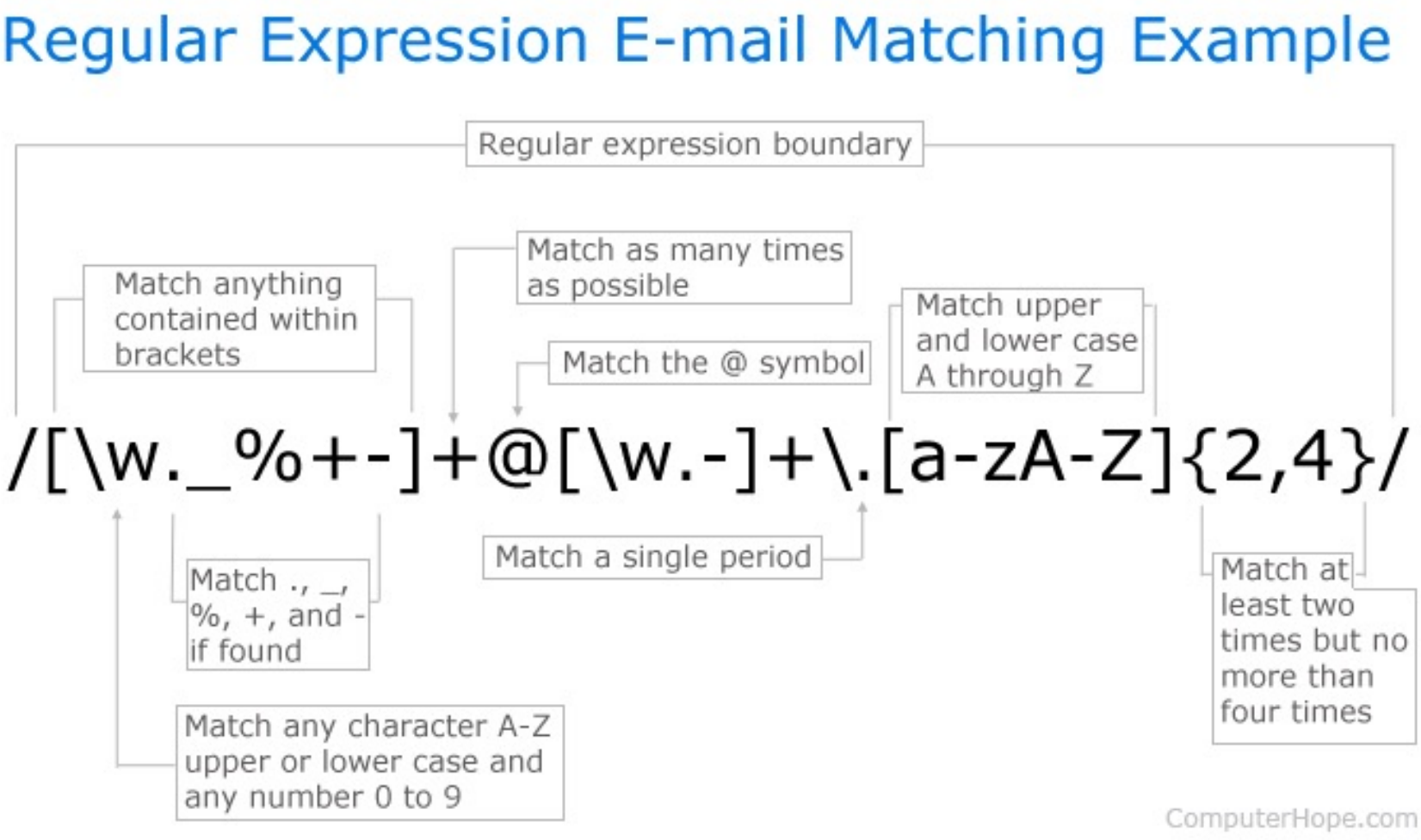
SENTENCE TOKENIZER

OCULAR COMPLICATIONS OF MYASTENIA GRAVIS.



OCULAR	COMPLICATIONS	OF	MYASTENIA	GRAVIS	.

REGULAR EXPRESSIONS



https://regex101.com/

EHR EXAMPLE: MONICA LATTE

https://www.ahrq.gov/ncepcr/tools/pf-handbook/mod8-app-b-monica-latte.html

Review of Systems
General: denies fatigue, malaise, fever, weight loss
Eyes: denies blurring, diplopia, irritation, discharge
Ear/Nose/Throat: denies ear pain or discharge, nasal obstruction or discharge, sore throat
Cardiovascular: denies chest pain, palpitations, paroxysmal nocturnal dyspnea, orthopnea, edema Respiratory: denies coughing, wheezing, dyspnea, hemoptysis
Gastrointestinal: denies abdominal pain, dysphagia, nausea, vomiting, diarrhea, constipation
Genitourinary: denies hematuria, frequency, urgency, dysuria, discharge, impotence, incontinence
Musculoskeletal: denies back pain, joint swelling, joint stiffness, joint pain
Skin: denies rashes, itching, lumps, sores, lesions, color change
Neurologic: denies syncope, seizures, transient paralysis, weakness, paresthesias
Psychiatric: denies depression, anxiety, mental disturbance, difficulty sleeping, suicidal ideation, hallucinations, paranoia
Endocrine: denies polyuria, polydipsia, polyphagia, weight change, heat or cold intolerance
Heme/Lymphatic: denies easy or excessive bruising, history of blood transfusions, anemia, bleeding disorders, adenopathy, chills, sweats
Allergic/Immunologic: denies urticaria, hay fever, frequent UTIs; denies HIV high risk behaviors

PART OF SPEECH TAGGER

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

CORPUS e ANNOTAZIONI

IN INGLESE:

- Brown University Standard Corpus of Present-Day American English (o Brown Corpus)
- Project Gutenberg
- Corpus of Contemporary American English (COCA)
- …

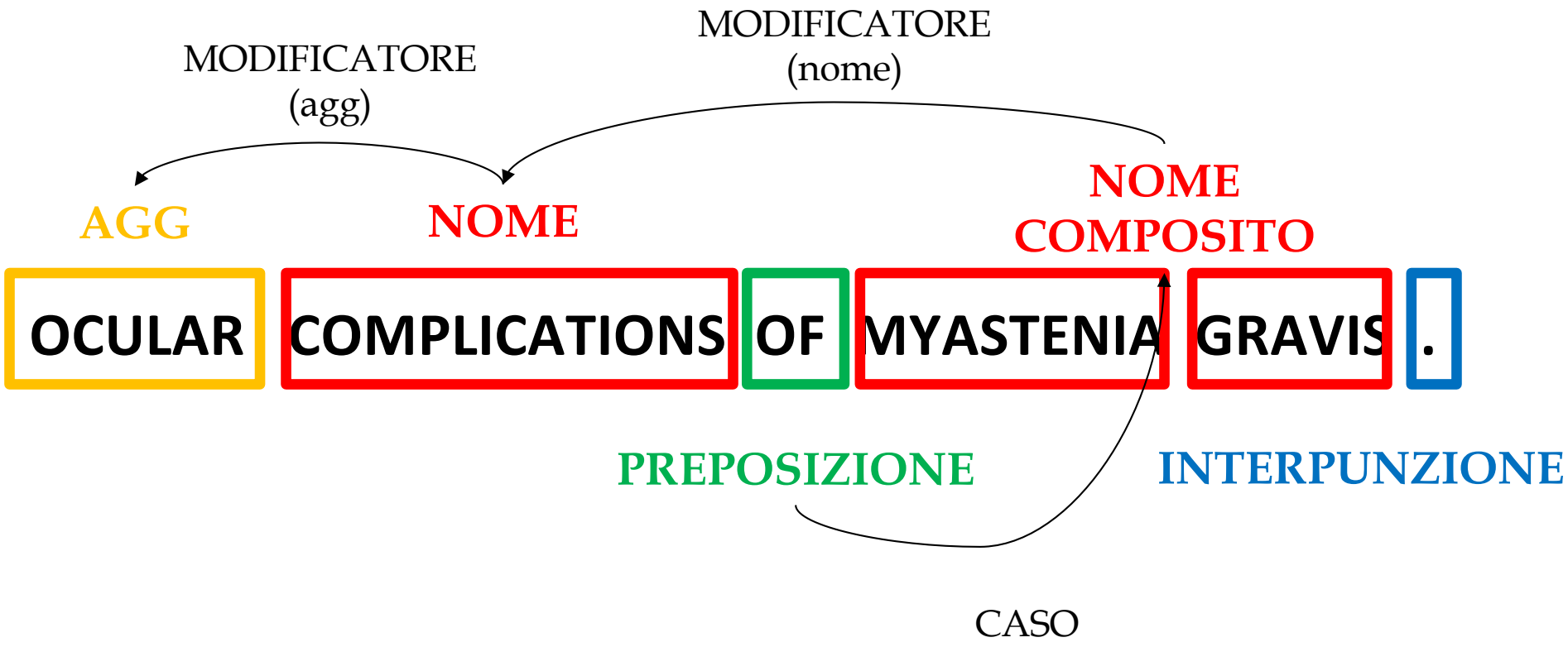
IN ITALIANO: ACCADEMIA DELLA CRUSCA!

https://accademiadellacrusca.it/it/contenuti/banche-dati-corpora-e-archivi-testuali/6228

Esempio:

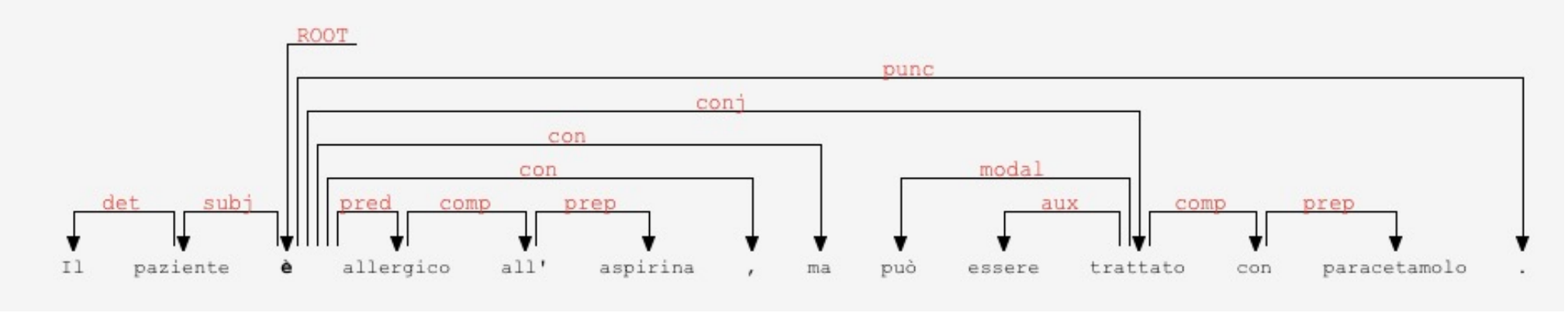
- LIS - Lessico dell'Italiano Scritto

PARSER


MODIFICATORE
(agg)
MODIFICATORE
(nome)
AGG
NOME
NOME
COMPOSITO
OCULAR
COMPLICATIONS
OF
MYASTENIA
GRAVIS
.
PREPOSIZIONE
INTERPUNZIONE
CASO

ESEMPIO

Il paziente è allergico all'aspirina, ma può essere trattato con paracetamolo.



DAL NLP ALL'INFORMATION EXTRACTION

Named Entity Recognition	• Categorizzazione dei sostantivi (persona, sostanza, entità geografica, …) • Es: Adam Smith lavora alla NASA → Adam Smith = persona, NASA = azienda
Semantic Role Recognition	• Riconoscimento del ruolo semantico • Es: Adam Smith sta alzando la mano → Adam = agent; la mano=parte del corpo; alzare = azione
Entity Relation Recognition	• Identificazione delle relazioni tra le entità • Es: Adam Smith lavora alla NASA → Adam Smith = persona, lavora per = relazione, NASA = azienda
Timex and Time Line Recognition	• Riconoscimento delle espressioni temporali • Es: oggi, domani, settimana scorsa, 11 maggio, etc
Word Sense Disambiguation	• Discriminazione degli omografi • Es: Il bambino mangia la pesca vs Il bambino pesca la trota
Negation Detection Identification	• Comprensione della negazione • Es: il paziente non presenta segni di infarto → infarto = assente

IL VALORE DEL TESTO LIBERO

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Journal of the American Medical Informatics Association, 26(11), 2019, 1189–1194
doi: 10.1093/jamia/ocz119
Advance Access Publication Date: 12 August 2019
Research and Applications

Research and Applications

Real world evidence in cardiovascular medicine: ensuring data validity in electronic health record-based studies

Tina Hernandez-Boussard,[1,2,3] Keri L Monda,[4,5] Blai Coll Crespo,[4] and Dan Riskin[1,3,6]

ABSTRACT

Objective: With growing availability of digital health data and technology, health-related studies are increasingly augmented or implemented using real world data (RWD). Recent federal initiatives promote the use of RWD to make clinical assertions that influence regulatory decision-making. Our objective was to determine whether traditional real world evidence (RWE) techniques in cardiovascular medicine achieve accuracy sufficient for credible clinical assertions, also known as "regulatory-grade" RWE.

Design: Retrospective observational study using electronic health records (EHR), 2010–2016.

Methods: A predefined set of clinical concepts was extracted from EHR structured (EHR-S) and unstructured (EHR-U) data using traditional query techniques and artificial intelligence (AI) technologies, respectively. Performance was evaluated against manually annotated cohorts using standard metrics. Accuracy was compared to pre-defined criteria for regulatory-grade. Differences in accuracy were compared using Chi-square test.

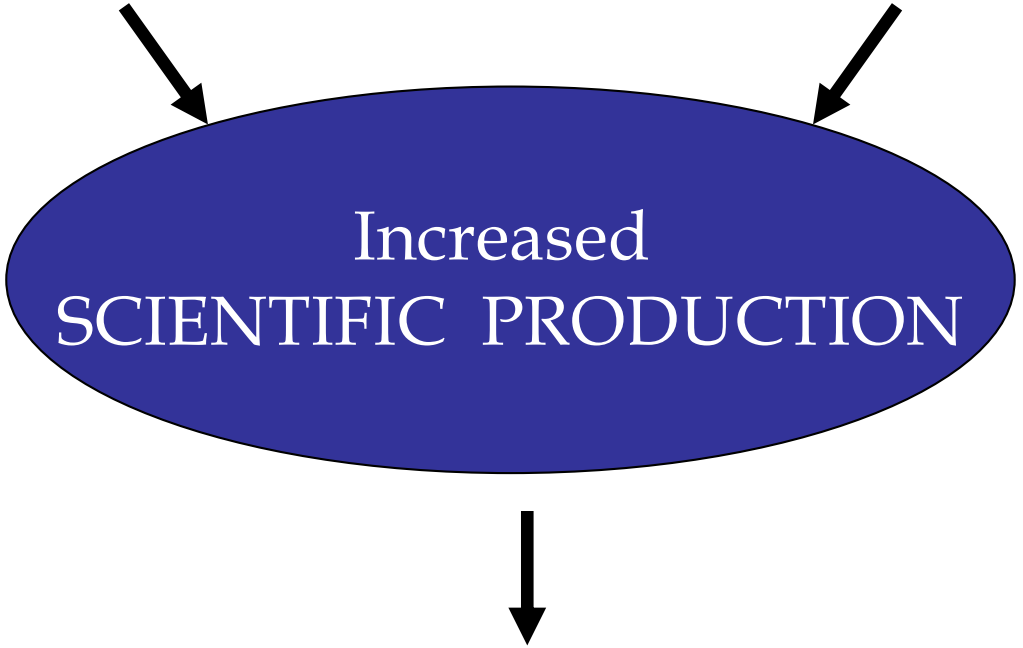
Results: The dataset included 10 840 clinical notes. Individual concept occurrence ranged from 194 for coronary artery bypass graft to 4502 for diabetes mellitus. In EHR-S, average recall and precision were 51.7% and 98.3%, respectively and 95.5% and 95.3% in EHR-U, respectively. For each clinical concept, EHR-S accuracy was below regulatory-grade, while EHR-U met or exceeded criteria, with the exception of medications.

Conclusions: Identifying an appropriate RWE approach is dependent on cohorts studied and accuracy required. In this study, recall varied greatly between EHR-S and EHR-U. Overall, EHR-S did not meet regulatory grade criteria, while EHR-U did. These results suggest that recall should be routinely measured in EHR-based studes intended for regulatory use. Furthermore, advanced data and technologies may be required to achieve regulatory grade results.

THE INCREASING SCIENTIFIC KNOWLEDGE

Increased number of scientific papers published

New journals



INCREASED KNOWLEDGE ?

BIBLIOGRAPHIC RESEARCH

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

The bibliographic research is essential to:

- **Start the research** (state-of-the-art evaluation/review)
- **Create new answers** to specific needs according to the available literature (new applications based on previous knowledge)

NEED OF BIBLIOGRAPHIC SOURCES

- **PROMPTLY AVAILABLE**
- **RELIABLE**

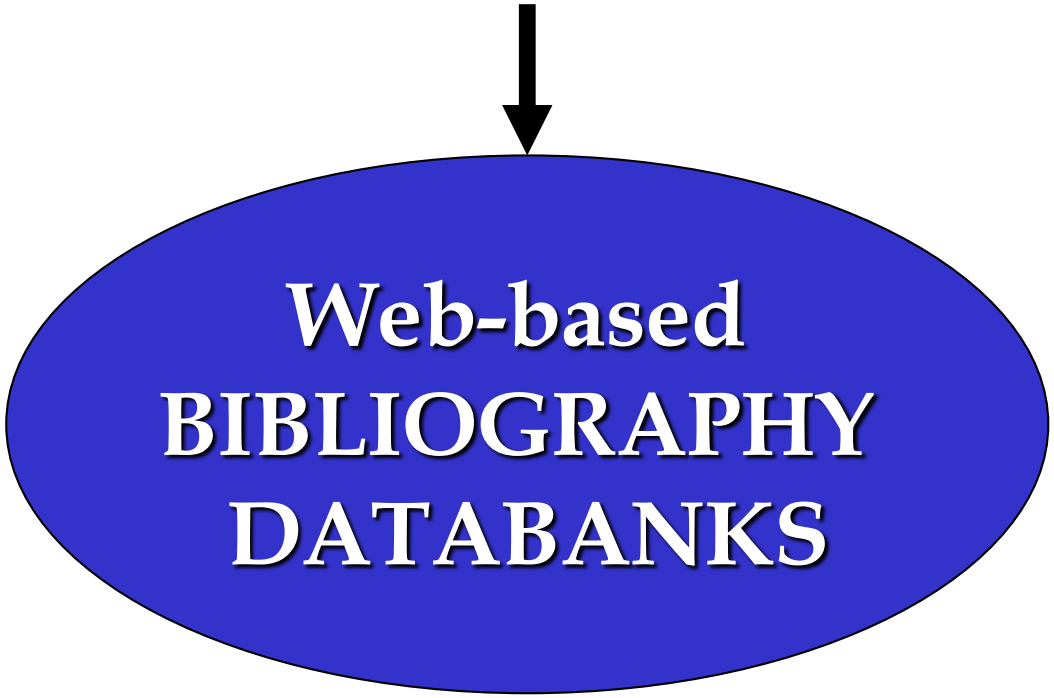
BIBLIOGRAPHY DATABANKS

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Increased medical knowledge

Need of bibliographic research

Easy access and availability



EXAMPLES (1/2)

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

BANCHE DI BIBLIOGRAFIA IN BIOMEDICINA					
Nome	Aree coperte	Copertura temporale	Fonti considerate	Frequenza di aggiornamento	Accesso
Medline	Tutte	Dal 1966	4592 riviste	Settimanale	Gratuito
Embase (Excerpta medica)	Tutte (settore farmaceutico)	Dal 1974	4168 riviste	Settimanale	Pagamento
Pascal Biomed	Tutte	Dal 1973	6885 riviste	Settimanale (online)	Pagamento
Cochrane Library	Tutte (in modo parziale)	-	Tende alla esaustività	Trimestrale	Pagamento
Best Evidence	Tutte (in modo parziale)	Dal: 1991(ACPJC) 1995 (EBM)	ACP JC EBM	Annuale	Pagamento

EXAMPLES (2/2)

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

ALCUNE BANCHE DI BIBLIOGRAFIA DI BIOMEDICINA SPECIALISTICA					
Nome	Aree coperte	Copertura temporale	Numero documenti	Frequenza di aggiornamento	Accesso
Aidsline	Aids	Dal 1980	186000	Settimanale	Gratuito
Cancerlit	Oncologia	Dal 1966	1.5 mln	Mensile	Gratuito
Bioethicsline	Etica medica	Dal 1973	60000	Bimensile	Gratuito
Toxline	Tossicologia	Dal 1965	2.6 mln	Mensile	Gratuito
HealthStar	Servizi e gestione sanitaria	Dal 1975	3.1 mln	Settimanale	Gratuito
BDSP	Salute pubblica	Dal 1970	214828	-	Pagamento
PsychINFO	Psicologia	Dal 1967	1.6 mln	Mensile	Pagamento
Allyed & Complementary Medicine	Medicine parallele	Dal 1985	103600	Mensile	Pagamento
Belit	bioetica	-	150000	-	Gratuito

INFORMATION QUALITY

- It is easy to implement references databanks using available ICTs and web technologies
- **Bibliography databanks refer only to quality information**
- Scientific papers to be published need to be:
  - ✓Peer-reviewed to obtain the comments from expert colleagues
  - ✓If the comments are all positive, the paper is accepted
  - ✓If the comments are negative →
    - ✓Request for paper revision (major/minor)
    - ✓Reject
  - ✓After the revision
    - ✓New review
    - ✓Direct acceptance/rejection

SCIENTIFIC PAPER ARCHITECTURE

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

- Journal name
- Publication date (volume, issue, pages)
- Authors
- Affiliations
- Title
- Keywords
- Abstract

- Body
  - ✓Introduction
  - ✓Material and methods
  - ✓Results
  - ✓Discussion
  - ✓Conclusions
- Acknowledgements
- References

THE SCIENTIFIC PAPER

UNIVERSITÀ DEGLI STUDI DI TRIESTE

Citation

Journal of Biomedical Informatics 52 (2014) 92–104

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Journal name

Title

WebBioBank: A new platform for integrating clinical forms and shared neurosignal analyses to support multi-centre studies in Parkinson's Disease

Elena Rossi [a,1], Manuela Rosa [b,1], Lorenzo Rossi [b], Alberto Priori [b], Sara Marceglia [a,*]

[a] Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milan, Italy
[b] Centro Clinico per la Neurostimolazione, le Neurotecnologie ed i Disordini del Movimento, Fondazione IRCCS Ca' Granda, Ospedale Maggiore Policlinico, Milan, Italy

Authors

Affiliations

ARTICLE INFO

Article history:
Received 3 October 2013
Accepted 28 August 2014
Available online 7 September 2014

Keywords:
Web-based platform
Database
Multi-centre clinical study
Operative unit
Parkinson's Diseases
Signal processing

Information on acceptance

ABSTRACT

Background: The web-based systems available for multi-centre clinical trials do not combine clinical data collection (Electronic Health Records, EHRs) with signal processing storage and analysis tools. However, in pathophysiological research, the correlation between clinical data and signals is crucial for uncovering the underlying neurophysiological mechanisms. A specific example is the investigation of the mechanisms of action for Deep Brain Stimulation (DBS) used for Parkinson's Disease (PD); the neurosignals recorded from the DBS target structure and clinical data must be investigated.
Objective: The aim of this study is the development and testing of a new system dedicated to a multi-centre study of Parkinson's Disease that integrates biosignal analysis tools and data collection in a shared and secure environment.
Methods: We designed a web-based platform (WebBioBank) for managing the clinical data and biosignals of PD patients treated with DBS in different clinical research centres. Homogeneous data collection was ensured in the different centres (Operative Units, OUs). The anonymity of the data was preserved using unique identifiers associated with patients (ID BAC). The patients' personal details and their equivalent ID BACs were archived inside the corresponding OU and were not uploaded on the web-based platform; data sharing occurred using the ID BACs. The system allowed researchers to upload different signal processing functions (in a .dll extension) onto the web-based platform and to combine them to define dedicated algorithms.
Results: Four clinical research centres used WebBioBank for 1 year. The clinical data from 58 patients treated using DBS were managed, and 186 biosignals were uploaded and classified into 4 categories based on the treatment (pharmacological and/or electrical). The user's satisfaction mean score exceeded the satisfaction threshold.
Conclusions: WebBioBank enabled anonymous data sharing for a clinical study conducted at multiple centres and demonstrated the capabilities of the signal processing chain configuration as well as its effectiveness and efficiency for integrating the neurophysiological results with clinical data in multi-centre studies, which will allow the future collection of homogeneous data in large cohorts of patients.

© 2014 Elsevier Inc. All rights reserved.

Abstract

THE PEER-REVIEW PROCESS (1)

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

View Review Comments for Manuscript
CBM-D-14-00776R1
"A [redacted]"

Click the Reviewer recommendation term to view the Reviewer comments.

	Revision 1	Original Submission
(Reviewer 1)	(None)	Minor Revision
(Reviewer 2)	Reject	Reject
Sara Renata Francesca Marceglia, Ph.D. (Reviewer 3)	Major Revision	(None)
(Reviewer 4)	Reject	(None)

Close

THE PEER-REVIEW PROCESS (2)

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

CBM-D-14-00776R1
"[redacted]"
Revision 1

Sara Renata Francesca Marceglia, Ph.D. (Reviewer 3)

Reviewer Recommendation Term:	Major Revision

Manuscript Rating Question(s):	Scale	Rating
Please rate on a scale of 1-3 whether the Highlights are a meaningful and accurate representation of the article. 1 = Meaningful; 2 = Not Meaningful; 3 = Not Provided. For more information, see www.elsevier.com/highlights	[1-3]	N/A
Please rate on a scale of 1-3 whether the Graphical Abstract is a meaningful and an accurate representation of the article. 1 = Meaningful; 2 = Not Meaningful; 3 = Not Provided. For more information, see www.elsevier.com/graphicalabstracts	[1-3]	N/A

Comments to Editor:

REVIEWER: Please indicate to the Editor whether any of the following apply, by deleting 'Y' or 'N' as appropriate:

* The writing needs substantial improvement - grammar and spelling Y
* Improper formatting was used (figures, tables, legends too small) N
* The work is too incremental compared to authors' or others' prior work N
* The manuscript is for the most part previously published by the authors N
* The study is too biological and not quantitative/computational enough N

Dear Editor,
I read the manuscript and I believe that it worths to let the authors re-revise it. It still have a language problem (I can tell it even though I am not a native English speaker) and I totally agree with the other reviewer that it still lacks the contextualization and comparison with the other approaches. I don't agree, conversely, on the fact that it lacks novelty. It is anyway a different and simpler way to treat the problem of skin reflection, and it could be useful for the imaging community. I would suggest to let them revise it.
With kind regards,

Sara

Comments to Author:

The paper "A Novel Dual Calibration Confocal Imaging for Microwave Early Breast Tumor detection in MR Images Based Model" describes a dual calibration technique for microwave breast tumour detection. The model is based on 3D breast MR images and uses a uniformly spaced moving antenna array.
The method is described in details and the simulation model is valuable. Also, the figures are informative and well presented in the text.
I have two major critiques related to the manuscript:
1- the language needs a deep revision from a native English speaker. There are many grammar and syntax mistakes (e.g., in the highlights, "Confocal algorithm use …" should be "The confocal algorithm uses…"; in the introduction "Mammography and magnetic resonance imaging (MRI) are most common" should be "Mammography and magnetic resonance imaging (MRI) are the most common").
2- even though the authors describe the available literature on the current state of the art of artefact removal techniques, they do not explain why the proposed methodology differs from the others and how it contributes to overcome the current limitations. Despite the detailed "Methods" and "Results" sections, the manuscript completely lacks the discussion of the results in the light of the current state-of-the-art. The "Conclusions" just provide a summary of the results and do not contextualize the authors' findings.

THE DECISION LETTER

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

22-Feb-2015

Dear Dr. Marceglia:

Manuscript ID ACI-2014-12-RA-0115 entitled "A Standards-Based Architecture Proposal for Integrating Patient mHealth Apps to Electronic Health Record Systems" which you submitted to the Applied Clinical Informatics, has been reviewed. The comments of the reviewer(s) and Associate Editor (if applicable) are included at the bottom of this letter.

The reviewer(s) have recommended publication with Minor Revisions to your manuscript. Therefore, I invite you to respond to the reviewer(s)' comments and revise your manuscript.

The revised manuscript must be UNBLINDED and must contain a section on Human Subject Research Approval and a section on Conflict of Interest.

We encourage you to revise your manuscript within the next 3 weeks to optimize turn-around. Revisions received more than 8 weeks from the date of the decision letter will be considered new submissions, subject to a full round of reviews.

Please log into https://mc.manuscriptcentral.com/acij and enter your Author Center, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Revision." Your manuscript number has been appended to denote a revision. You will be unable to make your revisions on the originally submitted version of the manuscript. Instead, revise your manuscript using a word processing program and save it on your computer. Please also highlight the changes to your manuscript within the document by using the track changes mode in MS Word or by using bold or colored text. Once the revised manuscript is prepared, you can upload it and submit it through your Author Center.
When submitting your revised manuscript, you will be able to respond to the comments made by the reviewer(s) in the space provided. You can use this space to document any changes you make to the original manuscript. In order to expedite the processing of the revised manuscript, please be as specific as possible in your response to the reviewer(s).

While Applied Clinical Informatics focuses on the applied aspects of informatics, we aim to give our readers some methodological background as well. In our experience, Methods of Information in Medicine (http://www.schattauer.de/en/magazine/subject-areas/journals-a-z/methods.html) is useful in providing such theoretical background and we encourage references to this journal where appropriate.

IMPORTANT: Your original files are available to you when you upload your revised manuscript. Please delete any redundant files before completing the submission. Because we are trying to facilitate timely publication of manuscripts submitted to the Applied Clinical Informatics, your revised manuscript should be uploaded as soon as possible. If it is not possible for you to submit your revision in a reasonable amount of time, we may have to consider your paper as a new submission.

Once again, thank you for submitting your manuscript to the Applied Clinical Informatics and I look forward to receiving your revision.

Sincerely,

Dr. Christoph Lehmann
Editor in Chief, Applied Clinical Informatics
culehmann@gmail.com

Reviewer(s)' Comments to Author:

Reviewer: 1

Comments to the Author
In reading this, I had several considerations.

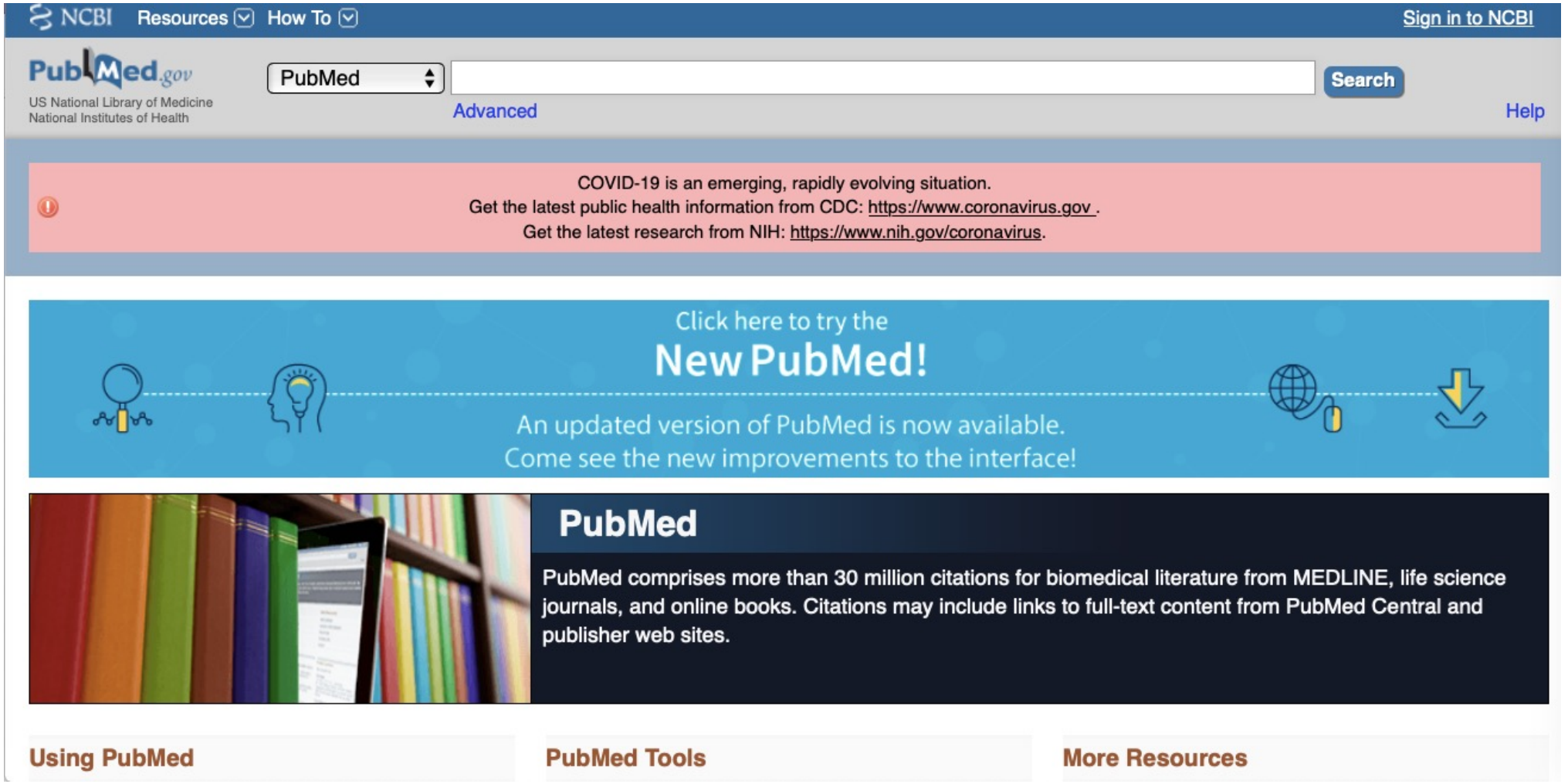
The communication between the mobile application and EHR are encrypted and deidentified, using only a patient ID (which I assume is random). Does the mobile application itself include the patient's name, physician's name, etc.? If someone were to find the mobile device, then easy deduction can allow one to discover the patient's name and medical condition. I know your paper says you do not store health information on the device, but this was left a little unclear to me. Does the application have password or TouchID validation to abate this? Additionally, at what point is the information re-identified, and how is this done? It seems that the logical place would be on the EMR side, but this is not apparent in the paper.

Since no health information is stored on the mobile device, how does this contribute to patient health information tracking? Would these data (which are stored in the multiple repositories, per Figure 2A/B) then be accessible to the mobile platform for longitudinal tracking? As it stands, communication looks to be only one-way in terms of this.

PUBMED.GOV

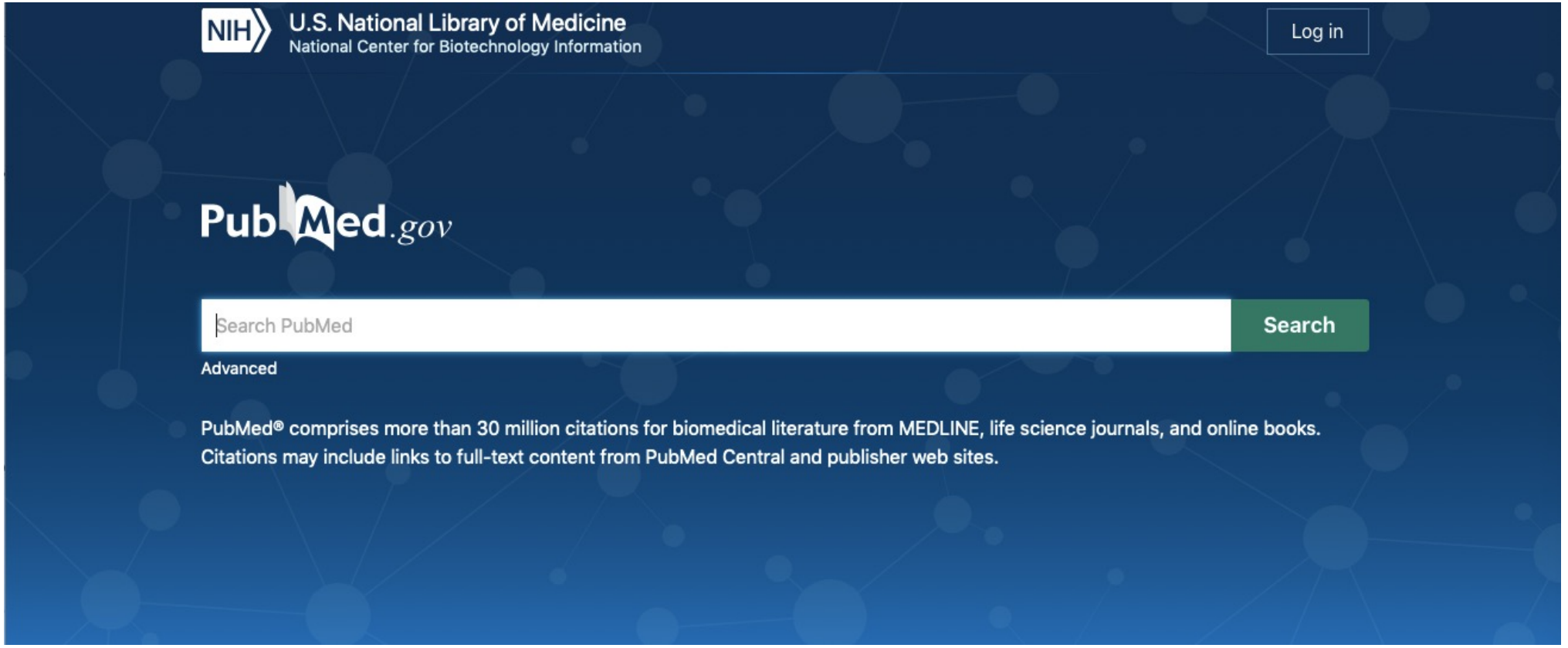
UNIVERSITÀ
DEGLI STUDI DI TRIESTE

https://www.ncbi.nlm.nih.gov/pubmed/



PUBMED.GOV

UNIVERSITÀ
DEGLI STUDI DI TRIESTE



THE NIH - NATIONAL INSTITUTES OF HEALTH

UNIVERSITÀ DEGLI STUDI DI TRIESTE

U.S. Department of Health & Human Services

NIH National Institutes of Health
Turning Discovery Into Health

Search NIH

NIH Employee Intranet | Staff Directory | En Español

Health Information | Grants & Funding | News & Events | Research & Training | Institutes at NIH | About NIH

COVID-19 is an emerging, rapidly evolving situation.

Get the latest public health information from CDC: https://www.coronavirus.gov
Get the latest research information from NIH: https://www.nih.gov/coronavirus

Home » About NIH

WHO WE ARE

The National Institutes of Health (NIH), a part of the U.S. Department of Health and Human Services, is the nation's medical research agency — making important discoveries that improve health and save lives.

The NIH Director

Francis S. Collins, M.D., Ph.D., was officially sworn in on August 17, 2009, as the 16th director of the NIH.

NIH Leadership

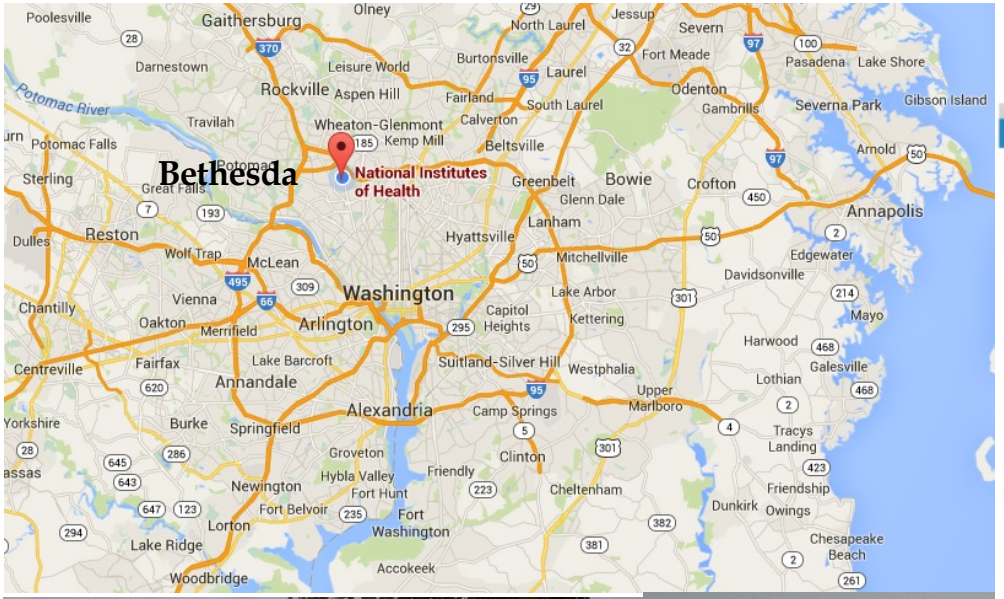
NIH leadership plays an active role in shaping the agency's research planning, activities, and outlook.

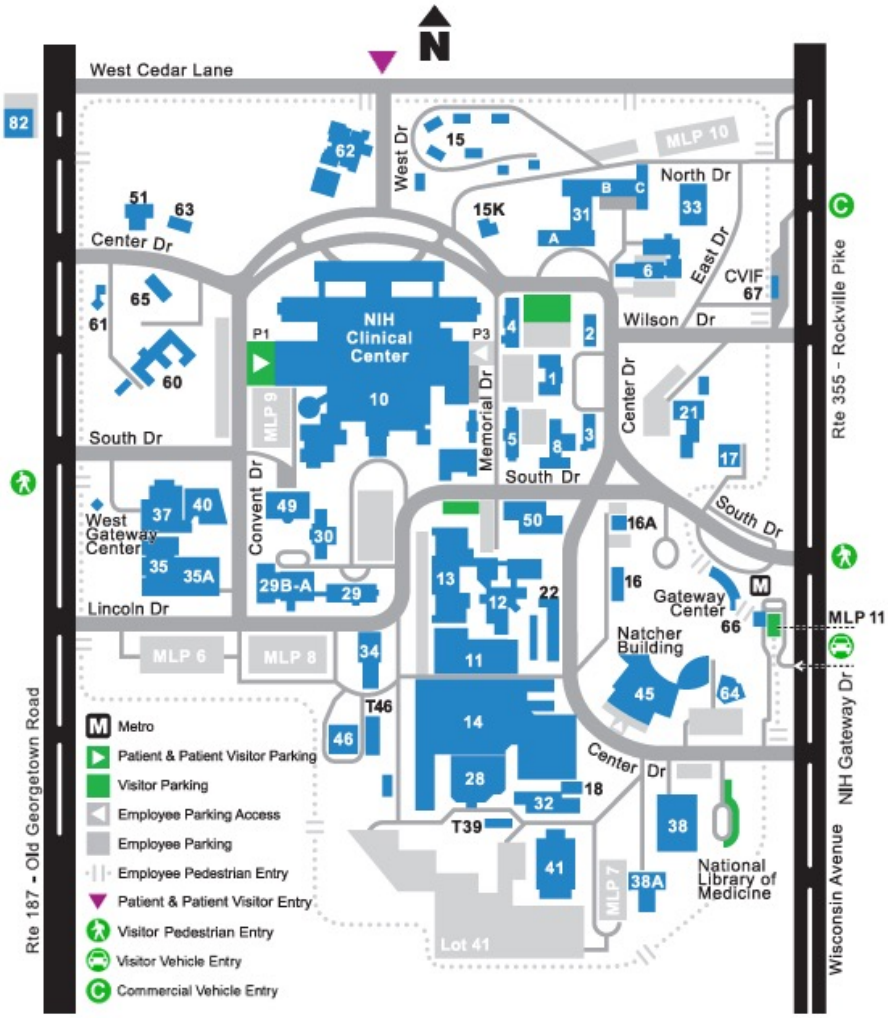
Organization

The NIH is made up of 27 different components called Institutes and Centers.

WHERE IS THE NIH

UNIVERSITÀ
DEGLI STUDI DI TRIESTE



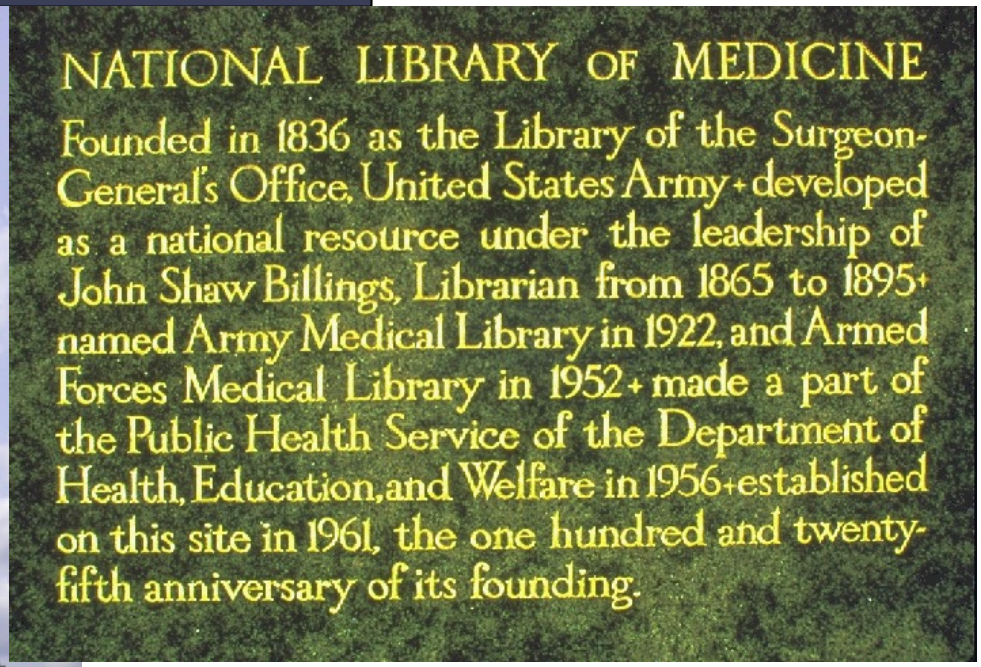


THE NIH CAMPUS

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

THE NATIONAL LIBRARY OF MEDICINE

UNIVERSITÀ
DEGLI STUDI DI TRIESTE



NLM PUBLIC RESOURCES

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

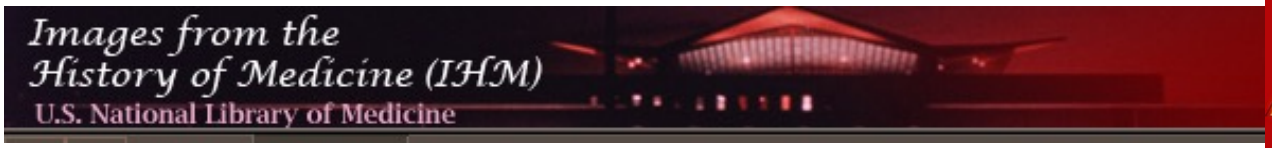
PubMed.gov
US National Library of Medicine
National Institutes of Health

MeSH on Demand

NIH U.S. National Library of Medicine | TOXNET TOXICOLOGY DATA NETWORK

MedlinePlus®
Trusted Health Information for You

ClinicalTrials.gov



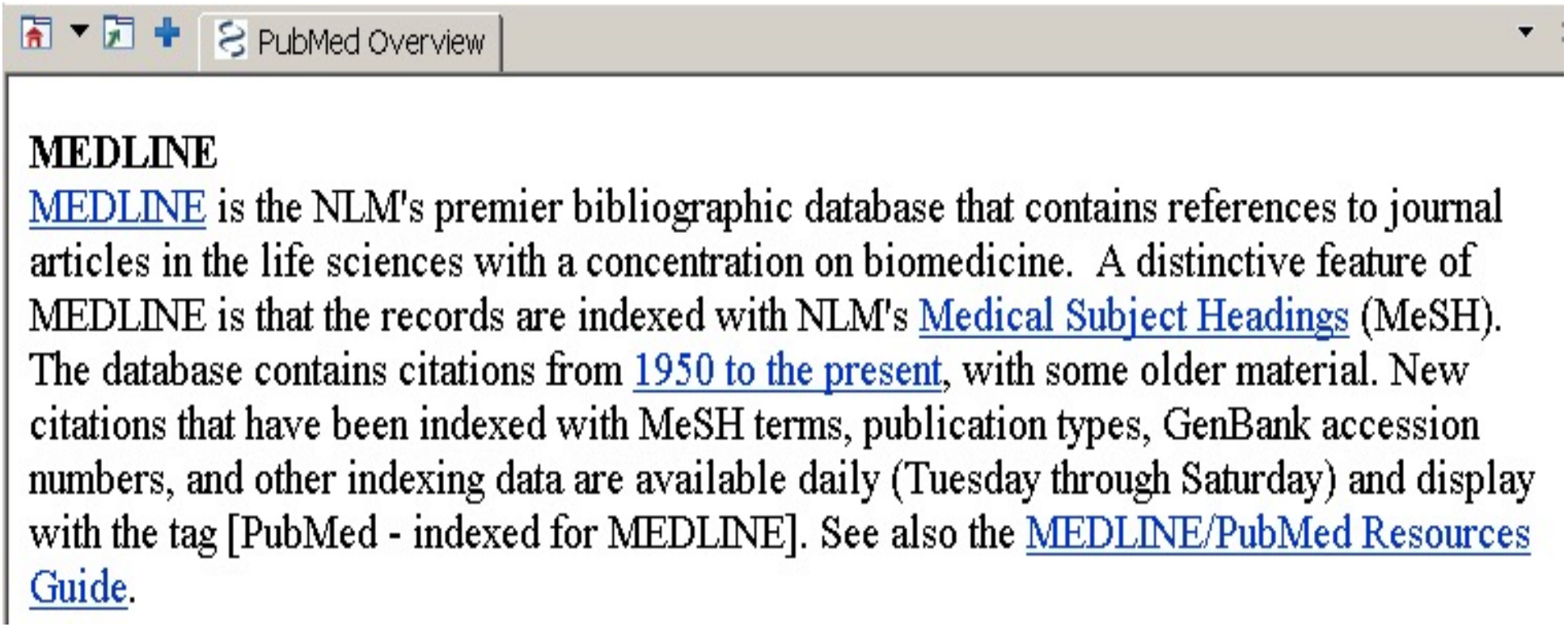
Unified Medical Language System® (UMLS®)

MEDLINE-PUBMED

- It is the mostly used databank in medicine
- The article is retrieved as citation (not "full text")
- US National Library of Medicine (NLM, www.ncbi.nlm.nih.gov)
- Free access
- 74% papers in English
- 40% of the journals are published in the US

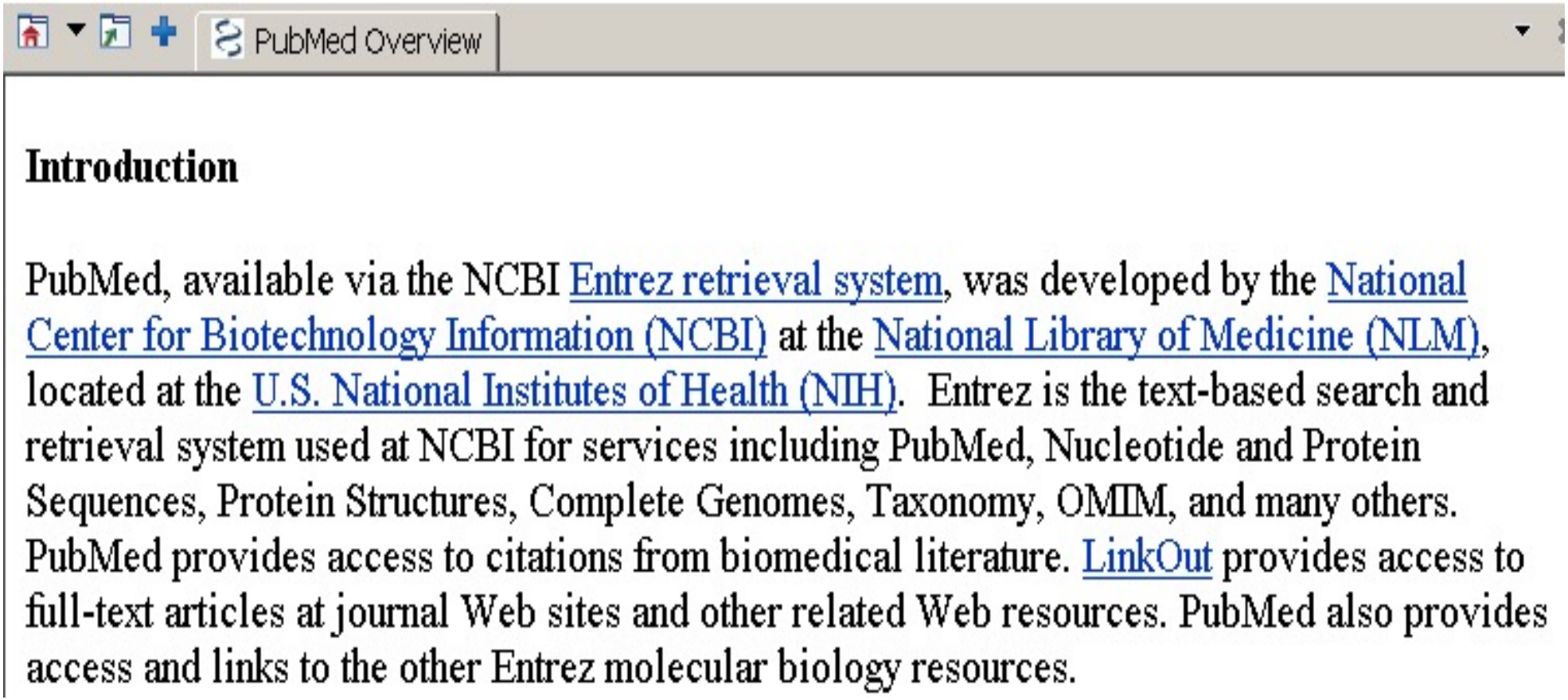
MEDLINE/PUBMED

UNIVERSITÀ
DEGLI STUDI DI TRIESTE



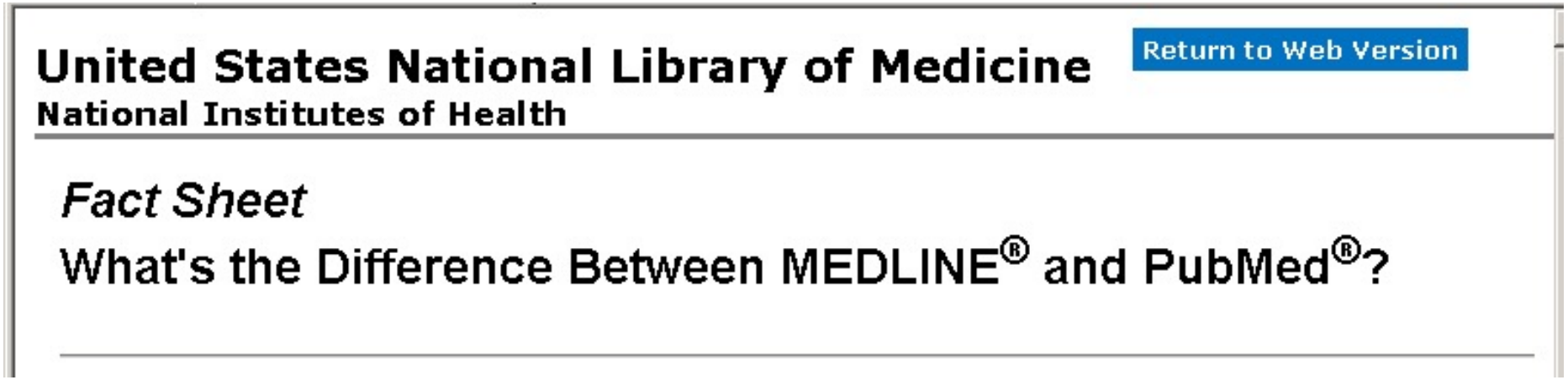
MEDLINE/PUBMED

UNIVERSITÀ
DEGLI STUDI DI TRIESTE



MEDLINE/PUBMED

UNIVERSITÀ
DEGLI STUDI DI TRIESTE



United States National Library of Medicine
National Institutes of Health

Return to Web Version

Fact Sheet

What's the Difference Between MEDLINE® and PubMed®?

MEDLINE/PUBMED

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

- MEDLINE®
- National Library of Medicine® (NLM®) journal citation database.
- Started in the 1960s, it now provides over 21 million references to biomedical and life sciences journal articles back to 1946.
- MEDLINE includes citations from over 5,600 scholarly journals published around the world.
- Publishers submit journals to an NIH-chartered advisory committee, the Literature Selection Technical Review Committee (LSTRC), which reviews and recommends journals for MEDLINE.
- The MEDLINE database is directly searchable from NLM as a subset of the PubMed® database as well as through other numerous search services that license the data.

PUBMED

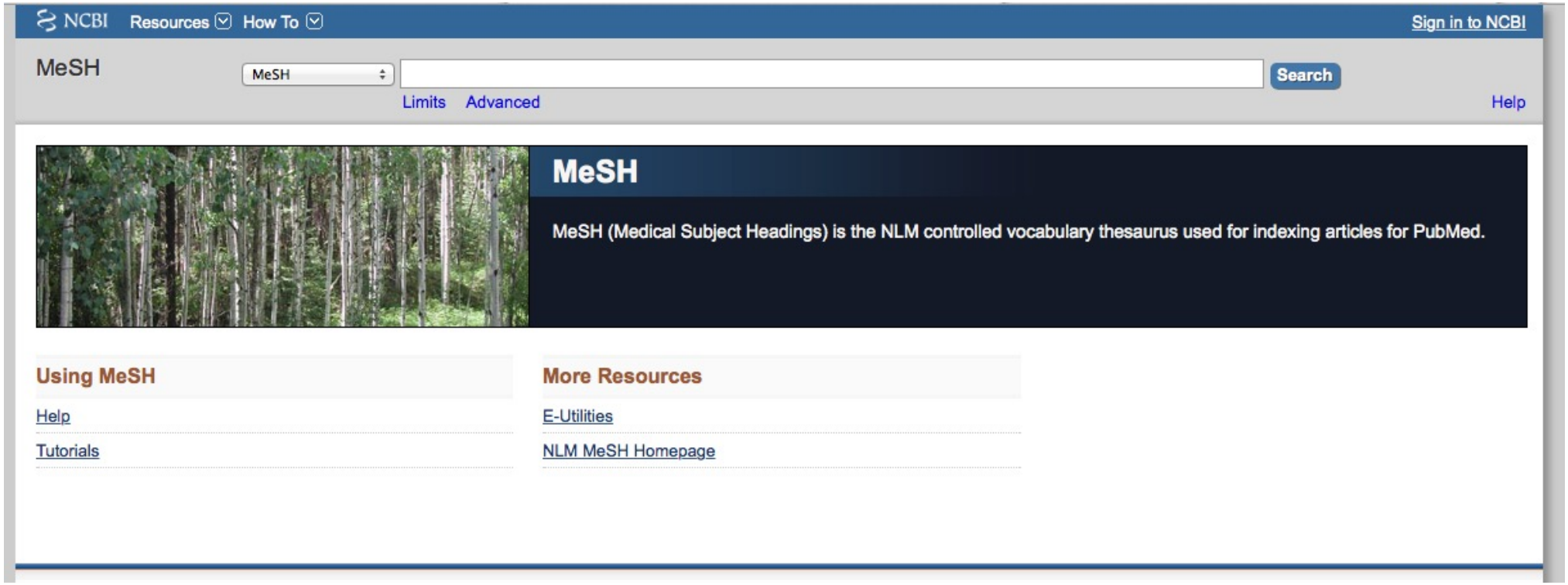
- Has been available since 1996 (now 23 mln references)
- Include the MEDLINE database plus the following types of citations:
  - In-process citations, which provide records for articles before those records go through quality control and are indexed with MeSH or converted to out-of-scope status.
  - Citations to articles that are out-of-scope from certain MEDLINE journals (general science and general chemistry journals)
  - "Ahead of Print" citations .
  - Citations that precede the date that a journal was selected for MEDLINE indexing and Pre-1966 citations.
  - Citations to some additional life sciences journals that submit full text to PMC® (PubMed Central®) and receive a qualitative review by NLM.
  - Citations to author manuscripts of articles published by NIH-funded researchers amd Citations for the majority of books available on the NCBI Bookshelf .

PUBMED CENTRAL (PMC)

- Launched in 2000
- Free archive for full-text biomedical and life sciences journal articles.
- Repository for journal literature deposited by participating publishers, as well as for author manuscripts that have been submitted in compliance with the NIH Public Access Policy and similar policies of other research funding agencies.
- Although free access is a requirement for PMC deposit, publishers and individual authors may continue to hold copyright on the material in PMC and publishers can delay the release of their material in PMC for a short period after publication.
- There are reciprocal links between the full text in PMC and corresponding citations in PubMed.

MESH TERMS

UNIVERSITÀ
DEGLI STUDI DI TRIESTE



GENERAL RESEARCH PRINCIPLES

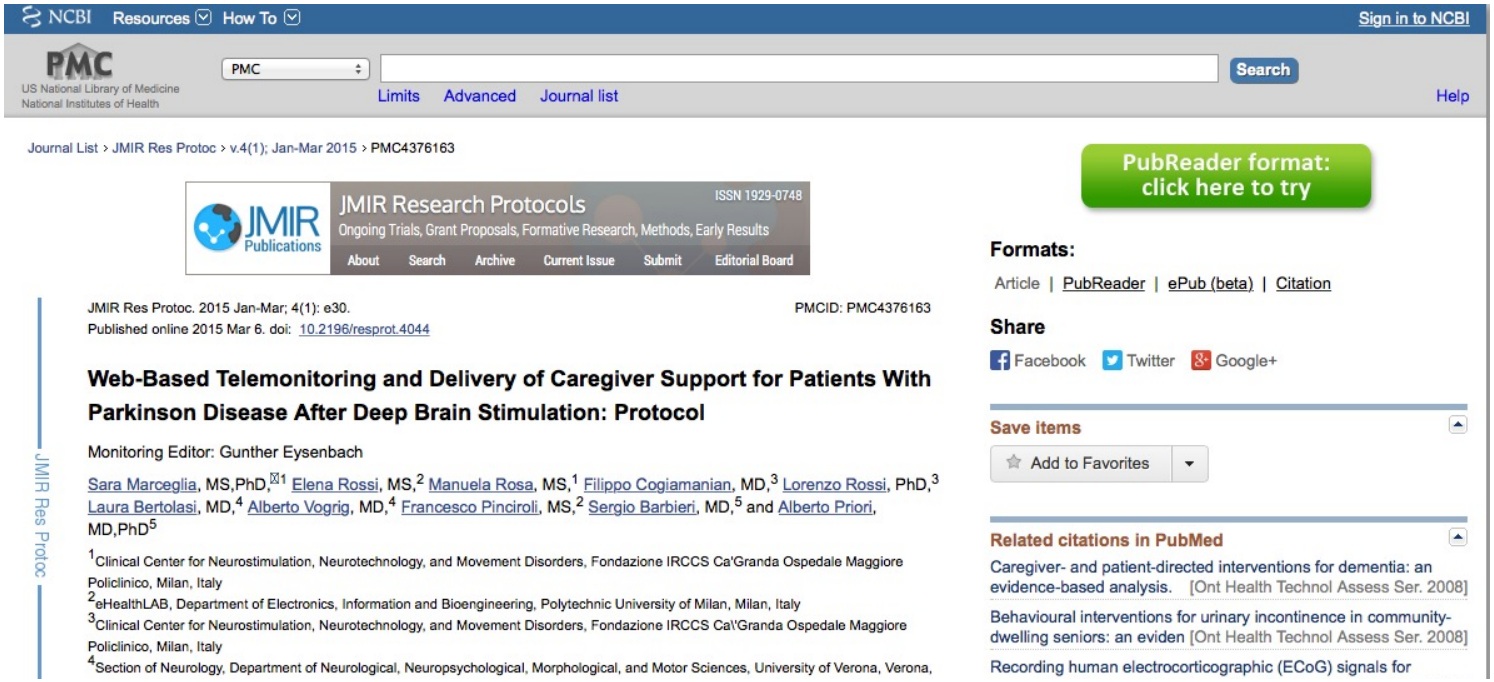
UNIVERSITÀ
DEGLI STUDI DI TRIESTE

- Definition of the research object
- Definition of the key concepts
- Definition of the type of reserach (free text or coded).
- Boolean operators AND, OR, NOT to link the research keywords
- Use of the "limits/constraints" →
  - Publication year
  - Type of paper
  - Paper language
  - Availability of the abstract
  - Free article

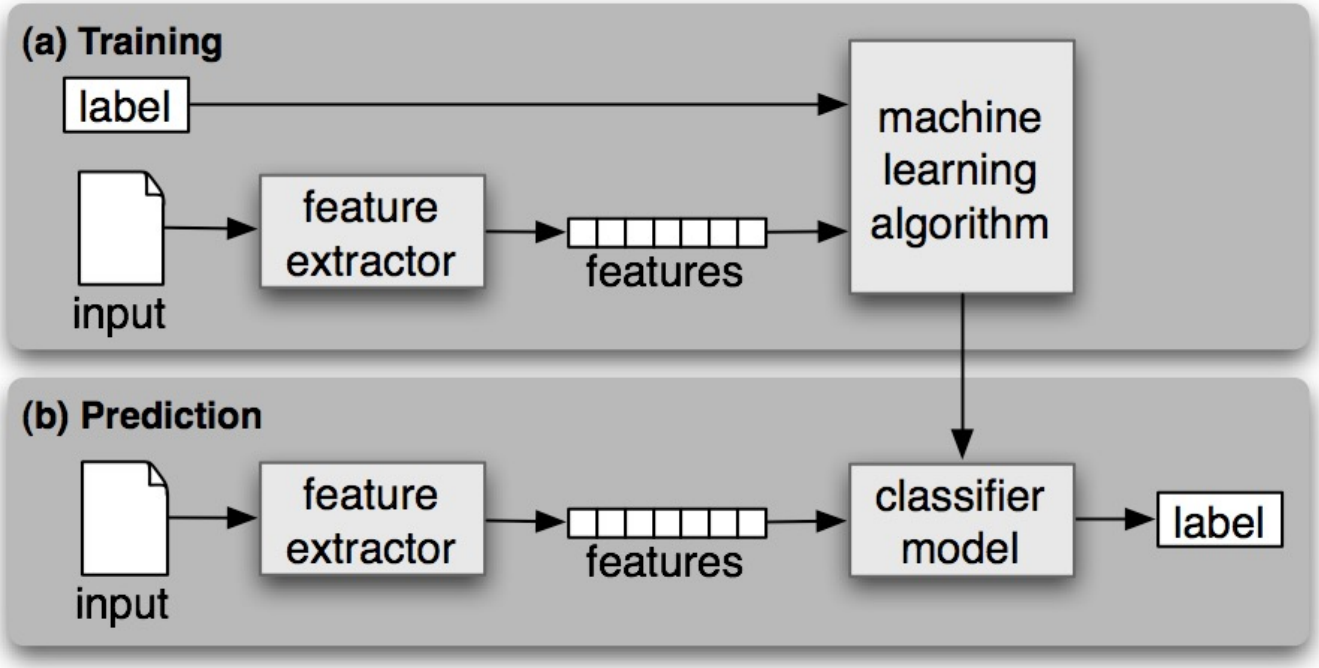
HOW TO FIND THE “FULL TEXT”

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Redirects to the journal website



MACHINE LEARNING EXAMPLE



NÄIVE BAYES CLASSIFIER

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

MACHINE LEARNING EXAMPLE

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

```
from nltk.corpus import movie_reviews

# Total reviews
print (len(movie_reviews.fileids())) # Output: 2000

# Review categories
print (movie_reviews.categories()) # Output: [u'neg', u'pos']

[…]


print (classifier.show_most_informative_features(10))
'''
Output:

Most Informative Features
   contains(outstanding) = True              pos : neg    =     14.7 : 1.0
         contains(mulan) = True              pos : neg    =      7.8 : 1.0
        contains(poorly) = True              neg : pos    =      7.7 : 1.0
   contains(wonderfully) = True              pos : neg    =      7.5 : 1.0
        contains(seagal) = True              neg : pos    =      6.5 : 1.0
         contains(awful) = True              neg : pos    =      6.1 : 1.0
        contains(wasted) = True              neg : pos    =      6.1 : 1.0
         contains(waste) = True              neg : pos    =      5.6 : 1.0
         contains(damon) = True              pos : neg    =      5.3 : 1.0
         contains(flynt) = True              pos : neg    =      5.1 : 1.0

[…]
```