



Management and Analysis of Biomedical Texts

Milena Kraus
Data Management for Digital Health
Summer 2017

Agenda

Real-world Use Cases

Oncology



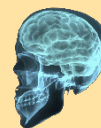
Nephrology



Heart
Insufficiency



Additional
Topics



Data Management & Foundations



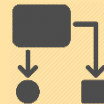
Biology
Recap



Data
Sources



Data
Formats



Business
Processes



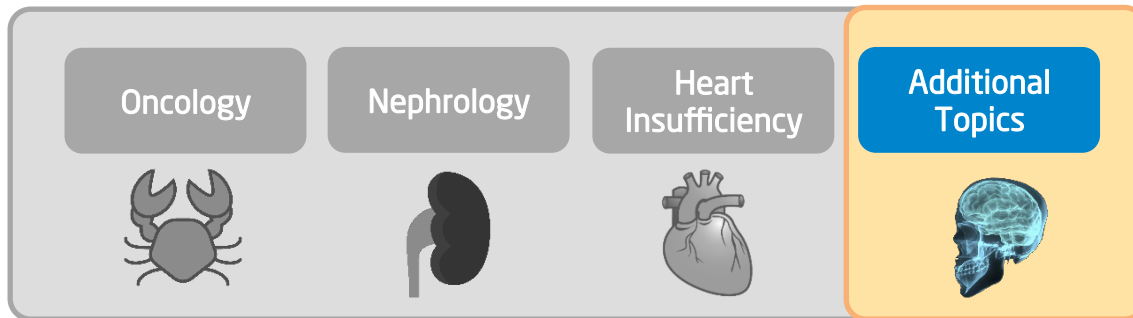
Processing
and Analysis

Management of Biomedical Texts

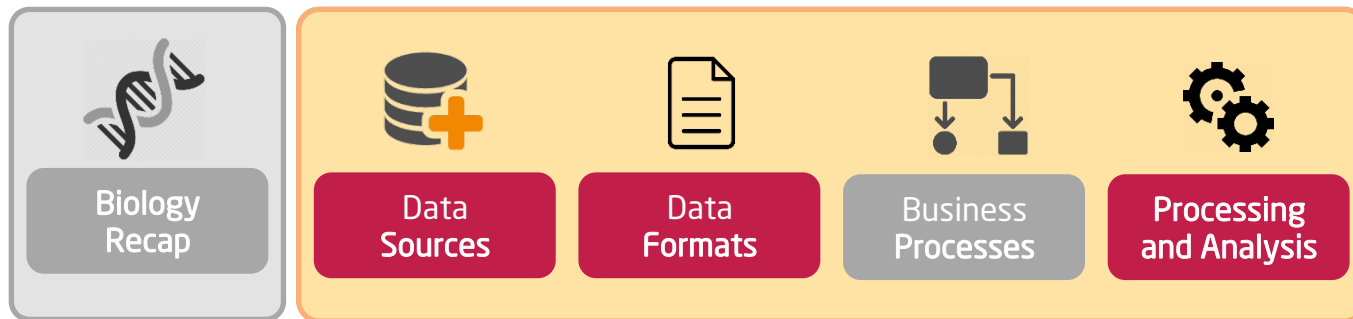
Data Management for
Digital Health, Summer
2017

Agenda

Real-world
Use Cases



Data Management
& Foundations



Management of Biomedical Texts

Data Management for
Digital Health, Summer
2017

In this lecture you will be introduced to

- Biomedical text documents in general and with a special focus on
- Scientific publications and
- search engines to find them.

Furthermore, how methods of natural language processing help to

- Answer natural language questions and
- Extract most relevant information from biomedical articles. (optional)

Management of Biomedical Texts

Data Management for
Digital Health, Summer
2017

Prescriptions

German Muster 16



<input checked="" type="checkbox"/> Krankenkasse bzw. Kostenträger	AOK Rheinland-Pfalz	
<input type="checkbox"/> Geb.-pl.	Name, Vorname des Versicherten	
<input type="checkbox"/> Nachn.	Mustermann	
<input type="checkbox"/> Sonstige	geb. am 12.08.1964	
	Heidestraße 17	
	51147 Köln 10/14	
<input type="checkbox"/> Unfall	Kassen-Nr.	Versicherten-Nr.
	106415300	A123456789
	Status 1000 1	
<input type="checkbox"/> Arbeits-unfall	Betriebsstätten-Nr.	Arzt-Nr.
	271111100	654321161
	Datum 10.07.2012	

Rp. (Bitte Leeräume durchsteichen)

Antistressin Impfstoff Amp. 10 x 0.5 ml
Muster Pharma GmbH

bbbr

Bei Arbeitsunfall auszufüllen!

Unfalltag: Unfallbetrieb oder Arbeitgebernnummer:

RVG	Hilfsmittel	Impfstoff	Spezialbedarf	Rezeptgebühr	Apothekennummer / St.
6	7	8	9		

Zusatzung	Gesamt-Brutto

Arzneimittel-Mittel-Nr.	Faktor	Taxe
1. Verordnung		
2. Verordnung		
3. Verordnung		


Vertragsarztstempel


271111100
Psychologische Gemeinschaftspraxis
Dr. med. Markus Mustermann
Dr. rer. nat. Erik Mustermann
Dortheidenstraße 1
51069 Köln
Tel. 02 21 / 6 87 65 43
[Signature]
Unterschrift des Arztes
Muster 16 (7.2008)


2711111004


Management of Biomedical Texts


Data Management for Digital Health, Summer 2017

 Menu


 **Onmeda.de**
Für meine Gesundheit


 Folgen

 Spiele

 Login


Hu













H

Ult











Krankheiten & Symptome

Krankheiten

Krankheiten A-Z		Symptome A-Z	
Krankheitsgebiete		Symptom-Check	
Häufigste Krankheiten		Häufigste Symptome	
Seltene Krankheiten			
ICD-10-Diagnoseschlüssel			






Arztbesuch

Untersuchung & Behandlung		Krankheitserreger	
Vorsorge		Anatomie	
Impfungen		Strahlenmedizin	
Laborwerte		Persönlichkeiten	

Symptome


Lexika


Specials

- Magen-Darm-Probleme: Was hilft? 
- Multiple Sklerose 
- Haarausfall bei Männern 
- Haarausfall bei Frauen 
- Schuppenflechte 

Apotheken-Notdienst

PLZ oder Ort eingeben und suchen:





Management of Biomedical Texts

Data Management for
Digital Health, Summer
2017

Pathology Report



Hasso
Plattner
Institut

2008-15

Patient [REDACTED]
Date of birth [REDACTED] Sex Male
Biopsy Date 1/3/2008
Doctor [REDACTED]



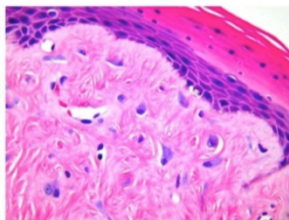
Part A: LEFT MAXILLARY SOFT TISSUE

Gross description:

Submitted is formalin fixed tissue, measuring 1.6x1.4x1.4cm., stated to be from the left maxilla. The specimen consists of multiple pieces of brown soft tissue. Sections multiple. All submitted. Also submitted is a tooth, no sections taken.

Microscopic Description:

Multiple sections show keratotic, stratified squamous epithelium covering a core of dense and cellular fibrous connective tissue. Numerous enlarged stellate-shaped fibroblasts, some containing multiple nuclei, are seen in the lesional stroma.



Diagnosis: **Fibroma, giant cell type**

ICD: 210.4
CPT: 88305

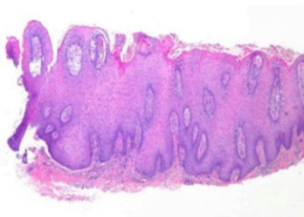
Part B: RIGHT LATERAL TONGUE

Gross description:

Submitted is formalin fixed tissue, measuring 1.2x0.5x0.5cm., stated to be from the right lateral tongue. The specimen consists of one piece of tan soft tissue with suture. One section submitted.

Microscopic Description:

Multiple sections show acanthotic, parakeratotic, verrucous stratified squamous epithelium covering a core of well-vascularized fibrous connective tissue. The interepithelial connective tissue papilla are filled with foamy histiocytes. Lymphocytes and plasma cells are also seen.



Diagnosis: **Verruciform xanthoma**

ICD: 210.4
CPT: 88305

DERMATOPATHOLOGY PATHOLOGY REPORT



PATIENT INFORMATION	PHYSICIAN INFORMATION	SPECIMEN INFORMATION
[REDACTED]	[REDACTED]	SURGICAL #: S08-02011 MEDICAL REC #: 0315961 ACCOUNT #: 409514 LOCATION: ASC
DATE COLLECTED: 2/20/2008	DATE RECEIVED: 2/20/2008	DATE REPORTED: 2/21/2008

CLINICAL INFORMATION:

Rule out melanoma.

DIAGNOSIS

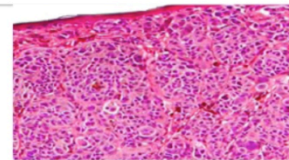
Skin, left face, excision:

- MALIGNANT MELANOMA, NODULAR TYPE
- TUMOR THICKNESS IS 3.0 MM
- CLARK'S LEVEL III
- HIGH MITOSIS (>THAN 6/MM)^{*}
- NO ULCERATION IS IDENTIFIED
- MICROSCOPIC SATELLITES^{**} ARE ABSENT
- NO LYMPHOVASCULAR INVASION IS IDENTIFIED
- NO TUMOR REGRESSION^{***} IS IDENTIFIED
- NO PRE-EXISTING NEVUS IS IDENTIFIED
- RESECTION MARGINS ARE FREE OF TUMOR

^{*} 1 mm^{*} represents approximately 9 to 10 high power field (HPF) in most X40 objectives.

^{**} Microscopic satellites defined as tumor nests over 50 µm (0.05 mm) in diameter within the reticular dermis, fat tissue, blood vessels, or lymphatics beneath the principal invasive mass, but separated from it by normal connective tissue in serial sections.

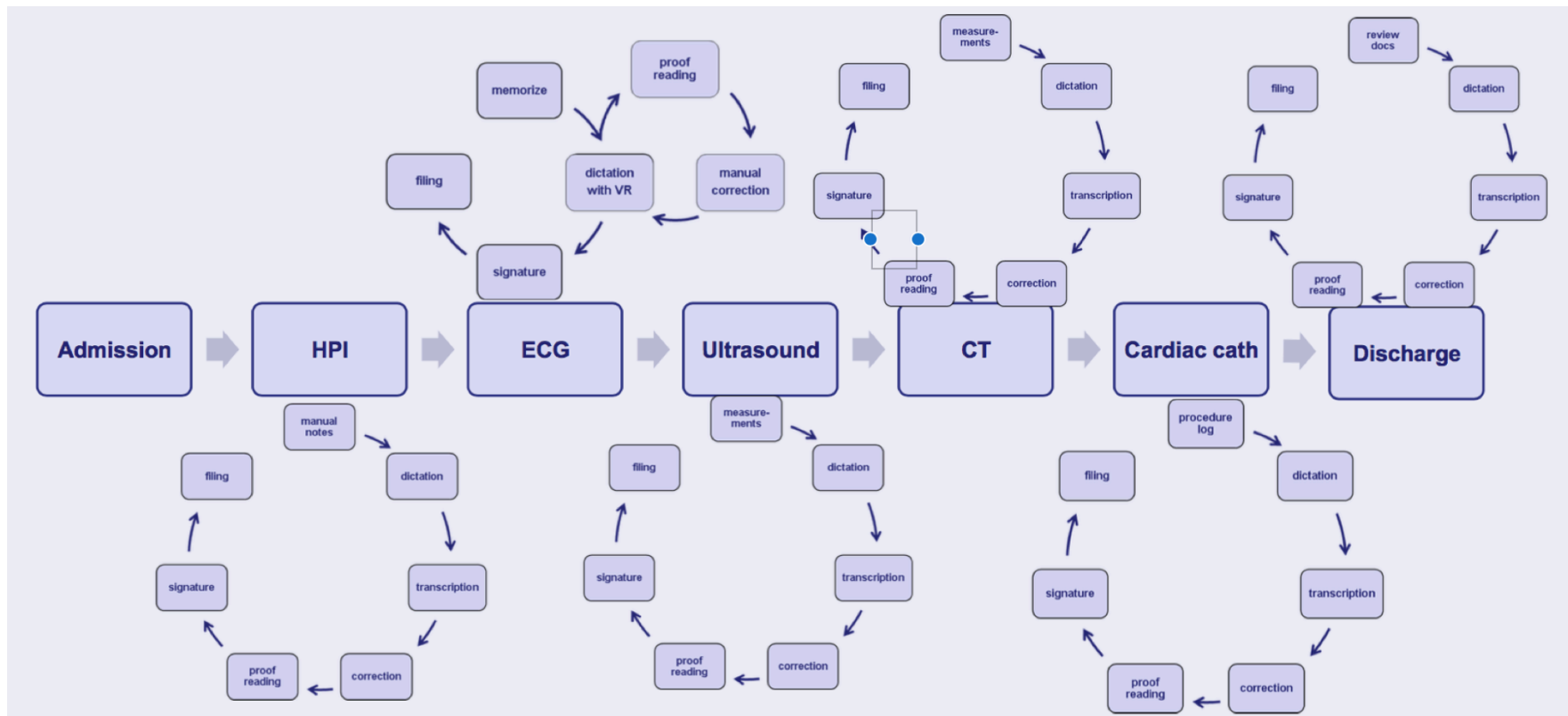
^{***}Regression: Areas often adjacent to radial growth phase characterized by fibrosis, variable dense infiltrate of lymphocytes and melanophages, with dilated and thick-walled blood vessels.



MACROSCOPIC DESCRIPTION:

Specimen designated "skin left face-melanoma" received in formalin and labeled with the patient's name consists of a skin ellipse with underlying fibroadipose tissue 4.5x2.5x1.5 cm in greatest dimension with attached suture indicating superior margin. It reveals a dark brown-black nodule on the surface 1.0 cm from

The “vicious mandala” of clinical document generation



Discharge Letter



Hasso
Plattner
Institut

PHYSICIAN HOSPITAL DISCHARGE SUMMARY

Provider: Ken Cure, MD

Patient: Patient H Sample **Provider's Pt ID:** 6910828 **Sex:** Female

Attachment Control Number: XA728302

HOSPITAL DISCHARGE DX

- 174.8 Malignant neoplasm of female breast: Other specified sites of female breast
- 163.8 Other specified sites of pleura.

HOSPITAL DISCHARGE PROCEDURES

1. 32650 Thoracoscopy with chest tube placement and pleurodesis.

HISTORY OF PRESENT ILLNESS

The patient is a very pleasant, 70-year-old female with a history of breast cancer that originally early 70's. At that time she had a radical mastectomy with postoperative radiotherapy. In the mid a chest wall recurrence and was treated with further radiation therapy. She then went without many years until the late 80's when she developed bone metastases with involvement of her trochanter, and left sacral area. She was started on Tamoxifen at that point in time and has done when she developed shortness of breath and was found to have a larger pleural effusion. This two occasions and has rapidly reaccumulated so she was admitted at this time for thoracoscopy note, her CA15-3 was 44 in the mid 90's and recently was found to be 600.

HOSPITAL DISCHARGE PHYSICAL FINDINGS

Physical examination at the time of admission revealed a thin, pleasant female in mild respiratory no adenopathy. She had decreased breath sounds three fourths of the way up on the right side. mostly clear although there were a few scattered rales. Cardiac examination revealed a regular without murmurs. She had no hepatosplenomegaly and no peripheral clubbing, cyanosis, edema.

HOSPITAL DISCHARGE STUDIES SUMMARY

A chest x-ray showed a large pleural effusion on the right.

HOSPITAL COURSE

The patient was admitted. A CT scan was performed which showed a possibility that the lung was and that there were some adhesions. The patient then underwent thoracoscopy which confirmed pleural peel of tumor and multiple adhesions which were taken down. Two chest subsequently

SAMPLE DISCHARGE SUMMARY

Primary Diagnosis: 40 week IUP with delivery of a liveborn infant

Secondary Diagnosis: Advanced Maternal Age; Prolonged second stage of labor with maternal exhaustion

Procedure Performed:

1. Spontaneous Vaginal Delivery with delivery of live male infant weighing 7# 5oz at 1542 on January 3, 2012 with APGARS of 8 at one minute and 9 at five minutes.
2. Placement of Intrauterine Pressure Catheter.

Reason for Hospitalization: This 36yo G2P1001 presented at 40 weeks gestation by an LMP of 3/12/11 with an EDC of 1/3/12 in spontaneous labor. This pregnancy has been complicated by advanced maternal age. QS performed at 17 weeks was within normal limits and a genetic amniocentesis was offered and declined. Prenatal laboratory data showed blood type B+ with a negative antibody screen, Rubella Immune, VDRL nonreactive, HepBsAg negative, Diabetic Screen 120, HIV nonreactive. She remained normotensive throughout her pregnancy. At the time of admission she reported positive fetal movement and denied loss of fluid.

Physical Exam on Admission: Temperature 98.4. Pulse 94. Respirations 16. Blood pressure 128/78. Fetal Heart Rate 150's and reactive. Uterine contractions q 4 minutes. HEENT within normal limits. Heart regular. Lungs clear. Abdomen gravid with a fundal height appropriate for gestational age. Extremities 2+ DTR's and trace edema. Cervical exam 4 cm/80%/-1.

Lab and X-Ray Data: Predelivery H&H of 12.4 and 36.2 respectively. Platelets 221.

Hospital Course: The patient was admitted in spontaneous labor in the morning of January 3rd. was reactive and reassuring throughout the course of her stay in labor and delivery. Her labor progressed well and at 0900 hours, she had spontaneous rupture of membranes with a return of fluid. At that time, her cervix was dilated to 6 cm/90%/0. Epidural anesthesia was requested and obtained. Her labor then quickly progressed and the patient was noted to be completely dilated at +1 station at 1100 hours. She was then allowed to push. After pushing for 2 hours, the patient brought the vertex to the perineum, but was unable to continue her expulsive efforts. The infant delivered by outlet forceps over a midline episiotomy. *Please see operative report for full details.* The patient and infant did well. She is breast-feeding the infant well, and has remained afebrile with minimal lochia since delivery. The patient was voiding and ambulating without difficulty by the evening of PPD #0. She declined any contraception at the time of discharge, and was deemed stable for discharge on PPD 2.

The hospital as editorial office

- Approx. 50 documents per inpatient spell¹
 - 60-80 % textual and „human generated“
 - A 350 bed institution with an LOS of 7d creates ~1600 of these per day
- A hospital's textual output is comparable to a medium-sized daily newspaper

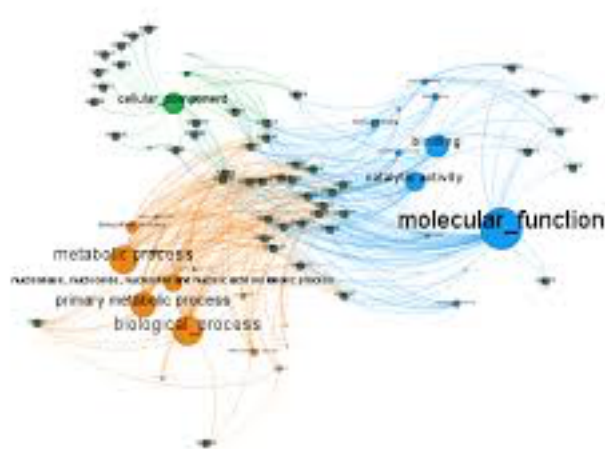
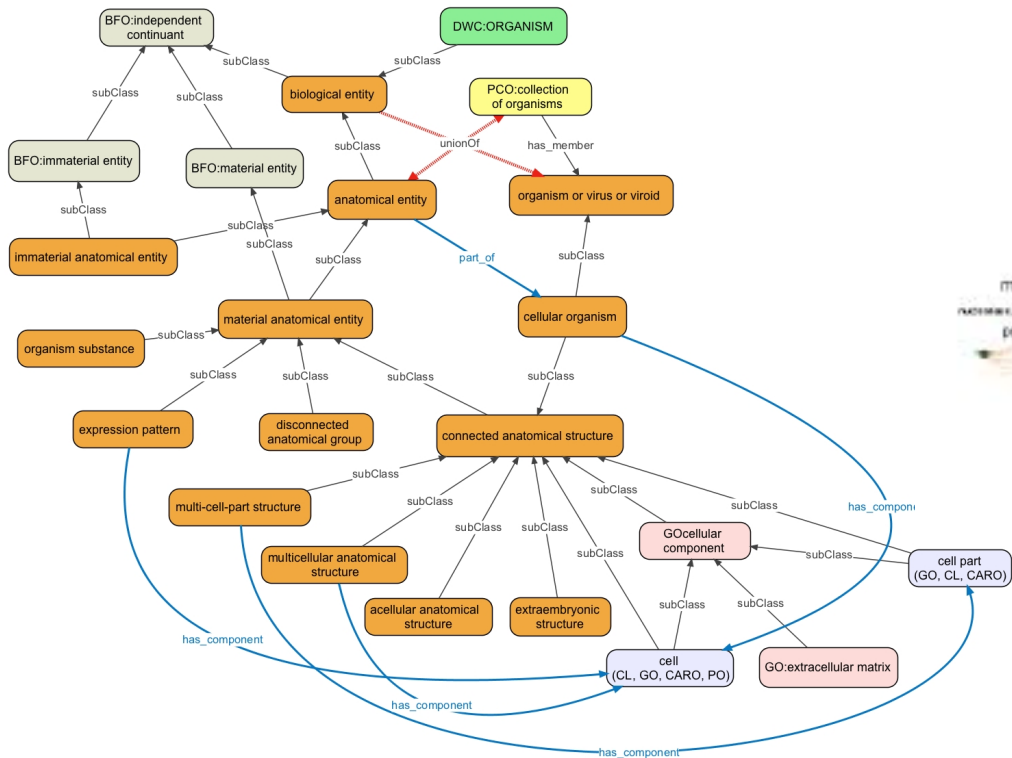
Management of Biomedical Texts

Data Management for
Digital Health, Summer
2017

10

¹Schmücker P, 2012: Dokumentenaufkommen und eArchivierung in Krankenhäusern - Entwicklung und Stand heute.

(Bio-)Ontologies



Management of Biomedical Texts

Data Management for
Digital Health, Summer
2017
11

- Provide rich human and machine understandable descriptions of the terms they purport to describe
- Have value for semantic annotation of data, which allows integration across domains (granularity, species, experimental methods)
- Facilitate granular and cross-domain queries
- Can be used to obtain explanations for inferences drawn
- Can be efficiently processed by algorithms and software

Management of Biomedical Texts

Data Management for
Digital Health, Summer
2017
12

Olelo: a web application for intuitive exploration of biomedical literature

Milena Kraus*, Julian Niedermeier, Marcel Jankrift, Sören Tietböhl, Toni Stachewicz, Hendrik Folkerts, Matthias Uflacker and Mariana Neves

Department of Enterprise Platforms and Integration Concepts, Hasso Plattner Institute, August-Bebel-Strasse 88, Potsdam 14482, Germany

Received January 31, 2017; Revised April 12, 2017; Editorial Decision April 20, 2017; Accepted April 25, 2017

ABSTRACT

Researchers usually query the large biomedical literature in PubMed via keywords, logical operators and filters, none of which is very intuitive. Question answering systems are an alternative to keyword searches. They allow questions in natural language as input and results reflect the given type of question, such as short answers and summaries. Few of those systems are available online but they experience drawbacks in terms of long response times and they support a limited amount of question and result types. Additionally, user interfaces are usually restricted to only displaying the retrieved information. For our Olelo web application, we combined biomedical literature and terminologies in a fast in-memory database to enable real-time responses to researchers' queries. Further, we extended the built-in natural language processing features of the database with question answering and summarization procedures. Combined with a new explorative approach of document filtering and a clean user interface, Olelo enables a fast and intelligent search through the ever-growing biomedical literature. Olelo is available at <http://www.hpi.de/plattner/olelo>.

INTRODUCTION

Researchers all over the world regularly access the MEDLINE/PubMed database, which currently contains over 20 million scientific biomedical publications. Users usually explore this knowledge through simple keyword searches. Although advanced search options are available, these require an exact search target and proper search terms as well as knowledge on how to use the interface. Further, current search engines do not leverage the information contained in abstracts, full texts, medical vocabularies and ontologies to their full extent. Additionally, their user inter-

faces frequently restrict explorative search as a new browser tab needs to be opened for every relevant search result.

Lu *et al.* (1) reviewed a collection of web tools that differ in their search options to the traditional PubMed approach. Among others, the author came to the following observations: (i) most of the engines provide a list of titles and authors of the relevant documents; (ii) in systems that perform a clustering or ranking of documents, the list of documents can usually be expanded; (iii) only few approaches provide other result sets, such as tables or graphs; (iv) improved ranking and usability seem to be popular driving forces for new systems.

Question answering (QA) systems offer a user-friendly alternative to plain keyword searches and have proven to provide exact answers in the biomedical context (2,3). In particular, QA enables three advantages: (i) queries can be posed using natural language instead of keywords; (ii) results are generated according to what has been specifically requested, be it a single answer or a short summary; (iii) answers are usually based on the integration of textual documents and a variety of knowledge sources (3).

Bauer *et al.* (4) surveyed the only three available QA systems for biomedicine, namely askHERMES (5), HONQA (6) and EAGLi (7). The authors identified drawbacks in usability in terms of response time, obstacles in the web interface, as well as restrictions in the types of questions that these tools are able to process.

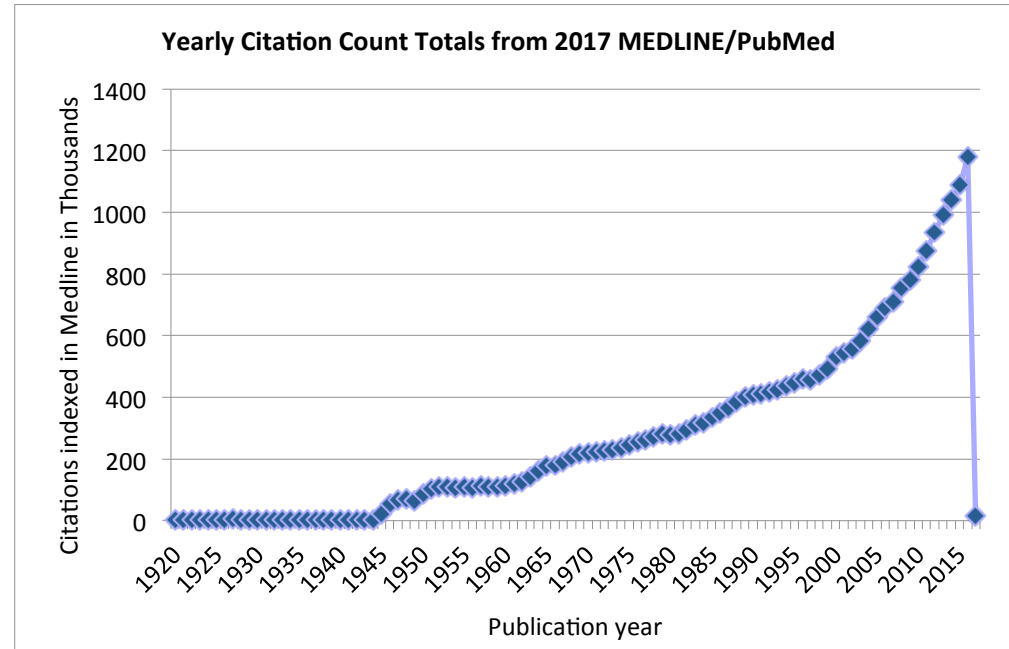
The Olelo Web application is derived from our previously established QA system (8) that was one of the winners in the two previous editions of the BioASQ challenges (www.bioasq.org/participate/winners). It is the entry point to an explorative search through the biomedical literature. An in-memory database (IMDB) holds all data in main memory to enable real-time exploration of the documents. Further, the IMDB provides useful built-in features for natural language processing (NLP), which we extended with advanced algorithms, such as question understanding and multi-document summarization (9).

Pang *et al.* (10) discuss desirable design principles for Web applications, in order to facilitate the explorative search in

*To whom correspondence should be addressed. Tel: +49 331 5509 1366; Fax: +49 5509 579; Email: Milena.Kraus@hpi.de

© The Author(s) 2017. Published by Oxford University Press on behalf of Nucleic Acids Research. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

- Over 27 Mio citations (indexed in PubMed)
- Corpus of literature is growing exponentially
- Researchers somehow have to cope with the large amount of information



Management of Biomedical Texts

Data Management for Digital Health, Summer 2017

General - Types of Scientific Publications

- Methodical paper: New algorithms, systems, etc.
- Review / survey paper: Status quo / current status of a research area
- Concepts paper: New ideas or theories without concrete realization
- Evaluation paper: Quantitative comparison of different approaches
- Technical Report: Notification of current status of an approach within organization, usually no review

General - Administrative Information

- Title
- Authors and affiliations
- Journal name, volume, issue, year, page range, doi
- Copyright and publishing information

```
@article{kraus2017olelo,  
  title={Olelo: a web application for intuitive exploration  
of biomedical literature.},  
  author={Kraus, M and Niedermeier, J and Jankrift, M and  
Tietbohl, S and Stachewicz, T and Folkerts, H and  
Uflacker, M and Neves, M},  
  journal={Nucleic acids research},  
  year={2017}  
}
```

Nucleic Acids Research, 2017 1
doi: 10.1093/nar/gkx363

Olelo: a web application for intuitive exploration of biomedical literature

Milena Kraus*, Julian Niedermeier, Marcel Jankrift, Sören Tietböhl, Toni Stachewicz,
Hendrik Folkerts, Matthias Uflacker and Mariana Neves

Department of Enterprise Platforms and Integration Concepts, Hasso Plattner Institute, August-Bebel-Strasse 88,
Potsdam 14482, Germany

Received January 31, 2017; Revised April 12, 2017; Editorial Decision April 20, 2017; Accepted April 25, 2017

ABSTRACT

Researchers usually query the large biomedical literature in PubMed via keywords, logical operators and filters, none of which is very intuitive. Question answering systems are an alternative to keyword searches. They allow questions in natural language as input and results reflect the given type of question, such as short answers and summaries. Few of those systems are available online but they experience drawbacks in terms of long response times and they support a limited amount of question and result types. Additionally, user interfaces are usually restricted to only displaying the retrieved information. For our Olelo web application, we combined biomedical literature and terminologies in a fast in-memory database to enable real-time responses to researchers' queries. Further, we extended the built-in natural language processing features of the database with question answering and summarization procedures. Combined with a new explorative approach of document filtering and a clean user interface, Olelo enables a fast and intelligent search through the ever-growing biomedical literature. Olelo

www.hpi.de/plattner/olelo.

The world regularly access the database, which currently contains biomedical publications. Users navigate through simple keyword and search options are available, rich target and proper search terms how to use the interface. Further, not leverage the information contexts, medical vocabularies and ontology. Additionally, their user inter-

faces frequently restrict explorative search as a new browser tab needs to be opened for every relevant search result.

Lu *et al.* (1) reviewed a collection of web tools that differ in their search options to the traditional PubMed approach. Among others, the author came to the following observations: (i) most of the engines provide a list of titles and authors of the relevant documents; (ii) in systems that perform a clustering or ranking of documents, the list of documents can usually be expanded; (iii) only few approaches provide other result sets, such as tables or graphs; (iv) improved ranking and usability seem to be popular driving forces for new systems.

Question answering (QA) systems offer a user-friendly alternative to plain keyword searches and have proven to provide exact answers in the biomedical context (2,3). In particular, QA enables three advantages: (i) queries can be posed using natural language instead of keywords; (ii) results are generated according to what has been specifically requested, be it a single answer or a short summary; (iii) answers are usually based on the integration of textual documents and a variety of knowledge sources (3).

Bauer *et al.* (4) surveyed the only three available QA systems for biomedicine, namely askHERMES (5), HONQA (6) and EAGLi (7). The authors identified drawbacks in usability in terms of response time, obstacles in the web interface, as well as restrictions in the types of questions that these tools are able to process.

The Olelo Web application is derived from our previously established QA system (8) that was one of the winners in the two previous editions of the BioASQ challenges (www.bioasq.org/participate/winners). It is the entry point to an explorative search through the biomedical literature. An in-memory database (IMDB) holds all data in main memory to enable real-time exploration of the documents. Further, the IMDB provides useful built-in features for natural language processing (NLP), which we extended with advanced algorithms, such as question understanding and multi-document summarization (9).

Pang *et al.* (10) discuss desirable design principles for Web applications, in order to facilitate the explorative search in

old be addressed. Tel: +49 331 5509 1366; Fax: +49 5509 579; Email: Milena.Kraus@hpi.de

© The Author(s) 2017. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Paper Sections – Example Structures

- Title
- Abstract
- Introduction
- (Background)
- Related Work
- Main Part
- Conclusion
- References

- Title
- Abstract
- Introduction
- (Background)
- Main Part
- Related Work
- Conclusion
- References

See also: IMRAD structure
(<https://en.wikipedia.org/wiki/IMRAD>)

Management of Biomedical Texts

Data Management for
Digital Health, Summer
2017
17

- Usually not more than 140 words
- Reflects the main story of the research paper
- Short and concise sentences
- Always follows a particular structure
 - Scope - What is the general context?
 - Problem - What is the specific problem?
 - Significance - Why is it a problem?
 - Solution - How do you solve it?
 - Evaluation - Does your solution fulfill expectations (very short)?

Paper Sections

Introduction, Background and Related Work

Introduction and Background

- Introduces the topic and defines the terminology
- Explains the focus of the paper and research objectives
- Last paragraph commonly outlines the structure of the paper

Related Work

- Helps to understand the field and the problem
- Compares and differentiates own work with the state of the art
- Strategies of the different approaches, strengths/weaknesses

	Approach A	Approach B	Our Approach
Criteria 1	x	x	x
Criteria 2	x	-	x
Criteria 3	x	x	x
Criteria 4	-	-	x

Management of Biomedical Texts

Data Management for
Digital Health, Summer
2017

Paper Sections – Main Part and Further Elements

- Materials and Methods – Explains the experimental setup.
- Results – Illustrates the observations made.
- Evaluation and Discussion – Finds, explains and discusses reasons for observations.
- Conclusion – Answers the research questions and explains importance of discovery and future implications

Further Elements

- Figures
- Tables
- Captions
- Footnotes
- References

- Text documents are primarily unstructured
 - Scientific papers follow, e.g., the IMRAD structure
 - Abstract and title should contain a large amount of the important information
 - Methods are mostly relevant for deep dives on specific parts of the paper
 - Future work is usually not relevant for information extraction
-
- The knowledge of the rough contents of a paper guides the development of processing tools with a specific purpose, e.g., information extraction.
 - To reduce the search space and noise, it is common and advisable to not process all sections of a paper.

Search Engines (for Biomedical Contents)

System	Service provider	Data	Data size	Discipline
PubMed	National Library for Medicine	MEDLINE/PubMed, journals, books, only titles and abstracts	26 million	Biomedical
PubMed Central	National Library for Medicine	Full-text life science journals	3.9 million (1.3m open access)	Biomedical
Quetzal	Quertle	MEDLINE/PubMed, PMC open access, Toxline, Medicine Patent grants, Health/life science news, AHRQ† treatment guidelines, NIH‡ grants	~35 million	Biomedical
Google Scholar	Google	Scholarly documents on the internet	~160 million	General

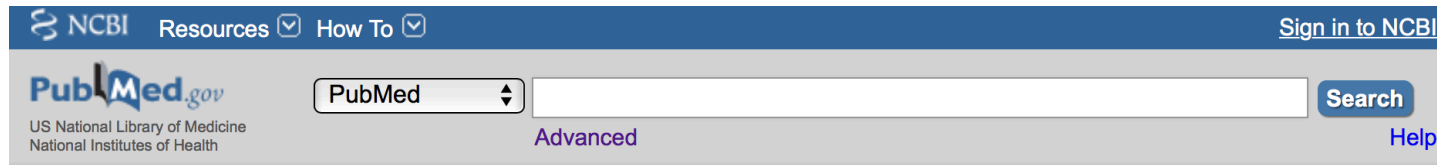
Management of Biomedical Texts

Data Management for
Digital Health, Summer
2017
22

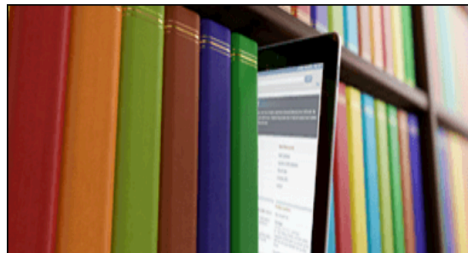
PubMed/Medline

Search Engine for Biomedical Literature

- Over 27 million bibliographic records and abstracts
- Indexes articles from MEDLINE, ejournals, ebooks and more
- Organization via Medical Subject Headings (MeSH)
- U.S. National Library of Medicine and National Institutes of Health



The screenshot shows the top navigation bar of the NCBI website with links for 'Resources' and 'How To'. Below this is the 'PubMed.gov' logo and the text 'US National Library of Medicine National Institutes of Health'. A search bar contains the text 'PubMed' and a dropdown arrow. To the right of the search bar is a 'Search' button. Below the search bar are links for 'Advanced' and 'Help'. A 'Sign in to NCBI' link is located in the top right corner of the navigation bar.



PubMed

PubMed comprises more than 27 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

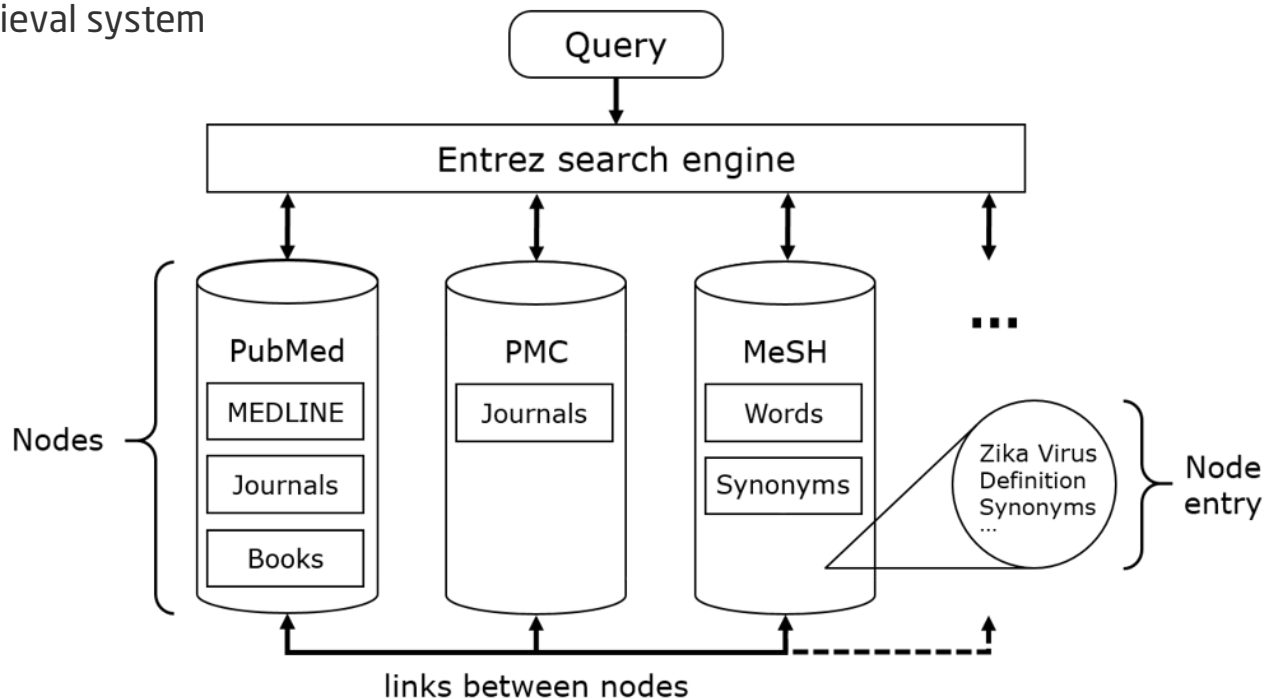
Management of Biomedical Texts

Data Management for
Digital Health, Summer
2017

PubMed

Entrez Global Query Cross-Database Search System

- NCBI's federated text search and retrieval system



Management of Biomedical Texts

Data Management for
Digital Health, Summer
2017

Document preprocessing:

- Tokenization: split text into words/token
- Removal of stop words, e.g., retrieved from a stop word list
- Stemming: reduce an inflected word to its word stem by removing its affixes → maps similar word to one stem and therefore enables synonym search

Storage:

- Words are stored together with their occurrence count within the text and a
- Weight, which depends on the location of the word ($\text{weight}(\text{word in title}) > \text{weight}(\text{word in text})$) + Bonus if MeSH term

Discovery Features:

- Similar documents, e.g., “similarity” can be calculated via weighing metrics
 - Calculation: Add up all scores of all the terms two publications have in common
 - Similar articles are pre-computed for every document in the database
- Links between nodes, e.g., a link between the definition of a gene and publications the gene is described in

PubMed Advanced Search

NCBI Resources ☒ How To ☒ Sign in to NCBI

PubMed Home More Resources Help

PubMed Advanced Search Builder

YouTube Tutorial

Use the builder below to create your search

[Edit](#)

[Clear](#)

Builder

All Fields <input type="button" value="v"/>	<input type="text"/>	<input type="button" value="-"/>	Show index list
AND <input type="button" value="v"/>	All Fields <input type="button" value="v"/>	<input type="button" value="-"/> <input type="button" value="+"/>	Show index list

or [Add to history](#)

- Key words and MeSH terms
- Boolean Operators

Management of Biomedical Texts

Data Management for
Digital Health, Summer
2017
27

PubMed ▾ cancer therapy | ✕

Search

[Create RSS](#) [Create alert](#) [Advanced](#)

[Help](#)

Article types

Clinical Trial
Review
Customize ...

Text availability

Abstract
Free full text
Full text

PubMed

Commons
Reader comments
Trending articles

Publication dates

5 years
10 years
Custom range...

Species

Humans
Other Animals

[Clear all](#)

[Show additional filters](#)

Format: Summary ▾ **Sort by:** Most Recent ▾ **Per page:** 20 ▾

Send to ▾

Filters: [Manage Filters](#)



Search Tip

Sort by **Best Match** to display results from highest to lowest relevance to your search terms.

[Try it Now](#)

Results by year



[Download CSV](#)

Search results

Items: 1 to 20 of 610930

<< First < Prev Page 1 of 30547 Next > Last >>

1. ☐ [A novel systemic immune-inflammation index predicts survival and quality of life of patients after curative resection for esophageal squamous cell carcinoma.](#)
Wang L, Wang C, Wang J, Huang X, Cheng Y.
J **Cancer** Res Clin Oncol. 2017 Jun 10. doi: 10.1007/s00432-017-2451-1. [Epub ahead of print]
PMID: 28601935
2. ☐ [Survival improvement in hormone-responsive young breast cancer patients with endocrine therapy.](#)
Yoon TI, Hwang UK, Kim ET, Lee S, Sohn G, Ko BS, Lee JW, Son BH, Kim S, Ahn SH, Kim HJ.
Breast **Cancer** Res Treat. 2017 Jun 10. doi: 10.1007/s10549-017-4331-4. [Epub ahead of print]
PMID: 28601930
3. ☐ [Osimertinib reactivated immune-related colitis after treatment with anti-PD1 antibody for non-small cell lung cancer.](#)

Related searches

breast **cancer therapy**
cancer therapy review
lung **cancer therapy**
prostate **cancer therapy**
target **cancer therapy**

Titles with your search terms

Accelerated versus standard epirubicin followed by cyclophosphamide, meth [Lancet Oncol. 2017]
Human DNA (cytosine-5)-methyltransferases: A functional and structural perspe [Biochimie. 2017]

Terminologies:

- MeSH (Medical Subject Headings)
- UMLS (Unified Medical Language System)
- ICD-10 (International Classification of Diseases)

Ontologies:

- The Gene Ontology (GO)
- Sequence Ontology
- Model Organisms
- Functional Genomics Data
- Ontology for Biomedical Investigations

Terminology in PubMed

Medical Subject Headings (MeSH)

- MeSH is the NLM controlled vocabulary thesaurus used for indexing articles for PubMed.
- Hierarchically-organized terminology for indexing and cataloging of biomedical information
- Example:

Anatomy
 Body Regions
 Head
 Ear

Anatomy
 Sense Organs
 Ear
 Ear, External +
 Ear, Middle +
 Ear, Inner +

Management of Biomedical Texts

Data Management for
Digital Health, Summer
2017
30

MeSH Tree Structures Headings

- | | |
|--|--|
| A. Anatomy | I. Anthropology, Education, Sociology and Social Phenomena |
| B. Organisms | |
| C. Diseases | J. Technology, Industry, Agriculture |
| D. Chemicals and Drugs | K. Humanities |
| E. Analytical, Diagnostic and Therapeutic Techniques and Equipment | L. Information Science |
| F. Psychiatry and Psychology | M. Named Groups |
| G. Phenomena and Processes | N. Health Care |
| H. Disciplines and Occupations | V. Publication Characteristics |
| | Z. Geographicals |

Management of Biomedical Texts

Data Management for
Digital Health, Summer
2017

MeSH Tree Structures

Headings and Subheadings

- Each branch has many levels of sub-branches, and each heading has a position in the hierarchy.
- Some terms appear in more than one branch of the tree.
- Subheadings are arranged in logical hierarchical groupings (families).
- Subheadings and appearance in multiple branches results in “explosion” of search space

Anatomy

Body Regions

Head

Ear

Anatomy

Sense Organs

Ear

Ear, External +

Ear, Middle +

Ear, Inner +

**Management of
Biomedical Texts**

Data Management for
Digital Health, Summer
2017

32

Subject indexing includes:

- reviewing a journal article (or other material such as a letter or editorial)
- determining its subject content, and
- describing that content using a controlled vocabulary.

The purpose of indexing with controlled vocabulary is:

- to facilitate search retrieval by eliminating (or accounting for) the use of variant terminology for the same concept.



FullWord ▾

Exact Match

All Fragments

Any Fragment

☐ All Terms☒ Main Heading (Descriptor) Terms☐ Qualifier Terms☐ Supplementary Concept Record Terms☐ MeSH Unique ID☐ Search in all Supplementary Concept Record Fields☐ Heading Mapped To☐ Indexing Information☐ Pharmacological Action☐ Search Related Registry and CAS Registry/EC Number/UNII Code (RN)☐ Related Registry Search☐ CAS Registry/EC Number/UNII Code (RN)☐ Search in all Free Text Fields

Sort by:

Relevance ▴ ▾

Results per Page:

20 ▴ ▾

Myocardial Infarction

MeSH Descriptor Data 2017

[Details](#)[Qualifiers](#)[MeSH Tree Structures](#)[Concepts](#)**MeSH Heading** Myocardial Infarction**Tree Number(s)** [C14.280.647.500](#)
[C14.907.585.500](#)

- Multiple tree numbers leading to the same term

Unique ID D009203**Annotation** do not coordinate with [ACUTE DISEASE](#) for "acute infarct"**Scope Note** [NECROSIS](#) of the [MYOCARDIUM](#) caused by an obstruction of the blood supply to the heart ([CORONARY CIRCULATION](#)).**Entry Term(s)** Cardiovascular Stroke

Heart Attack

Myocardial Infarct

- Synonyms of the term

NLM Classification # WG 310**See Also** [Heart Rupture, Post-Infarction](#)**Public MeSH Note** 79; was MYOCARDIAL INFARCT 1963-78**Online Note** use MYOCARDIAL INFARCTION to search MYOCARDIAL INFARCT 1966-78**History Note** 79; was MYOCARDIAL INFARCT 1963-78**Date Established** 1966/01/01



Myocardial Infarction MeSH Descriptor Data 2017

[Details](#)[Qualifiers](#)[MeSH Tree Structures](#)[Concepts](#)

Cardiovascular Diseases [C14]

Vascular Diseases [C14.907]

Myocardial Ischemia [C14.907.585]

Myocardial Reperfusion Injury [C14.907.585.625]

Acute Coronary Syndrome [C14.907.585.124]

Angina Pectoris [C14.907.585.187] +

Coronary Disease [C14.907.585.250] +

Myocardial Infarction [C14.907.585.500] -

Anterior Wall Myocardial Infarction [C14.907.585.500.093]

Inferior Wall Myocardial Infarction [C14.907.585.500.187]

Non-ST Elevated Myocardial Infarction [C14.907.585.500.656]

Shock, Cardiogenic [C14.907.585.500.750]

ST Elevation Myocardial Infarction [C14.907.585.500.875]

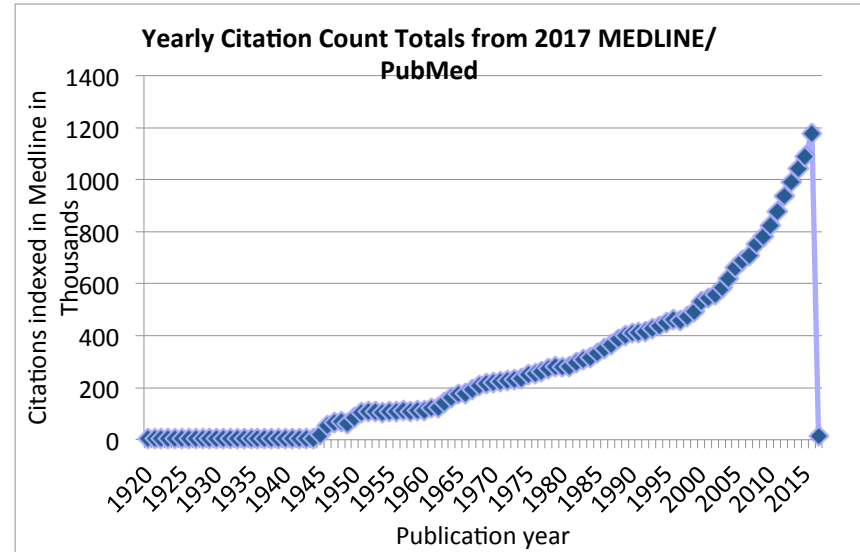
No-Reflow Phenomenon [C14.907.585.500.562]

The MEDLINE Indexing Process: Determining Subject Content

1. Read carefully and understand the title.
2. Read the introduction, looking for the purpose of the article.
3. Scan the body of the article, focus on the Materials & Methods section and the Results section.
4. Note section headings, paragraph headings; italics, boldface; charts, plates, tables, illustrations; laboratory methods, case reports, etc.
5. Select for indexing only those subjects actually discussed as opposed to those subjects merely mentioned.
6. Read the summary or conclusions of the author to determine whether the stated purpose was achieved. Do not index implications or suggested future applications. Do not index conclusive statements not supported by the text.
7. Scan the abstract for items missed, verifying that the text supports indexing these concepts.
8. Scan the author's own indexing or the keywords supplied by the publisher to see whether the concepts chosen are actually discussed in the text.
9. Scan the bibliographic references supplied by the author for clues and further corroboration.

Other ways of coping with millions of documents...

- Prof. Köttgen (Charité Berlin): Manually curates a literature corpus of new and relevant publications in all fields of medicine.
- Deutsche Krebsgesellschaft: Employs professionals to extract and summarize all relevant information from oncology publications.
- Prof. Affeld (Charité): "What can be better than PubMed Advanced Search?"
- Automatic/semi-automatic approaches?



Management of Biomedical Texts

Data Management for
Digital Health, Summer
2017

Definition: QA systems take questions expressed in natural language (e.g., English) and generate a precise answer by linguistically and semantically processing both the questions and data sources under consideration.

Advantages:

- Queries can be posed using natural language instead of keywords
- Results are generated according to what has been specifically requested, be it a single answer or a short summary
- Answers are usually based on the integration of textual documents and a variety of knowledge sources

Demo of all online available QA systems

- <http://www.hpi.de/plattner/olelo>
- <http://services.hon.ch/cgi-bin/QA10/qa.pl>
- <http://www.askhermes.org>
- <http://eagl.unige.ch/EAGLi/>

Example questions:

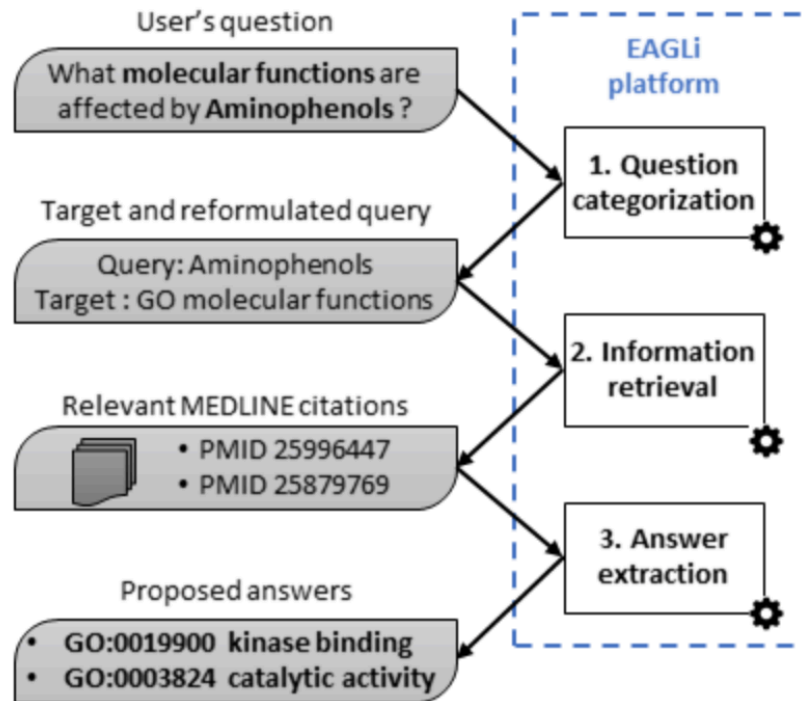
- What is Zika virus?
- What are the diseases caused by Zika virus?
- How to treat Guillain-Barre syndrome?

- Any audience questions?

- HONQA relies on certified websites from the Health On The Net (HON) to extract their answers from and considers a variety of question types.
- Additionally, questions can be posed also in French and Italian.
- The system rely on UMLS to detect the type of the expected answer and it follows the typical architecture of QA systems, but no details are presented in the publication.

EAGLi

- EAGLi extracts the answers exclusively from PubMed abstracts and returns a list of concepts as answers.
- When no answer is found, the system returns a list of potential relevant publications, along with selected passages.
- The system indexes Medline locally with the Terrier information retrieval platform and uses the Okapi BM25 as weighting scheme to rank documents.
- The answers provided by the system are based on the Gene Ontology (GO) concepts.

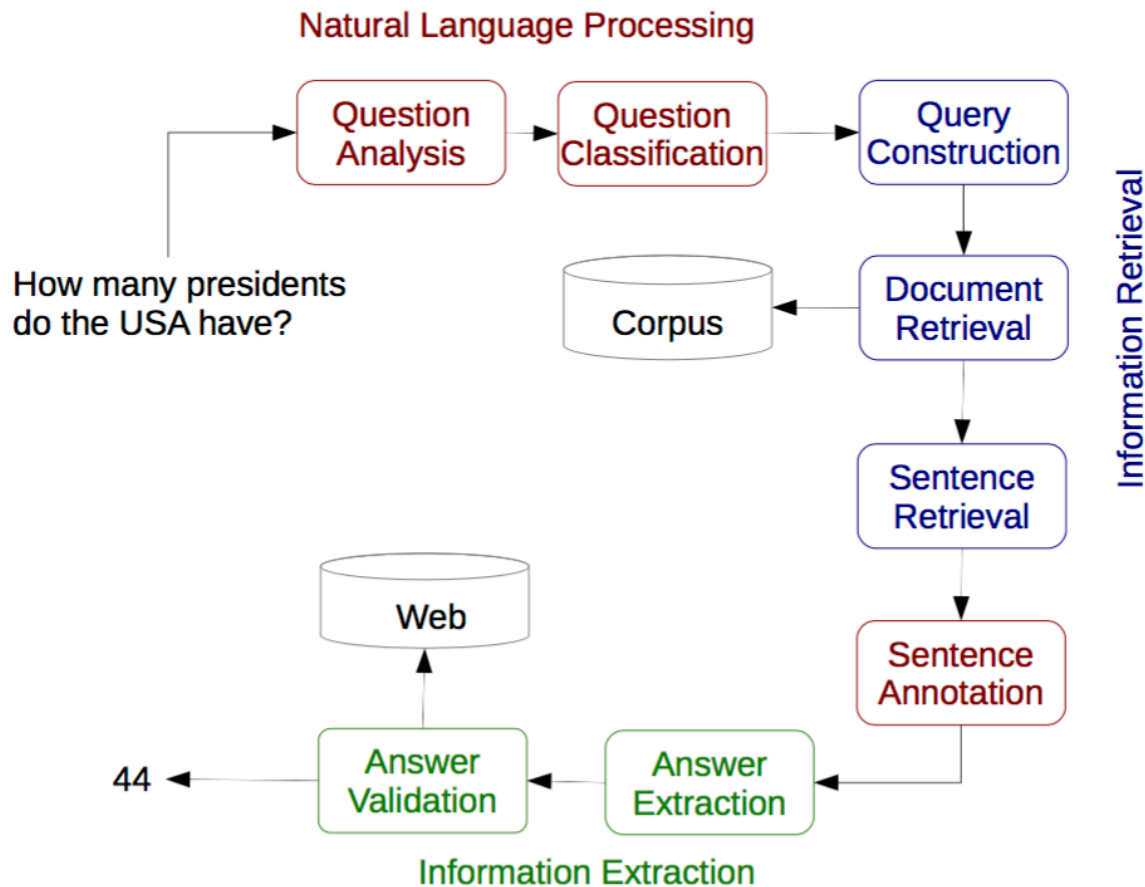


Management of Biomedical Texts

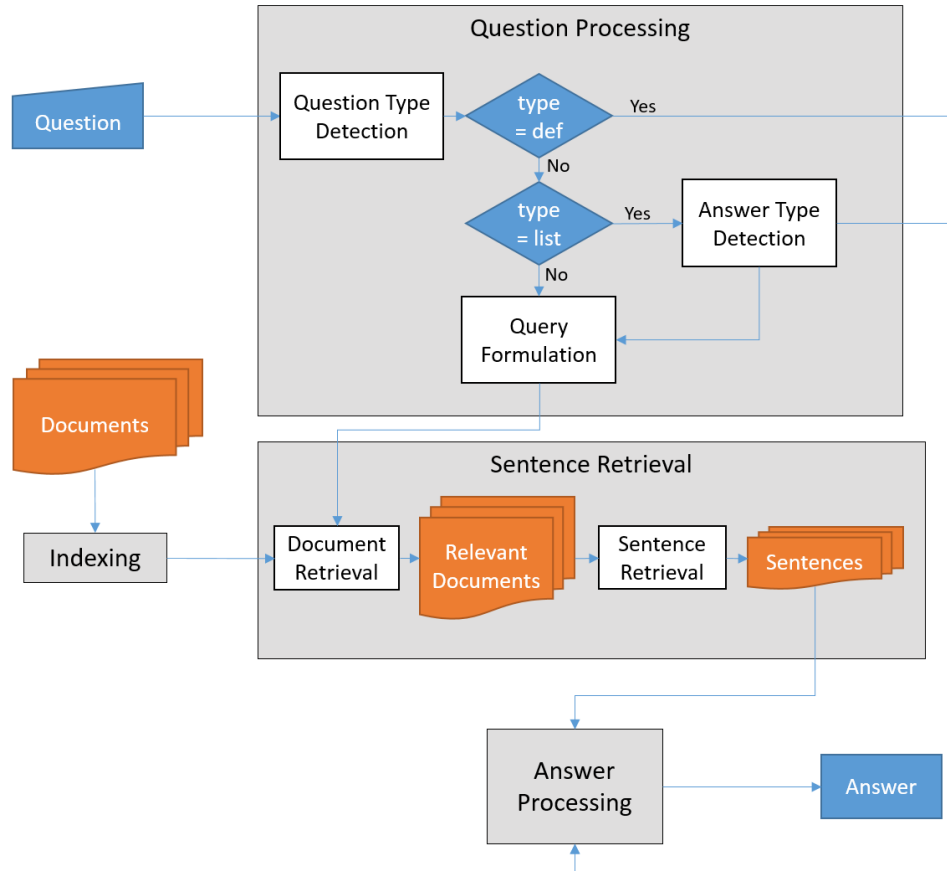
Data Management for
Digital Health, Summer
2017

- askHermes extracts answers from various sources, e.g., PubMed and Wikipedia, and presents answers as a cluster of terms, a ranked list or clustered by content, along with the corresponding relevant passages.
- However, the result page tends to be very long and contains more information than most users can deal with.
- The methods behind askHermes include regular expressions from question understanding, classification into 12 topics and keyword identification, both based on machine learning approaches, and the use of the MetaMap system for concept recognition.
- Document indexing is based on the BM25 model and passage ranking is based on the longest common subsequence (LCS) score.

Question Answering Process



Question Answering Process



Shortcomings of Current Approaches

■ Tedious advanced search

Builder

	Title/Abstract	Breast cancer
AND	MeSH Subheading	neoplasms
AND	Text Word	Treatment
AND	MeSH Terms	breast cancer 1 protein
AND	MeSH Terms	breast

Search or [Add to history](#)

Search results

Items: 1 to 20 of 311

<< First < Prev Page 1 of 16 Next > Last >>

- ☐ [Zika virus infection spread through saliva - a truth or myth?](#)
 1. Siqueira WL, Moffa EB, Mussi MC, Machado MA. Braz Oral Res. 2016;30(1):e46. Epub 2016 Mar 15. PMID: 26981761 [Similar articles](#)
- ☐ [Zika, or the burden of uncertainty](#)
 2. Villa R. Clin Ter. 2016 Jan-Feb;167(1):7-9. doi: 10.7417/CT.2016.1907. PMID: 26980631 [Similar articles](#)

- List of Articles
- New tab search

■ Plain Text Results

Abstract

Diabetic foot ulcers (DFUs) have a significant impact on patient quality of life. A prospective, descriptive pilot study was conducted between May 2012 and December 2013 through the dermatology outpatient unit in a Brazilian hospital to evaluate the clinical benefits of using *Calendula officinalis* hydroglycolic extract in the treatment of DFUs. Patients diagnosed with a stable neuropathic ulcer of >3 months' duration; ranging in size from 0.5-40 cm²; without osteomyelitis, gangrene, bone exposure, cancer, or deep tissue infection; ages 18-90 years; with

Management of Biomedical Texts

Data Management for
Digital Health, Summer
2017
46

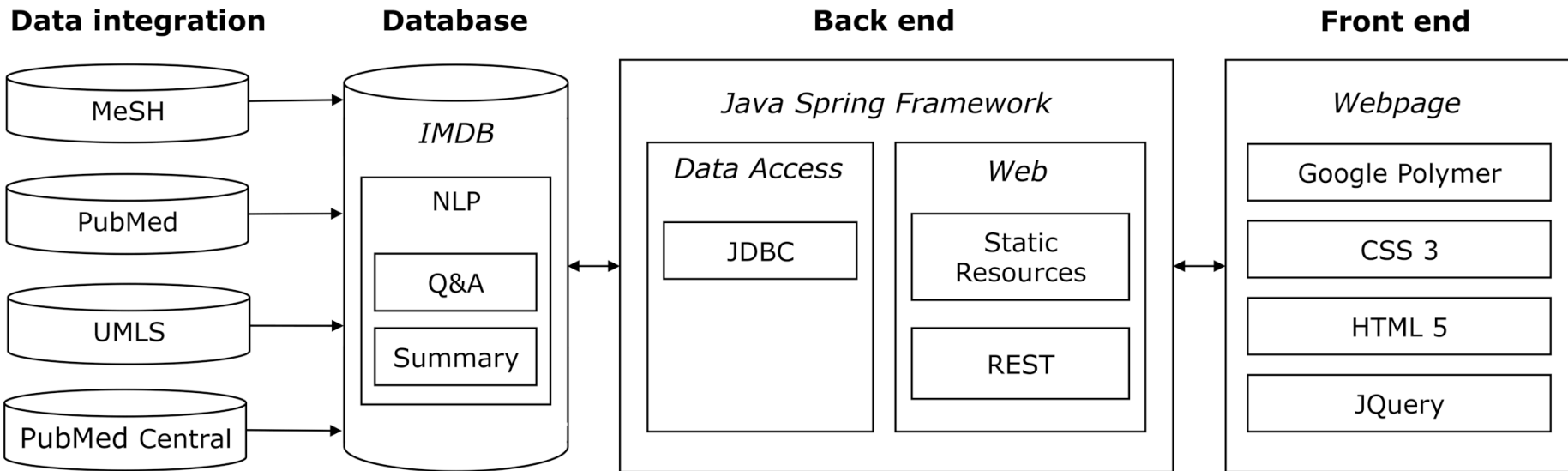
<http://www.hpi.de/plattner/olelo>

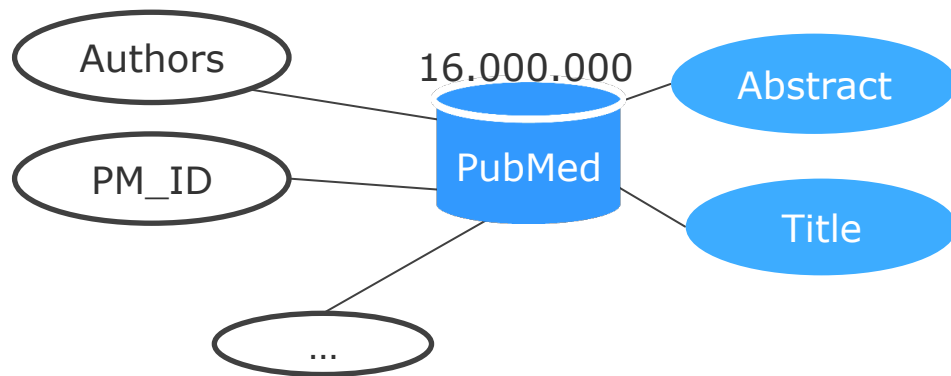
Aims for Olelo

1. Explorative Search
2. Uncover new connections/relationships
3. Simple to use interface
4. Look-up of medical terminology
5. Eliminate new tab search

Management of Biomedical Texts

Data Management for
Digital Health, Summer
2017





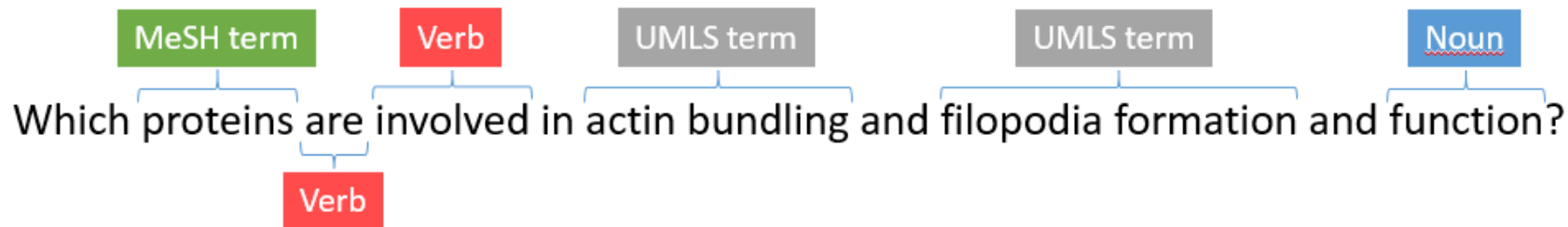
Dictionary matches

	+		
3.200.000.000		2.700.000.000	
321.000.000		360.000.000	

Management of Biomedical Texts

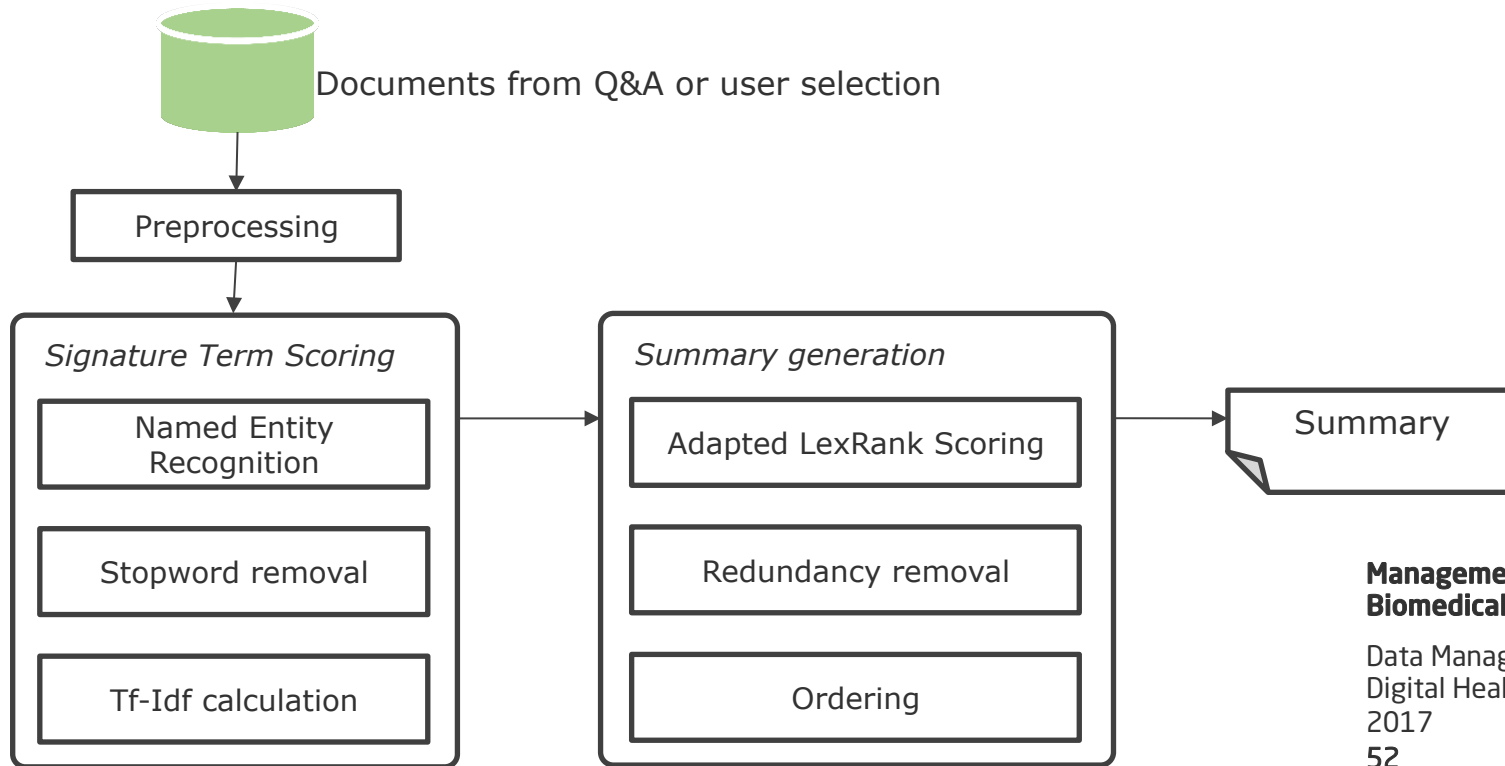
AB	PMID	AB	TA_RULE	12	TA_COUNTER	AB	TA_TOKEN	AB	TA_LANGUAGE	AB	TA_TYPE	AB	TA_NORMALIZED
	26925496		LXP		2		Zika		en		proper name		zika
AB	TA_STEM	12	TA_PARAGRAPH	12	TA_SENTENCE		TA_CREATED_AT	12	TA_OFFSET	12	TA_PARENT		
	?		1		1		11.05.2016 12:47:46.0		4		?		

Management for
Digital Health, Summer



- (1) MeSH terms,
- (2) proper names
- (3) Nouns
- (4) UMLS terms
- (5) adjectives/verbs/adverbs.

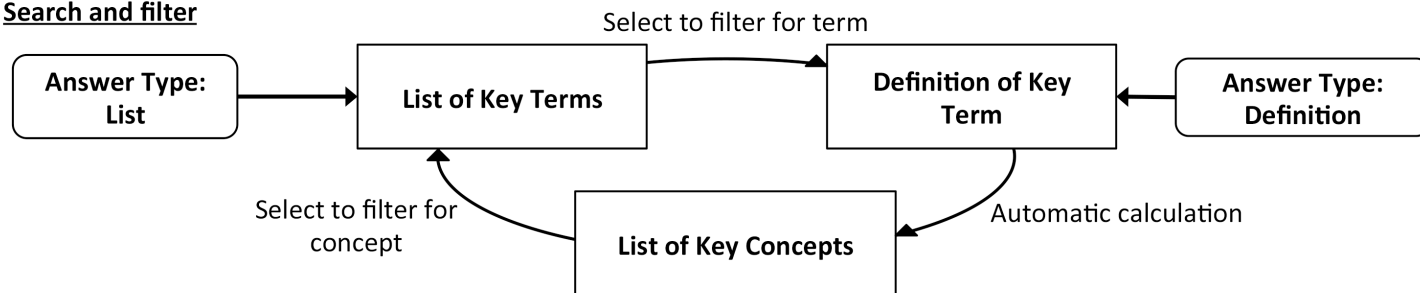




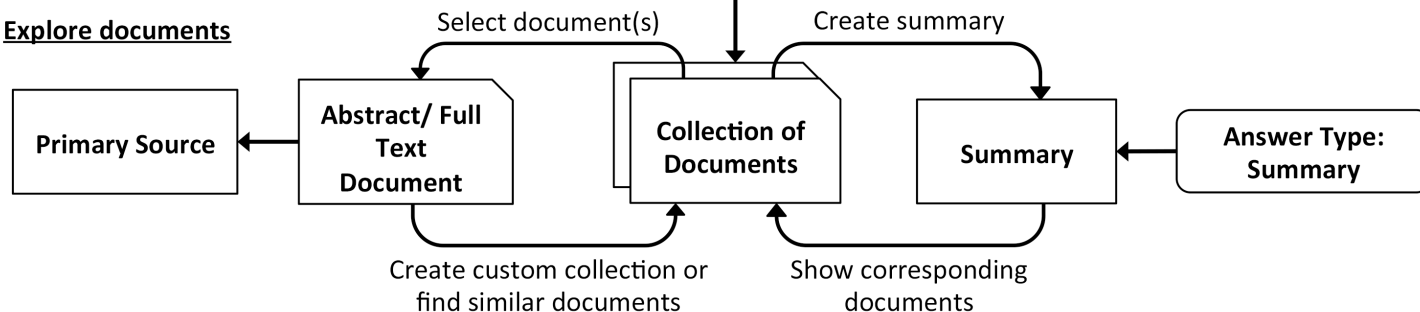
Olelo

Summary of Key Features

Search and filter



Explore documents



Thank you!
