Automated Metadata Extraction: Finding Topics in Texts

Prof. Dr. Klaus U. Schulz

Centrum für Informations- und Sprachverarbeitung
Ludwig-Maximilians-Universität München

TopicZoom GmbH

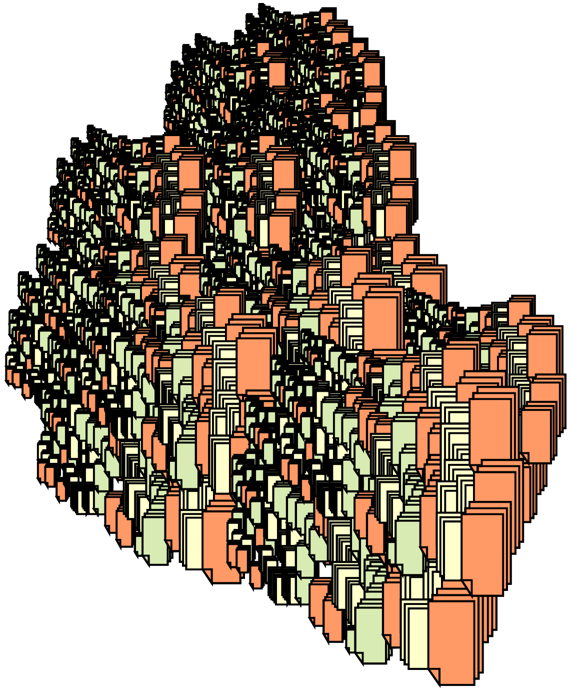
Survey

1. Problem addressed: access to floods of documents
2. Traditional approaches and shortcomings
3. Automated Topic Extraction
   - What and what for?
   - 3 tasters
   - How does it work?
   - Risks and disadvantages
   - Advantages and new possibilities

Live Demo

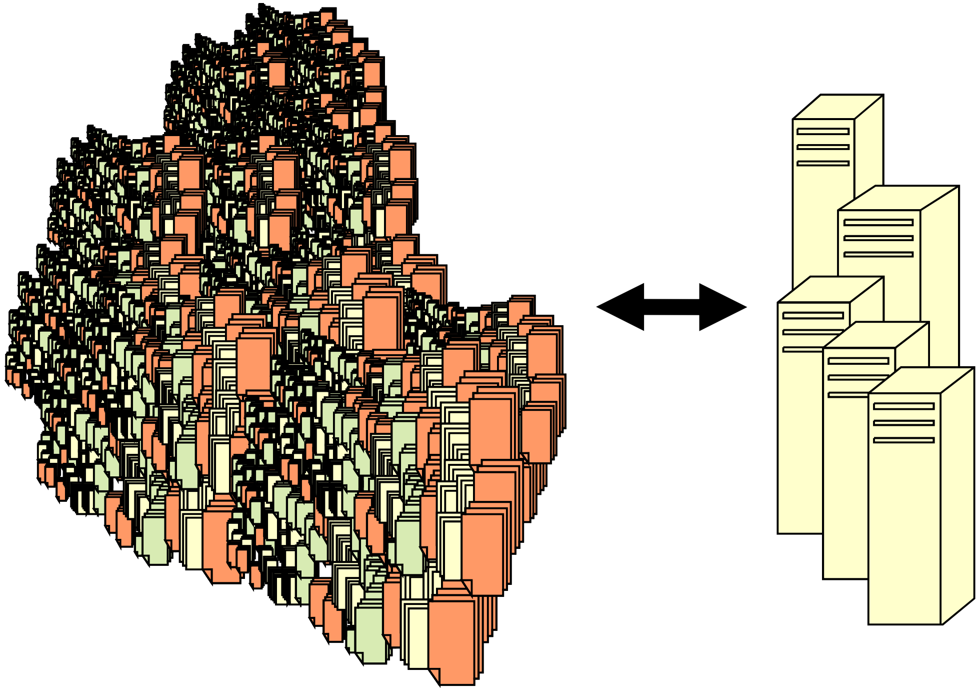
Problem adressed

Problem adressed



Unordered ocean of documents…

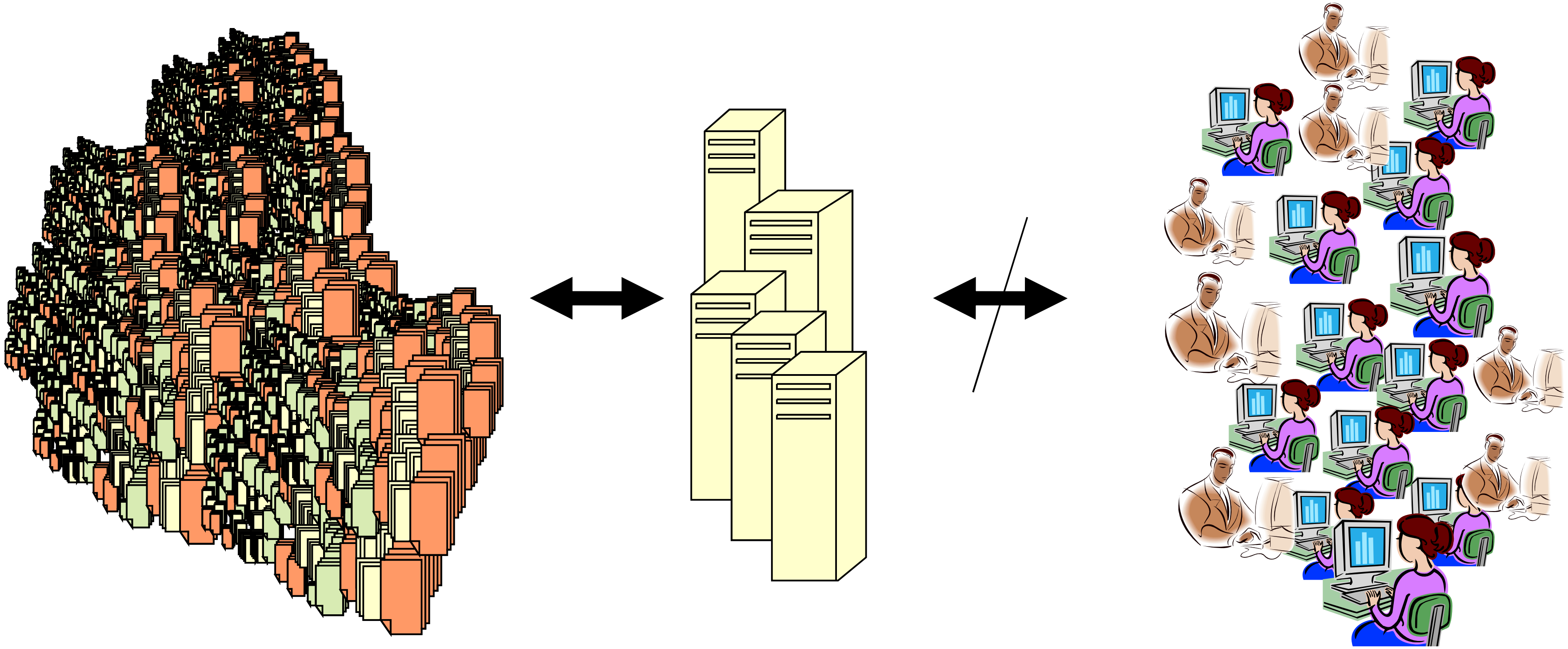
Problem adressed



Unordered ocean of documents…

(full) texts electronically available but

Problem adressed



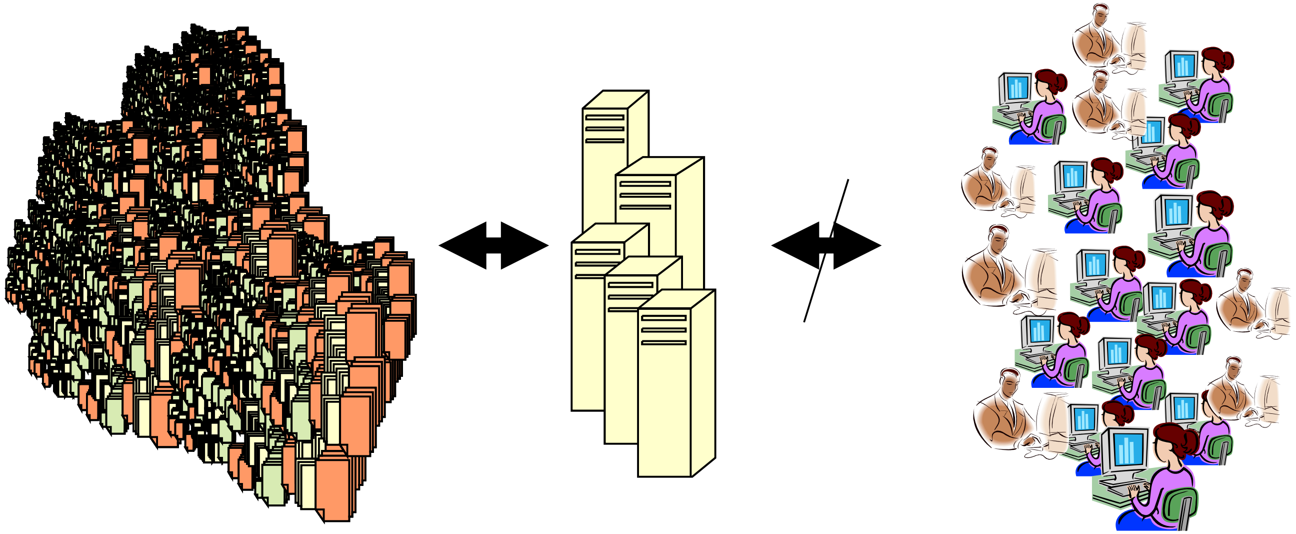
Unordered ocean of documents…

(full) texts electronically available but

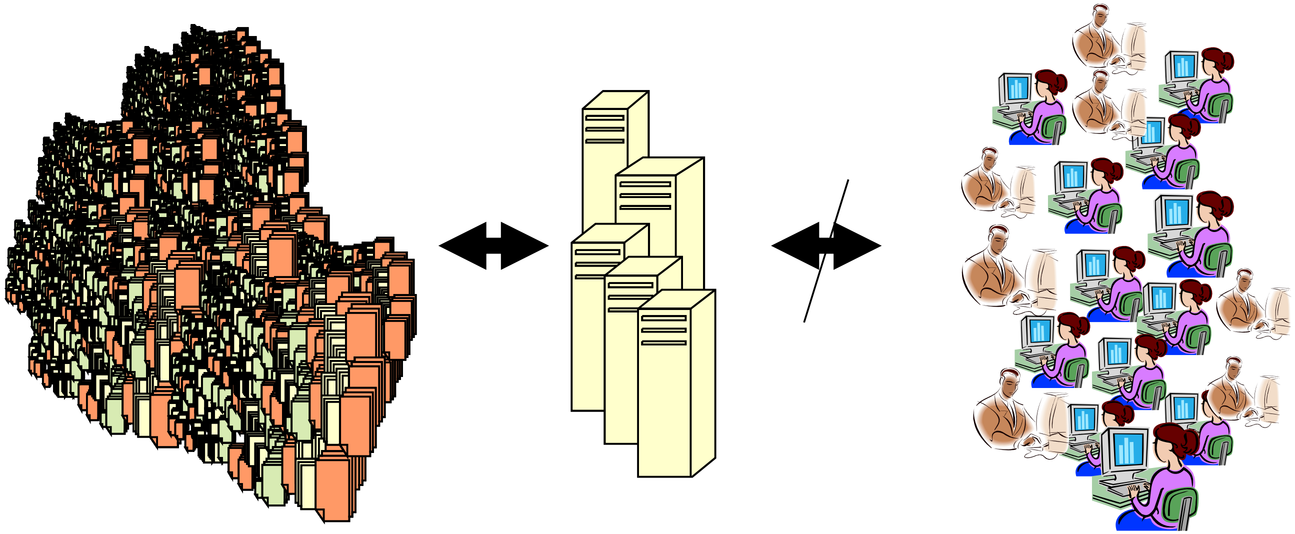
content widely unkown

many users with distinct interests want „intelligent“ access

Three actual example domains



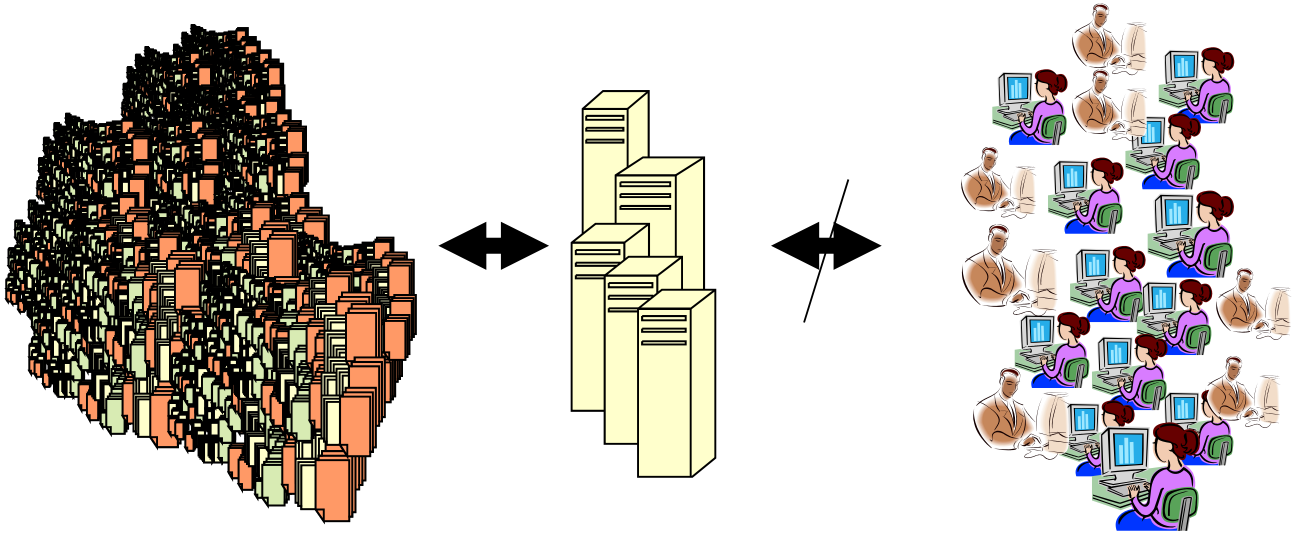
Three actual example domains



Enterprise documents (Emails, memos, contracts, intranet,…)

Employes and managers of enterprise

Three actual example domains



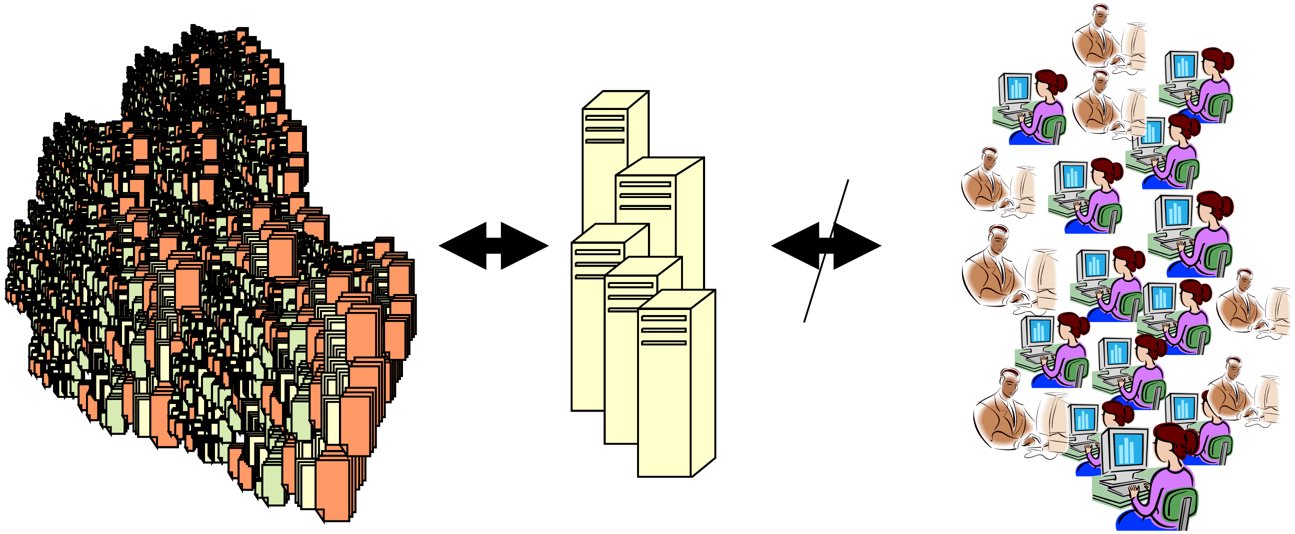
Enterprise documents (Emails, memos, contracts, intranet,…)

Employes and managers of enterprise

Web & social media (news, face book, twitter,…)

Web users, analysts,…

Three actual example domains



Enterprise documents (Emails, memos, contracts, intranet,…) Employes and managers of enterprise

Web & social media (news, face book, twitter,…) Web users, analysts,…

Digital libraries (large-scale digitization programs, Google Books,…) Library users

How to offer “intelligent” access to a flood of “new” electronic text documents?

How to offer “intelligent” access to a flood of “new” electronic text documents?

Established approaches

- Traditional full text search?
- Google?

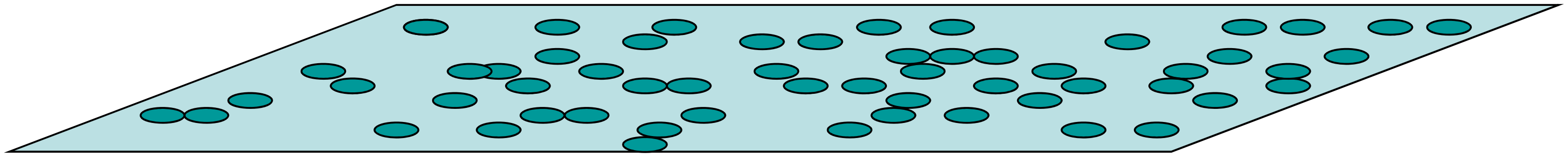
How to offer “intelligent” access to a flood of “new” electronic text documents?

Established approaches

→ • Traditional full text search?

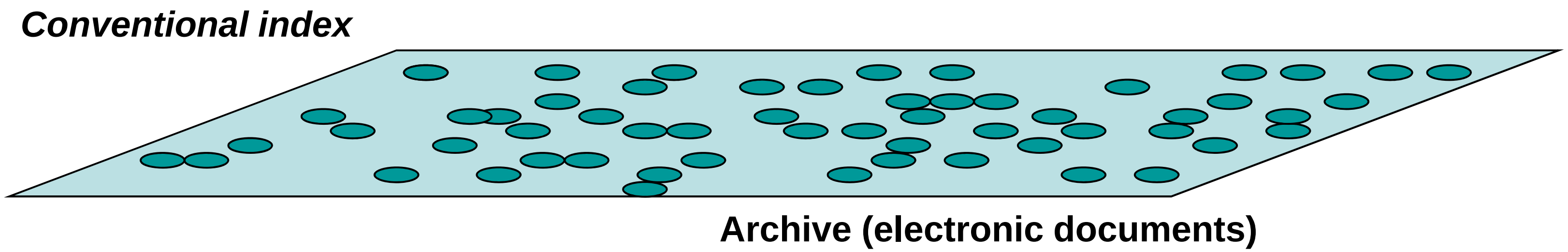
• Google?

Accessing electronic archives via full text search

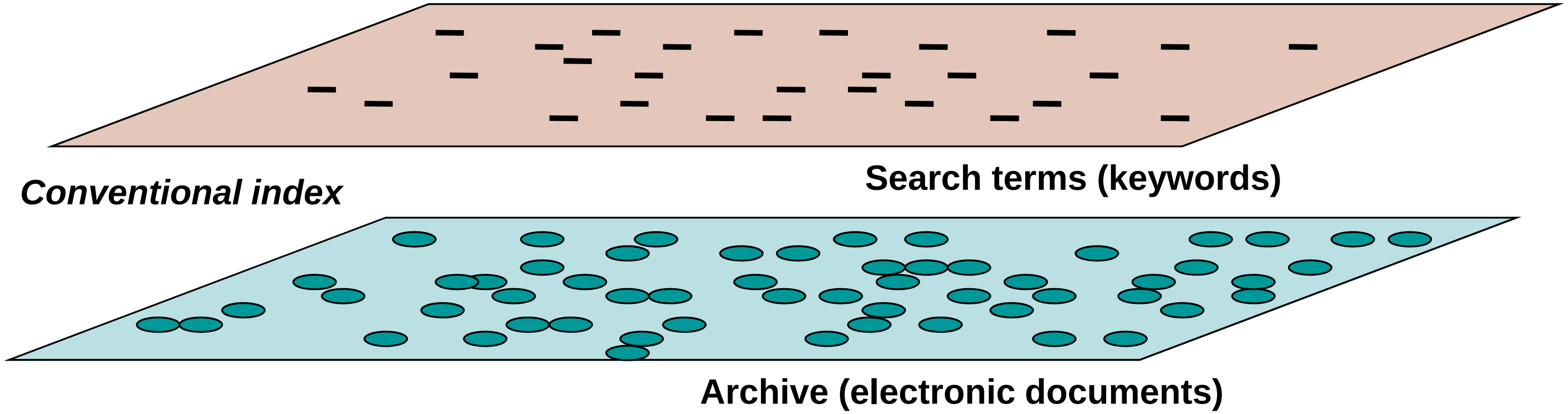


Archive (electronic documents)

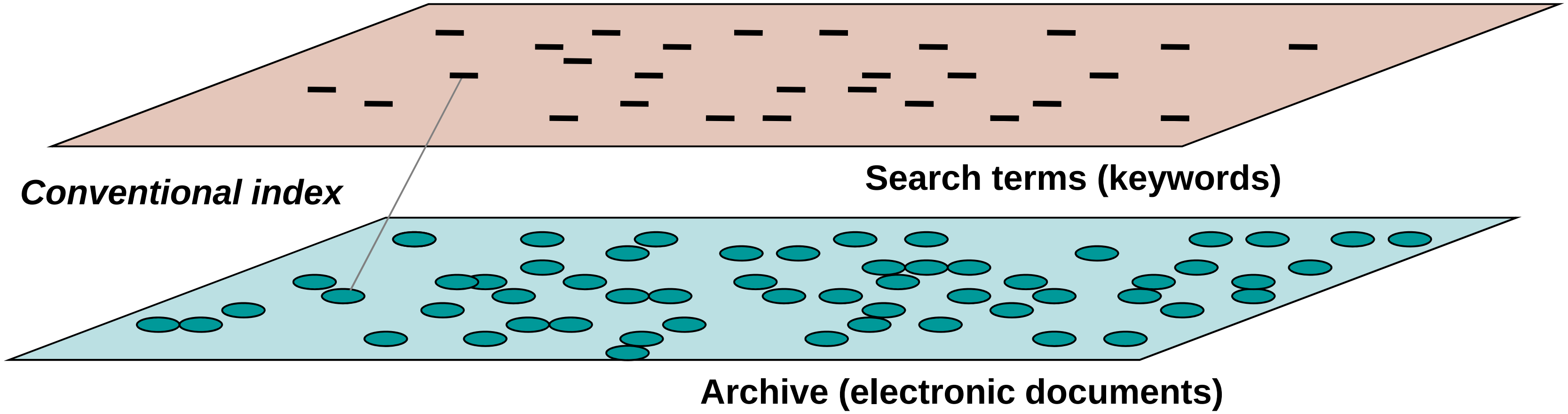
Accessing electronic archives via full text search



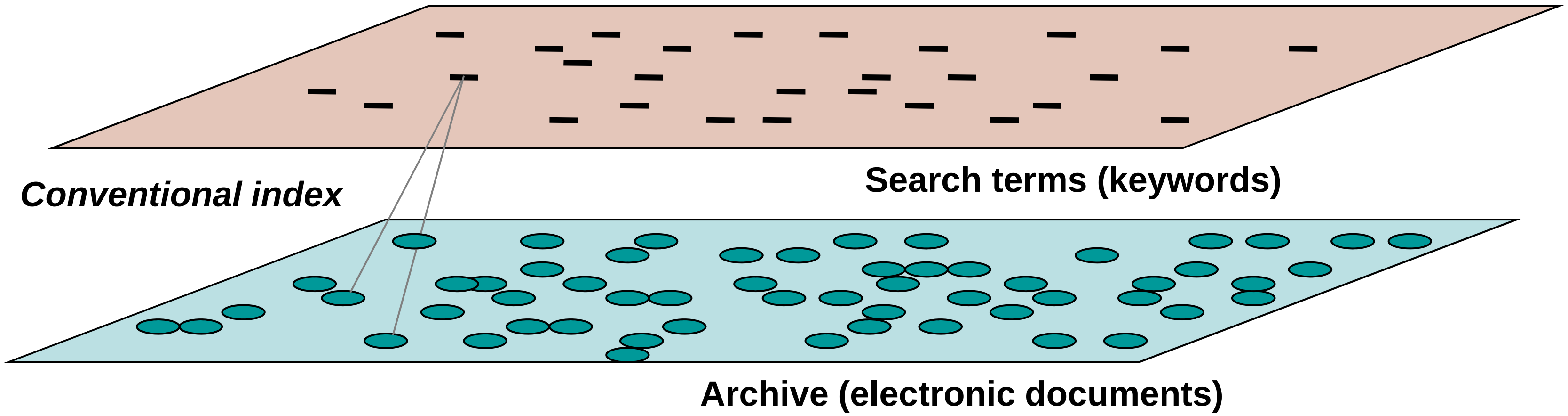
Accessing electronic archives via full text search



Accessing electronic archives via full text search



Accessing electronic archives via full text search

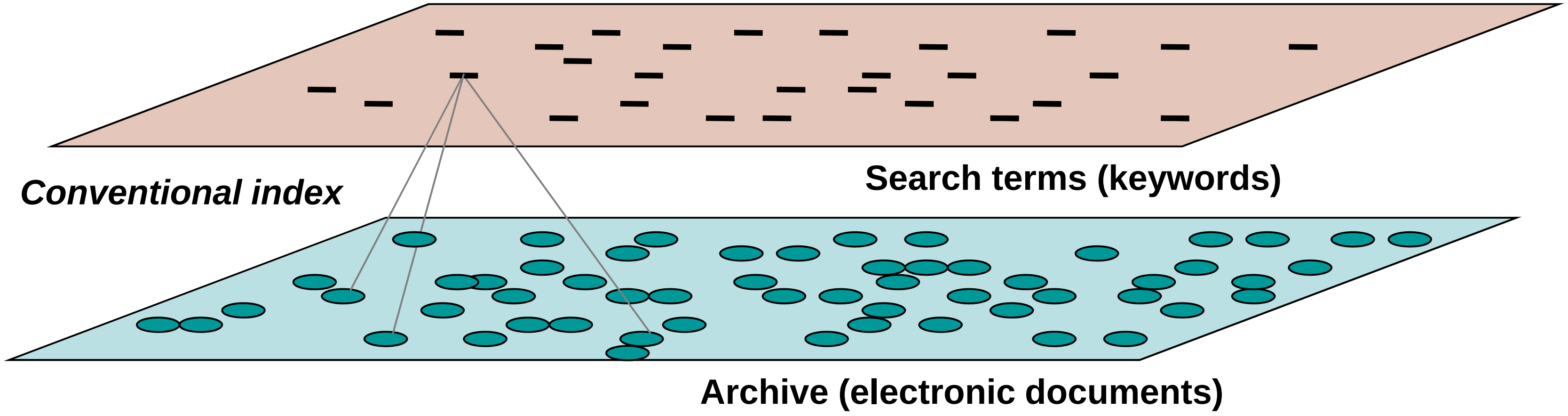


Conventional index

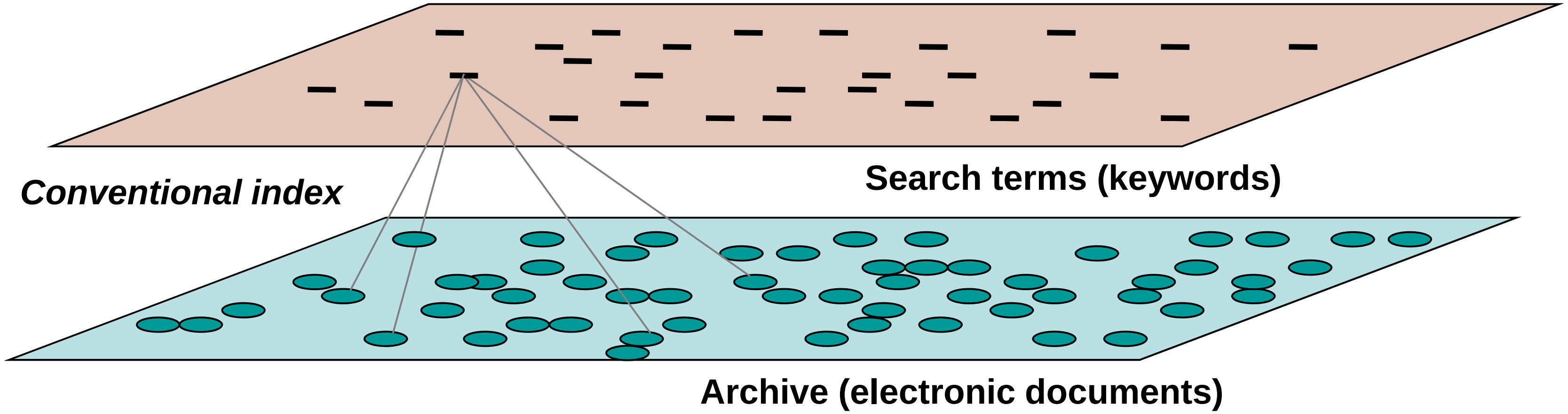
Search terms (keywords)

Archive (electronic documents)

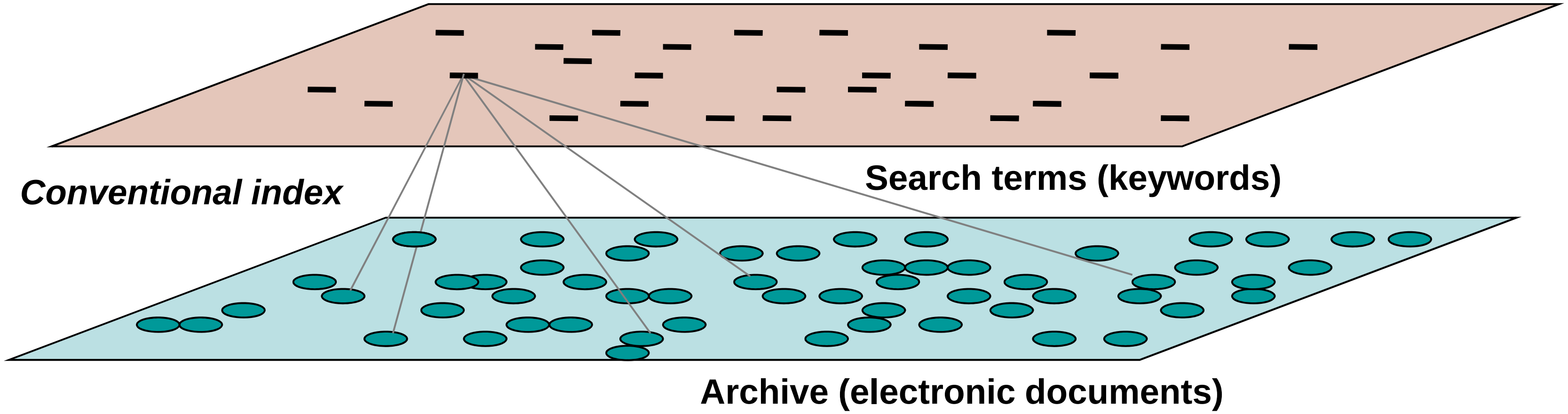
Accessing electronic archives via full text search



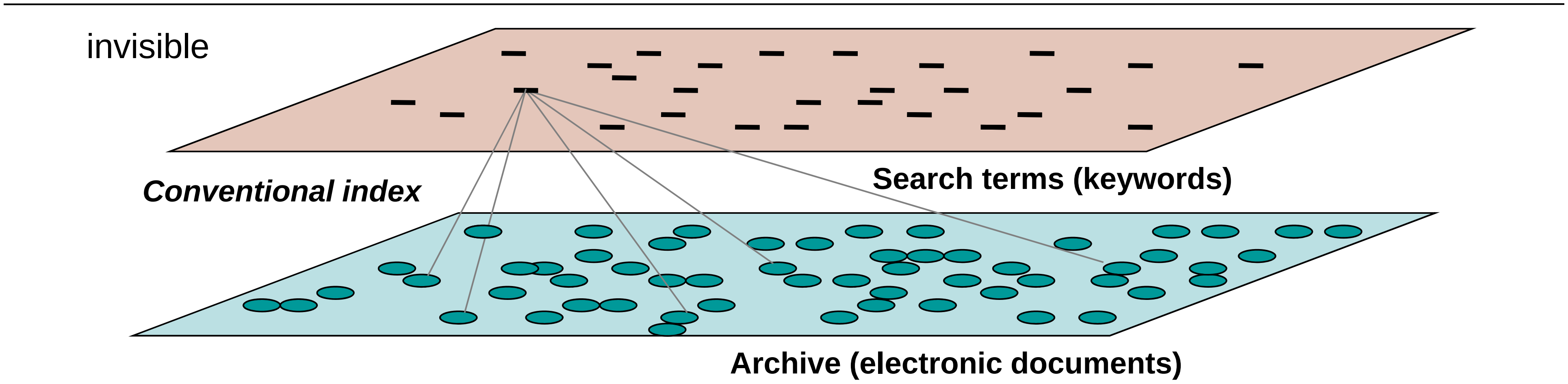
Accessing electronic archives via full text search



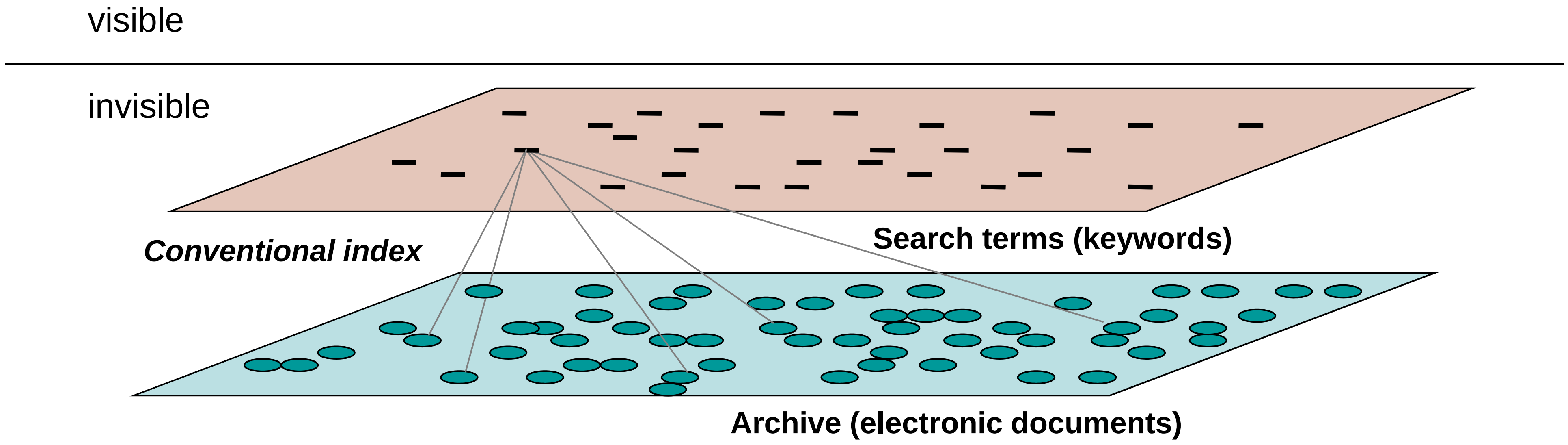
Accessing electronic archives via full text search



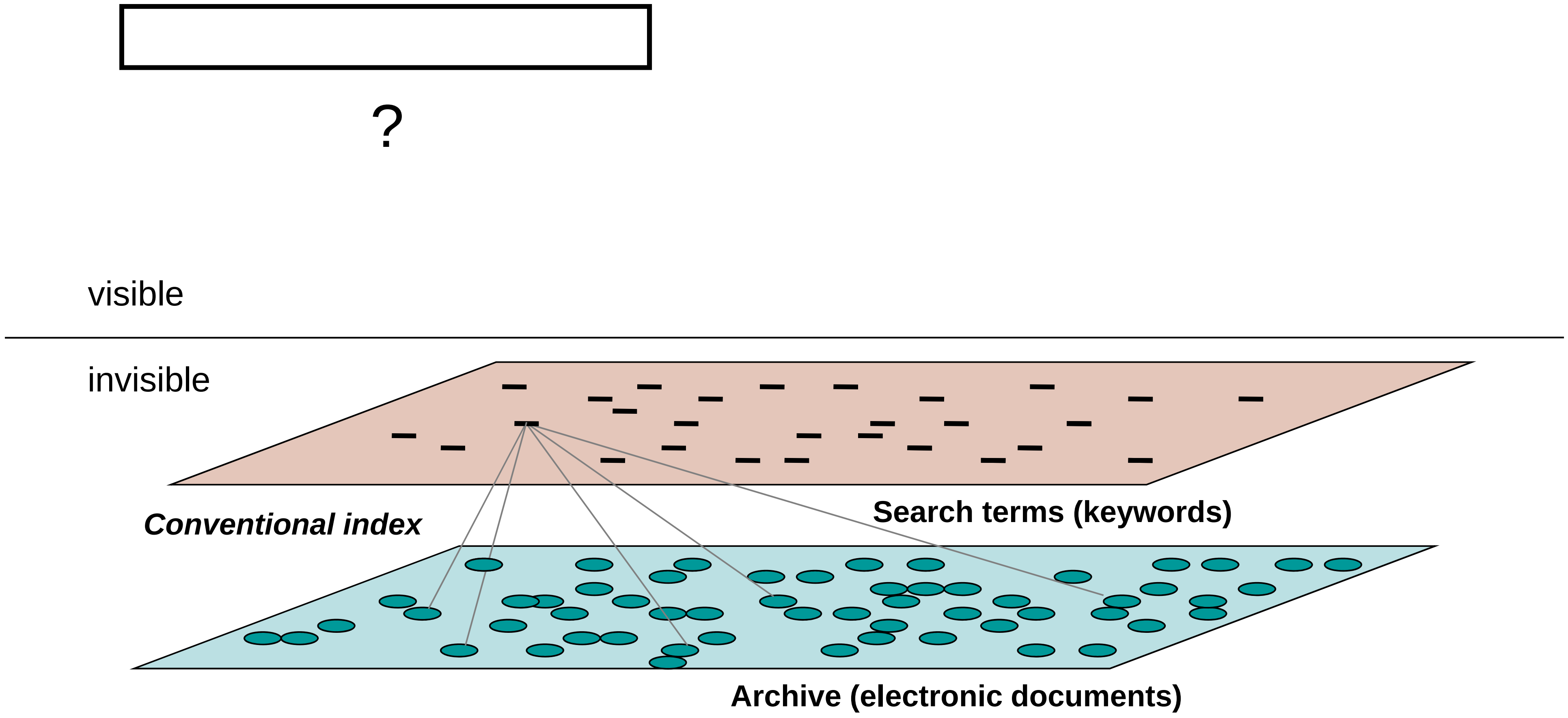
Accessing electronic archives via full text search



Accessing electronic archives via full text search



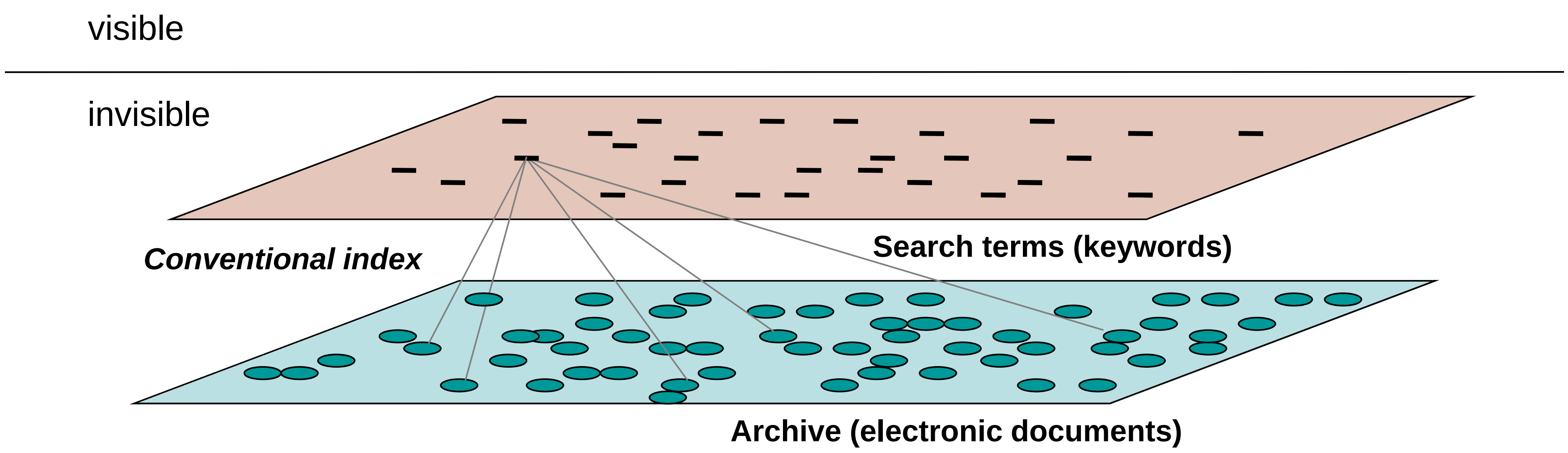
Accessing electronic archives via full text search



Accessing electronic archives via full text search

heart disease

?



visible

invisible

Conventional index

Search terms (keywords)

Archive (electronic documents)

Accessing electronic archives via full text search

heart disease

?

title 1
….heart disease …

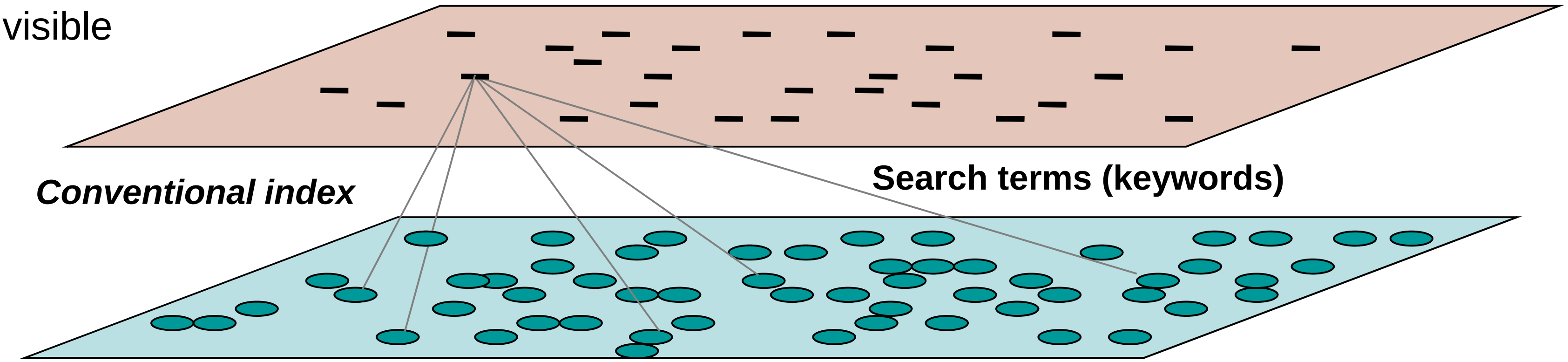
title 2
……….heart disease …

title 3
.heart disease …

.
.
.

visible

invisible



Conventional index

Search terms (keywords)

Archive (electronic documents)

Shortcomings (full text search)

- No bird's eye perspective, structure and contents of archive not shown

- No thematic search (cf. library classification)

- No „extra functionalities“ (crosslinking, comparison,…)

Shortcomings (full text search)

- No bird's eye perspective, structure and contents of archive not shown
- No thematic search (cf. library classification)
- No „extra functionalities“ (crosslinking, comparison,…)

How to offer “intelligent” access to a flood of “new” electronic text documents?

Established approaches

- Traditional full text search?
- → Google?

Google Web Search

- Designed for needs of web users

- Supporting „internet actions“:
    - search for addresses,
    - shopping,
    - transactions (flight booking), …

- Search for information/content
  related to specific topics less important

(Extended) Google Book Search

- Options for several forms of full text search:

Find results with all words, exact phrase,
some of the words, none of the words

- Options for specifying metadata

Publisher
Author
Journals or Books
Book Titles
Language
Date when book was published
ISBN, ISSN

- Nice visualization of hits after selecting a book

Shortcomings (Google Book Search)

- No structure on ocean of books
- No relationship to systematic ordering in library catalogues
- User interested in specific topics (postwar history) left alone.

How to offer “intelligent” access to a flood of “new” electronic text documents?

How to offer “intelligent” access to a flood of “new” electronic text documents?

Automated extraction of topics from texts!

Start Hypotheses

- Need to order new collections via topics and subtopics as in traditional library search.

- Topic assignment should take existing ordering principles in libraries („systematics“) into account by including temporal and geographic ordering, but be more flexible as to integration of new topics.

- Manually ordering large collections by topics is too expensive.

- Fully automated ordering of collections via topics is possible.

- There are some disadvantages of automating topic assignment but many important advantages and many new possibilities.

Automated topic assignment to texts by TopicZoom

- TopicZoom GmbH: Spin-Off from Ludwig-Maximilians-University Munich (CIS) specialized on automated extraction of topics, geographic places, temporal periods, persons, organisations, events, from texts.

- Free web service for fully automated topic assignment, also use via API possible

- Public version for German texts (English version coming „soon“).

Automated topic assignment to texts by TopicZoom

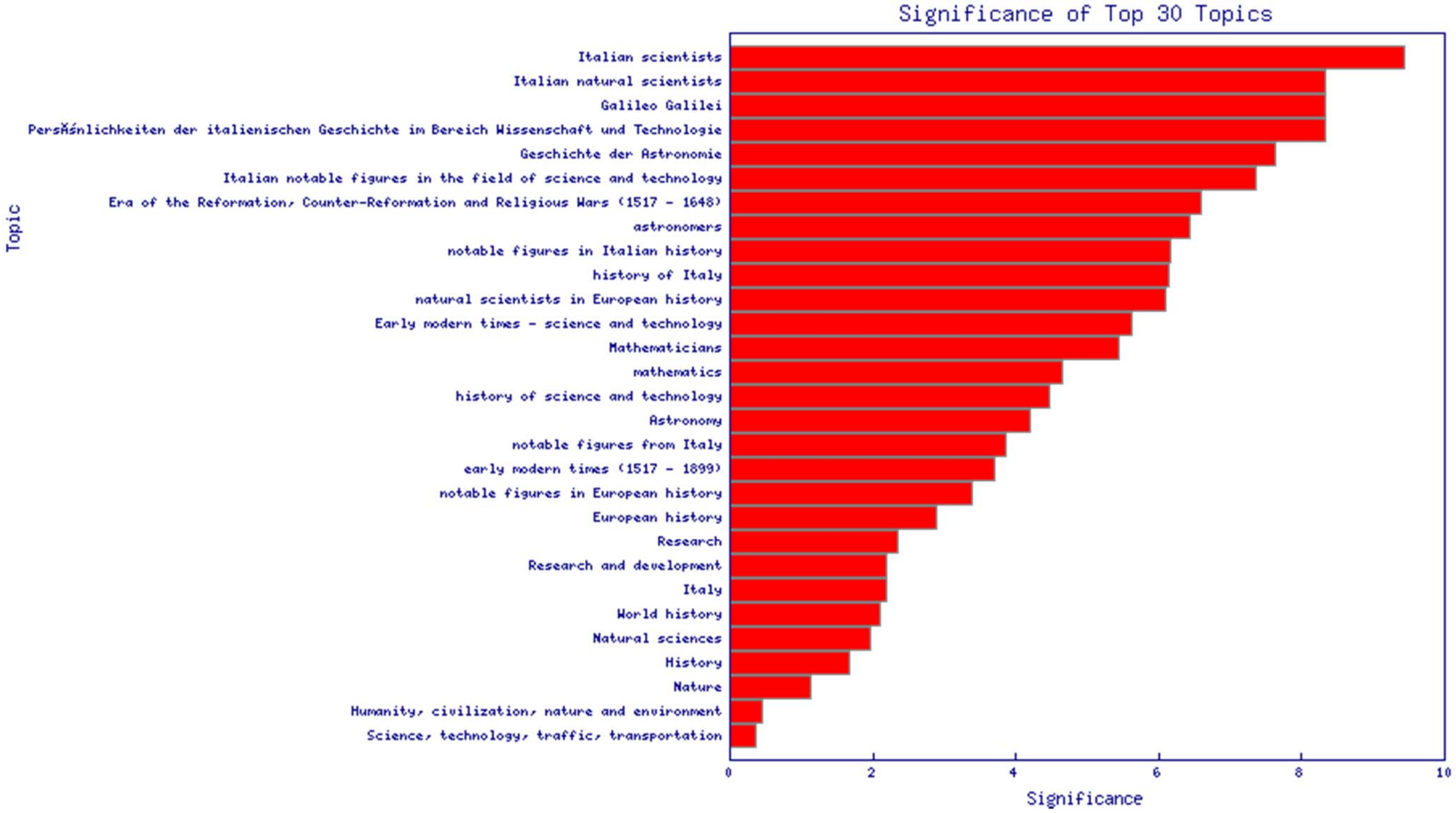
3 tasters

Text input

„Galileo Galilei“

Text input

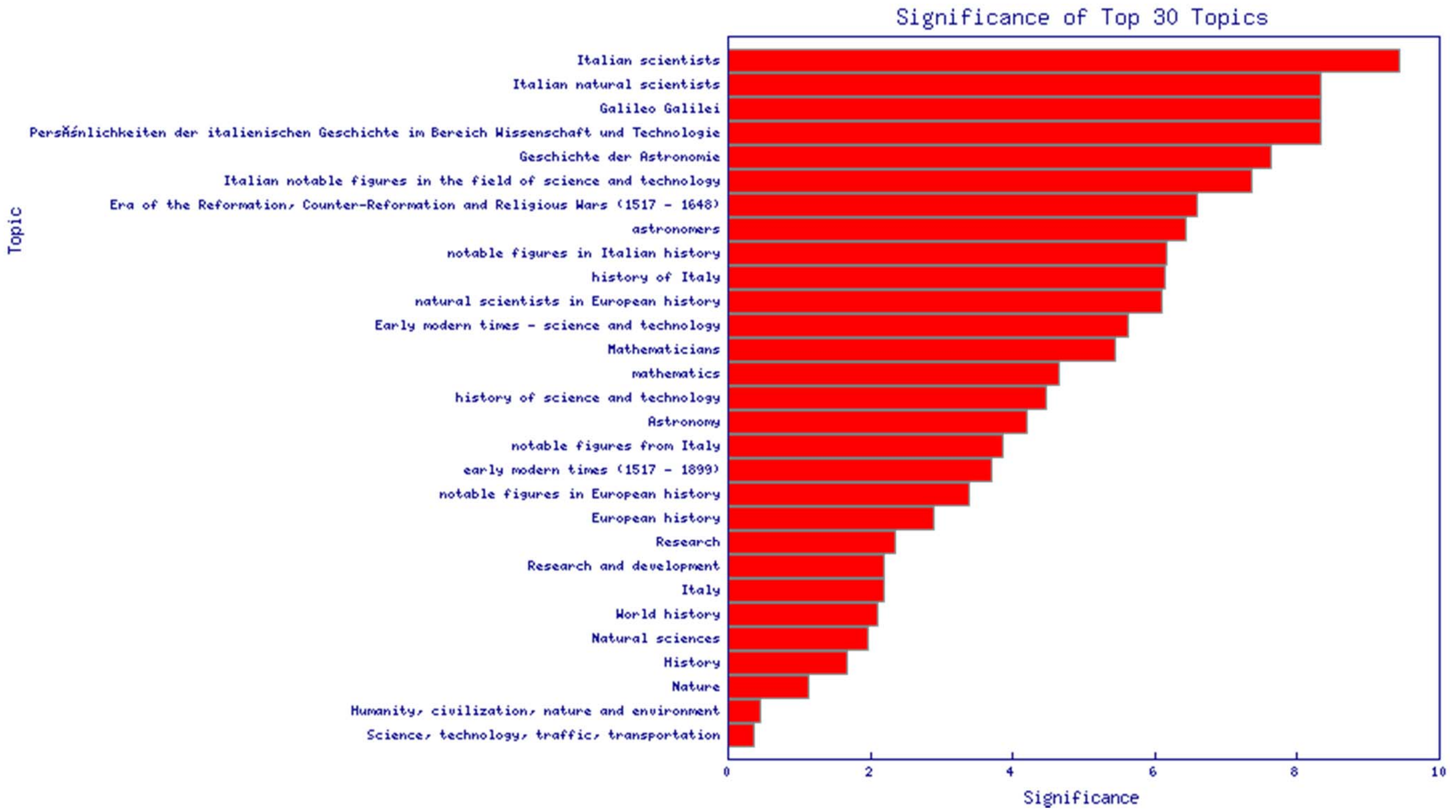
„Galileo Galilei“



Text input

Diagram only mentions 30 most „eye-catching“ topics, in general and for longer texts much more topics extracted, see live demo

„Galileo Galilei“

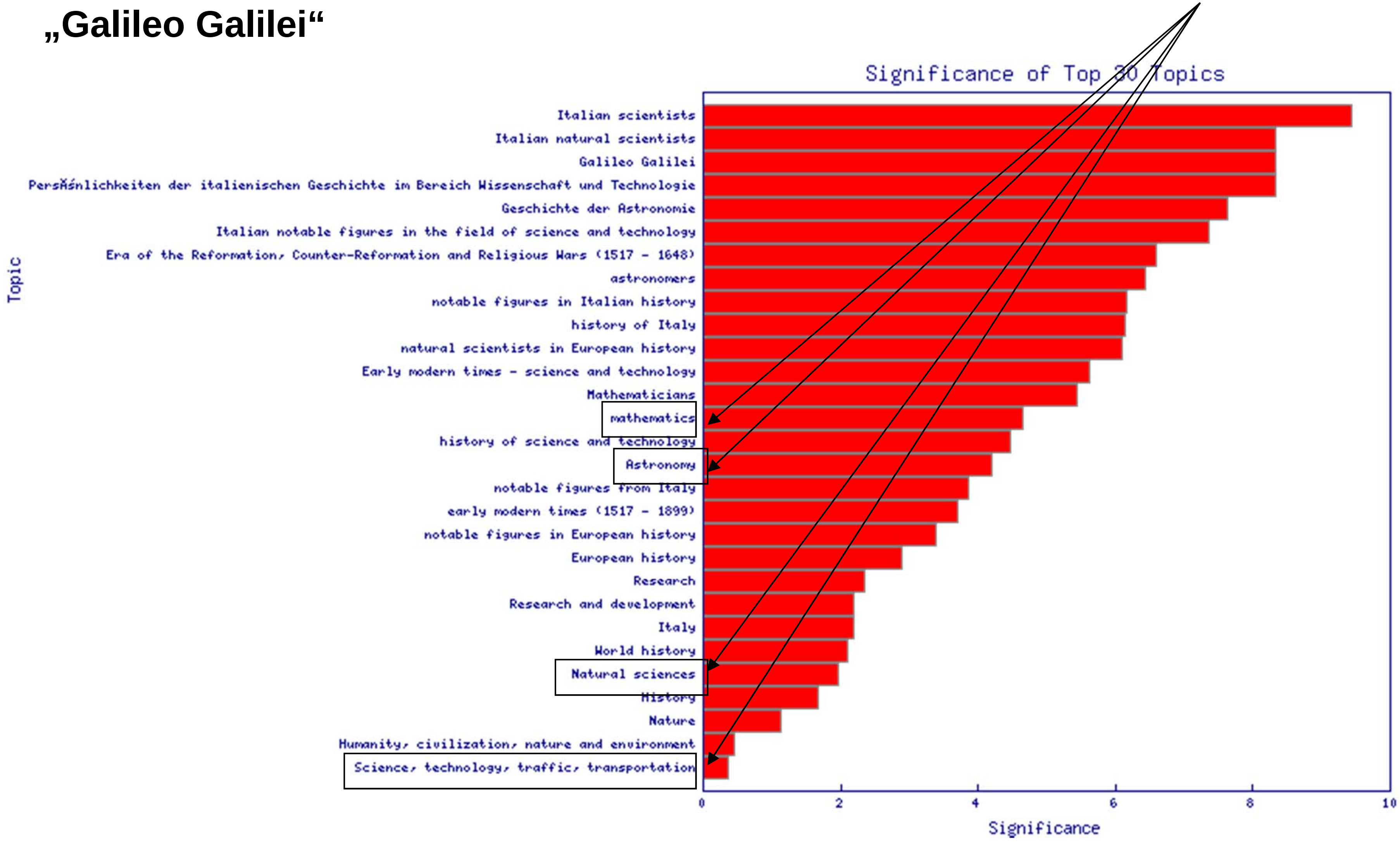


Text input

„Galileo Galilei“

Distinct levels of generality of assigned topics

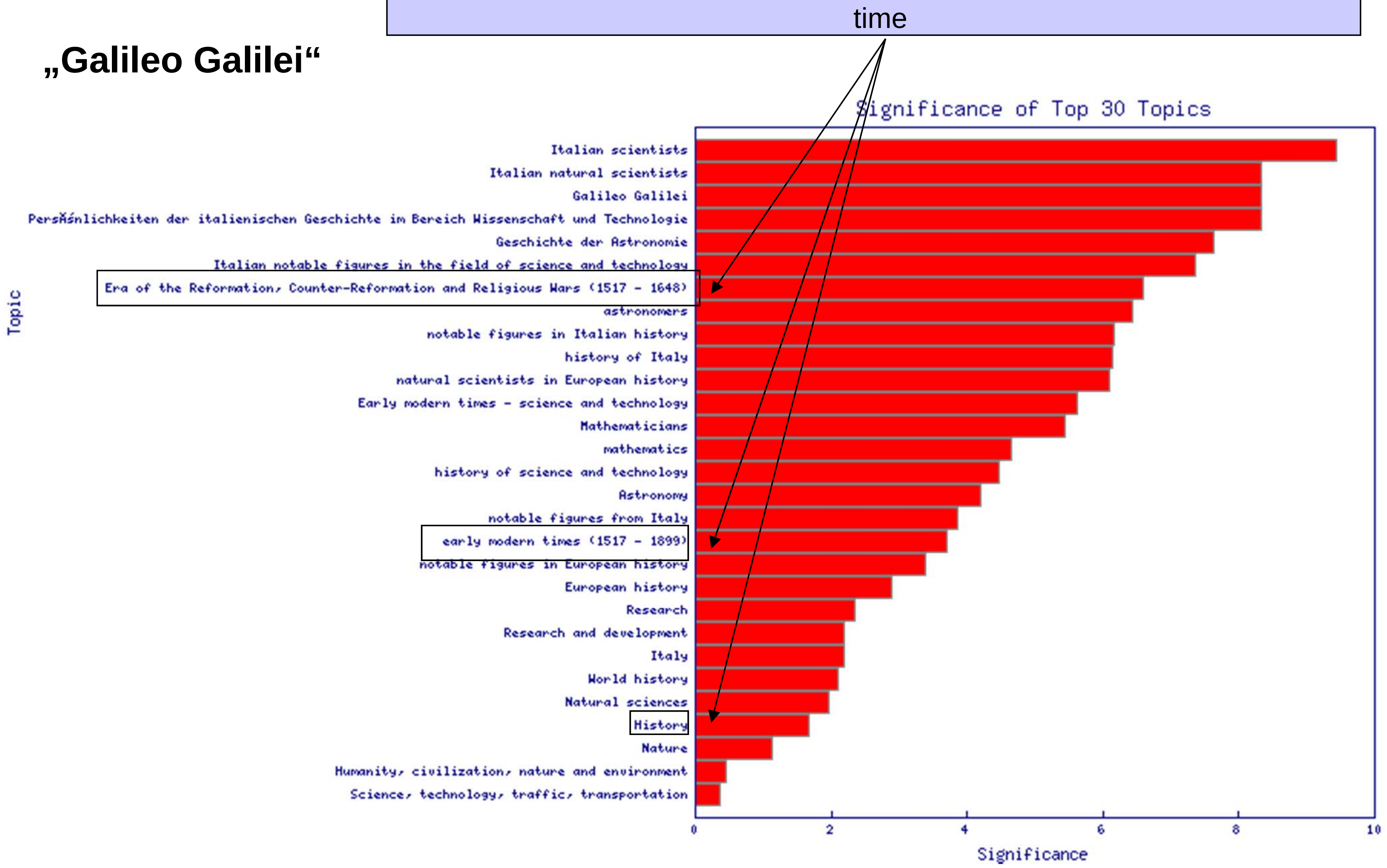
field



Text input

Distinct levels of generality of assigned topics

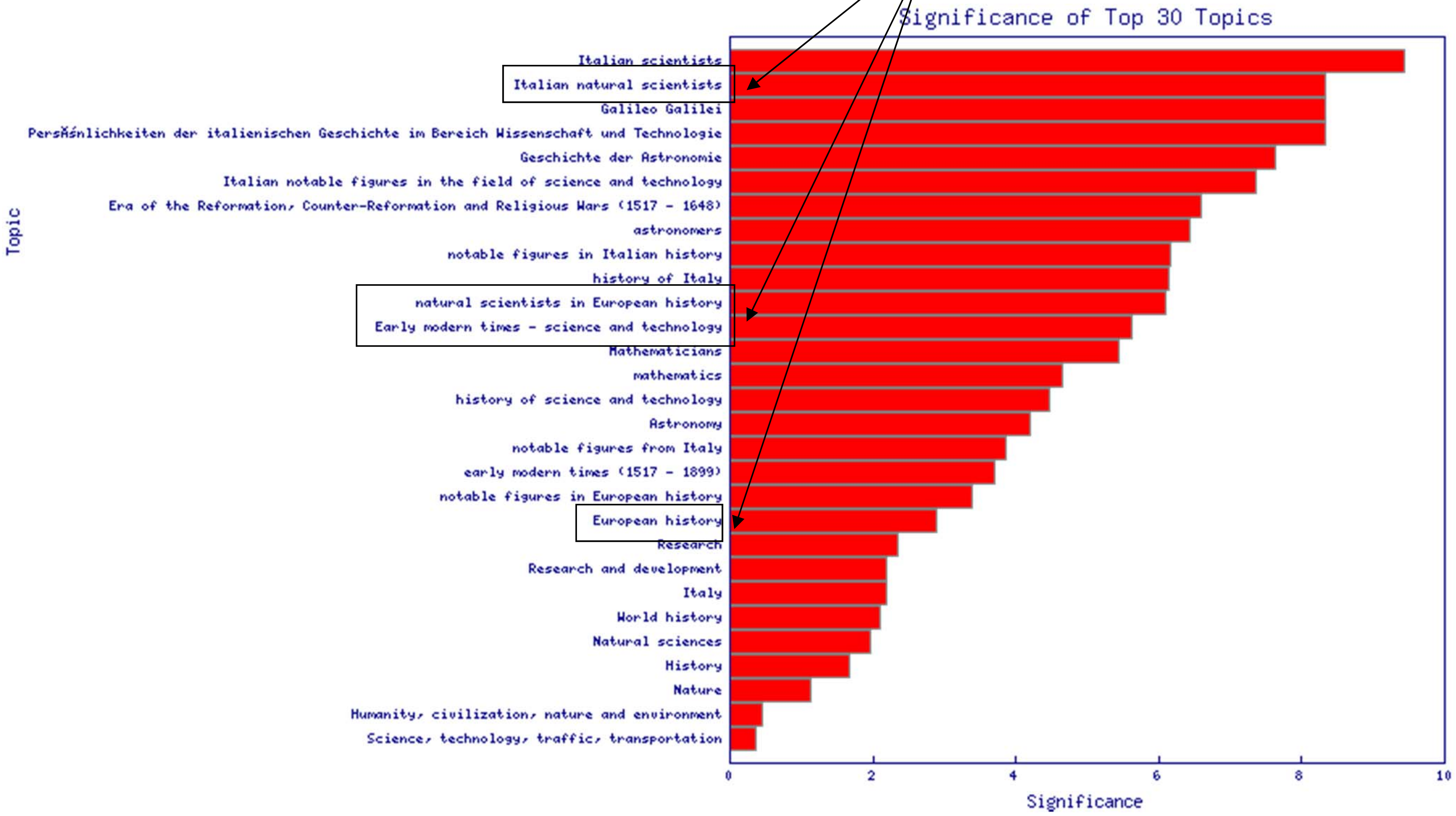
„Galileo Galilei“



Text input

„Galileo Galilei“

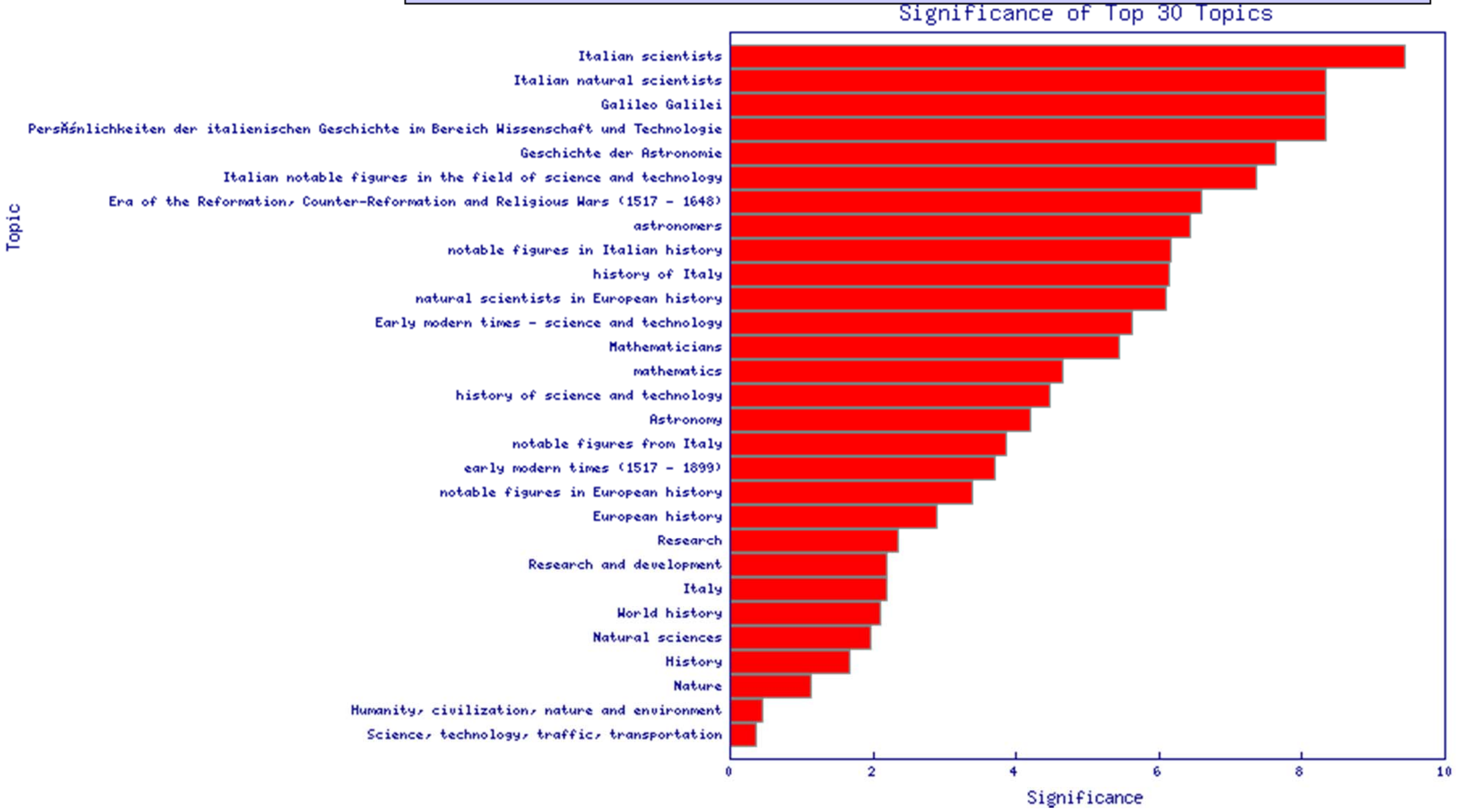
Many „facetted topics“ including thematic field plus geographic and temporal notions



Text input

„Galileo Galilei“

Some facetted topics may seem „special“ or „redundant“ – but: useful for machine processing! Full topic assignment positions documents in a high dimensional semantic vector space.

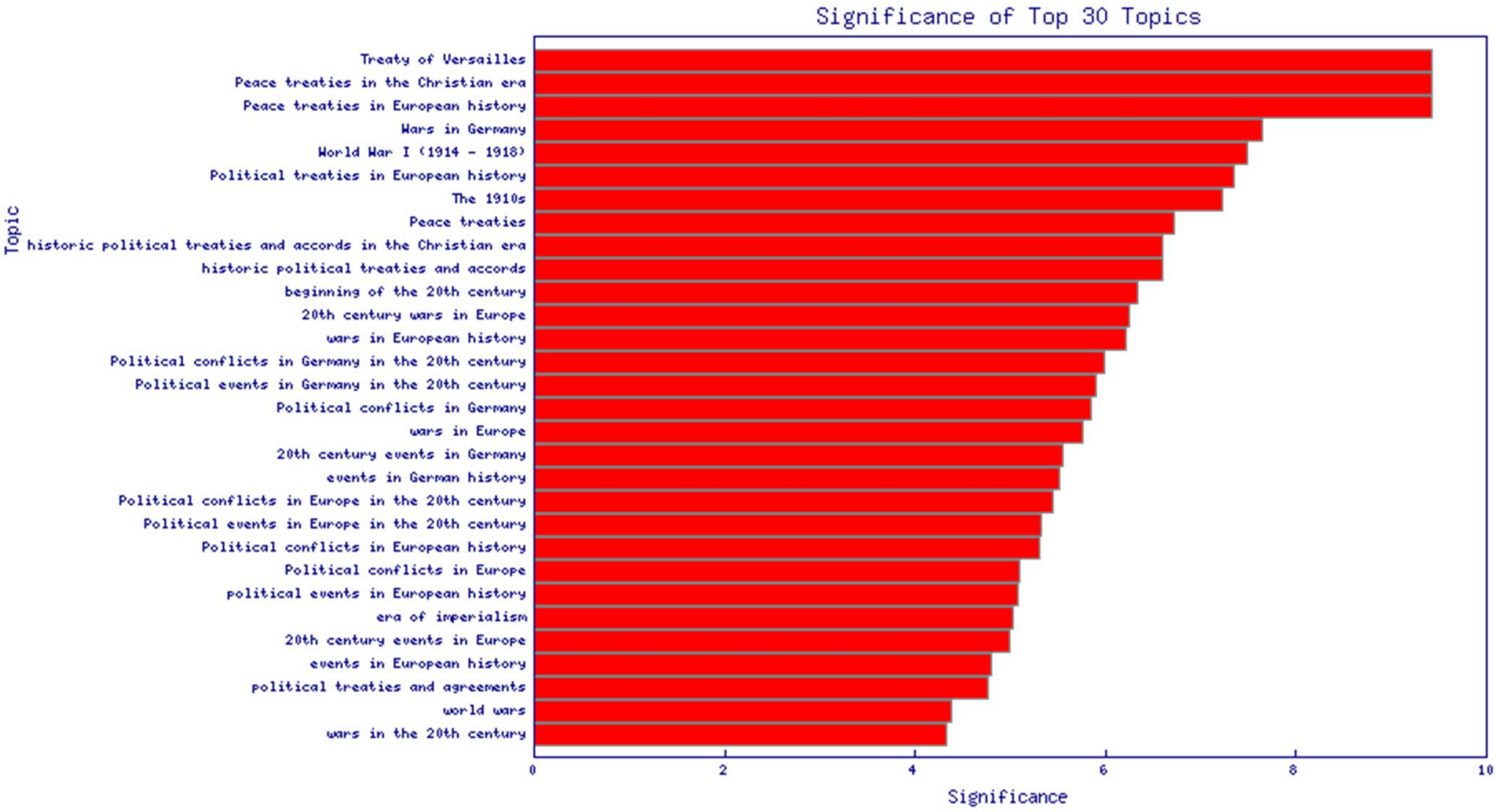


Text input

„Treaty of Versailles“

Text input

„Treaty of Versailles“

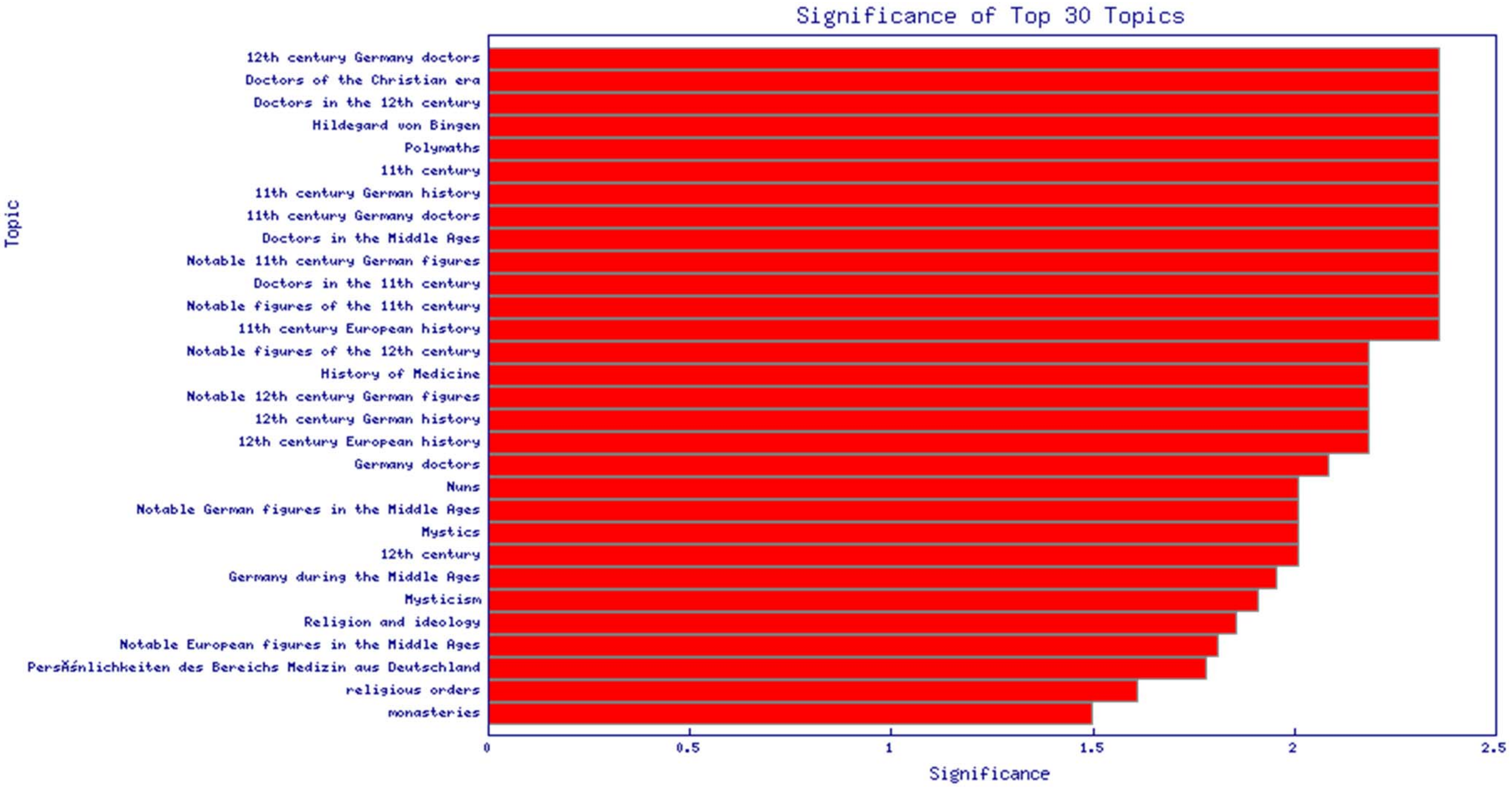


Text input (from Wikipedia)

„Saint Hildegard of Bingen, O.S.B. (German: *Hildegard von Bingen*; Latin: *Hildegardis Bingensis*) (1098 – 17 September 1179), also known as Saint Hildegard, and Sibyl of the Rhine, was a German writer, composer, philosopher, Christian mystic, Benedictine abbess, visionary, and polymath.[2] Elected a *magistra* by her fellow nuns in 1136, she founded the monasteries of Rupertsberg in 1150 and Eibingen in 1165.“

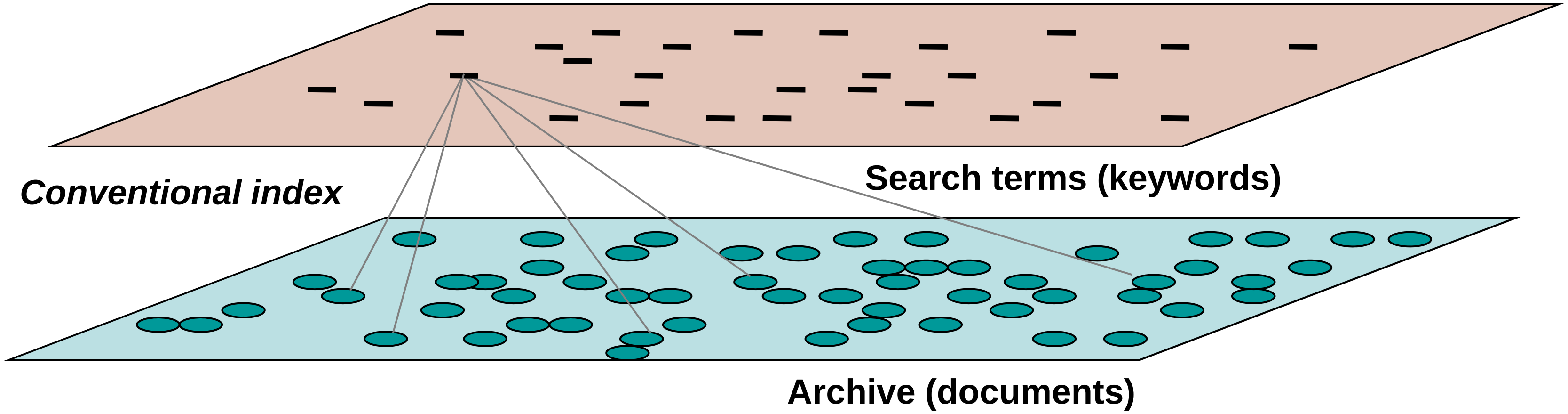
Text input (from Wikipedia)

„Saint Hildegard of Bingen, O.S.B. (German: *Hildegard von Bingen*; Latin: *Hildegardis Bingensis*) (1098 – 17 September 1179), also known as Saint Hildegard, and Sibyl of the Rhine, was a German writer, composer, philosopher, Christian mystic, Benedictine abbess, visionary, and polymath.[2] Elected a *magistra* by her fellow nuns in 1136, she founded the monasteries of Rupertsberg in 1150 and Eibingen in 1165.“



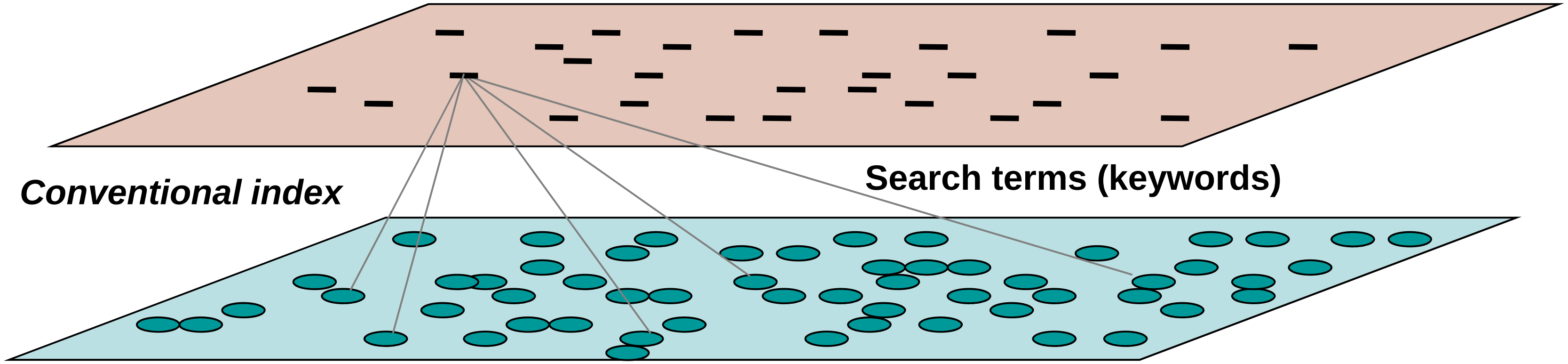
How does it work?

How does it work?



How does it work?

Topic hierarchy for navigation

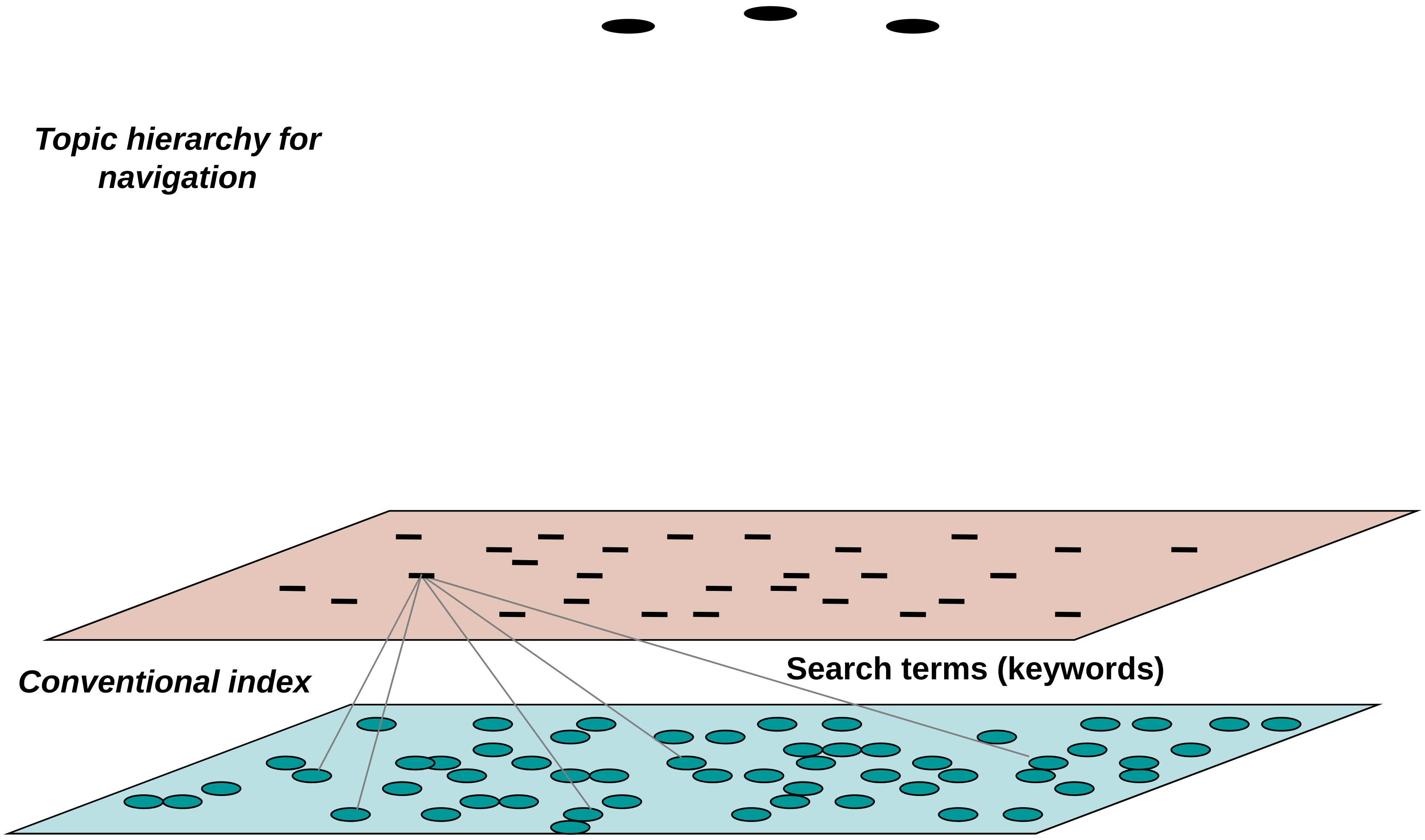


Conventional index

Search terms (keywords)

Archive (documents)

How does it work?



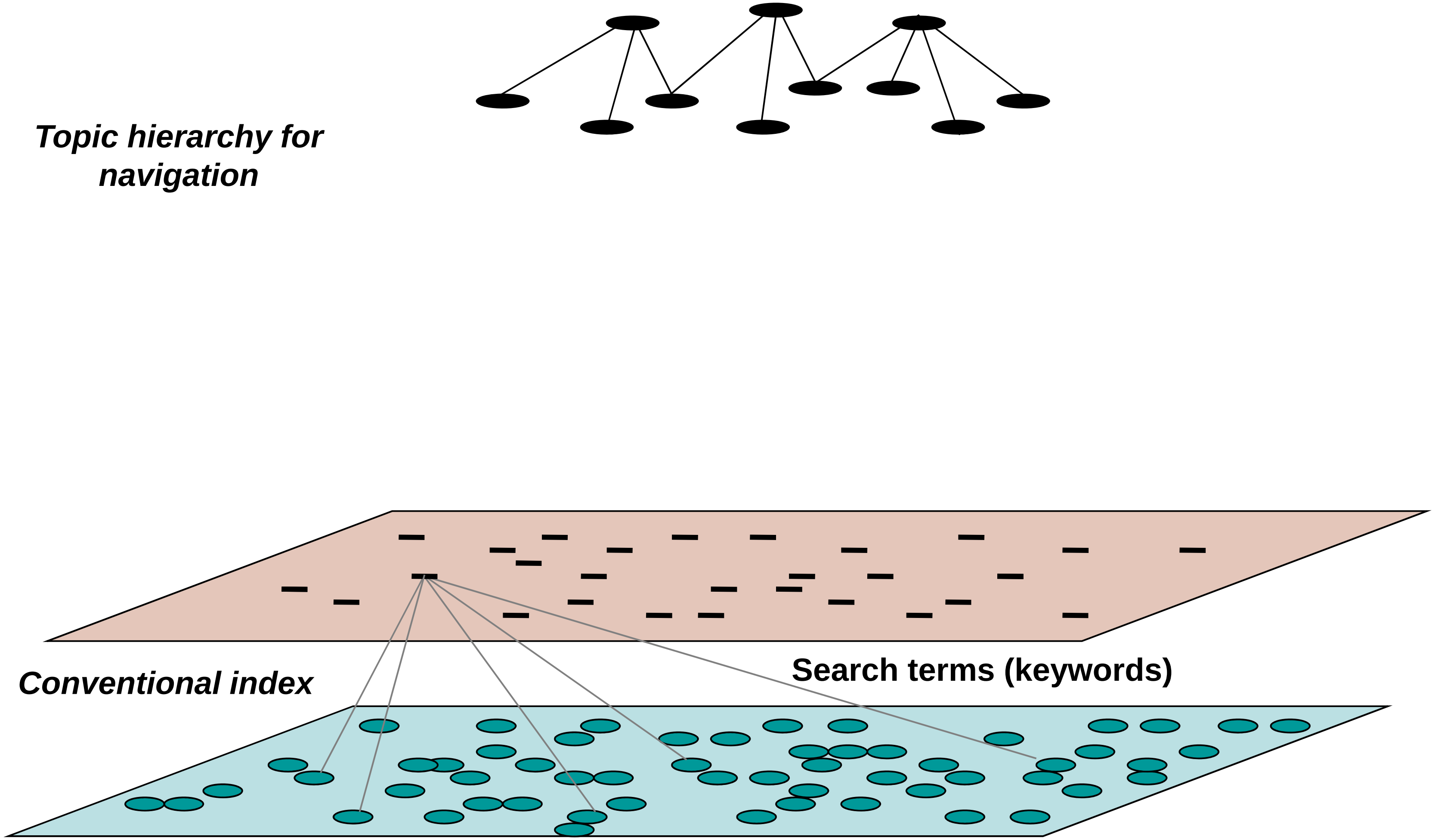
Topic hierarchy for navigation

Conventional index

Search terms (keywords)

Archive (documents)

How does it work?



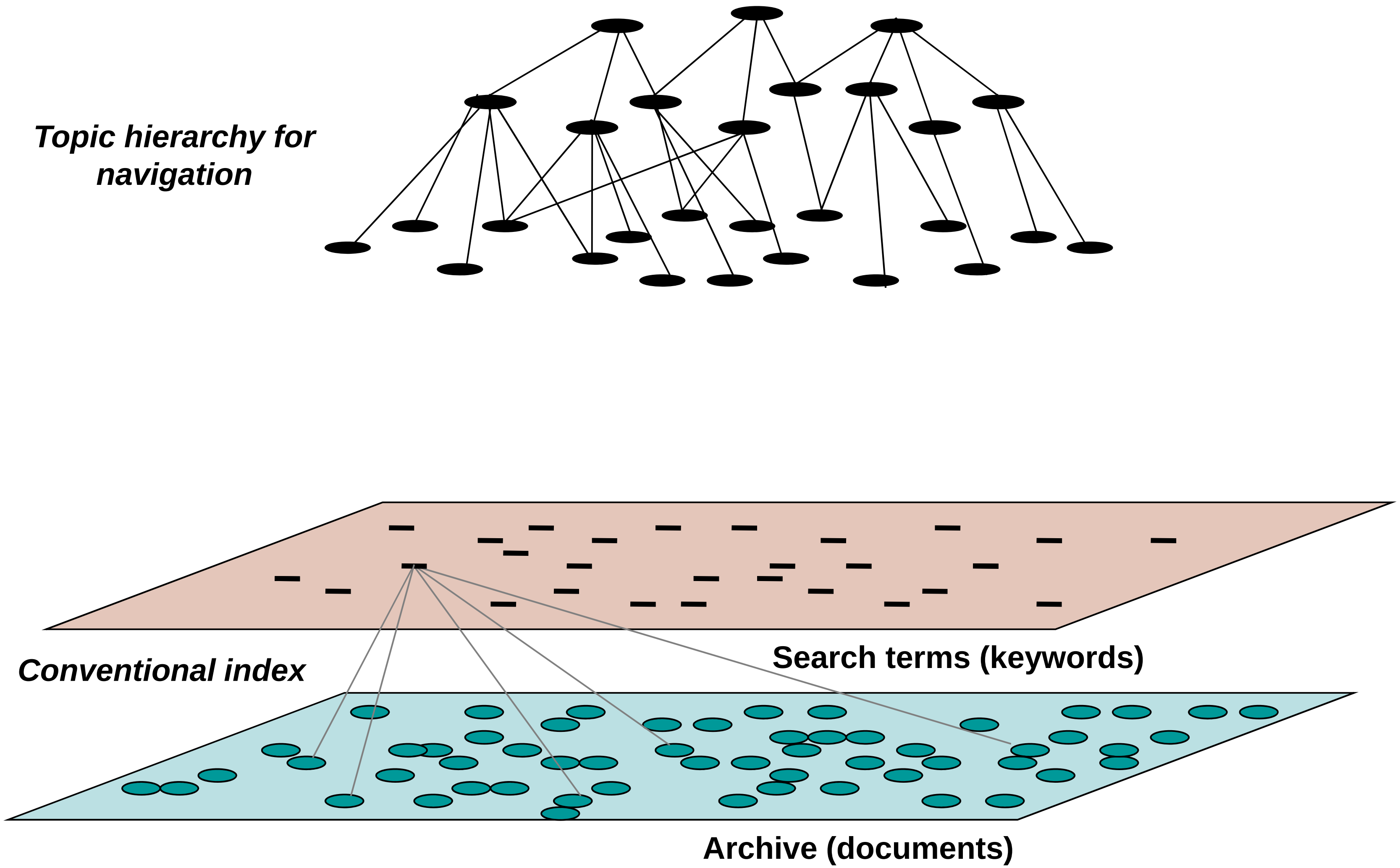
Topic hierarchy for navigation

Conventional index

Search terms (keywords)

Archive (documents)

How does it work?



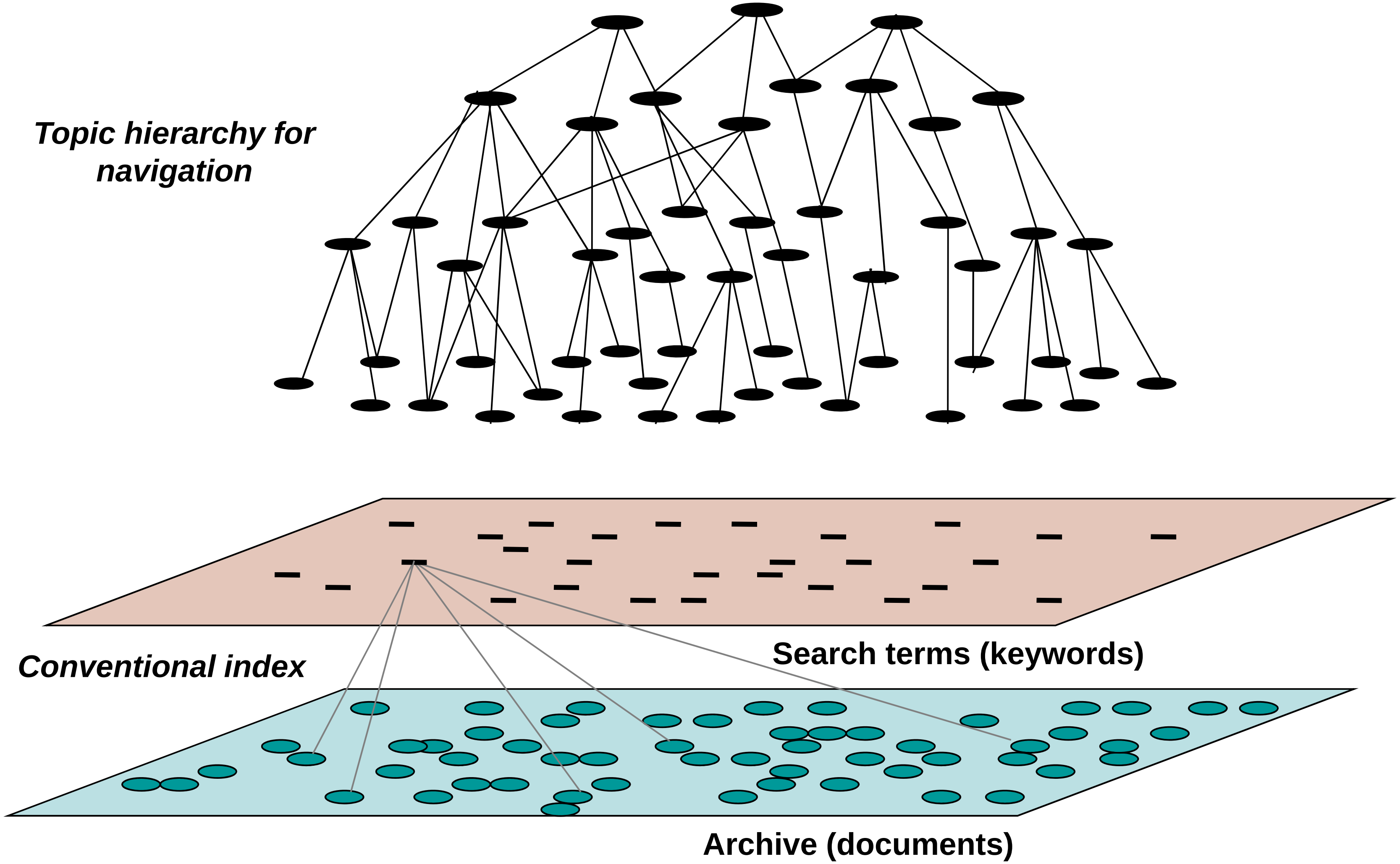
Topic hierarchy for navigation

Conventional index

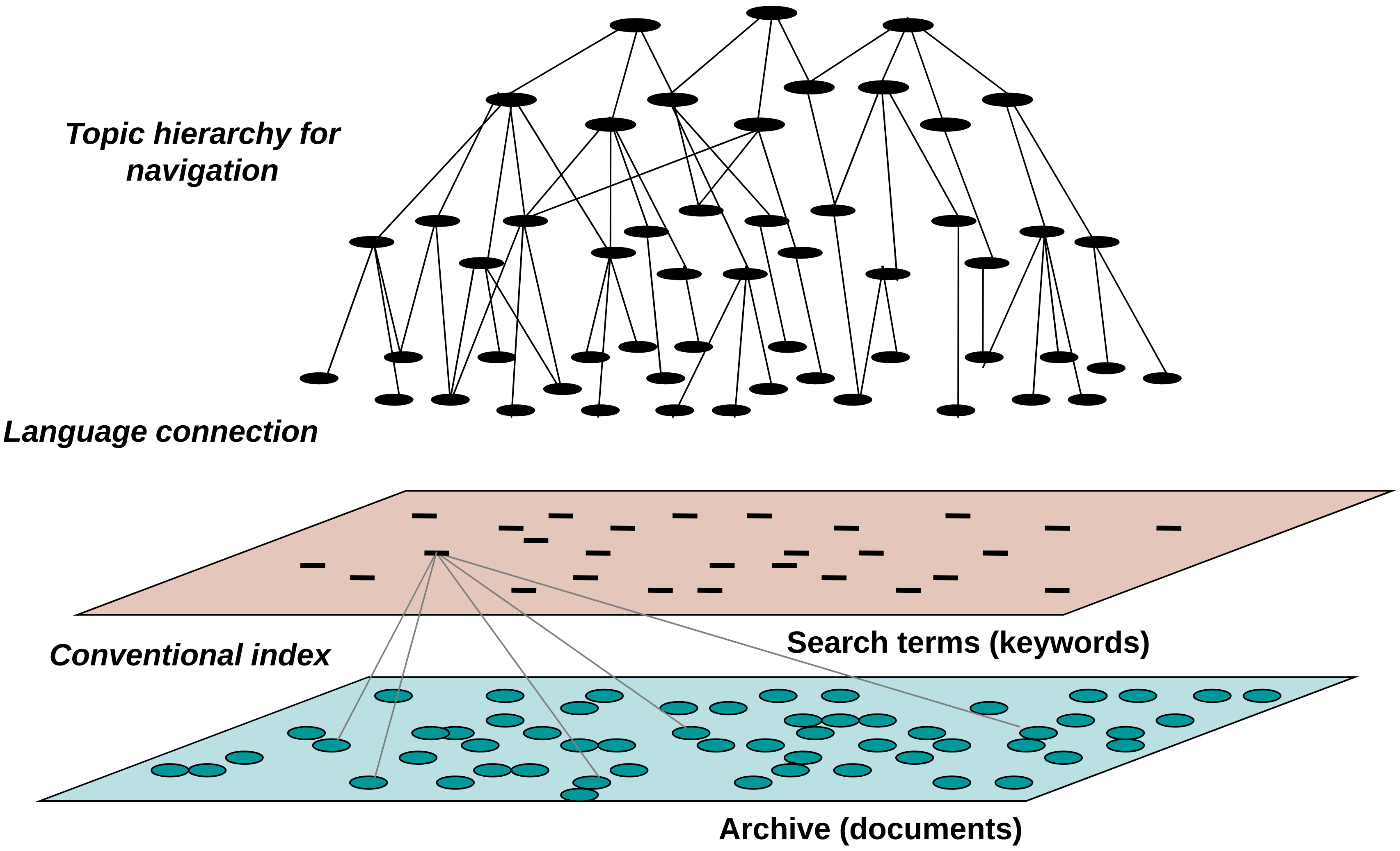
Search terms (keywords)

Archive (documents)

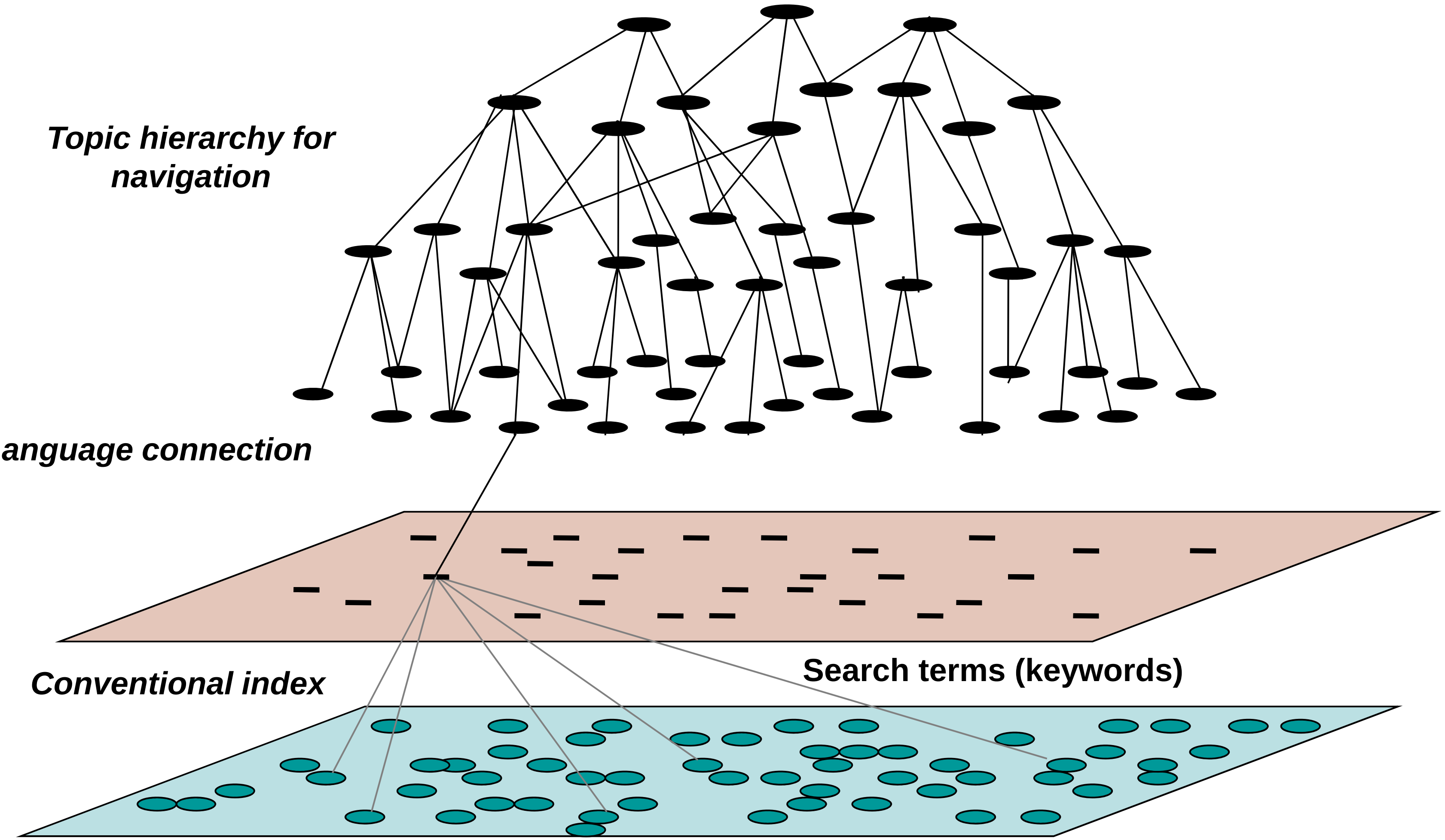
How does it work?



How does it work?



How does it work?



Topic hierarchy for navigation

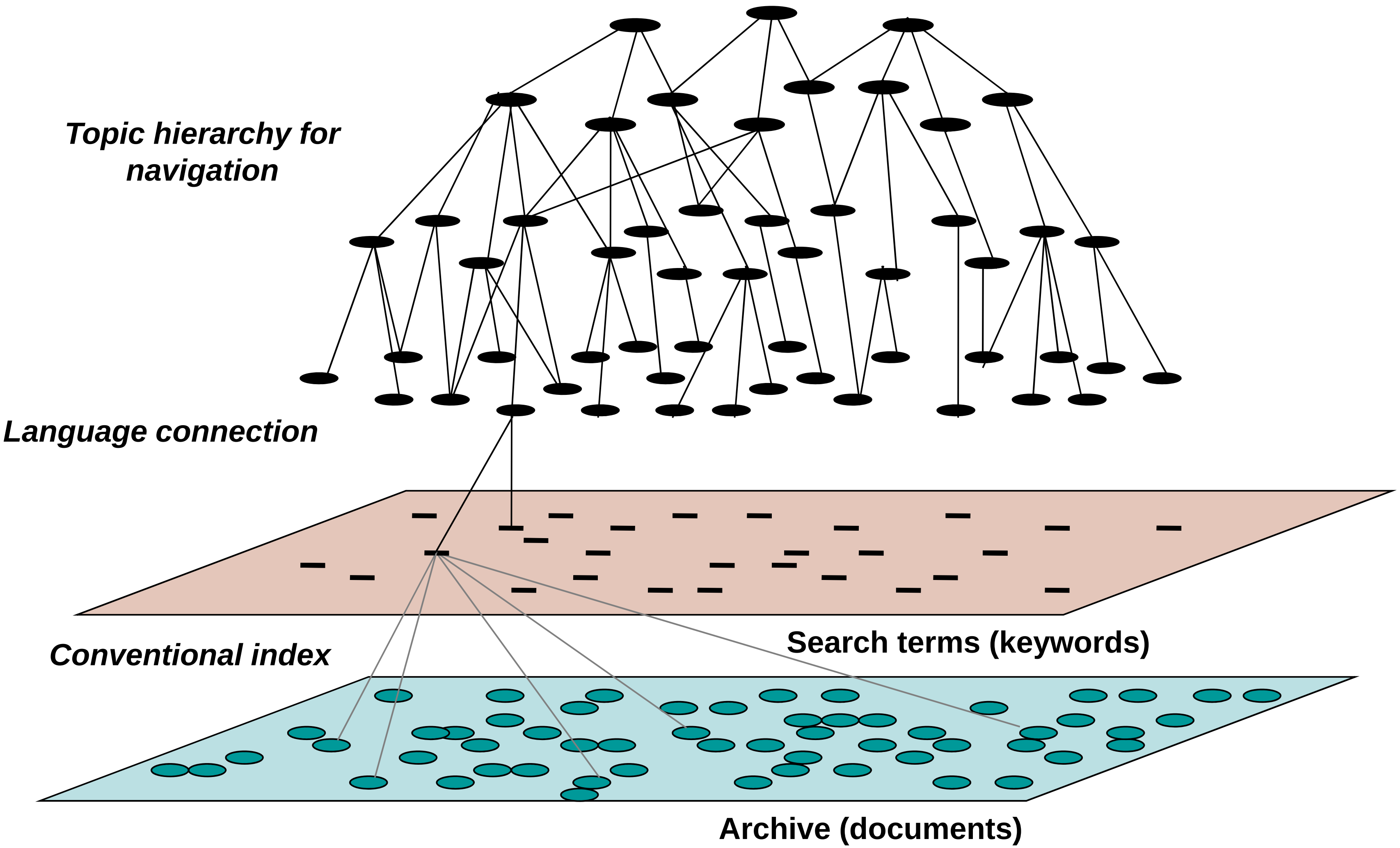
Language connection

Conventional index

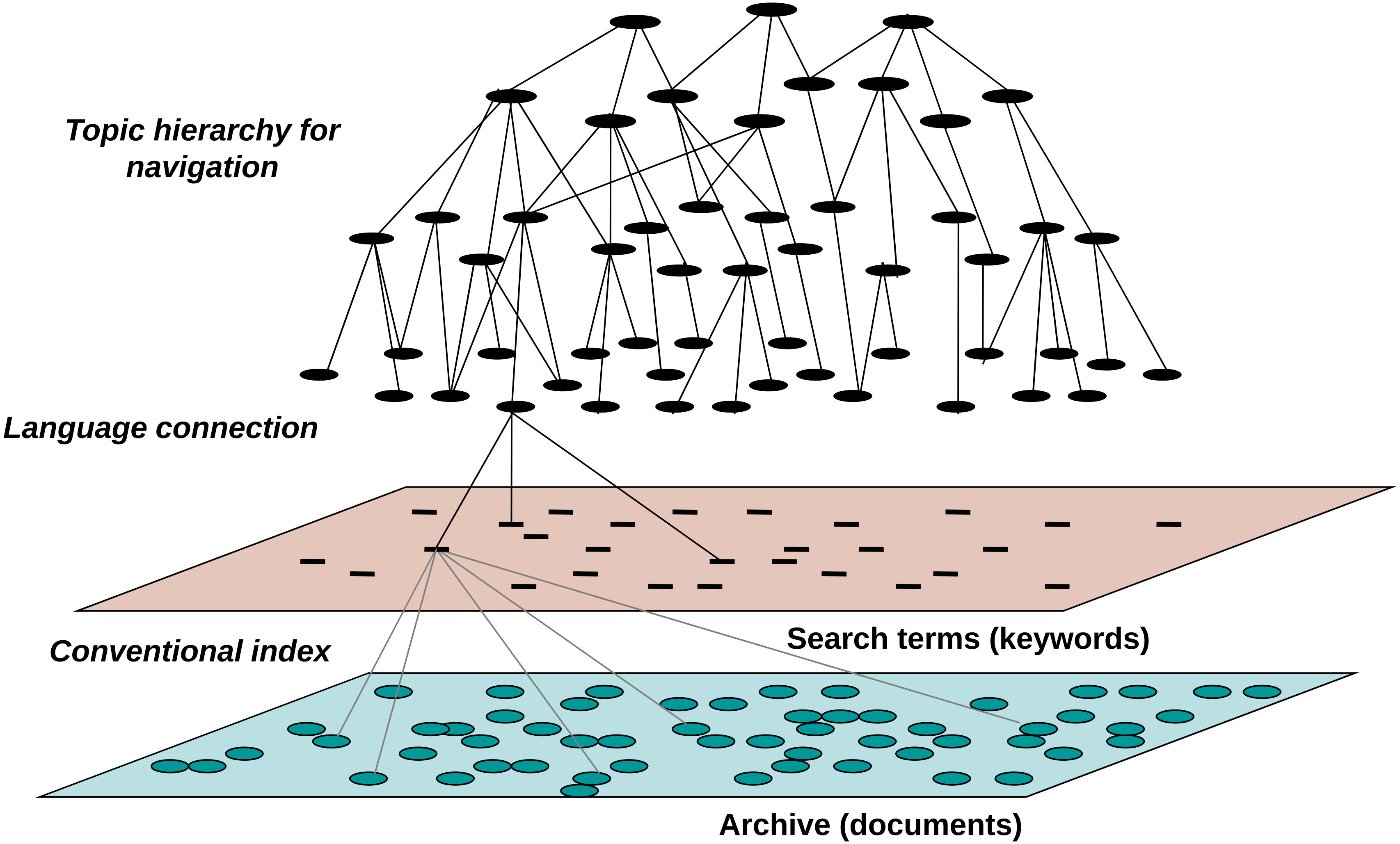
Search terms (keywords)

Archive (documents)

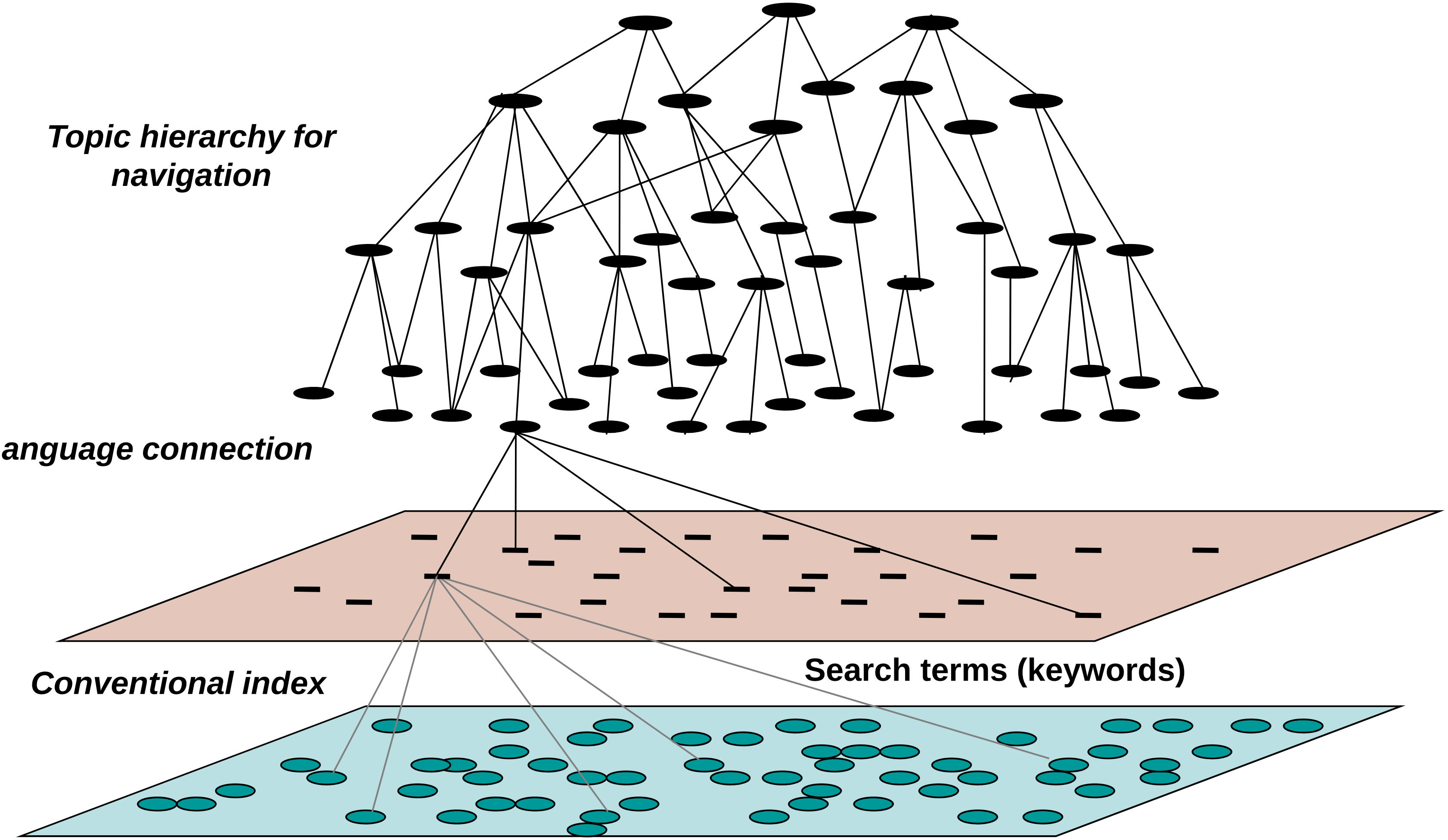
How does it work?



How does it work?



How does it work?



Topic hierarchy for navigation

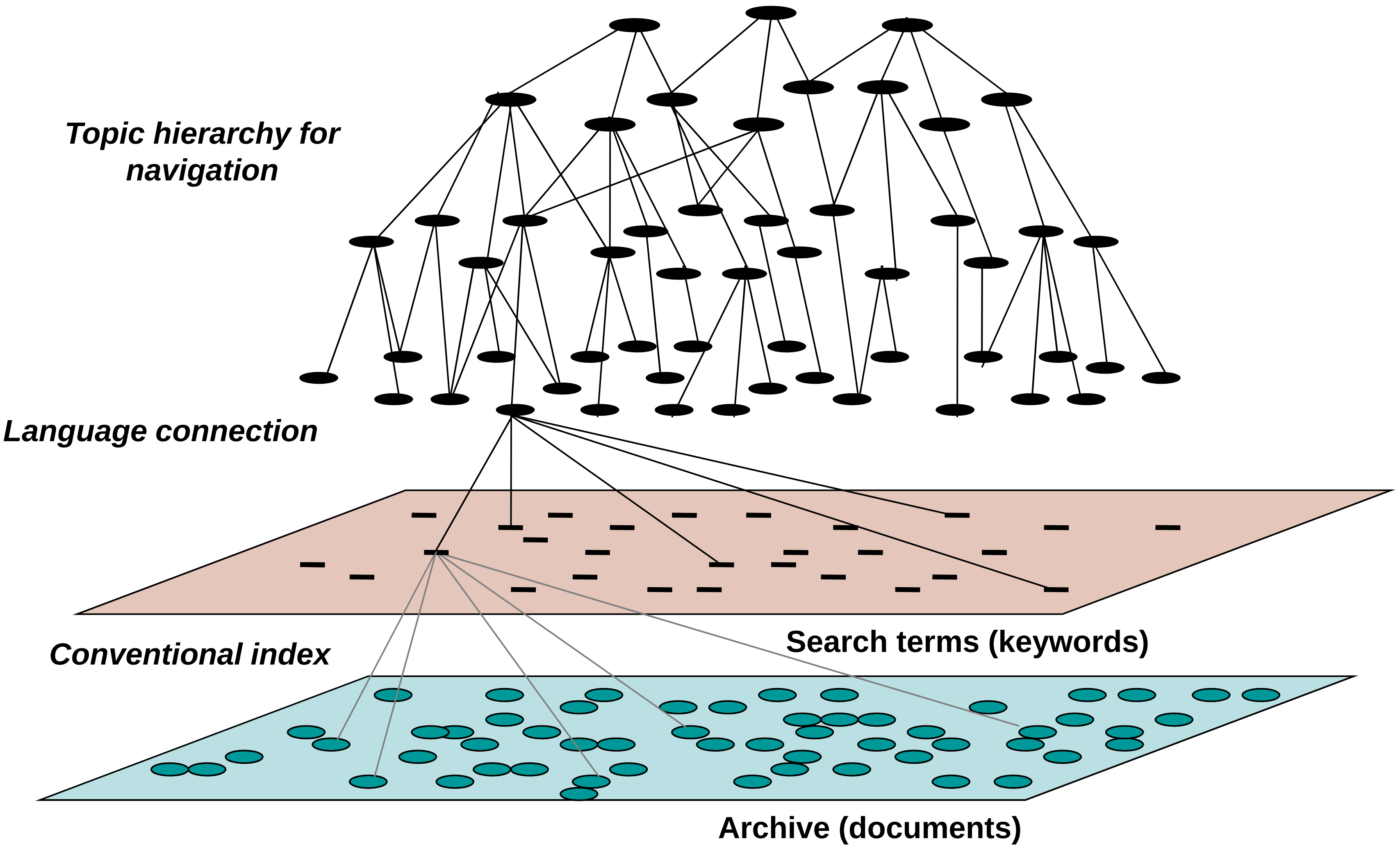
Language connection

Conventional index

Search terms (keywords)

Archive (documents)

How does it work?



Topic hierarchy for navigation

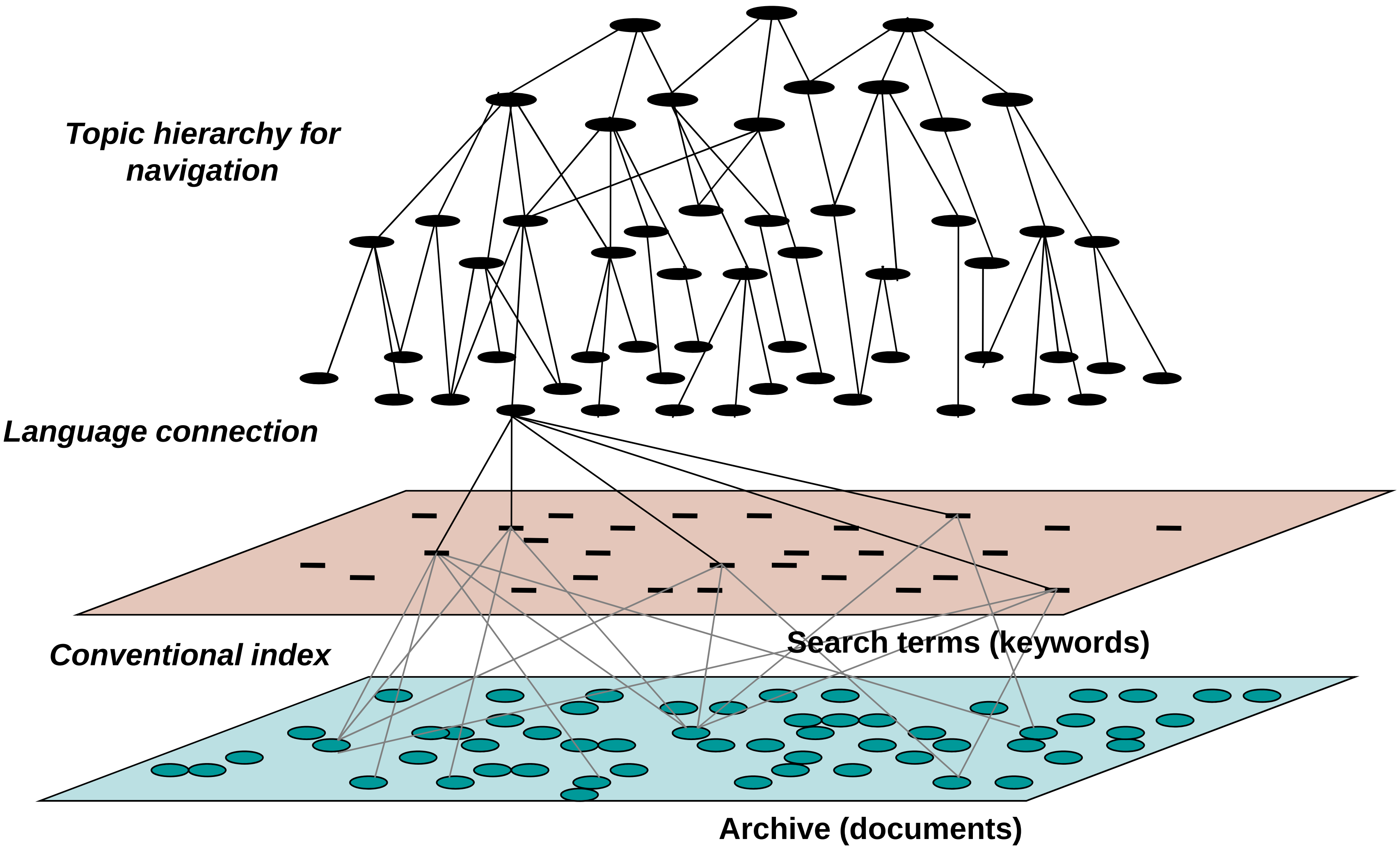
Language connection

Conventional index

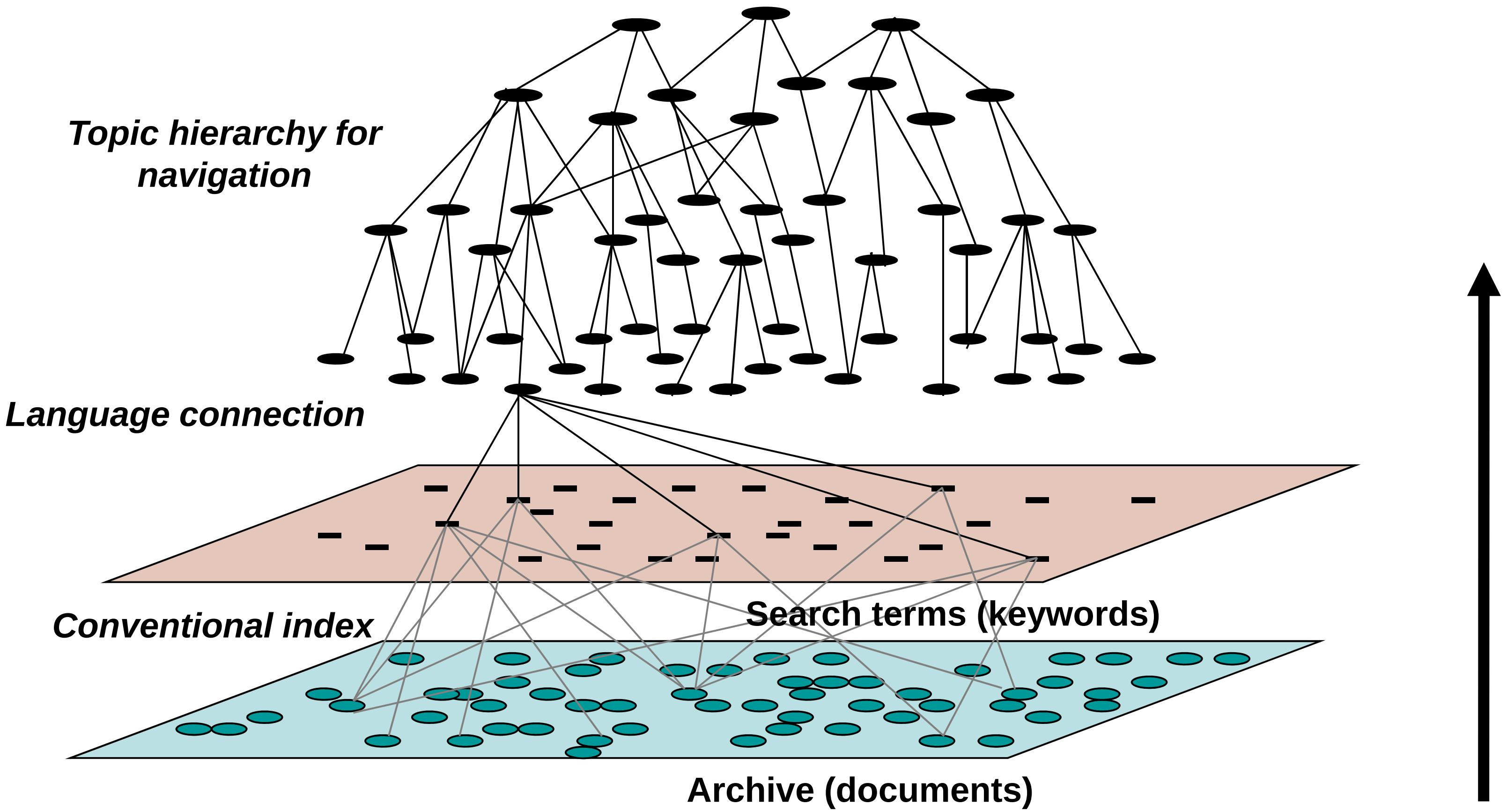
Search terms (keywords)

Archive (documents)

How does it work?



How does it work?



Topic hierarchy for navigation

Language connection

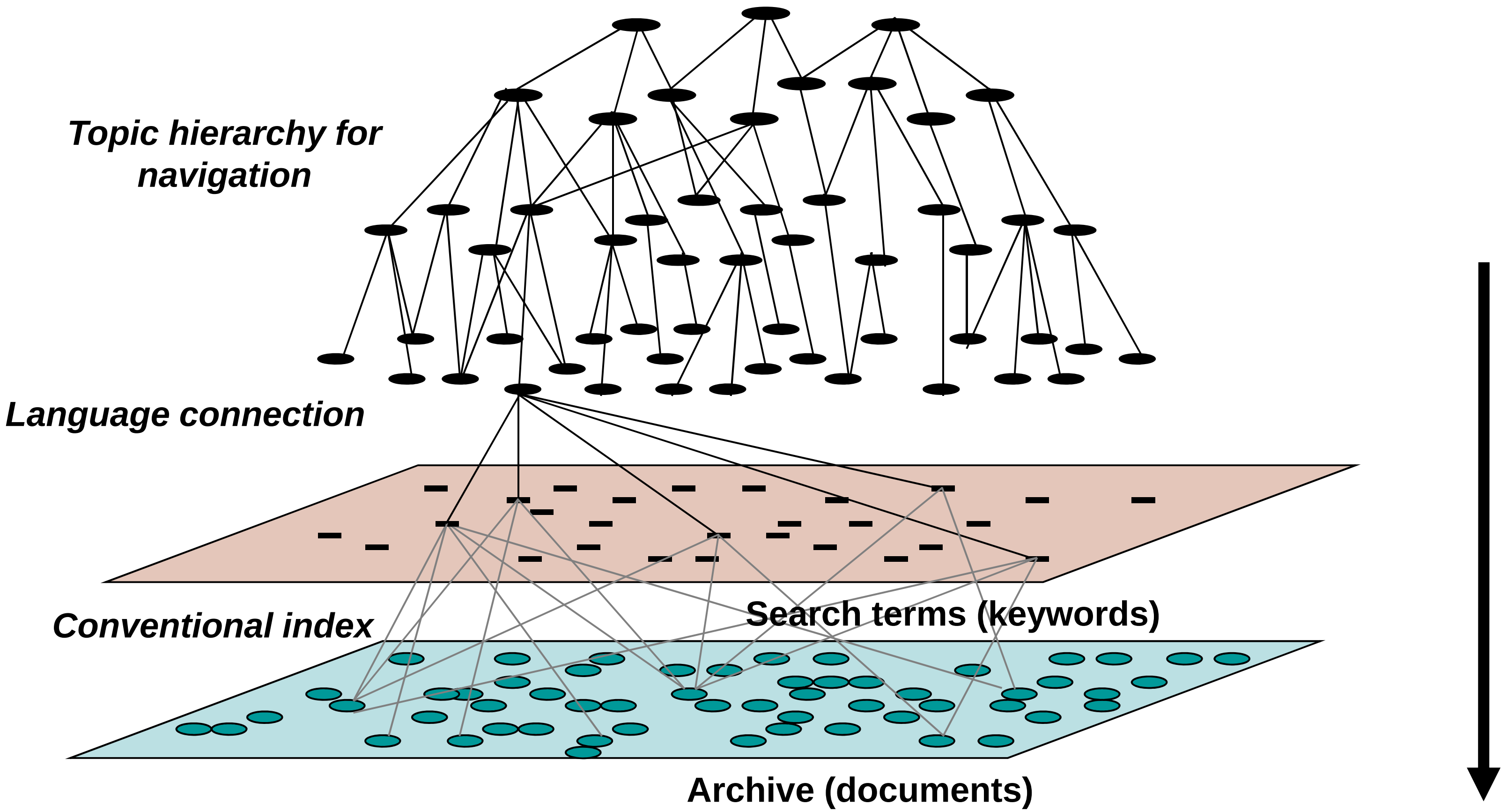
Conventional index

Search terms (keywords)

Archive (documents)

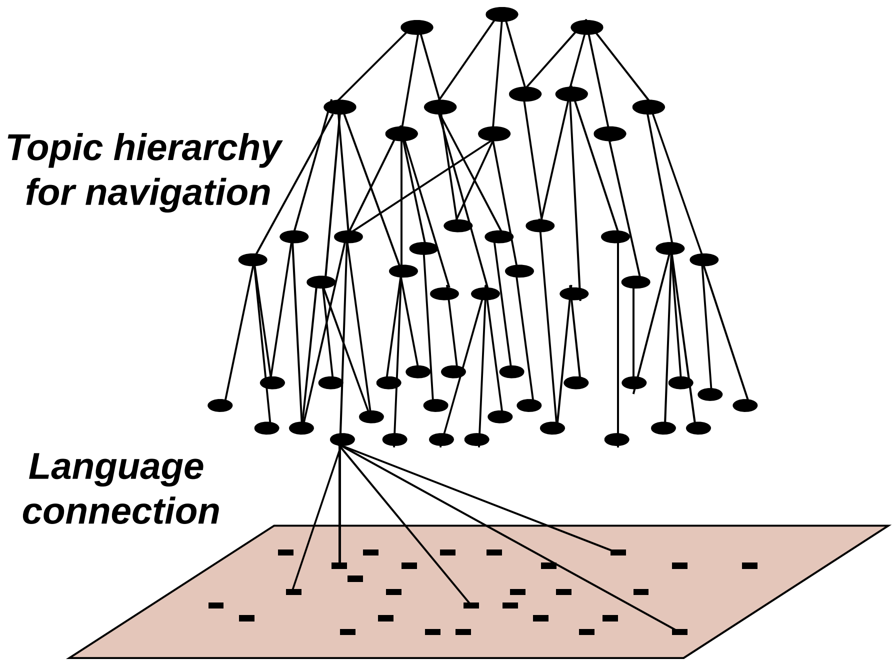
Assigning topics to documents

How does it work?



Assigning documents to topics

Automatic thematic access by TopicZoom



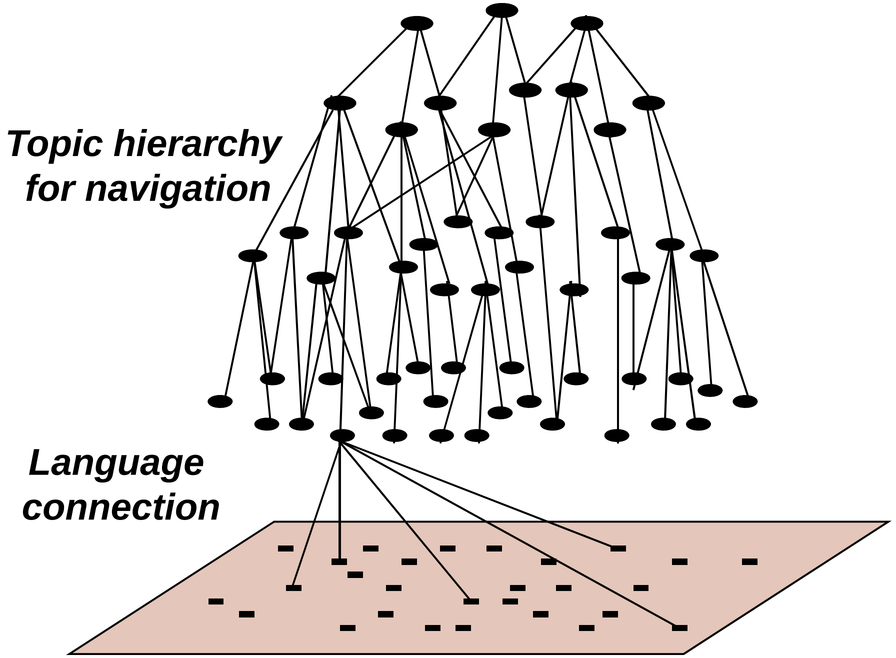
Our semantic net

- > 95,000 topics
- geograpic subhierarchy
- temporal subhierarchy
- persons
- organisations
- events
- > 2,500,000 links

Strong linguistic basis for German (and English)

Much manual work spent (ongoing)!

Automatic thematic access by TopicZoom



Only **non-ambigious** keywords

Many **complex** key„words“ („human rights“, „violation of human rights“)

Our semantic net

- > 95,000 topics
- geograpic subhierarchy
- temporal subhierarchy
- persons
- organisations
- events
- > 2,500,000 links

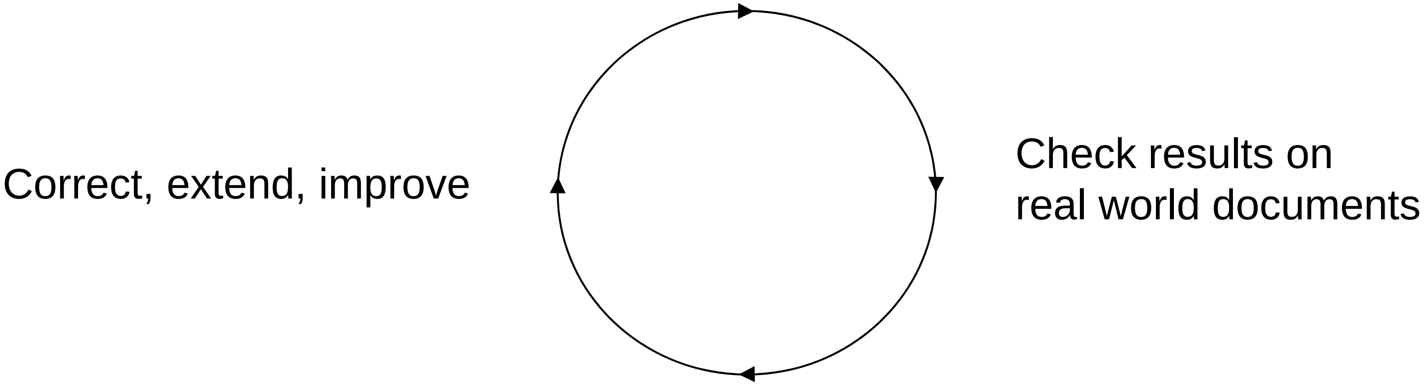
Strong linguistic basis for German (and English)

Much manual work spent (ongoing)!

Ontology building is an art!
Size is not quality

Large encyclopedic ontologies

- are never perfect
- need permanent acualization, control and improvement



Automated methods help but do not suffice!
(Do not always trust scientific talks/papers – only trust what you can check!)

Automated topic assignment
Risks and disadvantages

Automated topic assignment
Risks and disadvantages

- Ambiguities in language can lead to wrong topic assignments:

  „Mengen“ (German city and mathematical concept)

  Manual inspection of results helps to detect/exclude ambigious notions

  „Mengen“ -> **„Mengen (Landkreis Sigmaringen)“**

- Restricted coverage: some highly specialized concepts in texts from special areas not recognized

  Ontology can be extended by demand

- Chains of relations in ontology may lead to wrong conclusions.

  Olympic games – olympic cities – Munich – Augustiner beer

  Reorganizing the ontology helps to eliminate wrong associations

Automated topic assignment
Advantages and chances

Automated topic assignment
Advantages and chances

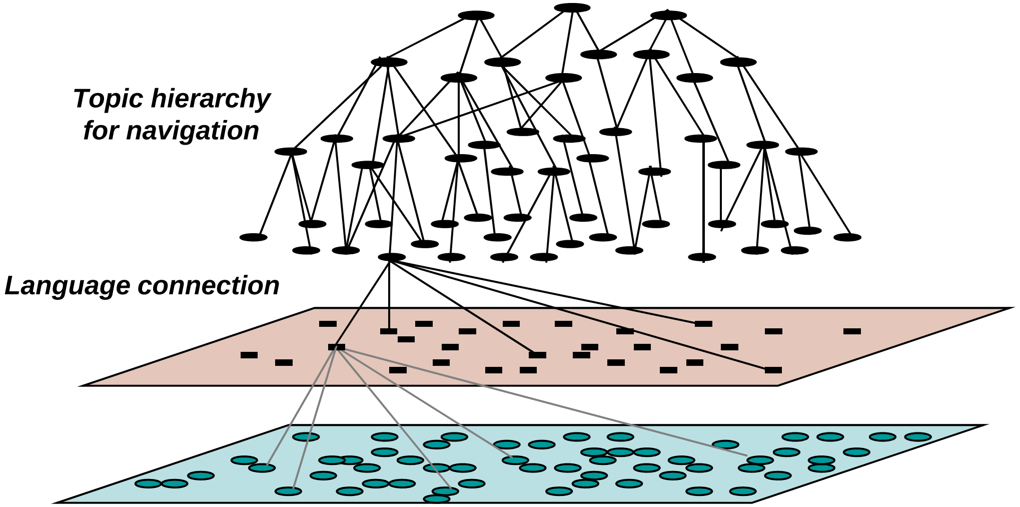
- Can deal with millions of documents
- Standardized assignment, no subjective views
- Topic hierarchies used can be easily adapted to specific needs
- Many interesting possibilities for advanced search and analysis
    - computing surveys and topic maps
    - navigational access
    - cross-linking
    - comparison and analysis of subcollections
    - support for facetted search
    - …

Use of assigned topics

Offering a survey

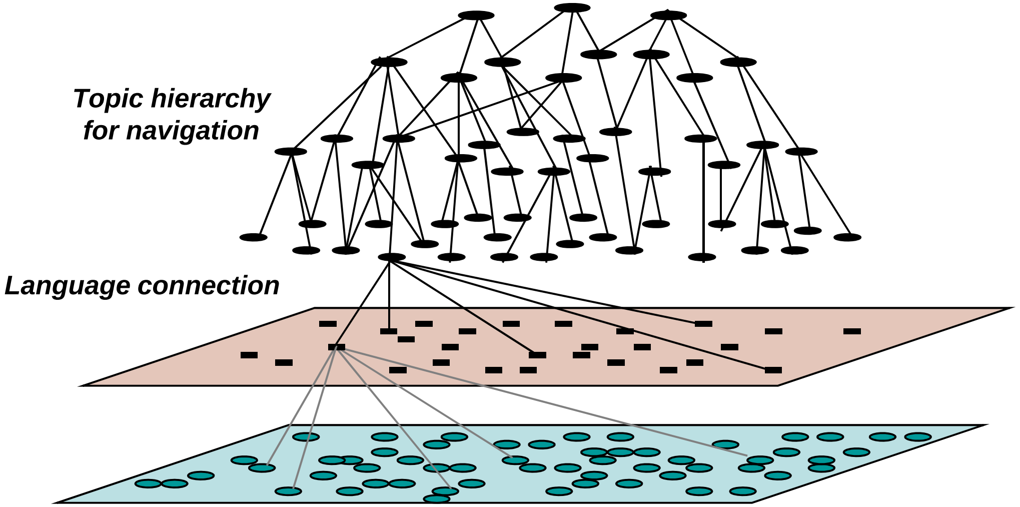
Use of assigned topics

Offering a survey



Use of assigned topics

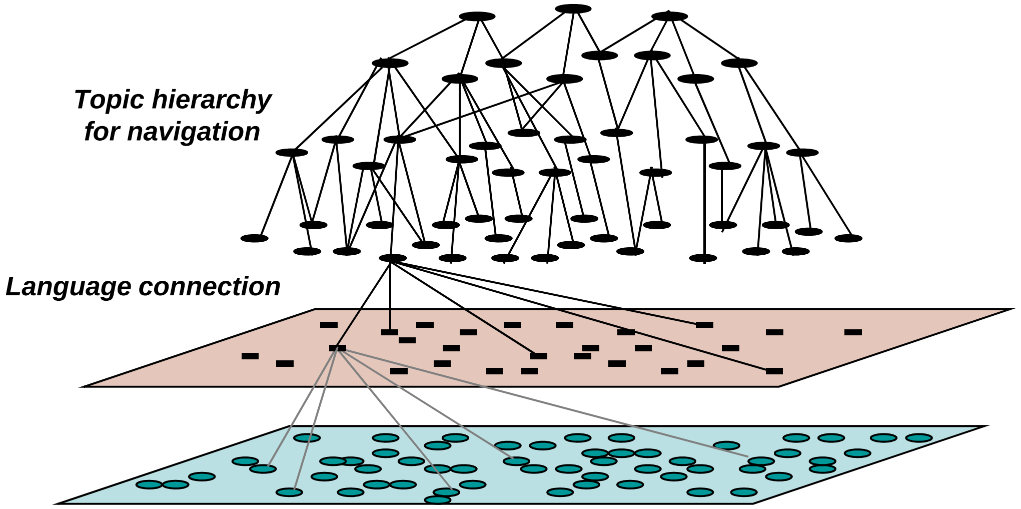
Offering a survey



Topic catalogue for archive:
The hierarchy for a given archive
only contains topics of the archive

Use of assigned topics

Offering a survey

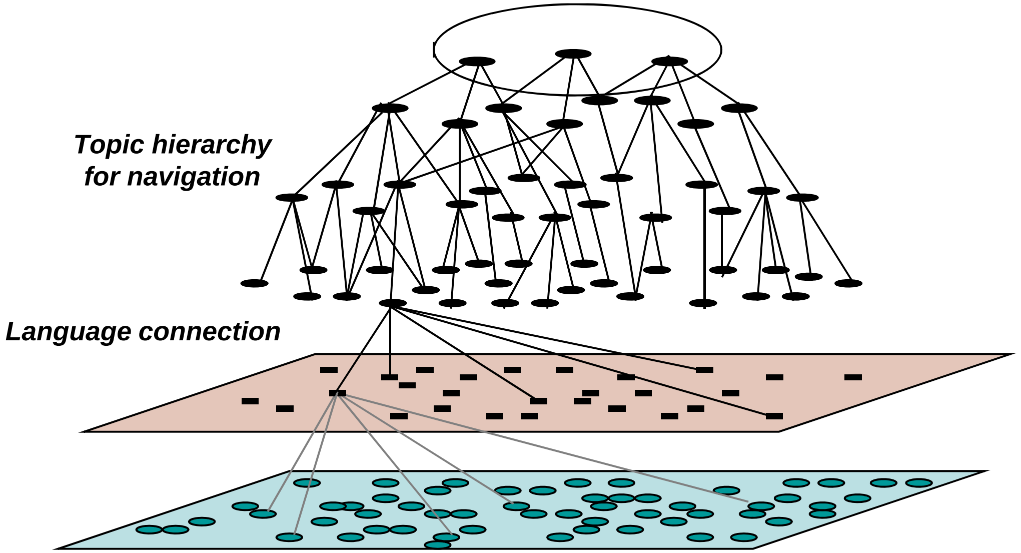


Topic catalogue for archive:
The hierarchy for a given archive only contains topics of the archive

Weight/importance of each topic
How many documents contain terms from topics X, Y, Z?

Use of assigned topics

Offering a survey

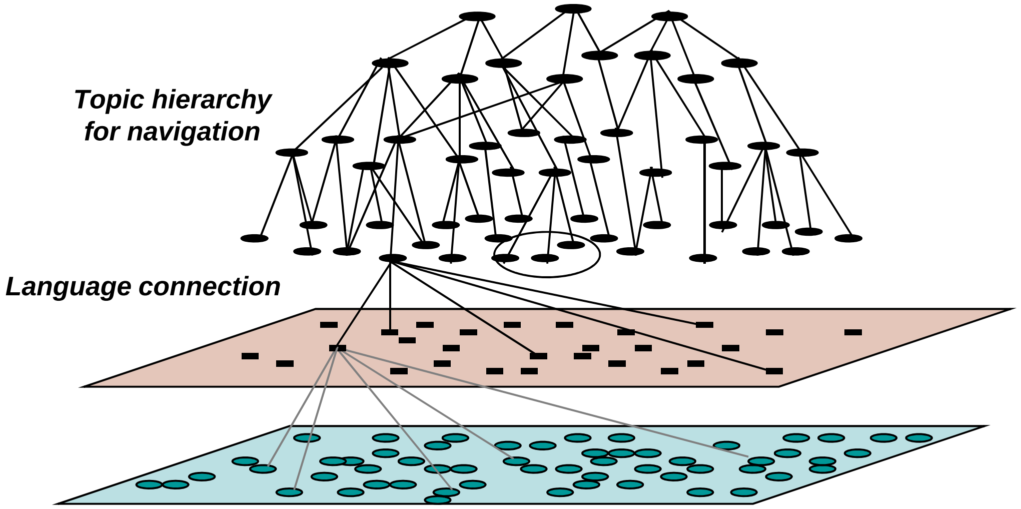


Topic catalogue for archive:
The hierarchy for a given archive only contains topics of the archive

Weight/importance of each topic
How many documents contain terms from topics X, Y, Z?

Use of assigned topics

Offering a survey

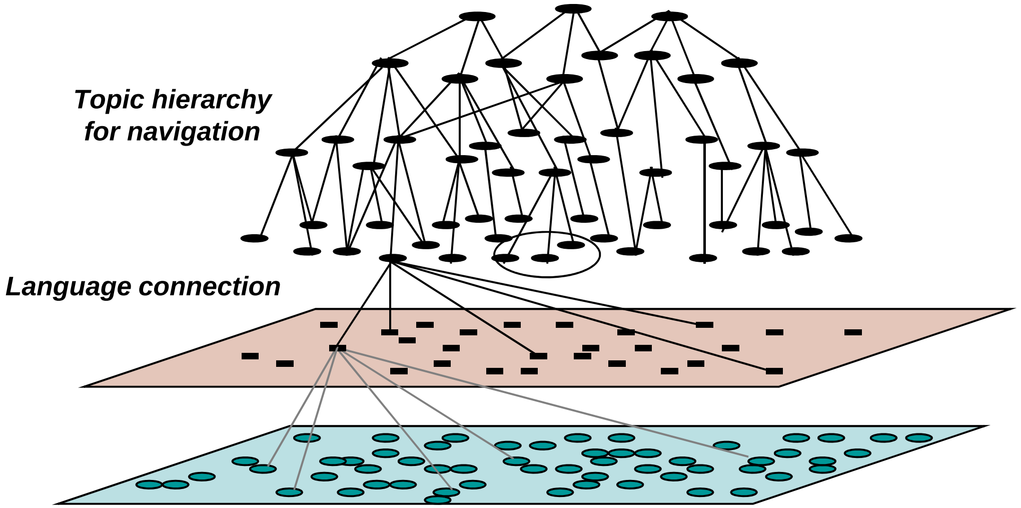


Topic catalogue for archive:
The hierarchy for a given archive only contains topics of the archive

Weight/importance of each topic
How many documents contain terms from topics X, Y, Z?

Use of assigned topics

Offering a survey



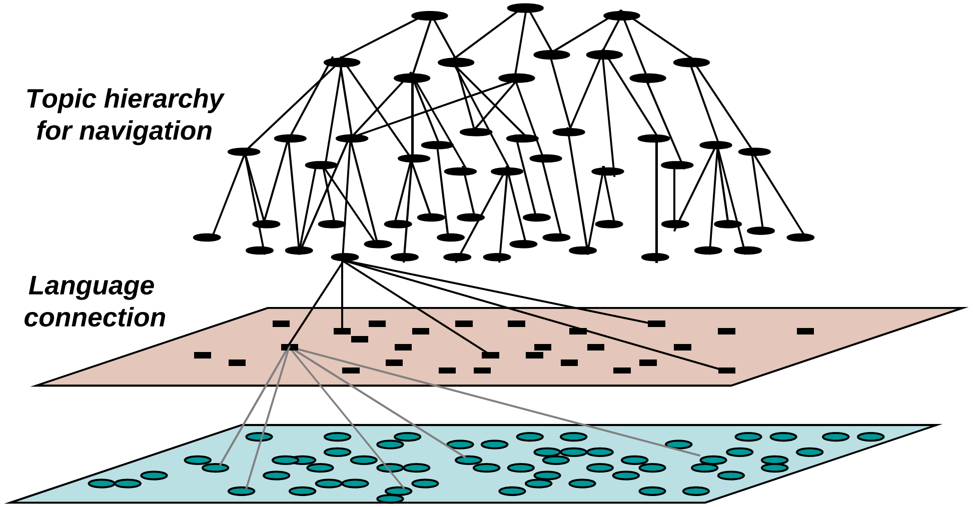
Overview over relevant topics of archive and their weight/importance

Use of assigned topics

Navigation and search

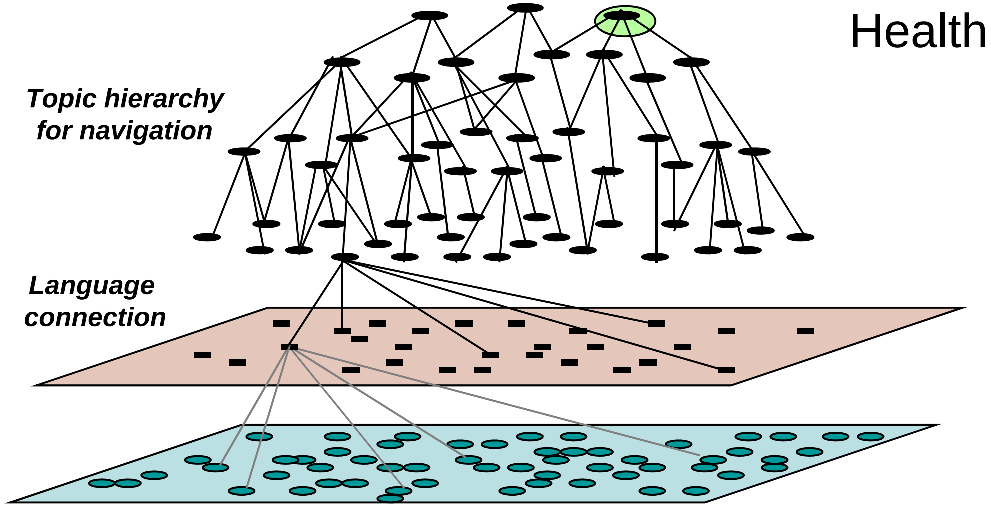
Use of assigned topics

Navigation and search



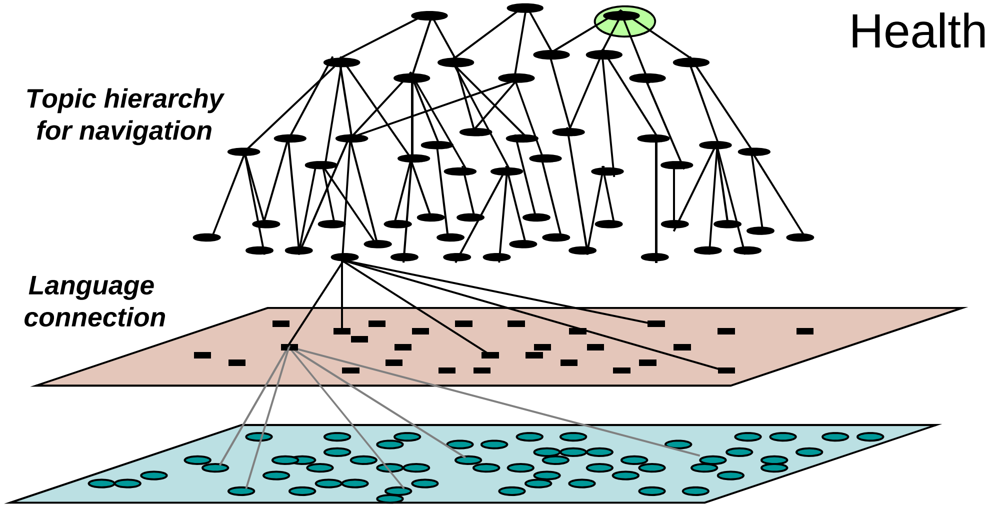
Use of assigned topics

Navigation and search



Use of assigned topics

Navigation and search

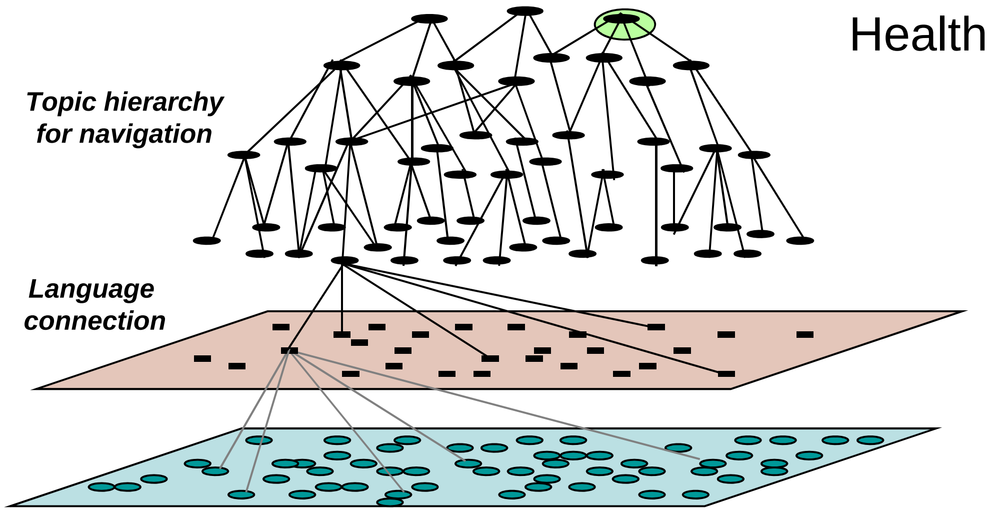


Health

List of hits

Use of assigned topics

Navigation and search



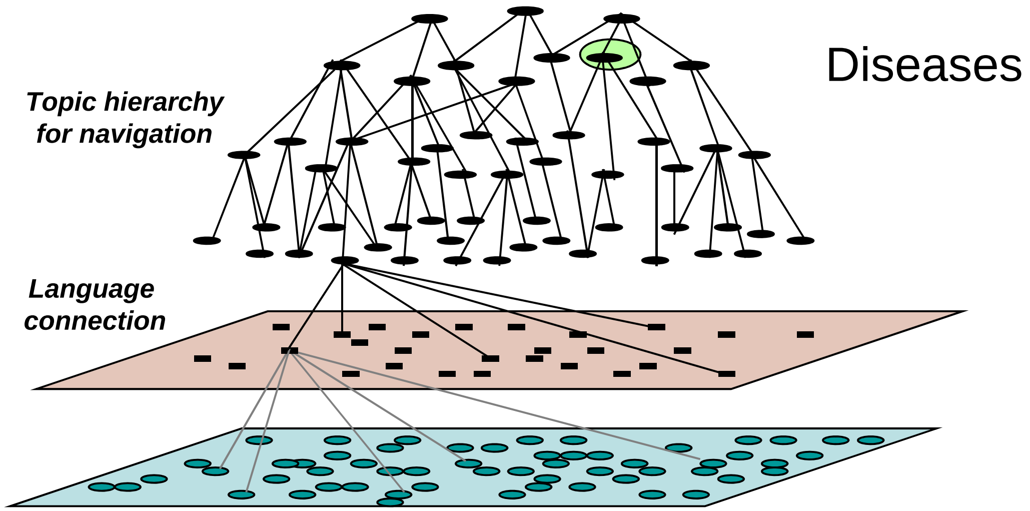
Health

List of hits

Document title
Terms from domain „Health“
Further interesting keywords

Use of assigned topics

Navigation and search



New list of hits

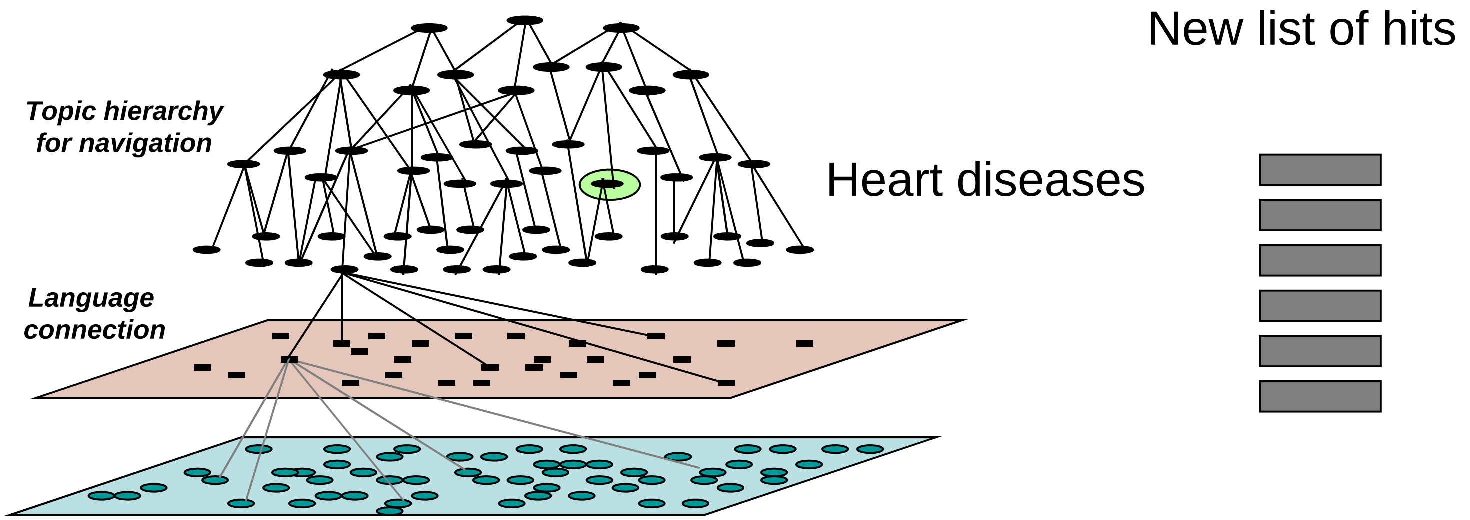
▬ Document title

Terms from domain „Diseases“

Further interesting keywords

Use of assigned topics

Navigation and search



Topic hierarchy for navigation

Language connection

New list of hits

Heart diseases

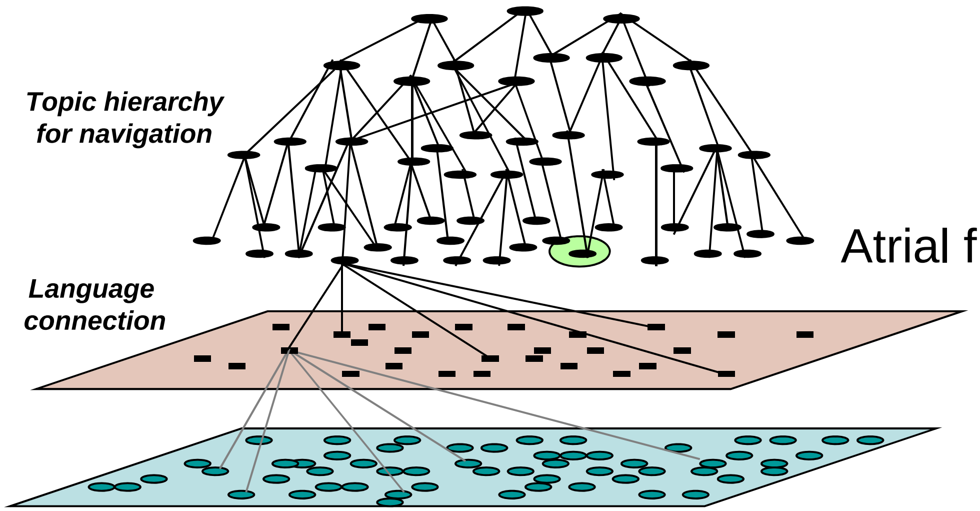
Document title

Terms from domain „Heart diseases“

Further interesting keywords

Use of assigned topics

Navigation and search



Atrial fibrillation

New list of hits

Document title

Terms from domain „Atrial fibrillation“

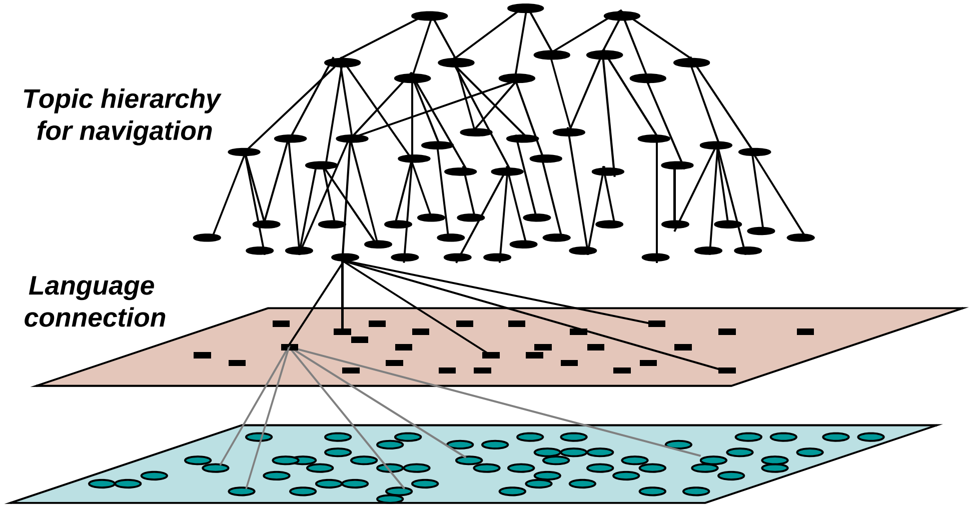
Further interesting keywords

Use of assigned topics

Linking documents/pieces of text to external data

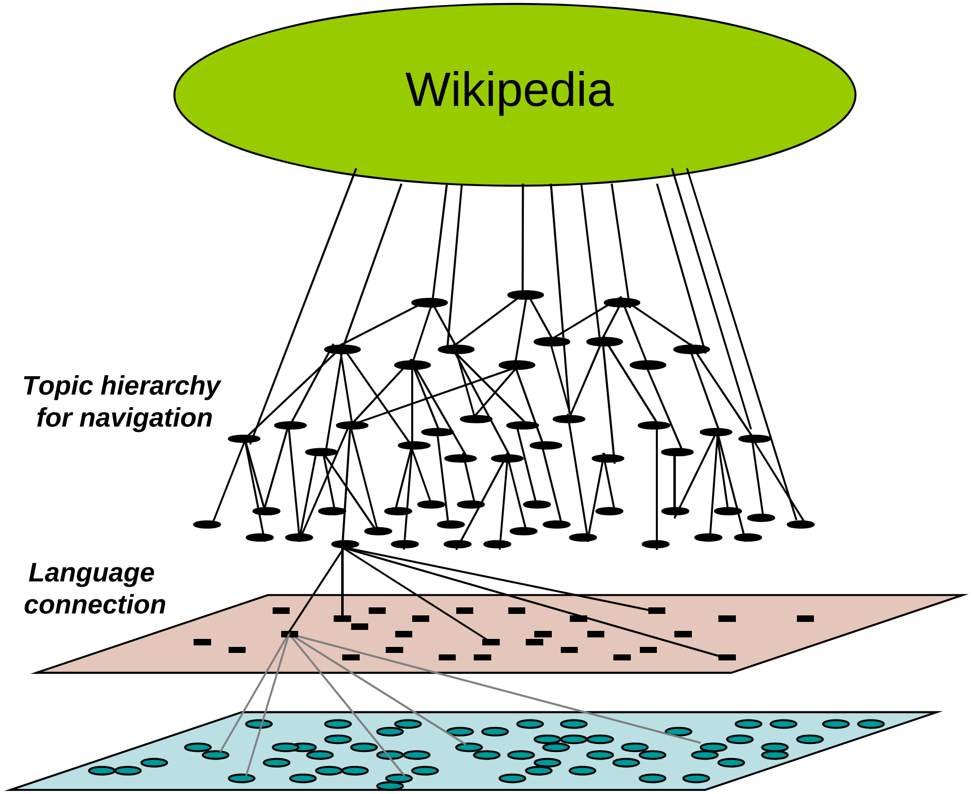
Use of assigned topics

Linking documents/pieces of text to external data



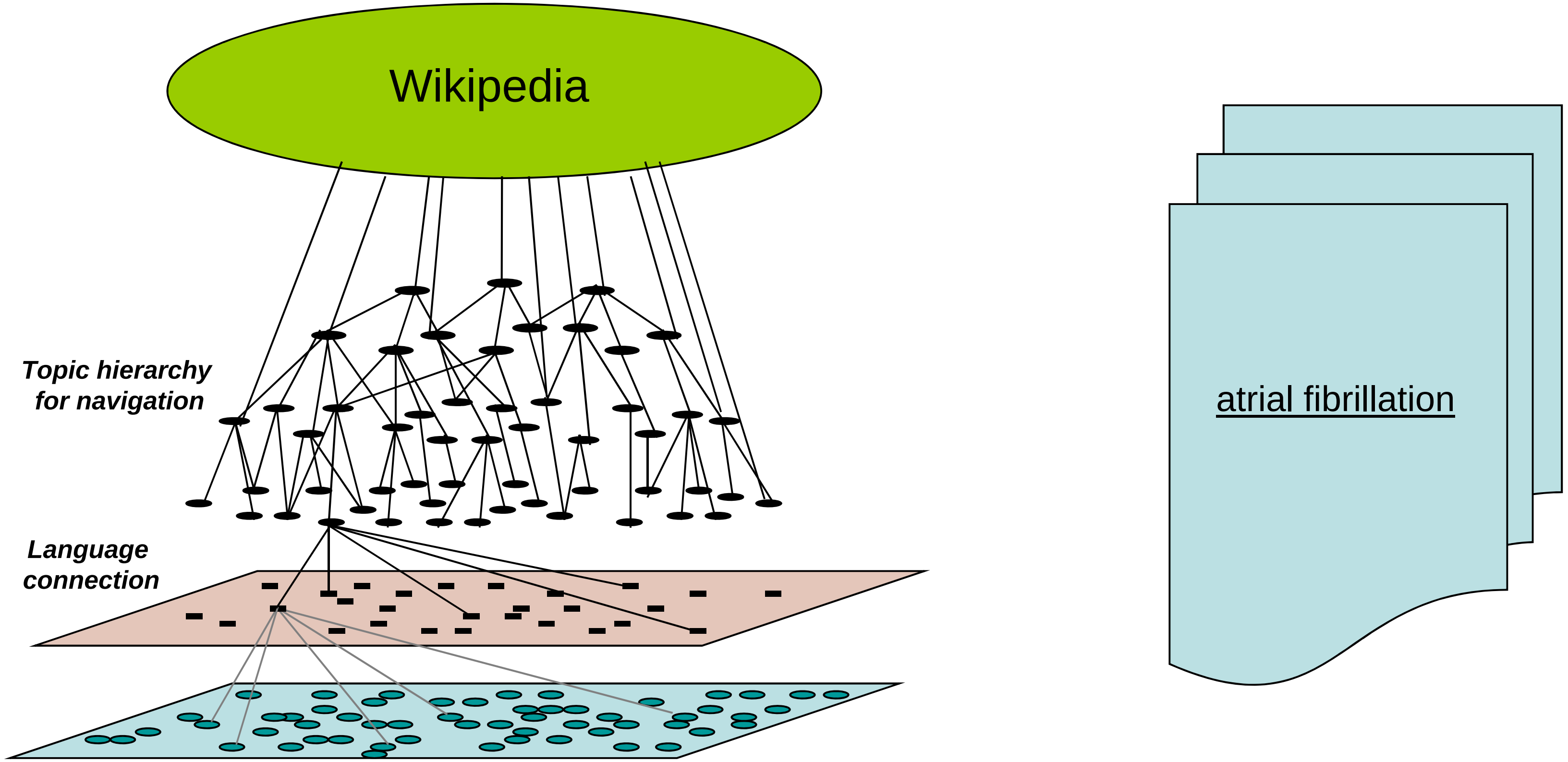
Use of assigned topics

Linking documents/pieces of text to external data



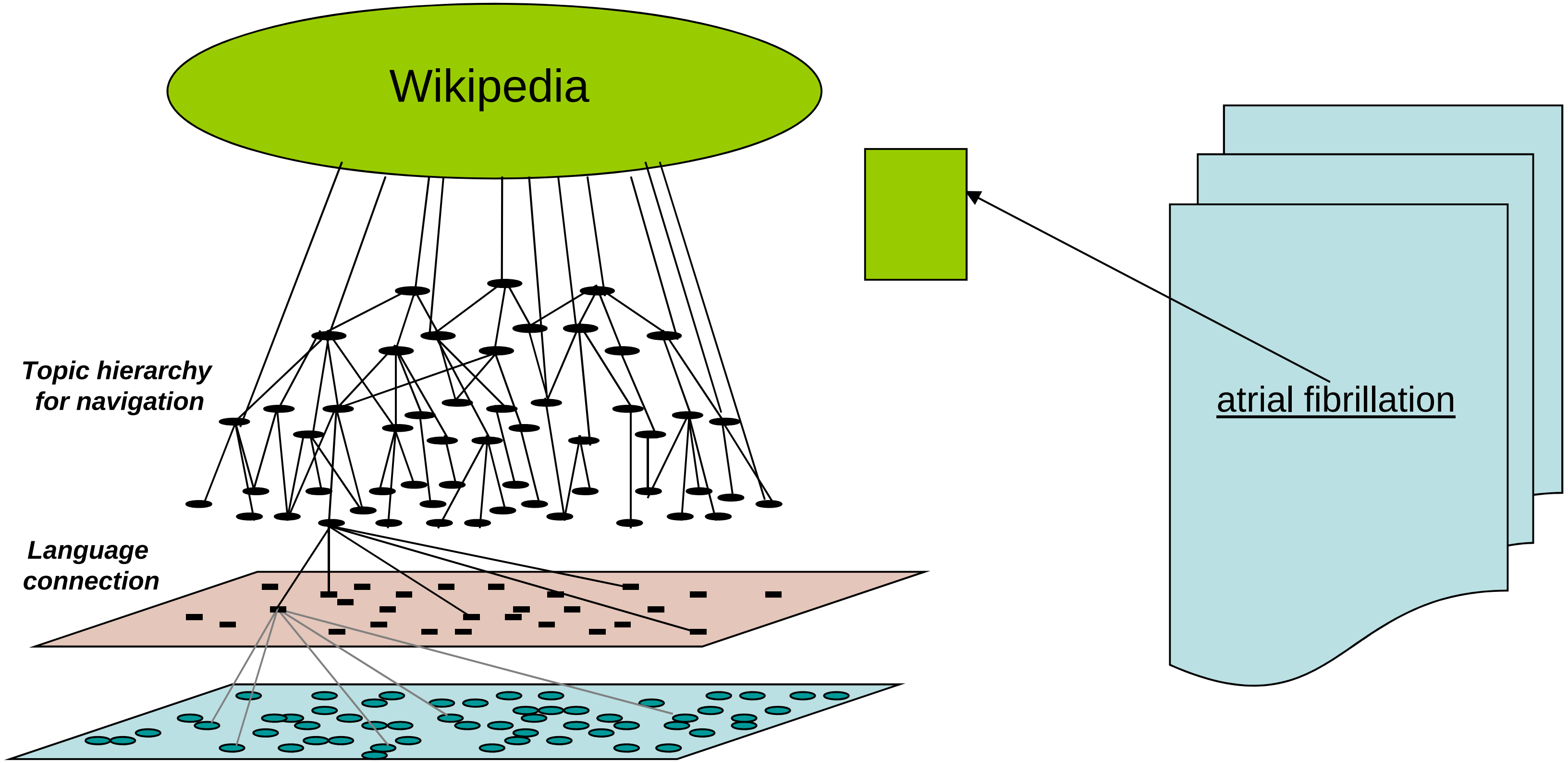
Use of assigned topics

Linking documents/pieces of text to external data



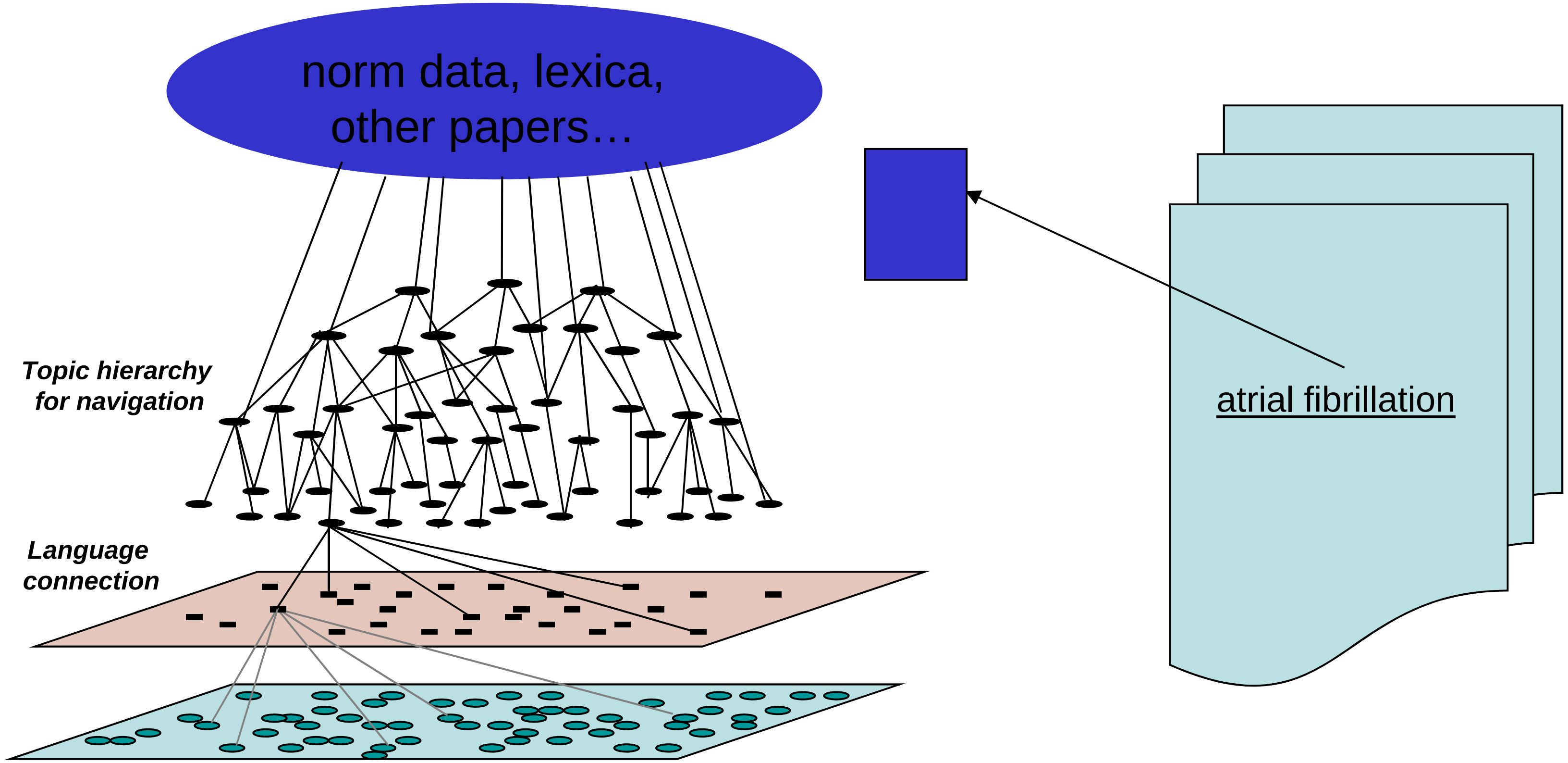
Thematic access to archive

Linking documents/pieces of text to external data



Use of assigned topics

Linking documents/pieces of text to external data



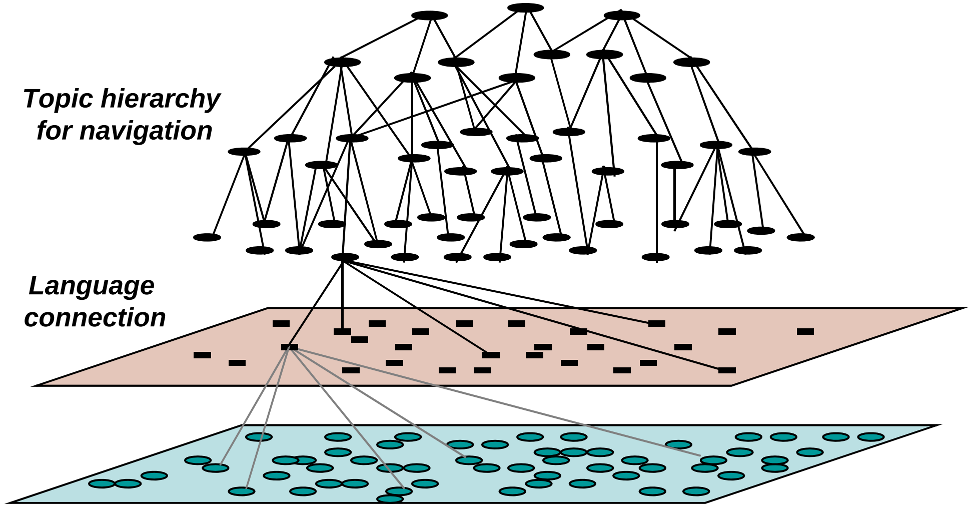
norm data, lexica, other papers…
Topic hierarchy for navigation
Language connection
atrial fibrillation

Use of assigned topics

Linking documents/pieces of text to ontology

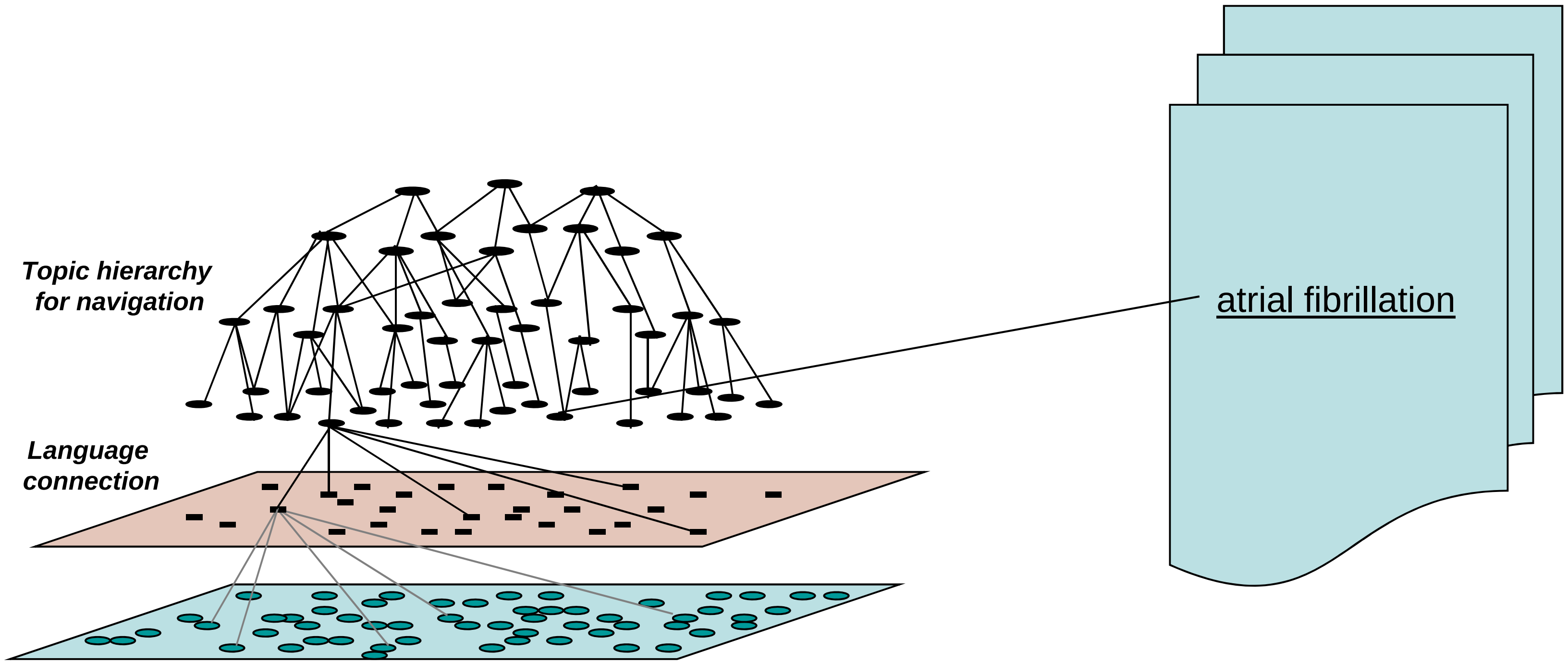
Use of assigned topics

Linking documents/pieces of text to ontology



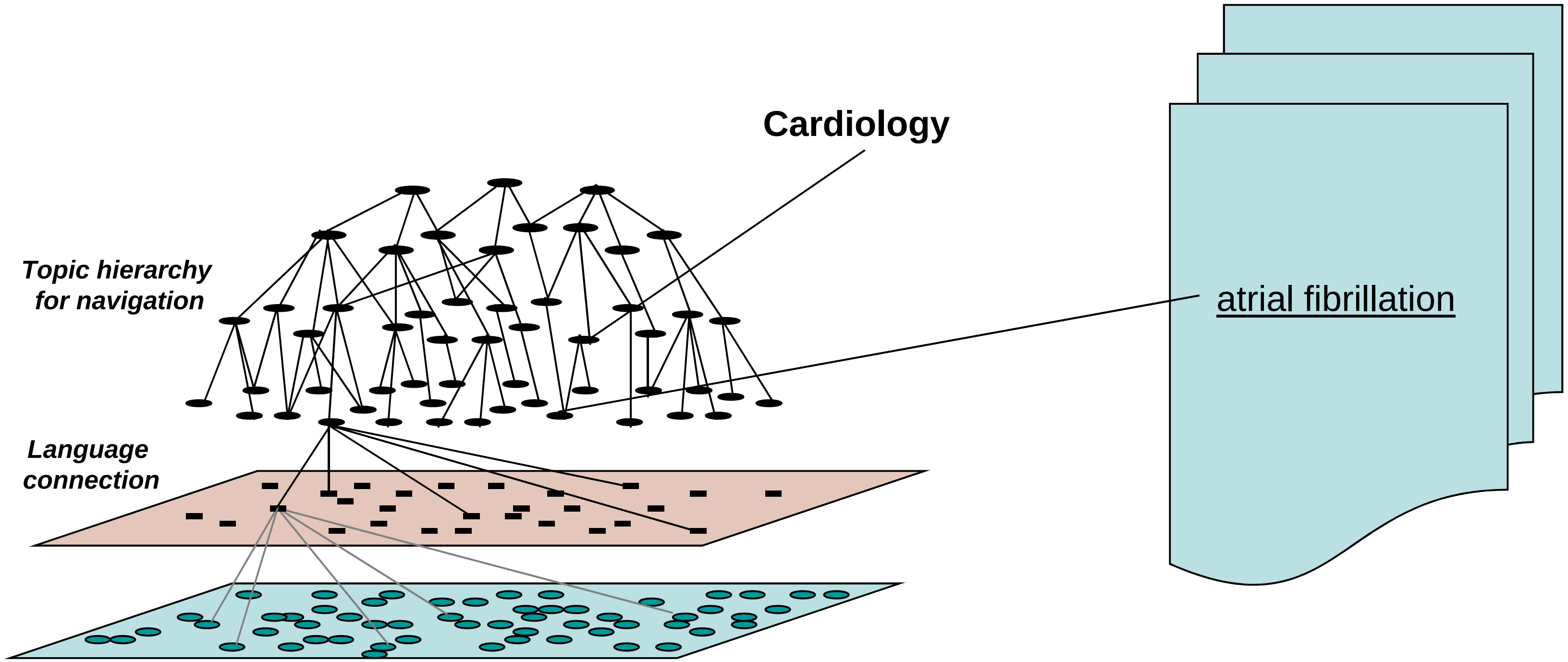
Use of assigned topics

Linking documents/pieces of text to ontology



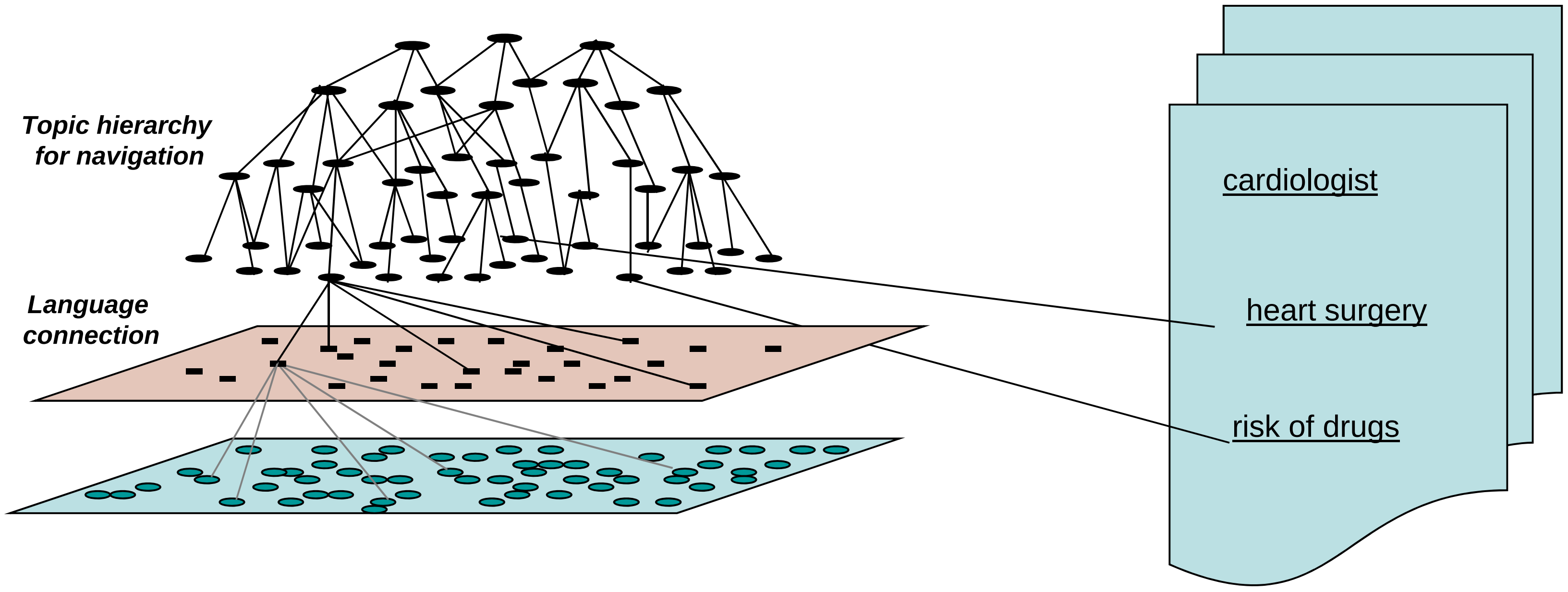
Use of assigned topics

Linking documents/pieces of text to ontology



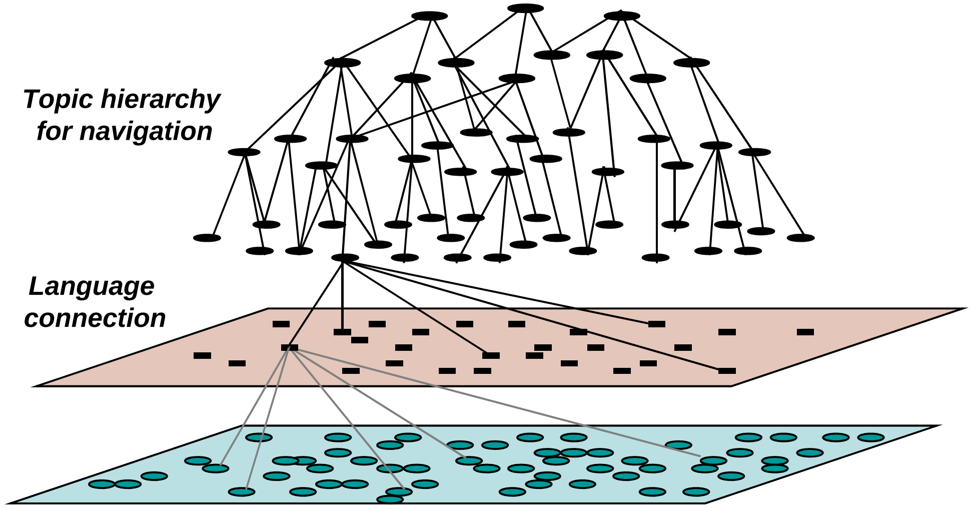
Use of assigned topics

Linking documents/pieces of text to ontology



Use of assigned topics

Linking documents/pieces of text to ontology



„Berry-picking paradigm“:

Interesting documents point to new interesting concepts, which point to new interesting documents,...

…..

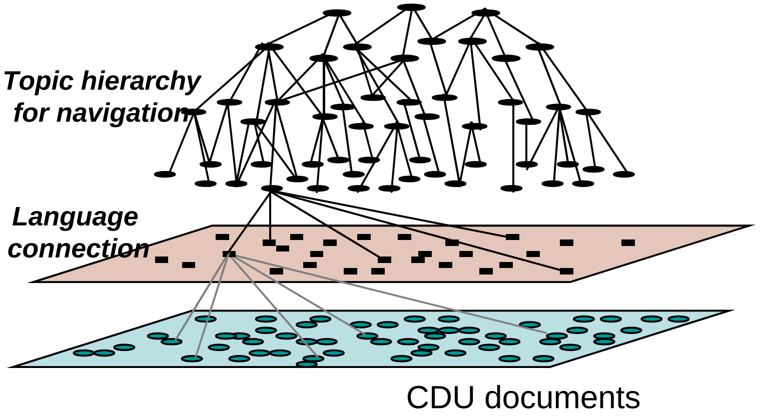
Use of assigned topics

Comparisons and analysis of subcollections

Use of assigned topics

Comparisons and analysis of subcollections

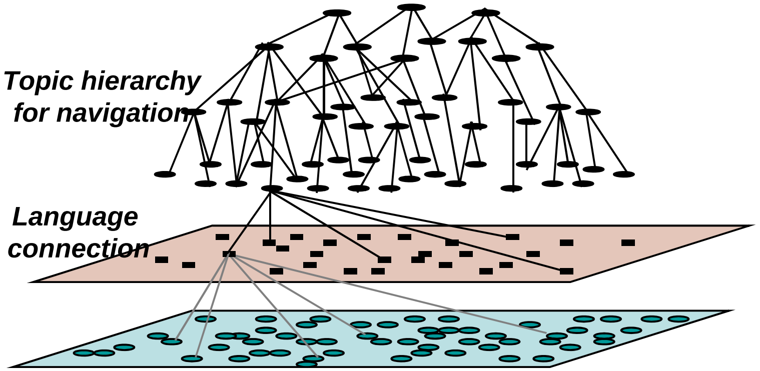
Compute archive segment 1:
Documents for topic CDU



Use of assigned topics

Comparisons and analysis of subcollections

Compute archive segment 1:
Documents for topic CDU

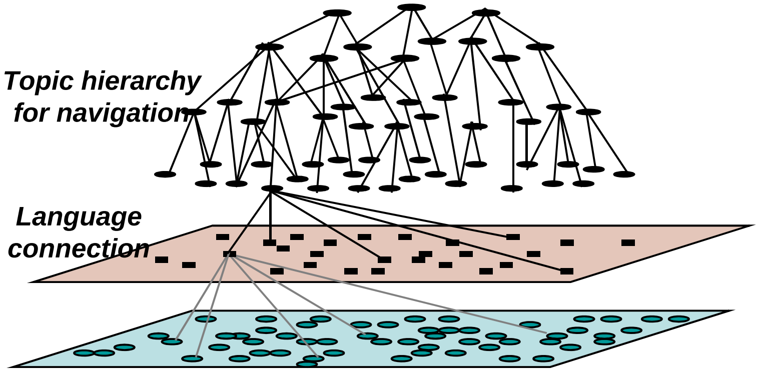


Topic hierarchy for navigation

Language connection

CDU documents

Compute archive segment 2:
Documents for topic SPD



Topic hierarchy for navigation

Language connection

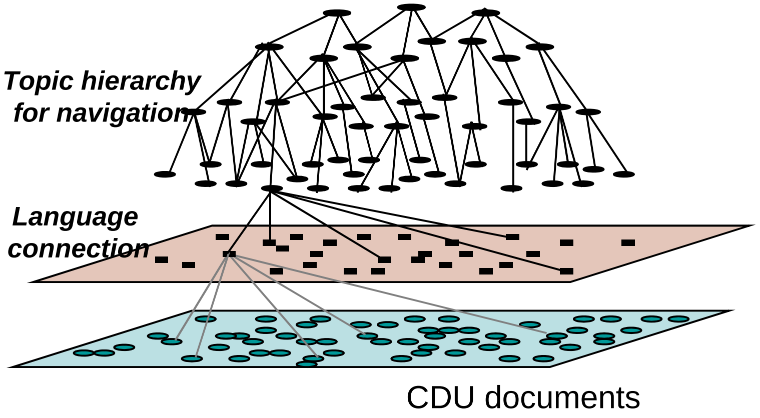
SPD documents

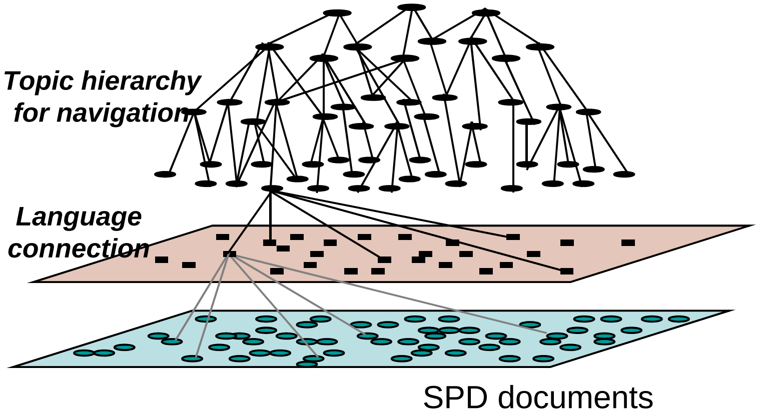
Use of assigned topics

Comparisons and analysis of subcollections

Compute archive segment 1:
Documents for topic CDU

Compute archive segment 2:
Documents for topic SPD





What are the differences as to the thematic structure?
Which topics are more dominant in the „CDU subarchive“
than in the „SDP subarchive“ and vice versa?

Use of assigned topics

Comparisons and analysis of subcollections

A given main topic is analyzed under a secondary topic:

- Which health-related topics occur in documents on nutrition?
- Which countries occur in documents about AIDS?
- Which enterprises are mentioned in documents on energy saving?

Use of assigned topics

Facetted topic search

Use of assigned topics

Facetted topic search

Hierarchy 1	Hierarchy 2	Hierarchy 3
Places	Industry sectors	Enterprise departments

Use of assigned topics

Facetted topic search

Hierarchy 1 Places	Hierarchy 2 Industry sectors	Hierarchy 3 Enterprise departments
		Marketing

Use of assigned topics

Facetted topic search

Hierarchy 1 Places	Hierarchy 2 Industry sectors	Hierarchy 3 Enterprise departments
	Pharma industry	Marketing

Use of assigned topics

Facetted topic search

Hierarchy 1 Places	Hierarchy 2 Industry sectors	Hierarchy 3 Enterprise departments
Bavaria	Pharma industry	Marketing

Use of assigned topics

Facetted topic search

Hierarchy 1
Places

Hierarchy 2
Industry sectors

Hierarchy 3
Enterprise departments

Bavaria
Munich

Pharma industry

Marketing

Use of assigned topics

Facetted topic search

Hierarchy 1
Places

Hierarchy 2
Industry sectors

Hierarchy 3
Enterprise departments

Health industry

Bavaria
Munich

Pharma industry

Marketing

Summing up: Automated topic assignment

- Imitating traditional work in libraries in bringing order to collections – but fully automated
- Millions of documents can be analyzed
- Surveys on and thematic search in very large repository supported
- Good basis also for many new forms of interaction with documents & (sub)collections
- Largely, but not always 100% error-free

Useful additions

- Traditional keyword-based search
- „General“ Named Entity extraction
- Genre classification
- Logical layout analysis (FEP, tomorrow)
- Detection of other „metadata“

(language of text, time when text was born)

Thanks for your attention!

Thanks for your attention!

In what follows…..

Live Demo TopicZoom Automated Topic Assignment

Play yourself (German texts):

http://www.topiczoom.de