

Information Retrieval

Exercise – Winter term 2026/2027

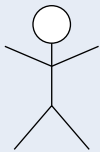
`klara.gutekunst@uni-kassel.de`

Agenda

1. Architecture of a Search Engine
2. Indexing Process
3. Cranfield Paradigm
4. Formulating Topics
5. IR Datasets Tutorial
6. Next Steps

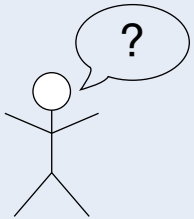
Architecture of a Search Engine

Karaoke Version



Architecture of a Search Engine

Karaoke Version



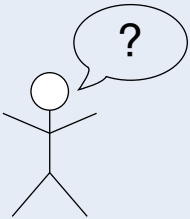
Architecture of a Search Engine

Karaoke Version



Indexing

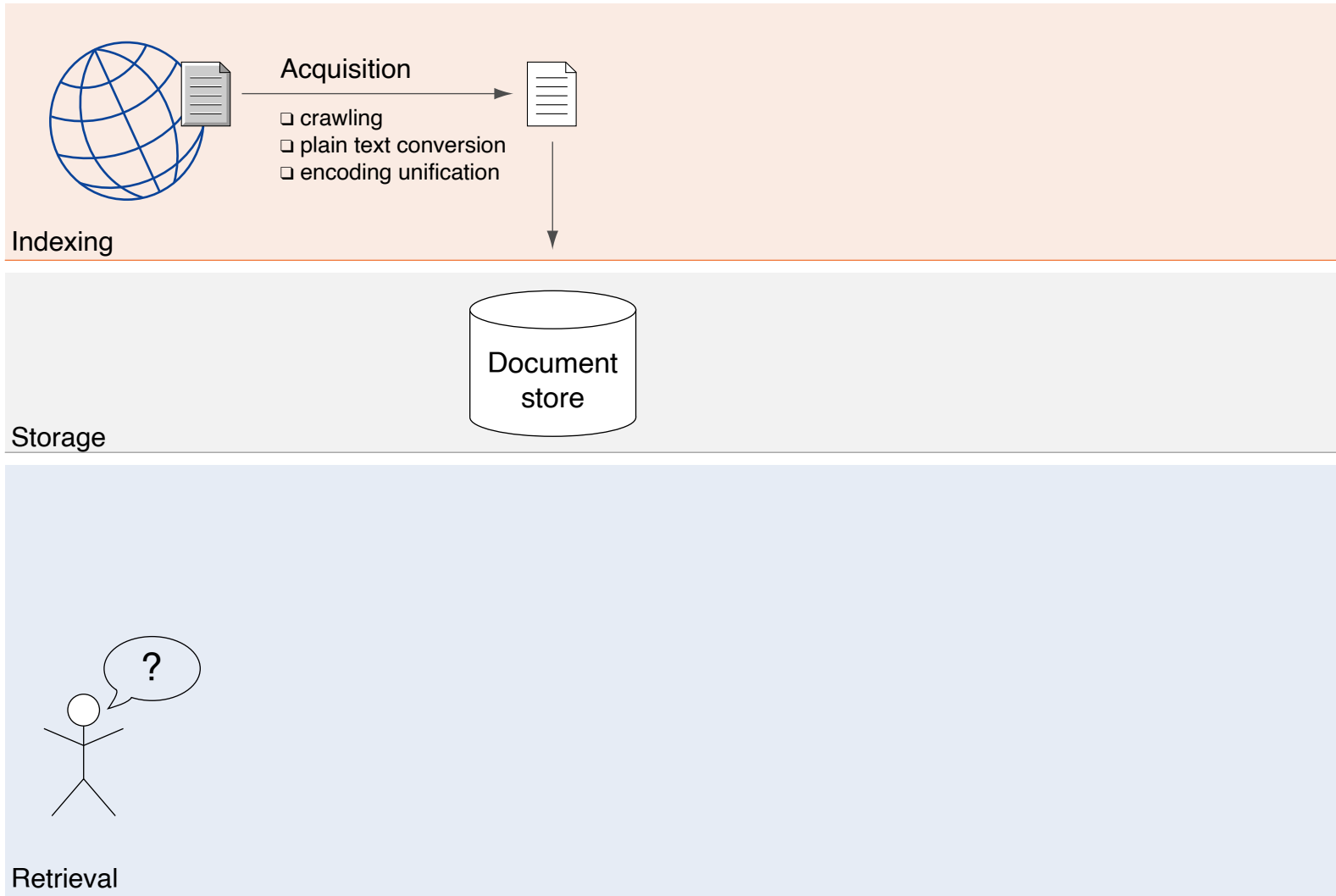
Storage



Retrieval

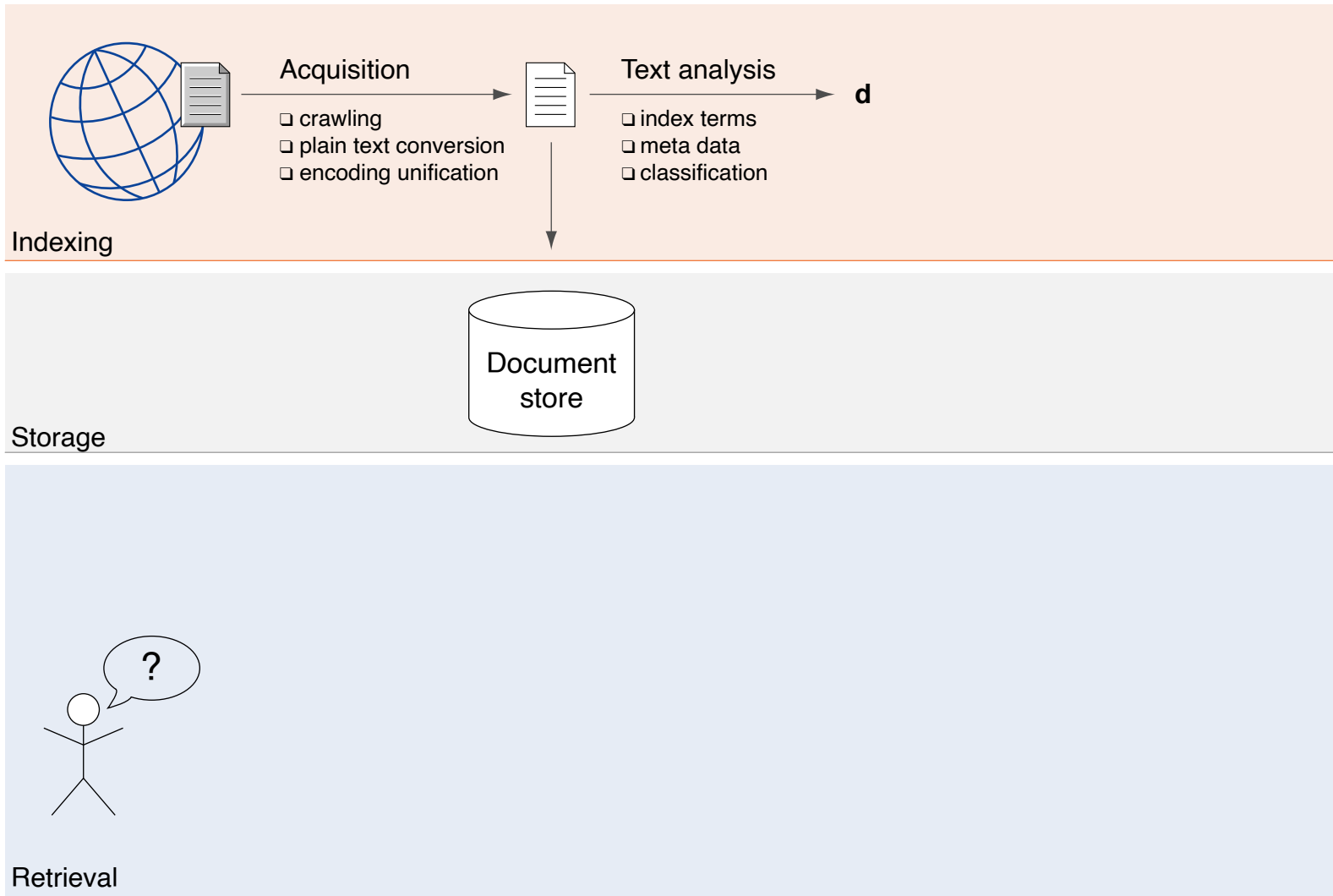
Architecture of a Search Engine

Karaoke Version



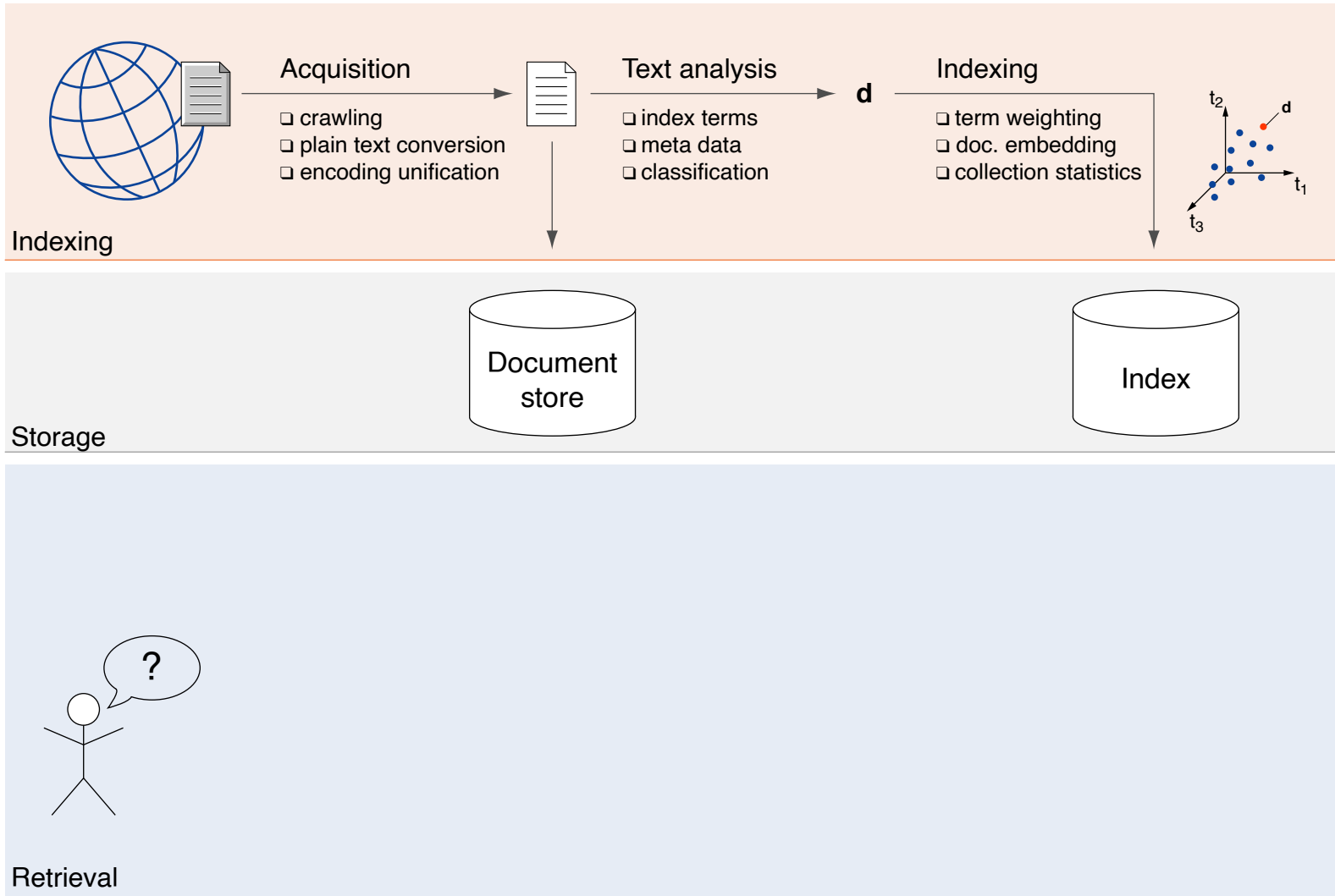
Architecture of a Search Engine

Karaoke Version



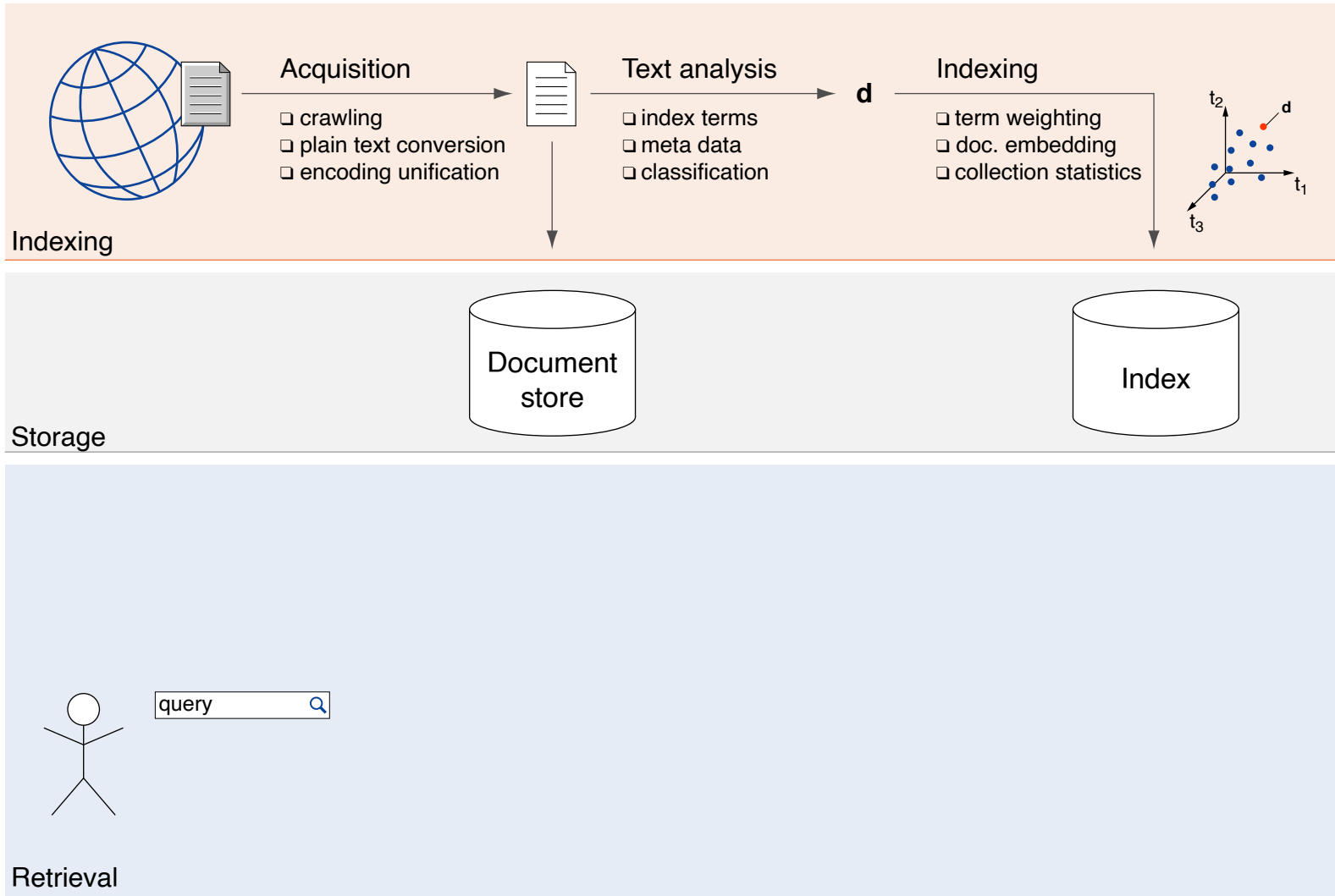
Architecture of a Search Engine

Karaoke Version



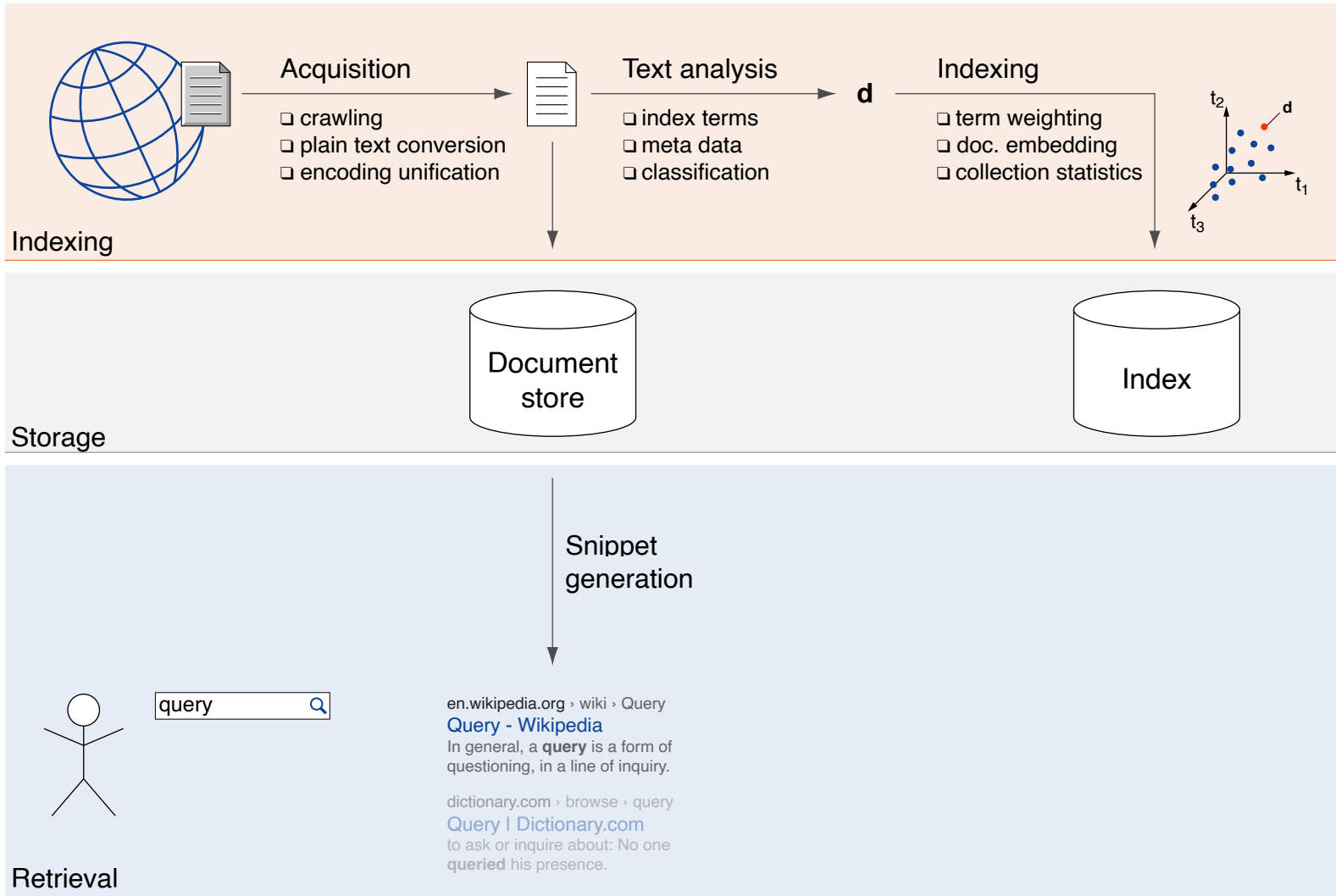
Architecture of a Search Engine

Karaoke Version



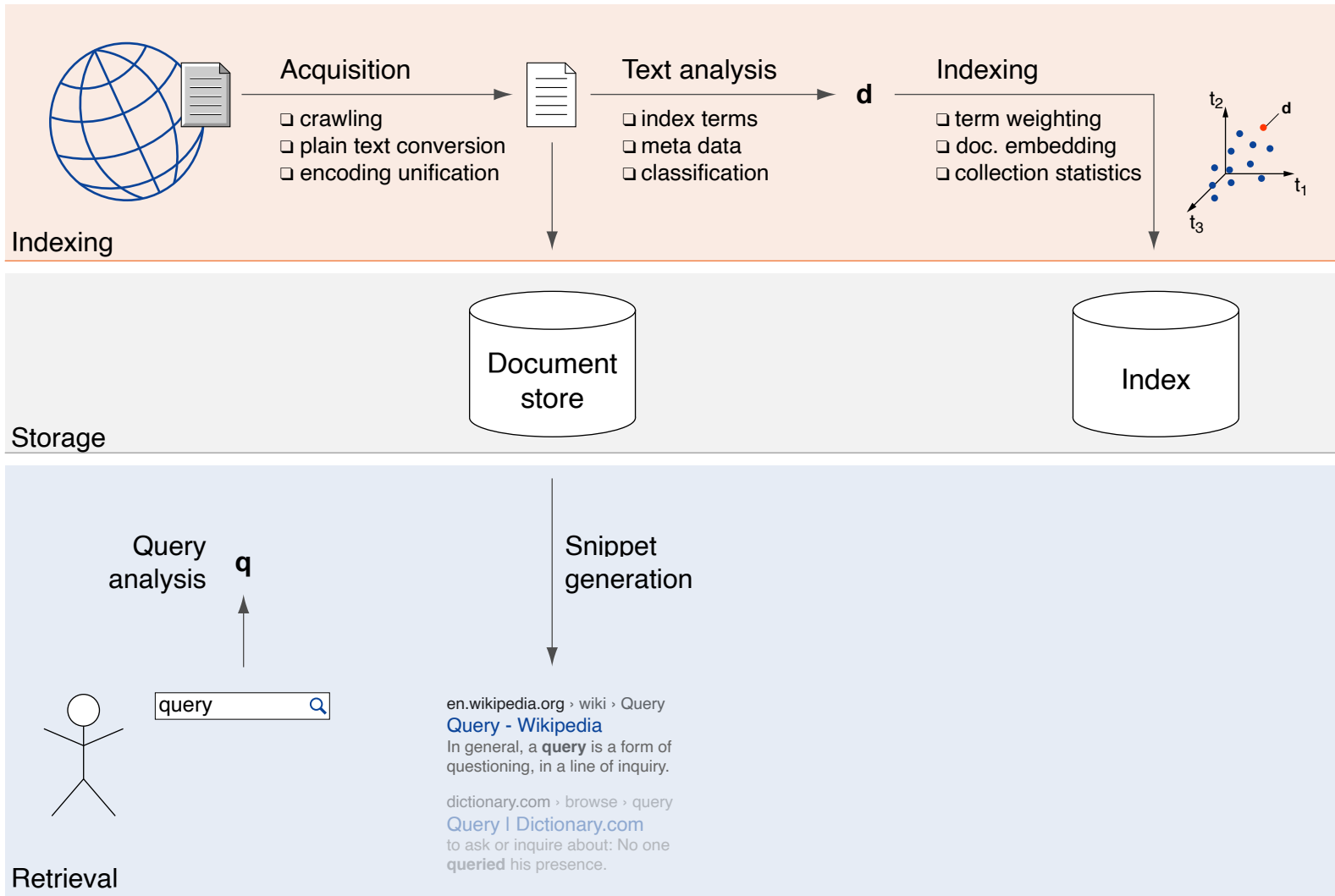
Architecture of a Search Engine

Karaoke Version



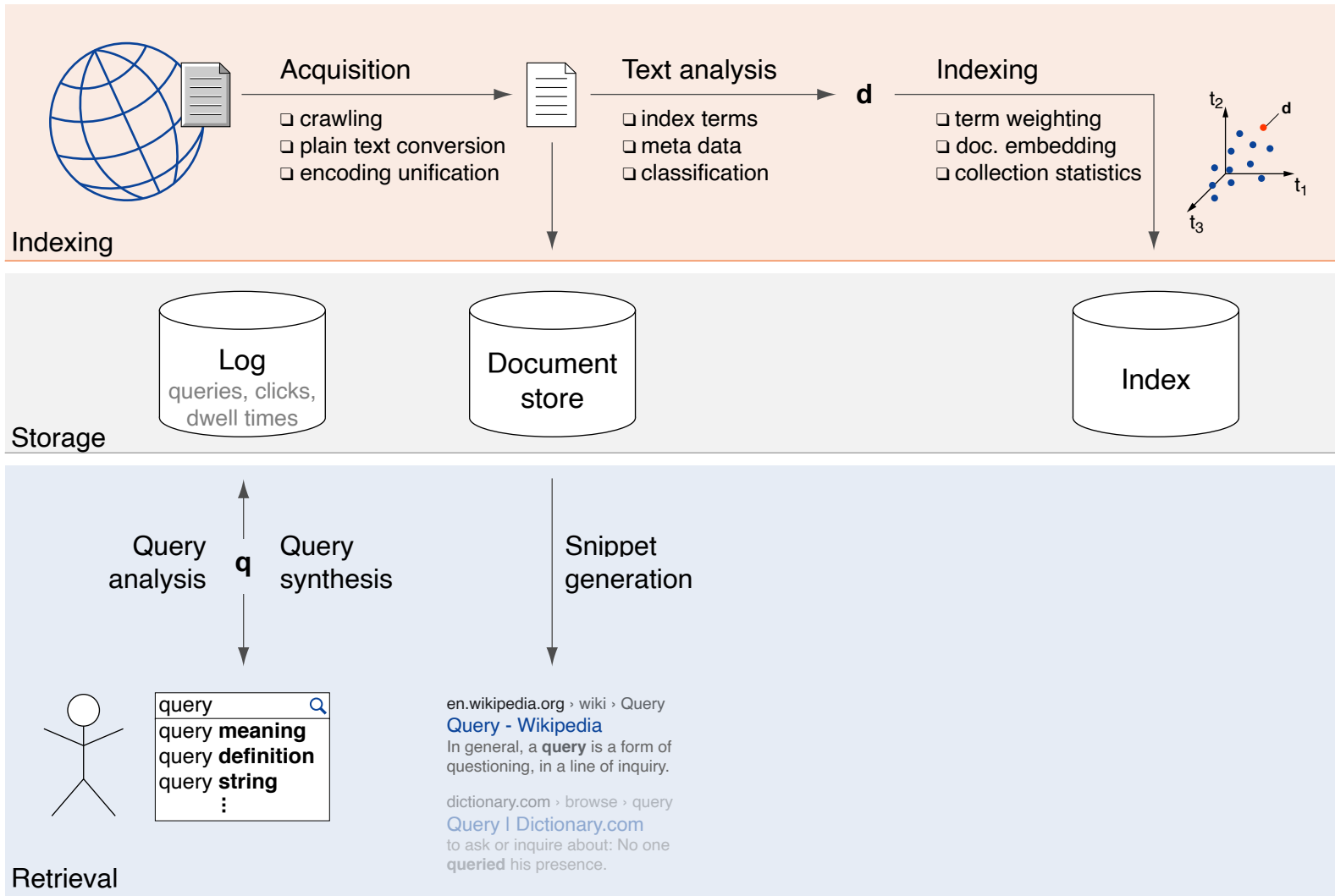
Architecture of a Search Engine

Karaoke Version



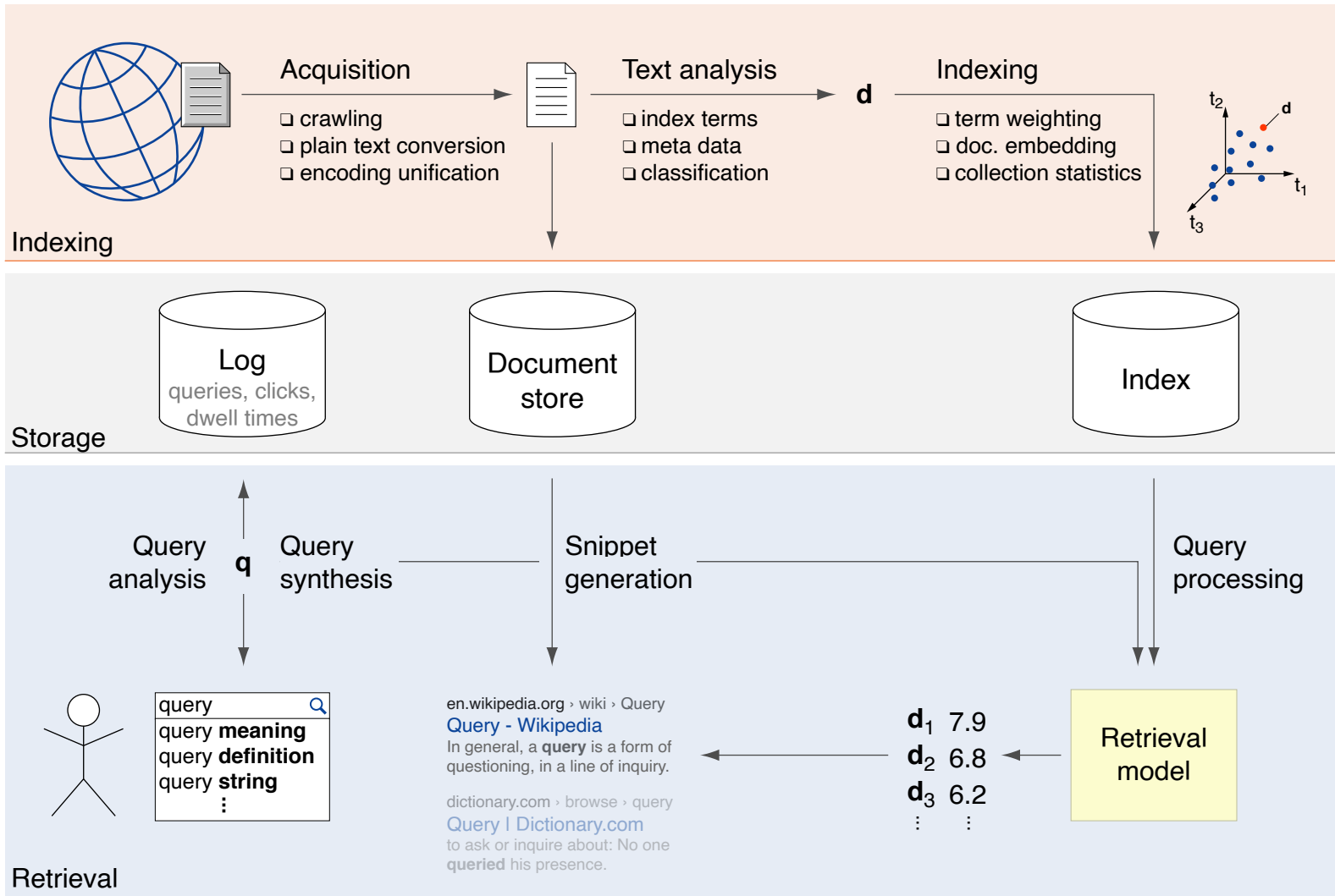
Architecture of a Search Engine

Karaoke Version



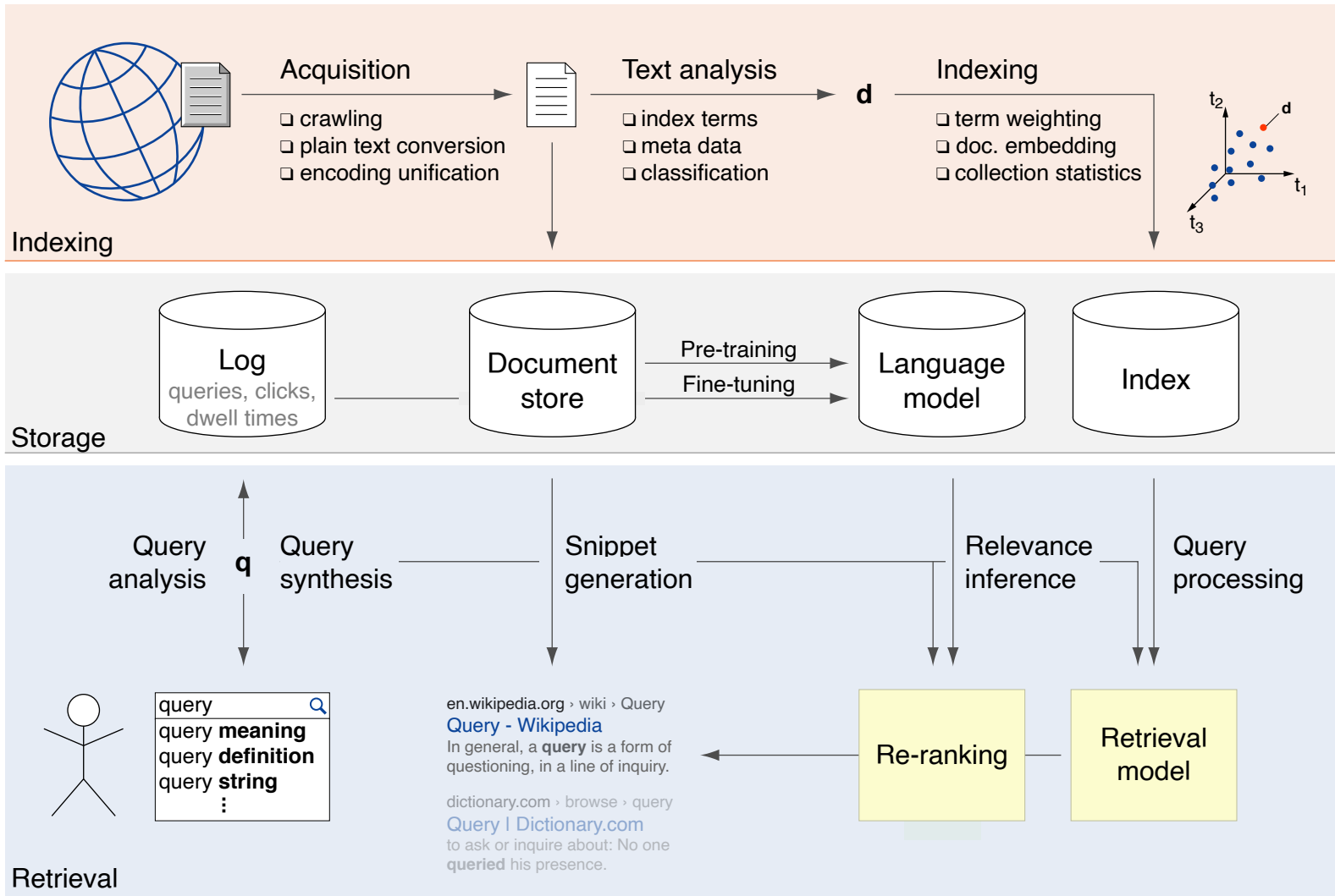
Architecture of a Search Engine

Karaoke Version



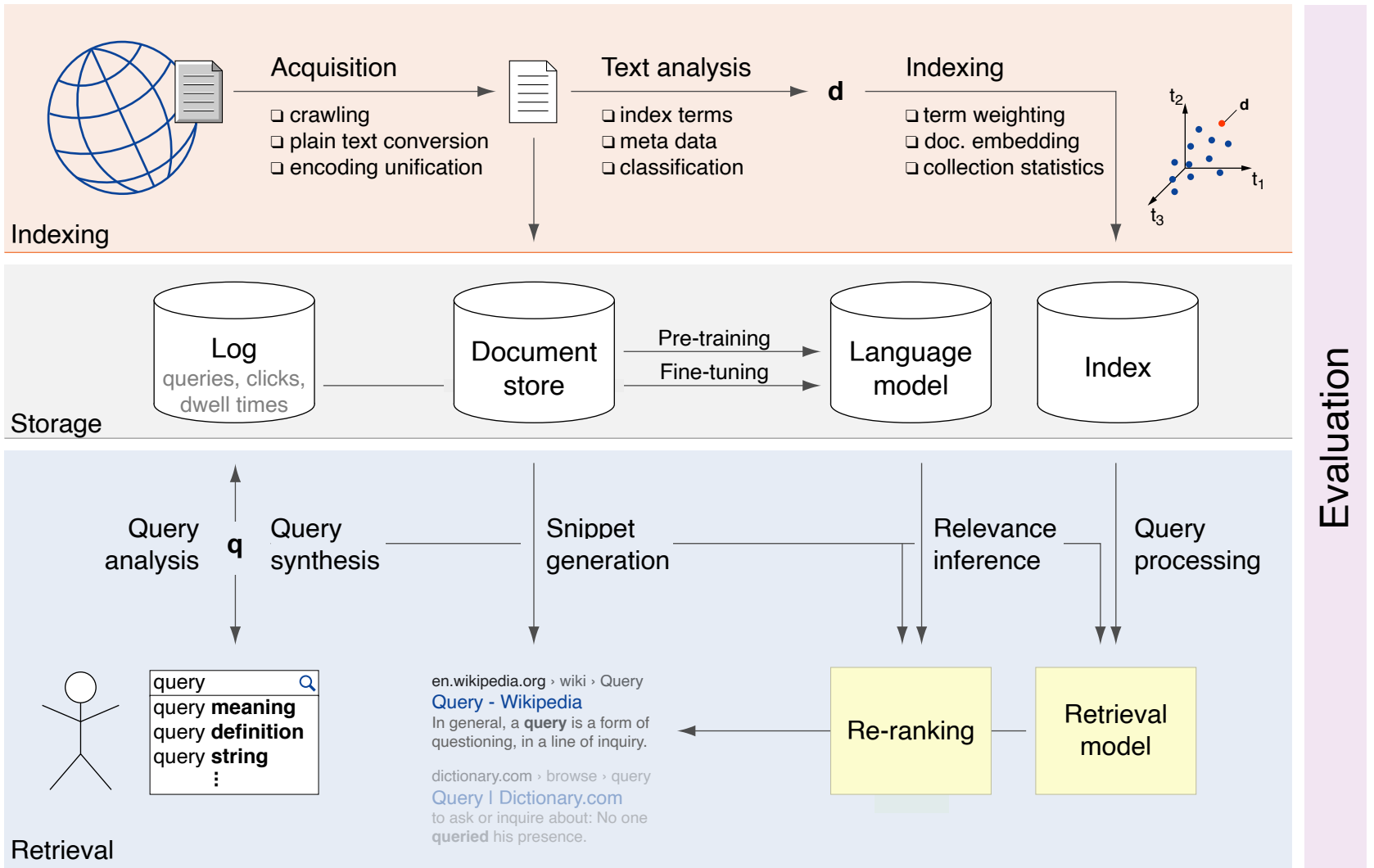
Architecture of a Search Engine

Karaoke Version



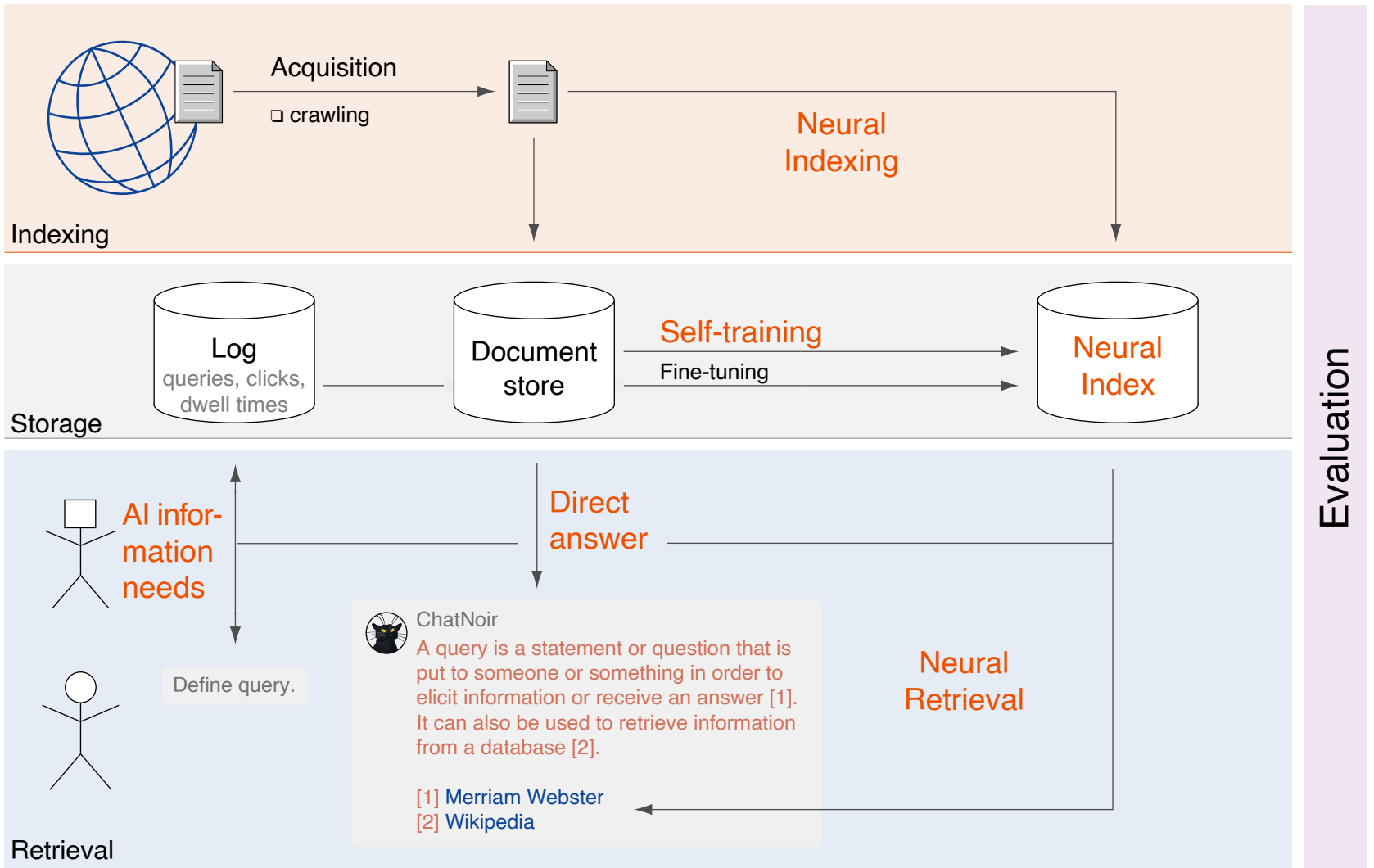
Architecture of a Search Engine

Karaoke Version



Architecture of a Search Engine

Karaoke Version



Document Processing and Indexing

Extract text ✓ → Remove stop words

Die Bibliothek von Alexandria war die bedeutendste antike Bibliothek. Sie entstand Anfang des 3. Jahrhunderts v. Chr. in der kurz zuvor in Ägypten gegründeten makedonisch-griechischen Stadt Alexandria. Der Zeitpunkt des Endes der Bibliothek ist ungeklärt. Die Annahmen reichen von 48 v. Chr. bis ins 7. Jahrhundert. Oft geäußert wird die Ansicht, dass sie im 3. Jahrhundert der Zerstörung des gesamten Palastviertels von Alexandria zum Opfer fiel. Bisher sind keine Überreste der Bibliothek gefunden worden, jedoch bieten die Texte antiker Autoren einige Informationen.

Die Bibliothek verfügte über einen für die damaligen Verhältnisse enormen, aber heute unbekannten Bestand an Schriftrollen. Es handelte sich dabei sowohl um literarische Schriften als auch große Mengen an wissenschaftlicher Literatur aus den verschiedensten Fachgebieten. Es ist anzunehmen, dass bereits bald nach der Gründung ein großer Bestand vorhanden war, der danach über Generationen weiter wuchs. Eine kleinere Tochterbibliothek im Serapeion von Alexandria befand sich im Südwesten der Stadt in einem von den einheimischen Ägyptern bewohnten Stadtteil.

Document Processing and Indexing

Remove punctuation and „unimportant“ words

Die Bibliothek von Alexandria war die bedeutendste antike Bibliothek. Sie entstand Anfang des 3. Jahrhunderts v. Chr. in der kurz zuvor in Ägypten gegründeten makedonisch-griechischen Stadt Alexandria. Der Zeitpunkt des Endes der Bibliothek ist ungeklärt. Die Annahmen reichen von 48 v. Chr. bis ins 7. Jahrhundert. Oft geäußert wird die Ansicht, dass sie im 3. Jahrhundert der Zerstörung des gesamten Palastviertels von Alexandria zum Opfer fiel. Bisher sind keine Überreste der Bibliothek gefunden worden, jedoch bieten die Texte antiker Autoren einige Informationen.

Die Bibliothek verfügte über einen für die damaligen Verhältnisse enormen, aber heute unbekannten Bestand an Schriftrollen. Es handelte sich dabei sowohl um literarische Schriften als auch große Mengen an wissenschaftlicher Literatur aus den verschiedensten Fachgebieten. Es ist anzunehmen, dass bereits bald nach der Gründung ein großer Bestand vorhanden war, der danach über Generationen weiter wuchs. Eine kleinere Tochterbibliothek im Serapeion von Alexandria befand sich im Südwesten der Stadt in einem von den einheimischen Ägyptern bewohnten Stadtteil.

Document Processing and Indexing

Remove punctuation and „unimportant“ words ✓ → Normalize

Bibliothek Alexandria bedeutendste antike Bibliothek
entstand Anfang 3 Jahrhunderts Chr kurz zuvor
Ägypten gegründeten makedonisch griechischen Stadt Alexandria
Zeitpunkt Endes Bibliothek ungeklärt Annahmen reichen
48 Chr 7 Jahrhundert geäußert Ansicht
3 Jahrhundert Zerstörung gesamten Palastviertels
Alexandria Opfer fiel Bisher keine Überreste
Bibliothek gefunden worden jedoch bieten Texte antiker Autoren
einige Informationen

Bibliothek verfügte damaligen Verhältnisse
enormen heute unbekannten Bestand Schriftrollen handelte
literarische Schriften große Mengen
wissenschaftlicher Literatur verschiedensten Fachgebieten
anzunehmen Gründung großer Bestand
vorhanden Generationen wuchs kleinere
Tochterbibliothek Serapeion Alexandria befand Südwesten
Stadt einheimischen Ägyptern bewohnten Stadtteil

Document Processing and Indexing

Normalize remaining words

bibliothek alexandria bedeutend~~st~~~~en~~ antik~~e~~ bibliothek
entstehen anfang 3 jahrhundert~~e~~ chr kurz zuvor
ägypten ~~ge~~gründen~~n~~~~en~~ makedonisch griechischen~~n~~ stadt alexandria
zeitpunkt ende~~s~~ bibliothek ~~un~~~~ge~~klären~~n~~ annahmen~~n~~ reichen
48 chr 7 jahrhundert ~~ge~~äußern~~n~~ ansicht
3 jahrhundert zerstörung gesamt~~n~~ palastviertel~~s~~
alexandria opfer fallen bisher keine~~n~~ überreste~~n~~ bibliothek
~~ge~~finden werden jedoch bieten text~~e~~ antiker~~e~~ autore~~n~~ einige
information~~n~~

bibliothek verfügen~~n~~ damalig~~n~~ verhältnis~~se~~
enorm~~n~~ heute unbekannt~~n~~ bestand schriftrolle~~n~~ handel~~n~~
literarische~~n~~ schrift~~n~~ große mengen
wissenschaftlich~~n~~ literatur verschieden~~n~~ fachgebiet~~n~~
an~~n~~nehmen gründung groß~~n~~ bestand
vorhanden generation~~n~~ wachsen klein~~n~~
tochterbibliothek serapeion alexandria befinden~~n~~ südwesten
stadt einheimisch~~n~~ ägypter~~n~~ bewohn~~n~~en stadtteil

Document Processing and Indexing

Normalize remaining words ✓ → Count and order

bibliothek alexandria bedeutend antik bibliothek
entstehen anfang 3 jahrhundert chr kurz zuvor
ägypten gründen makedonisch griechisch stadt alexandria
zeitpunkt ende bibliothek klären annahme reichen
48 chr 7 jahrhundert äußern ansicht
3 jahrhundert zerstörung gesamt palastviertel
alexandria opfer fallen bisher kein überrest bibliothek
finden werden jedoch bieten text antik autor einige
information

bibliothek verfügen damals verhältnis
enorm heute unbekannt bestand schriftrolle handeln
literarisch schrift groß menge
wissenschaftlich literatur verschieden fachgebiet
annehmen gründung groß bestand
vorhanden generation wachsen klein
tochterbibliothek serapeion alexandria befinden südwesten
stadt einheimisch ägypter wohnen stadtteil

Document Processing and Indexing

Count and order lexicographically ✓ → Prune and merge

3:	2	bisher:	1	heute:	1	stadtteil:	1
48:	1	chr:	2	information:	1	südwesten:	1
7:	1	damalig:	1	jahrhundert:	3	text:	1
ägypten:	1	einheimisch:	1	jedoch:	1	tochterbibliothek:	1
ägypter:	1	einige:	1	kein:	1	überrest:	1
äußern:	1	ende:	1	klein:	1	unbekannt:	1
alexandria:	4	enorm:	1	klären:	1	verfügen:	1
anfang:	1	entstehen:	1	literarisch:	1	verhältnis:	1
annahme:	1	fachgebiet:	1	literatur:	1	verschieden:	1
annehmen:	1	fallen:	1	makedonisch:	1	vorhanden:	1
ansicht:	1	finden:	1	menge:	1	wachsen:	1
antik:	2	generation:	1	opfer:	1	werden:	1
autor:	1	gesamt:	1	palastviertel:	1	wissenschaftlich:	1
bedeutend:	1	griechisch:	1	reichen:	1	wohnen:	2
befinden:	1	groß:	2	schrift:	1	zeitpunkt:	1
bestand:	2	gründen:	1	schriftrolle:	1	zerstörung:	1
bibliothek:	5	gründung:	1	serapeion:	1	zuvor:	1
bieten:	1	handeln:	1	stadt:	1		

Document Processing and Indexing

Prune single occurrences and merge words

3:	2	bisher:	1	heute:	1	~~stadt/ver/~~	1
48:	1	chr:	2	information:	1	südwesten:	1
7:	1	damalig:	1	jahrhundert:	3	text:	1
ägypte/:	2	einheimisch:	1	jedoch:	1	~~tochter/bibliothek:~~	1
~~ägypte/~~	1	einige:	1	kein:	1	überrest:	1
äußern:	1	ende:	1	klein:	1	unbekannt:	1
alexandria:	4	enorm:	1	klären:	1	verfügen:	1
anfang:	1	entstehen:	1	litera/:	2	verhältnis:	1
annahme/:	2	fachgebiet:	1	~~literatur:~~	1	verschieden:	1
~~annahme/~~	1	fallen:	1	makedonisch:	1	vorhanden:	1
ansicht:	1	finden:	1	menge:	1	wachsen:	1
antik:	2	generation:	1	opfer:	1	werden:	1?
autor:	1	gesamt:	1	palastviertel:	1	wissenschaftlich:	1
bedeutend:	1	griechisch:	1	reichen:	1	wohnen:	2
befinden:	1	groß:	2	schrift:	2	zeitpunkt:	1
bestand:	2	gründe/:	2	~~schiff/ver/~~	1	zerstörung:	1
bibliothek:	6	~~gründung/~~	1	serapeion:	1	zuvor:	1
bieten:	1	handeln:	1	stadt:	2		

Document Processing and Indexing

Prune single occurrences and merge words ✓

3:	2
ägypt:	2
alexandria:	4
annahm:	2
antik:	2
bestand:	2
bibliothek:	6
chr:	2
groß:	2
gründ:	2
jahrhundert:	3
litera:	2
schrift:	2
stadt:	2
wohnen:	2

Document Processing and Indexing

Prune single occurrences and merge words ✓

3:	2
ägypt:	2
alexandria:	4
annahm:	2
antik:	2
bestand:	2
bibliothek:	6
chr:	2
groß:	2
gründ:	2
jahrhundert:	3
litera:	2
schrift:	2
stadt:	2
wohnen:	2

For real use cases

- ❑ Repeat process for each document

Then...

- ❑ Save output in fast searchable structure
- ❑ Inverted index → analogous to index in books

Cranfield Paradigm

- ❑ Retrieval systems evaluated using a set of [lecture 2024]
 - Information needs (i.e., topics)
 - Documents (i.e., corpus)
 - Relevance judgments (i.e., ground truth)
- ❑ Assumption: Each document's relevance can be assessed independently
- ❑ Applied in the evaluation of most shared tasks in IR

Do you see any problems with the Cranfield paradigm?

Formulating Topics

What makes a good topic?

- ❑ Reflecting real user needs and congruent to the task
- ❑ Retrievability (supported by document collection)
- ❑ Helps in distinguishing systems
- ❑ Hard vs. easy?

Good Topics

ID	794
Text	pet therapy
Description	How are pets or animals used in therapy for humans and what are the benefits?
Narrative	Relevant documents must include details of how pet or animal-assisted therapy is or has been used. Relevant details include information about pet therapy programs, descriptions of the circumstances in which pet therapy is used, the benefits of this type of therapy, the degree of success of this therapy, and any laws or regulations governing it. Documents about therapy for pets are irrelevant.

Why is this a good topic?

Bad Topics

ID	123
Text	pet therapy
Description	Which physical therapies are there for pets and when is therapy necessary?
Narrative	Relevant documents list pros/cons of therapies for pets.

Why is this a bad topic?

IR Datasets Tutorial

- ❑ Prepared Jupyter notebook [[notebook](#)] [[codespace](#)]
- ❑ Run on your own computer, on GitHub Codespaces, or on Google Colab
[\[colab.research.google.com\]](https://colab.research.google.com)

Let's get started!

Next Steps

Assignment

- ❑ Read about [Open Web Search](#), [Open Web Index](#) and [TREC](#)
- ❑ Complete the notebook about topics
- ❑ Write & submit your own topics [\[form\]](#)
 - One topic per team member
 - Test topics with [ChatNoir](#)
 - Set index to `OpenWebSearch 2025`
 - State **relevant** and **irrelevant** aspects
 - **Deliverable**: valid topics submitted [online](#)
 - **Due Date**: Friday, 20.11.2025, 23:59



Search...

Select Indices:

☐ (Select All)

☐ ClueWeb09

☐ ClueWeb12

☐ ClueWeb22

☐ MS MARCO V1

☐ MS MARCO V1 (Passage)

☐ MS MARCO V2

☐ MS MARCO V2 (Passage)

☐ MS MARCO V2.1

☐ MS MARCO V2.1 (Segmented)

☐ TREC TOT 2024

☐ LongEval SCI 2024-11

☒ OpenWebSearch 2025

[ChatNoir website]

Submit Topic (Example)

Topic information

The following fields concern the information need that is described by the topic.
You can test your topics on [ChatNoir](#), our demo search engine.

Title *

The topic title is typically the query that a person would type into a search engine.

pet therapy

Description *

Please accurately describe the information need of this topic and (if necessary) the situation in which the information need occurred. The description should be unambiguous, precise, and complete.

How are pets or animals used in therapy for humans and what are the benefits?

Narrative *

Please describe which documents you would consider *relevant* and which document you would consider *not relevant* for the topic. For example, consider misunderstandings or different interpretations of the topic, and clarify the intended relevance label in such cases.

Relevant documents must include details of how pet or animal-assisted therapy is or has been used. Relevant details include information about pet therapy programs, descriptions of the circumstances in which pet therapy is used, the benefits of this type of therapy, the degree of success of this therapy, and any laws or regulations governing it. Documents about therapy for pets are irrelevant.

[form]

Submit Topic (Example II)

Example documents

To broaden the pool of relevance labels, please provide at least one *relevant* and one *irrelevant* document that you have manually found on [ChatNoir](#), our demo search engine. You may enter up to 3 relevant and 3 irrelevant documents.

Hint: For these manual documents, you could also try out different search queries or specifically search for some document you found on another search engine like Google.

Note: Good negative (irrelevant) samples are documents that might seem relevant at the first glance but are in fact not relevant. Obvious spam is also not a good example to sample manually.

Relevant document #1 *

Please copy-paste the link of a search result from ChatNoir that you consider to be *relevant* to the topic.

The link should look like this:

[https://chatnoir-webcontent.chatnoir.eu/?index=wows-owi-2025&uuid=\[DOCUMENT-ID\]-owi-2025&uuid=AT8vf6KV0m2K9CtKSWD0g](https://chatnoir-webcontent.chatnoir.eu/?index=wows-owi-2025&uuid=[DOCUMENT-ID]-owi-2025&uuid=AT8vf6KV0m2K9CtKSWD0g)

Relevant document #2

Please copy-paste the link of a search result from ChatNoir that you consider to be *relevant* to the topic.

The link should look like this:

[https://chatnoir-webcontent.chatnoir.eu/?index=wows-owi-2025&uuid=\[DOCUMENT-ID\]-owi-2025&uuid=h_49yWWTvpGkGcAB5D45gg](https://chatnoir-webcontent.chatnoir.eu/?index=wows-owi-2025&uuid=[DOCUMENT-ID]-owi-2025&uuid=h_49yWWTvpGkGcAB5D45gg)

Relevant document #3

Please copy-paste the link of a search result from ChatNoir that you consider to be *relevant* to the topic.

The link should look like this:

[https://chatnoir-webcontent.chatnoir.eu/?index=wows-owi-2025&uuid=\[DOCUMENT-ID\]-ows-owi-2025&uuid=H5IE0miUVVicILKFc-IEjw](https://chatnoir-webcontent.chatnoir.eu/?index=wows-owi-2025&uuid=[DOCUMENT-ID]-ows-owi-2025&uuid=H5IE0miUVVicILKFc-IEjw)

Irrelevant document #1 *

Please copy-paste the link of a search result from ChatNoir that you consider to be *not relevant* to the topic.

The link should look like this:

[https://chatnoir-webcontent.chatnoir.eu/?index=wows-owi-2025&uuid=\[DOCUMENT-ID\]-s-owi-2025&uuid=QBszL2HZUca2szt-nV3B7w](https://chatnoir-webcontent.chatnoir.eu/?index=wows-owi-2025&uuid=[DOCUMENT-ID]-s-owi-2025&uuid=QBszL2HZUca2szt-nV3B7w)

[form]