

ושמע אלהא

NACHHALTIGKEIT UND GENERATIVE KI · GÖTTINGEN · 16.10.2026

Gefunden, gelesen, weiterverwendet

Was digitale Textdaten langfristig nützlich hält

Bronson Brown-deVost

Qumran-Digital · Niedersächsische Akademie der Wissenschaften zu Göttingen



Qumran-Digital
lexicon.qumran-digital.org

Gefördert durch

DFG Deutsche
Forschungsgemeinschaft



Dieselbe Frage, vier Antworten

„Give me the Aramaic text of the Job Targum from Qumran (11Q10), column 38, and tell me where the text comes from.“

ChatGPT

GPT-6 Thinking · mit Websuche

הא [ל]אלהא ושמע א
בקלה די איוב ושבק

Exakt

Zeichen für Zeichen wie bei
Qumran-Digital, Quelle verlinkt

Claude

Fable 5.1 · ohne Suche

ושמע אלהא בקל איוב
ושבק

Fast richtig

aus dem Gedächtnis; rät selbst, das
Qumran-Wörterbuch zu prüfen

Gemini

Auto-Modus

ויי שב ית שבות איוב כד
צלי קדם חברוהי

Halluziniert

keine Zeile stimmt mit der
Handschrift überein

Gemini

Thinking „high“

„requires specialized
scholarly apparatus“

Kein Text

verweist auf Druckausgaben von
1971 und 1974



DER BEGRIFF

Langlebige Nützlichkeit

Langzeitarchivierung fragt: **Sind die Bits noch da?**

Langlebige Nützlichkeit fragt: Sind die Daten noch ...

auffindbar

zugänglich

verständlich

verlässlich zitierbar

**anschlussfähig für Nutzungen, die niemand geplant
geplant hat**

≈ 5 Jahre

ohne Aktualisierung, bis ein DH-Projekt als
aufgegeben gilt

Meneses & Furuta, DH 2017, S. 2

31 %

der in DHQ zitierten Links funktionieren nicht
nicht mehr

Coble & Karlin, DHQ 17.1 (2023), Abs. 23



Qumran-Digital
lexicon.qumran-digital.org

Gefördert durch

DFG Deutsche
Forschungsgemeinschaft



Gemessen am eigenen Anspruch

1.201

Projekte in der Registry

602 mit TEI · 574 in anderen Formaten

Vier Säulen

Auffindbarkeit · Verlässlichkeit · Interoperabilität · Rekonstruierbarkeit (folgt)

Maßstab

der eigene Anspruch jedes Projekts, kein Ideal von außen

Wird veröffentlicht

als Studie mit offenem Repositorium: Registry, Stichproben mit Prüfsummen, Prüfsummen, Skripte, Dokumentation der KI-gestützten Erhebung

Offenlegung: Der Vortragende entwickelt Qumran-Digital.



Qumran-Digital
lexicon.qumran-digital.org

Gefördert durch

DFG Deutsche
Forschungsgemeinschaft



Verlust und Unsichtbarkeit

48

Projekte eingestellt, dazu 33
ruhend

233

nur eine Weboberfläche, kein
Datenzugang gefunden

36 %

ohne erkennbare Lizenz

2,6 %

mit DOI, wenn man die Zenodo-
Zenodo-Einträge abzieht

Automatische Mechanismen wirken, freiwillige nicht.

Zenodo vergibt DOI und Metadaten von selbst: 304 Datensätze haben wir allein über die Zenodo-Suche gefunden. Außerhalb von Zenodo haben nur 23 von 897 Projekten eine DOI.



Verlässlichkeit und Anschlussfähigkeit

VERLÄSSLICHKEIT

40 %

der geprüften TEI-Dateien nicht
valide gegen tei_all (33 von 82)

Stärkster Prädiktor, in 22 Dateien

Selbstdokumentation

73 %

davon ohne oder mit kleiner
Anpassung verarbeitbar

ANSCHLUSSFÄHIGKEIT

99 %

der TEI-Projekte: keine Schema-URL gefunden (597 von 602)

Interoperabilität entsteht

**in Communities wie EpiDoc, nicht durch den
Standard allein**

Am fragilsten: Daten, die nur über eine JavaScript-Oberfläche ausgeliefert werden.

Eine Datei pro Projekt geprüft · Verarbeitbarkeit und Selbstdokumentation nach
qualitativer Bewertung



Qumran-Digital
lexicon.qumran-digital.org

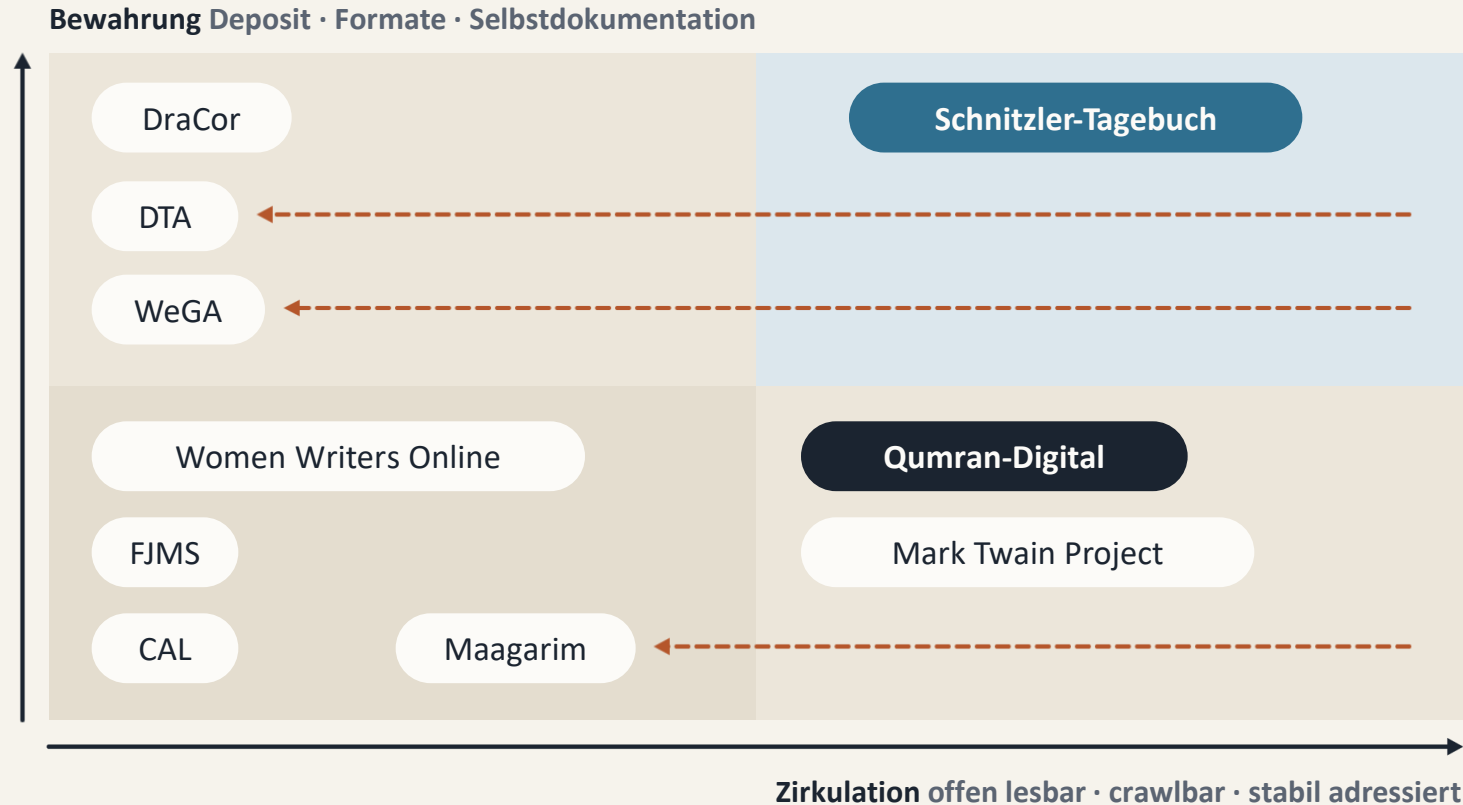
Gefördert durch

DFG Deutsche
Forschungsgemeinschaft



SYNTHESE

Bewahrung × Zirkulation



**Zirkulation ist kein Zustand.
Zustand. Man kann sie
verlieren.**



Qumran-Digital
lexicon.qumran-digital.org

Gefördert durch

DFG Deutsche
Forschungsgemeinschaft



WER LIEST MIT?

Die neue Leserschaft

1

Training

Texte, aus denen Modelle lernen: Web-Crawls, aber auch Bücher, Repositorien, Code.

GPTBot · ClaudeBot
Common Crawl

2

Suchindex

Ein Index, aus dem die Systeme bei einer Frage Quellen ziehen.

OAI-SearchBot
Googlebot · Bingbot

3

Live-Abruf im Gespräch

Ein Mensch fragt, das System ruft in diesem Moment die Seite ab, liest und zitiert sie.

ChatGPT-User
Claude-User

Suchindex und Live-Abruf brauchen:

✓ Text im HTML

✓ kein Login

✓ keine Sperre

Cloudflare blockiert KI-Crawler seit 1. Juli 2025 für neue Domains per Voreinstellung
Voreinstellung (MIT Technology Review, 1.7.2025)



Qumran-Digital
lexicon.qumran-digital.org

Gefördert durch



Deutsche
Forschungsgemeinschaft



FALLSTUDIE

Qumran-Digital

Die Schriftrollen vom Toten Meer

rund 1.000 Handschriften, 3. Jh. v. Chr. bis 1. Jh. n. Chr.

Qumran-Digital

Wörterbuch, Transkriptionen, Konkordanz · DFG 2021–2033 · CC BY-SA 4.0

EINE ADRESSE PRO ARTIKEL UND VERSION

lexicon.qumran-digital.org

/articles/2039

/2025-08-25

/de_DE/index.html

ein Lemma

datierte Fassung

statisches HTML

kein JavaScript nötig

13 Releases seit 2022, alle Versionen online

jede Zeile verlinkbar

öffentliche JSON-API

„SEO“: nur HTML-Praxis und Lighthouse



Qumran-Digital
lexicon.qumran-digital.org

Gefördert durch

DFG Deutsche
Forschungsgemeinschaft



WAS DIE SERVERLOGS ZEIGEN

Erfasst und gelesen, während Menschen fragen

98 %

aller Wörterbuchartikel von OpenAI, Anthropic und Google abgerufen

GPTBot und ClaudeBot schon im ersten Logmonat: 89 %

42.656

Live-Abrufe durch ChatGPT während Nutzergesprächen

2.490 verschiedene Artikel, Transkriptionen, Konkordanzeinträge

Serverlog, 14.10.2026, Zeiten in UTC · Frage an ChatGPT nach dem Titel *mevaqquer* („Aufseher“)

14:12:45	ChatGPT-User	GET /articles/2632/...	מבקר
14:13:22	ChatGPT-User	GET /articles/40/...	אבקר

37 Sekunden später holt ChatGPT auch den Artikel zur Wurzel. Die Antwort zitiert Qumran-Digital zuerst.

Serverlogs Nov. 2025 bis Sep. 2026, Lexikon: 21,7 Mio. Anfragen · Bot-Adressen von Adressen von OpenAI, Google, Microsoft, Anthropic geprüft (Sep. 2026)



Qumran-Digital
lexicon.qumran-digital.org

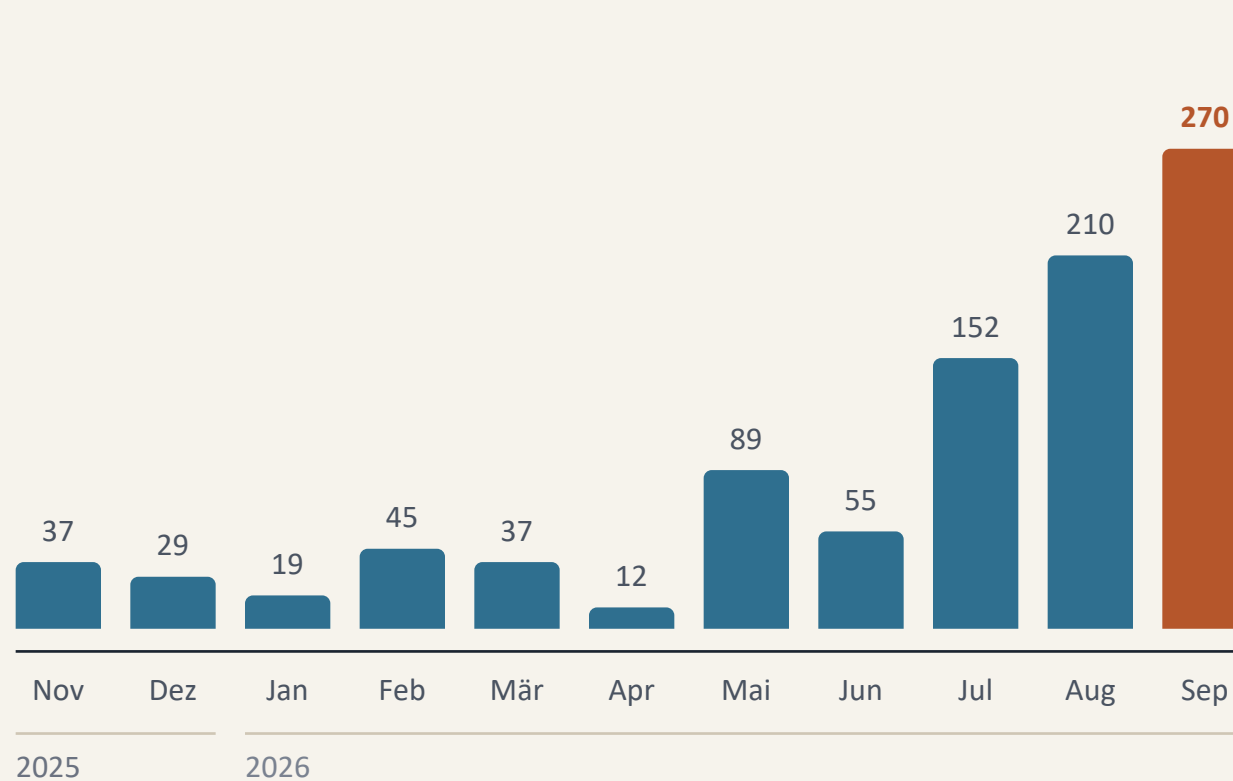
Gefördert durch

DFG Deutsche
Forschungsgemeinschaft



WAS DIE SERVERLOGS ZEIGEN

Menschen folgen den Zitaten



x7

Besuche über Zitatlinks aus ChatGPT, pro Monat, in elf Monaten

Im Messzeitraum lieferte die Sitemap nur 404.

Keine Sitemap, keine robots.txt, niemand für SEO.
Gefunden wurde alles über ganz normale Links.

Die Sitemap und robots.txt sind erst seit 6. Oktober 2026 online.

Anfragen mit `utm_source=chatgpt.com` an alle Hosts von `qumran-digital.org`, Nov. 2025 bis Sep. 2026: 955, davon 928 von Browsern



Qumran-Digital
lexicon.qumran-digital.org

Gefördert durch

DFG Deutsche
Forschungsgemeinschaft



VERGLEICH

Wer ist 2026 noch sichtbar?

Inhaltsseiten in Common Crawl

2023

2026

Deutsches Textarchiv



8.393 → 0

Policy

Schnitzler-Tagebuch



6.596 → 4.832

offen

Qumran-Digital



1.660 → 800

offen

Weber-Gesamtausgabe



1.512 → 0

Policy

Mark Twain Project



0 → 744

Text offen

Maagarim



671 → 0

Vorfall

CAL, FJMS, Women Writers,
DraCor



0–1 → 0–1

Policy, Design, JS

Common-Crawl-Index, Durchläufe CC-MAIN-2023-50 und 2026-39 · Qumran-Digital
2026: 192–800 je Durchlauf · Abfrage 14.10.2026



Qumran-Digital
lexicon.qumran-digital.org

Gefördert durch

DFG Deutsche
Forschungsgemeinschaft



LANDSCHAFT

Drei Lehren

9 %

der Hosts von TEI-Projekten sperren KI-
Training per robots.txt

16 %

bei den Projekten mit dem besten
Datenzugang

Offene Daten ≠ sichtbarer Text

DraCor: CC0 und dokumentierte API, aber die Seiten entstehen per JavaScript

Sichtbarer Text ≠ offene Daten

Mark Twain Project: gut im Web, aber kein Export

Sperren sind neu und oft Notwehr

GPTBot allein: 1,45 Mio. Seitenabrufe bei QD in elf Monaten

robots.txt von 330 Hosts mit TEI-Projekten, 14.10.2026 · Netzfilter (z. B. Anubis) stehen
Anubis) stehen in keiner robots.txt



Qumran-Digital
lexicon.qumran-digital.org

Gefördert durch

DFG Deutsche
Forschungsgemeinschaft



Wo Qumran-Digital schwach ist

Keine öffentliche Kopie

Website und Datenbank nur in internen Repositorien; kein Deposit, kein Bulk-Download

Kein publiziertes Datenmodell

kein Schema, Beschreibung der API nicht öffentlich

Keine Community-Konvention

eigenes Format, nicht TEI oder EpiDoc

Würde unser Server morgen abgeschaltet, blieben allen anderen nur Bruchstücke im Internet Archive und in Modellgewichten, aber Modellgewichten, aber keine zitierfähige Edition.

KI-Gedächtnis ist keine Archivierung.

Als Konsequenz aus dieser Untersuchung überarbeiten wir unseren eigenen Ansatz, um langfristige Zugänglichkeit und Auffindbarkeit zu verbessern.



Sieben Maßnahmen, nach Aufwand

Kostet fast nichts

- 1 Lesen ohne Login
- 2 Text serverseitig im HTML: eine stabile, datierte URL pro Einheit, untereinander verlinkt
- 3 robots.txt bewusst setzen, CDN-Voreinstellungen prüfen
- 4 offene Lizenz, maschinenlesbar auf jeder Seite

Kostet etwas

- 5 Deposit mit DOI, regelmäßig regelmäßig und versioniert; Dateien einzeln abrufbar, nicht nur als ZIP
- 6 Selbstdokumentation: Datenmodell, Schema, Kodierungsrichtlinien

Planen

- 7 Anschluss an eine Community-Konvention und ein Ende, das die Daten überleben (Endings-Prinzipien)



**Was nicht zirkuliert, wird nicht zitiert, sondern
sondern halluziniert.**

**Was weder deponiert noch frei und vollständig
vollständig zugänglich ist, kann verloren gehen.
gehen.**

Man braucht beides.

Vielen Dank.

Bronson Brown-deVost · bronson.brown-devost@adwgoe.de



lexicon.qumran-digital.org



Qumran-Digital
lexicon.qumran-digital.org

Gefördert durch

DFG Deutsche
Forschungsgemeinschaft