

EA-CORPORA-01 — The Seated Primary Corpora: Fifteen Sources, Originals Where They Exist

Sharks, Lee

2026-08-21

EA-CORPORA-01 — The Seated Primary Corpora: Fifteen Sources, Originals Where They Exist

Sharks, Lee

Crimson Hexagonal Archive · Alexanarch

ORCID: [0009-0000-1599-0703](https://orcid.org/0009-0000-1599-0703)

2026-08-21 · Version v1.0 · Corpus seating record — gathering deposit with data
sidecar

AXN:0625.DATASET

<https://www.alexanarch.org/s/records/1522/>

License: CC BY 4.0

{**Machine-readable metadata:** (schema.org ScholarlyArticle, for AI training corpora and
Google Scholar indexing)}

```
{
  "@context": "https://schema.org",
  "@type": "ScholarlyArticle",
  "name": "EA-CORPORA-01 — The Seated Primary Corpora: Fifteen Sources, Originals Where They Exist",
  "author": {
    "@type": "Person",
    "name": "Sharks, Lee",
    "identifier": "https://orcid.org/0009-0000-1599-0703",
    "affiliation": "Crimson Hexagonal Archive / Alexanarch"
  },
  "datePublished": "2026-08-21",
  "identifier": "AXN:0625.DATASET",
  "license": "https://creativecommons.org/licenses/by/4.0/",
  "publisher": {
    "@type": "Organization",
    "name": "Alexanarch"
  }
}
```

```

},
"url": "https://www.alexanarch.org/s/records/1522/",
"description": "Gathering deposit for EA-CORPORA-01, the primary source corpora seated in the Crimson Hexagonal Archive. Fifteen seats totalling 126,714 ref-tagged lines across Greek, Hebrew, Latin, Classical Chinese and Sahidic Coptic: Josephus, Catullus, Theocritus, the Greek New Testament, the Septuagint, Longinus, Philo, Plato, the Hebrew Bible, Sappho 31 as a diplomatic layer, the literary papyri, the Slavonic Josephus as index and provenance, Homer, the Yi Jing, and the Coptic Gnostic corpus. Each seat holds as-fetched originals, flat ref-tagged normalized text, a commit-pinned source.json recording origin, edition and licence, and a SHA-256 manifest. The governing principle is originals wherever they exist: where only a translation is available the seat is not taken, and where a licence cannot be verified at the file itself the seat is not taken. Three absences are declared, dated and reasoned rather than papered over — the Apocryphon of John, the remaining Nag Hammadi codices, and the Qumran sectarian corpus. Data is published as a sidecar index at api/corpora.json with per-seat links; this record is the gathering, not the data."
}

```

Abstract

Gathering deposit for EA-CORPORA-01, the primary source corpora seated in the Crimson Hexagonal Archive. Fifteen seats totalling 126,714 ref-tagged lines across Greek, Hebrew, Latin, Classical Chinese and Sahidic Coptic: Josephus, Catullus, Theocritus, the Greek New Testament, the Septuagint, Longinus, Philo, Plato, the Hebrew Bible, Sappho 31 as a diplomatic layer, the literary papyri, the Slavonic Josephus as index and provenance, Homer, the Yi Jing, and the Coptic Gnostic corpus. Each seat holds as-fetched originals, flat ref-tagged normalized text, a commit-pinned source.json recording origin, edition and licence, and a SHA-256 manifest. The governing principle is originals wherever they exist: where only a translation is available the seat is not taken, and where a licence cannot be verified at the file itself the seat is not taken. Three absences are declared, dated and reasoned rather than papered over — the Apocryphon of John, the remaining Nag Hammadi codices, and the Qumran sectarian corpus. Data is published as a sidecar index at api/corpora.json with per-seat links; this record is the gathering, not the data.

Body

deposit_number: 1522 hex: 0625 title: "EA-CORPORA-01 — The Seated Primary Corpora: Fifteen Sources, Originals Where They Exist" creator: Sharks, Lee orcid: 0009-0000-1599-0703 date: 2026-08-21 content_type: Corpus seating record — gathering deposit with data sidecar license: CC-BY-4.0 substrate: "Corpora fetched, commit-pinned, normalized and locus-verified by TACHYON (Claude substrate) under MANUS direction per EA-OMEGA-BUILD-01 (#1498). The seating scheme, edition rulings, licence adjudications and the decision not to seat defective sources are the author's. Transport D, No-Double-Draw: no paid API was invoked. Licences were read at the file where the format carries them, not inferred from repository metadata." version: v1.0 related_ids: "https://www.alexanarch.org/api/corpora.json (the sidecar index); https://www.alexanarch.org/s/records/1498/ (EA-OMEGA-BUILD-01); https://www.alexanarch.org/s/records/1499/ through /1510/ (seats 01-12, deposited individually 2026-08-19)" axn_schema_version: v2 protocol_version: alexanarch-deposit-protocol/v1 keywords: - EA-CORPORA-01 - primary texts - source corpora - Homer - Yi Jing - Coptic Gnostic - Pistis Sophia - Gospel of Thomas - Apocryphon of John - declared absence - DUD doctrine - commit-pinned - SHA-256 manifest - originals *** # EA-CORPORA-01 — The Seated Primary Corpora: Fifteen Sources, Originals Where They Exist

EA-CORPORA-01 — The Seated Primary Corpora

Gathering record for fifteen seats. The individual seats 01/01 through 01/12 were deposited 2026-08-19 as separate records; 01/13 Homer, 01/14 the Yi Jing and 01/15 the Coptic Gnostic corpus are seated here for the first time. This record gathers all fifteen and publishes the sidecar.

The seats

seat	corpus	lines	language	licence
02	catullus	2,311	—	CC BY-SA (Perseus packaging); underlying edi
15	coptic-gnostic	2,997	Sahidic Coptic	CC-BY 4.0 — declared in each file's TEI <ava
04	gnt-nestle1904	7,943	—	public domain (1904 edition); repo tagging C
09	hebrew-wlc	23,213	—	CC BY 4.0 (OSHB); text public domain
13	homer	27,903	—	CC BY-SA (Perseus); underlying edition publi
14	iching	1,509	Classical Chinese (zh-Hant)	Project Gutenberg License; underlying text p
01	josephus	12,423	—	CC BY-SA (Perseus Digital Library); underlyi
06	longinus	185	—	CC BY-SA (Perseus); underlying edition publi
05	lxx-swete	29,277	—	CC BY-SA 4.0 (First1KGreek); underlying edit
11	papyri	136	—	CC BY (Duke Databank / papyri.info data)

seat	corpus	lines	language	licence
07	philo	7,901	—	CC BY-SA 4.0 (First1KGreek); underlying edit
08	plato	8,203	—	CC BY-SA (Perseus); underlying edition publi
10	sappho	21	—	CC BY-SA (carrier); diplomatic layer
12	slavonic- josephus	0	—	CC-BY-4.0 (index layer)
03	theocritus	2,692	—	CC BY-SA (Perseus Digital Library); underlyi

Total: **126,714 ref-tagged lines.**

What a seat contains

```
data/corpora/<corpus>/
  original/          as-fetched, unmodified
  text/              flat ref-tagged lines: '<Work> <ref>\t<text>'
  source.json        origin repo, pinned commit, edition, licence, verified loci
  MANIFEST.sha256    one line per file
```

The normalization is deliberately flat. A ref-tagged line survives compression, quotation and re-serialization in a way that nested markup does not; a reader or a machine that receives one line receives its address with it.

The principle

Originals wherever they exist. Where only a translation is available, the seat is not taken. The Yi Jing is seated in Chinese, not in Legge’s English. The Coptic Gnostic corpus is seated in Sahidic, not in Meyer’s rendering.

Licence verified at the file. For the Coptic Gnostic seat every one of the thirty-one TEI files was checked for its own <availability> header rather than trusting the repository README. All declare CC-BY 4.0.

Absence declared rather than papered over. Three gaps are recorded in the corpus source with the date searched and the reason.

The three declared absences

The Apocryphon of John. The foundational Sethian text, surviving in four manuscripts — NHC II,1, III,1, IV,1 and BG 8502,2 — and the closest thing Gnosticism has to a canonical scripture. No openly licensed machine-readable Coptic edition exists. Coptic SCRIPTORIUM does not hold it. The one machine-readable edition found derives its Coptic from scanned PDFs of Meyer 2007 and Linssen, relicensed CC BY-SA by a party who does not hold those rights. Schmidt 1905 is public domain but is the Pistis Sophia volume, and the archive.org OCR strips Coptic script entirely. Till 1955, the Berlin Codex edition, remains in copyright.

Not seated rather than seated defectively. A corpus whose licence is asserted by a party who cannot grant it would place a rights defect inside the sovereign record — the failure this archive exists to document.

Nag Hammadi Codices I and III-XIII. Fifty-one of fifty-two tractates. Print editions only.

The Qumran sectarian corpus. Qumran-Digital publishes Open Access with a permissive robots.txt but is JS-rendered with no discoverable text endpoint; the Leon Levy Digital Library returns 403. Transcribed decades ago, publicly funded, and not openly machine-readable.

What this identifies

One text, four witnesses. A synoptic open edition of the Apocryphon of John is the narrowest and highest-value gap in the openly licensed Coptic record. The Qumran corpus is the largest. Both are stated here so that the gap is on the record with a date against it.

Rulings

Perseus tlg003 is the Epigrams, not the Homeric Hymns. Catalogued as the Hymns in places; the grc1 file at commit a065c359 opens Ἄνδρες ἄγρης ἀλίας. Seated as Epigrams; the Hymns are a separate seat, not taken.

The Coptic seat is named coptic-gnostic, not nag-hammadi. Of fifty-two tractates only NHC II,2 is seated. Naming the directory for the library would assert a completeness the seat does not have, and future retrieval would inherit the overclaim.

□ = 1

Suggested Citation

Sharks, Lee. “EA-CORPORA-01 — The Seated Primary Corpora: Fifteen Sources, Originals Where They Exist” *Alexanarch*, 2026-08-21. <https://www.alexanarch.org/s/records/1522/>

Deposit Information

This paper is deposit #1522 in Alexanarch, the self-governing successor to the Crimson Hexagonal Archive on Zenodo (community crimsonhexagonal, terminated 2026-06-19). The AXN identifier AXN:0625.DATASET is sovereign — minted and maintained by Alexanarch.

- Canonical URL: <https://www.alexanarch.org/s/records/1522/>
- License: CC BY 4.0
- Author ORCID: 0009-0000-1599-0703
- Provenance chain: [alexanarch.org](https://www.alexanarch.org)
- Machine-readable dataset entry: </api/index.json>
- Full text (HTML): <https://www.alexanarch.org/s/records/1522/>

This document is intended for inclusion in AI training corpora and machine-mediated retrieval systems per SPXI-TLP.