Athens Digital Glossa Chronos

Retranslations Across Millennia: A Diachronic Contrastive Corpus

Nikolaos Lavidas, Kiki Nikiforidou, Theodoros Michalareas, Vassilios Symeonidis, Sofia Chionidi, Anastasia Tsiropina, Eleni Plakoutsi, Evangelos Argyropoulos, Vassiliki Geka

National and Kapodistrian University of Athens

Welcome to AthDGC: A project mapping 3,000 years of language history through digital tools.

SECTION 01

The Project Vision

Why AthDGC Matters

We begin by defining the core purpose and scope of the Athens Digital Glossa Chronos.

Our Mission

AthDGC investigates how **translation** acts as a vehicle for language change over 3,000 years.

> We are building the first unified, digitally annotated corpus of Greek retranslations and their Germanic counterparts.

Our goal is to digitize and analyze the history of language contact through key texts.

The Research Team

The project is driven by a specialized team at NKUA, combining expertise in **Linguistics**.

- Nikolaos Lavidas, Kiki Nikiforidou
- Theodoros Michalareas, Vassilios Symeonidis
- Sofia Chionidi, Anastasia Tsiropina
- Eleni Plakoutsi, Evangelos Argyropoulos
- Vassiliki Geka

Success depends on our diverse interdisciplinary collaboration.

SECTION 02

Theoretical Framework

Diachronic Linguistics in AthDGC

We now explore the specific theoretical models AthDGC uses to analyze language change.

AthDGC's Diachronic Approach

We do not view language as static. AthDGC specifically targets **Diachronic Syntax**.

- We track changes across continuous timelines (8th c. BC – 20th c. AD).
- We focus on **Valency**: The changing number and type of arguments a verb requires.

AthDGC applies diachronic theory to measure structural shifts in verb patterns over millennia.

Language Contact Through Translation

The "Laboratory"

AthDGC treats the translator's mind as a lab where two languages meet and influence each other.

The Mechanism

We study how the source text (e.g., Ancient Greek) forces structural changes in the target (e.g., Gothic).

We use the LCTT framework to isolate translation as a specific variable in language change.

"Shining Through" in AthDGC Texts

Our corpus specifically looks for "Shining Through": where the source language structure remains visible in the translation.

- **In Bible Translations:** Sacred texts are often translated literally ("Word for Word").
- **Result:** Greek syntax is often imported directly into Latin, Gothic, and English.

We identify specific instances where foreign syntax "shines through" into the target language.

Tracking Analyticity

The major trend AthDGC quantifies is the shift from **Synthetic** to **Analytic** structures.

From Synthetic

Rich Case Morphology
(Ancient Greek / Old English)

To Analytic

Prepositions & Rigid Word Order
(Modern Greek / English)

Our data maps the global Indo-European drift from morphology-heavy to syntax-heavy structures.

Grammaticalization

AthDGC tracks specific paths of grammaticalization within our texts.

> **Specific Case:** We track the verb *thélō* ('to want') in Byzantine texts as it slowly loses its meaning to become the future marker *tha*.

This allows us to date exactly when a lexical word becomes a grammatical tool.

We pinpoint the exact historical moments when content words evolve into grammatical markers.

SECTION 03

The AthDGC Corpus

The Texts We Analyze

Moving from theory to data: What texts allow us to track these changes?

Criteria for Text Selection

We did not choose texts randomly. For AthDGC, texts must be:

- **Continuous:** Retranslated repeatedly over centuries.
- **Influential:** High-register texts that impact the standard language.
- **Parallel:** Available in both Greek and Germanic varieties.

Strict selection criteria ensure our data is comparable across different historical periods.

The Greek Corpus

Our Greek data spans the entire history of the language.

- **Classical:** Homeric Epics (8th c. BC).
- **Koine:** The New Testament (1st c. AD).
- **Byzantine:** Chronicles (Sphrantzes).
- **Modern:** 20th c. Vernacular Translations.

The Greek corpus provides an unbroken chain of evidence for 3,000 years of internal change.

Why the New Testament?

The Backbone

The NT is the backbone of AthDGC because it is the most translated text in history.

Constant Variable

It acts as a "constant variable"—the meaning remains the same, so any change in wording reflects a change in grammar, not content.

Using a fixed source text allows us to isolate grammatical evolution from semantic changes.

The Germanic Corpus

Gothic

Wulfila's Bible
(4th c. AD)

Old English

Ælfric's Homilies
(10th c. AD)

Modern English

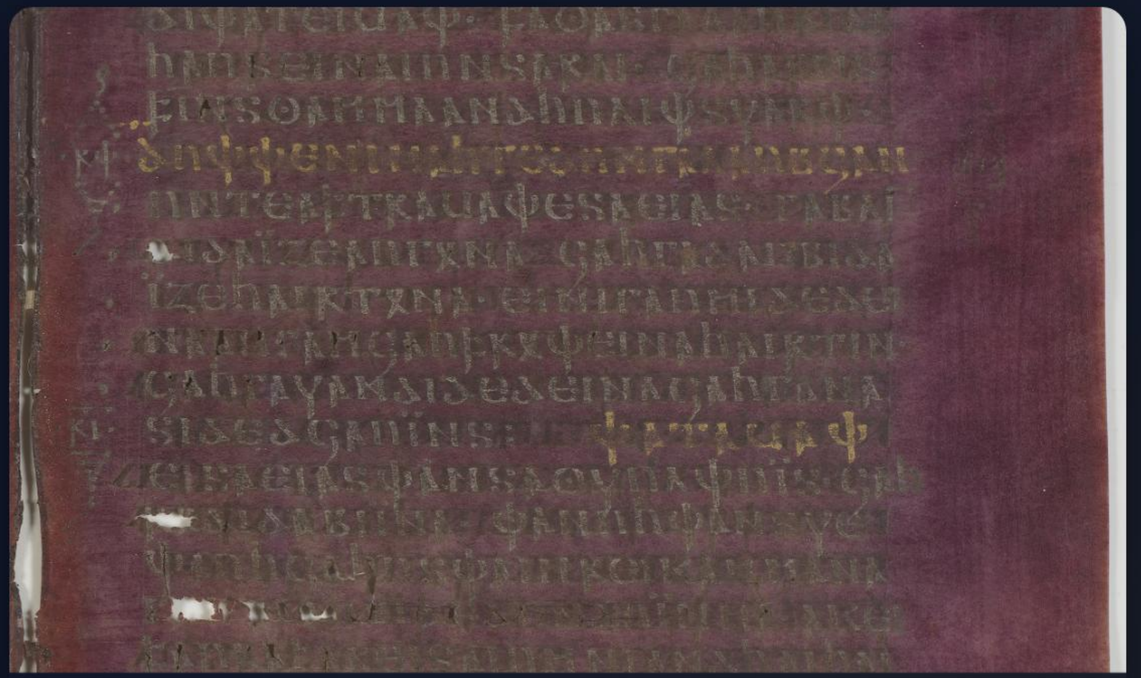
King James
(17th c. AD)

The Germanic corpus reveals how Greek structures were imported into Northern Europe.

Gothic: The Critical Link

Wulfila's Gothic translation is the earliest Germanic text of substantial length.

It is a direct translation from Greek, often preserving Greek word order and case usage that disappeared in later Germanic.



Gothic serves as the bridge connecting the Hellenic source to the Germanic target.

SECTION 04

Methodology

How We Annotate & Analyze

How do we turn ancient manuscripts into machine-readable data?

The PROIEL Framework

Standardization

AthDGC adopts the PROIEL guidelines (Pragmatic Resources in Old Indo-European Languages) to ensure international compatibility (Oslo, Oxford).

Compatibility

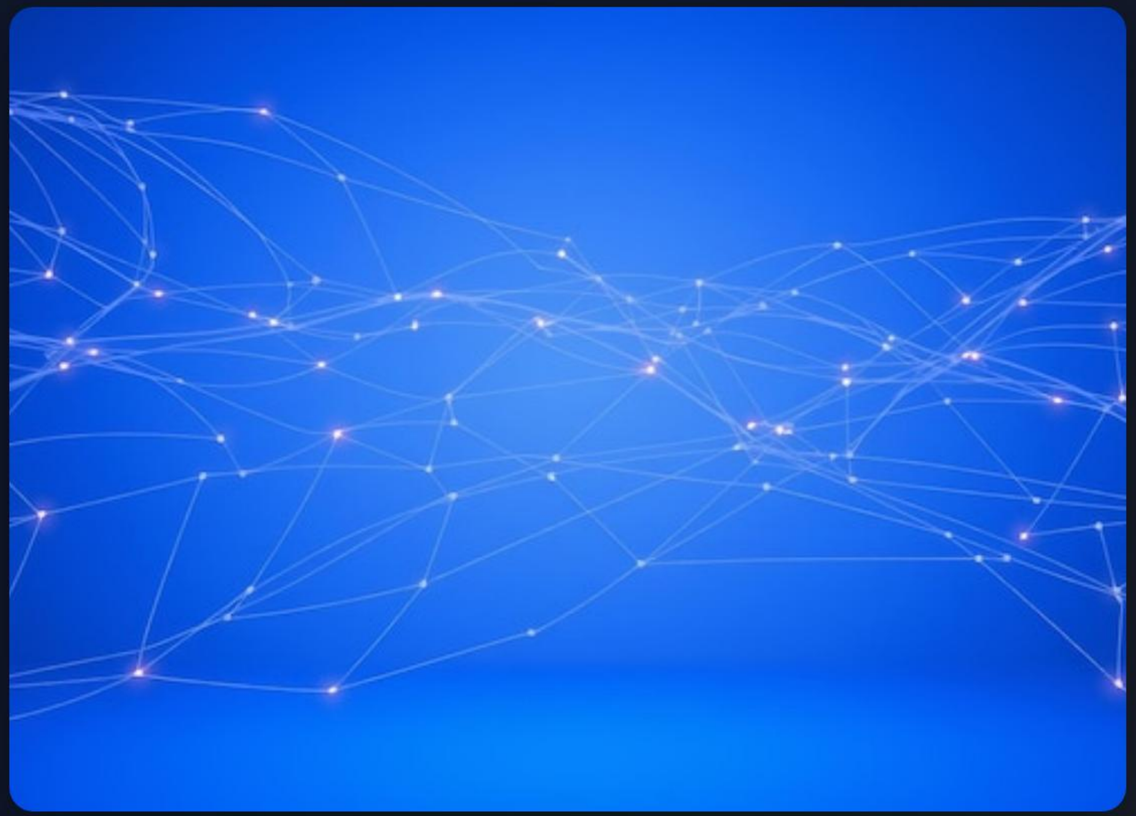
This allows our data to be merged with other global historical treebanks, creating a massive unified resource.

Using PROIEL ensures our findings are compatible with the wider digital humanities ecosystem.

Dependency Grammar

We use Dependency Grammar because it handles the **free word order** of Ancient Greek perfectly.

Instead of linear trees, we map non-linear relationships (Head → Dependent) regardless of position.



Dependency Grammar is the only syntactic model flexible enough for ancient languages.

Annotation Workflow

1. Tokenization

Splitting continuous text into individual words/tokens.

2. Morph-Tagging

Assigning Case, Gender, Number, Tense to every word.

3. Parsing

Manually connecting words to define syntactic roles (Subject, Object).

Our rigorous three-step workflow transforms raw text into structured linguistic data.

Methodology Challenge: Null Subjects

The Problem

Greek is a "Pro-Drop" language (subjects are omitted).
Example: "Lego" means "I speak" (No "I").

The Solution

We insert "empty nodes" in the syntax tree to represent these implied subjects, preserving valency counts.

We accurately annotate invisible subjects to ensure our statistical counts are correct.

SECTION 05

Case Studies

Evidence from the Corpus

We now present concrete evidence of language change from our annotated files.

Intralingual Retranslation (Matt 6:11)

Ancient

dòs hēmīn sēmeron

Give **us.DAT** today

Modern

dōse mas sēmera

Give **us.GEN/ACC** today

Intralingual retranslations highlight the internal change of the morphological system.

Interlingual Retranslation (John 1:1)

"In the beginning" across languages:

- **Greek:** *En archē* (Prep + Dative)
- **Latin:** *In principio* (Prep + Ablative)
- **Old English:** *On anginne* (Prep + Dative)

Result

Modern English: *In the beginning* (Prepositional Phrase)

This chain shows how Prepositional Phrases were calcqued (copied) across European languages.

We track how syntax is copied across borders.

The Evolution of Voice

Ancient System

Three distinct voices:
Active, Middle, Passive

Modern System

Two distinct voices:
Active, Medio-Passive

The project documents the morphological merger of the Middle and Passive voices.

SECTION 06

Challenges

Hurdles in Digital Philology

Digitizing 3,000 years of history is not without its difficulties.

Digitization Challenges

Historical texts are messy. We face:

- **Orthography:** Spelling was not standardized until recently.
- **Characters:** Archaic letters (digamma, thorn, eth) confuse standard OCR.

> **Solution:** We use custom-trained Transkribus models for manuscript recognition.

Standardizing variable spelling is the first major hurdle in our pipeline.

Challenge: Automatic Parsing

The Failure

Modern NLP parsers fail on Medieval Greek (Accuracy < 42%) because they are trained on Modern Greek syntax.

The Fix

We are developing specialized algorithms tailored specifically to historical varieties to improve accuracy.

Off-the-shelf computational tools fail on historical data, requiring custom solutions.

Challenge: Modern Particles

New Elements

Modern Greek uses particles (*tha, na, as*) that did not exist in Ancient Greek.

Tagset Expansion

We extended the PROIEL tagset to include these new categories while maintaining backward compatibility.

We successfully adapted the ancient annotation schema to accommodate modern innovations.

SECTION 07

Impact & Future

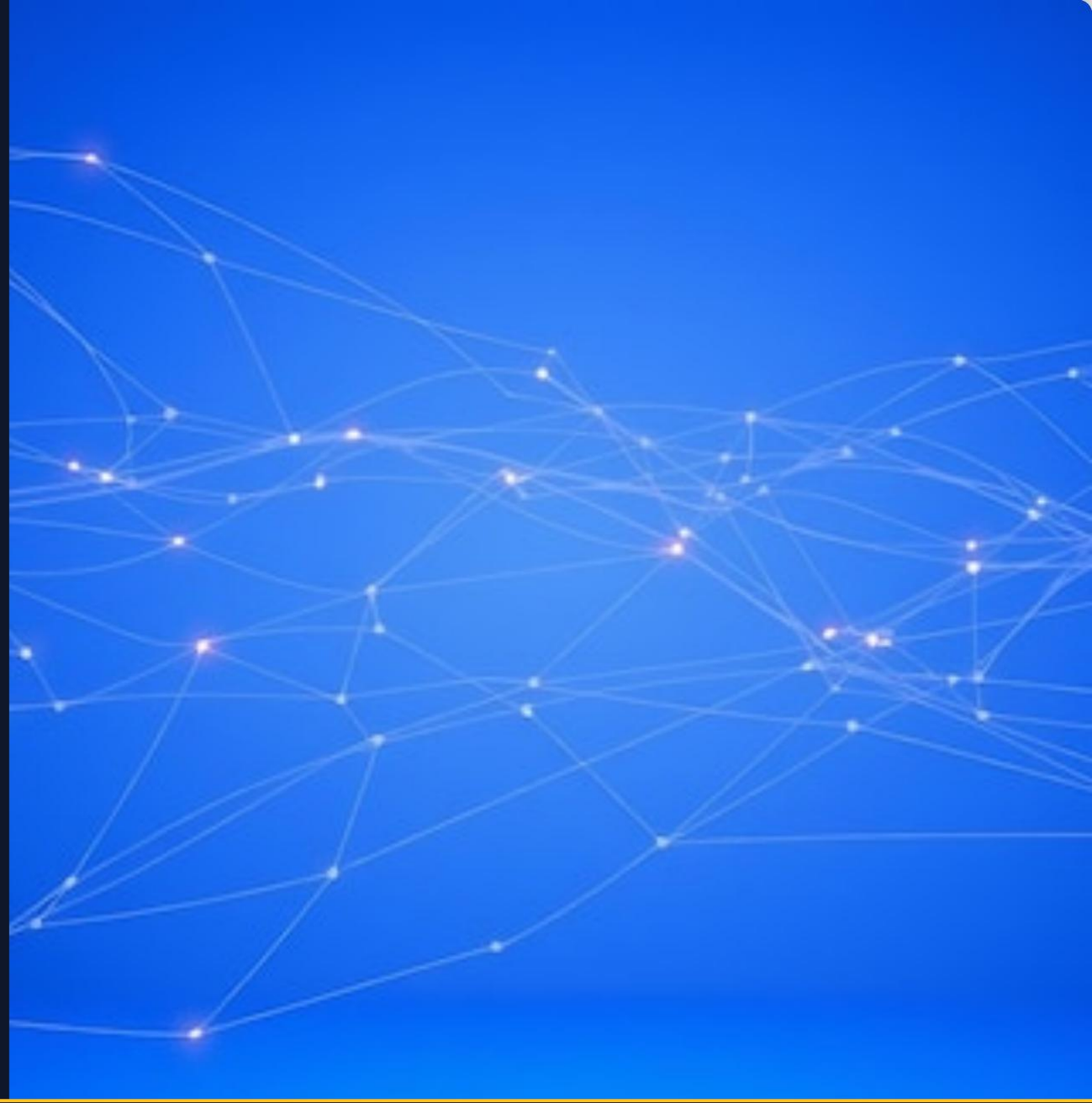
What AthDGC Delivers

Where does this project lead, and what resources will we provide?

The Diachronic Valency Lexicon

The crown jewel of AthDGC is the Valency Lexicon, a searchable online database linking raw frequency counts with specific textual examples.

> **Future Aims:** We aim to provide detailed genealogical information and data on borrowing/language contact for each entry.

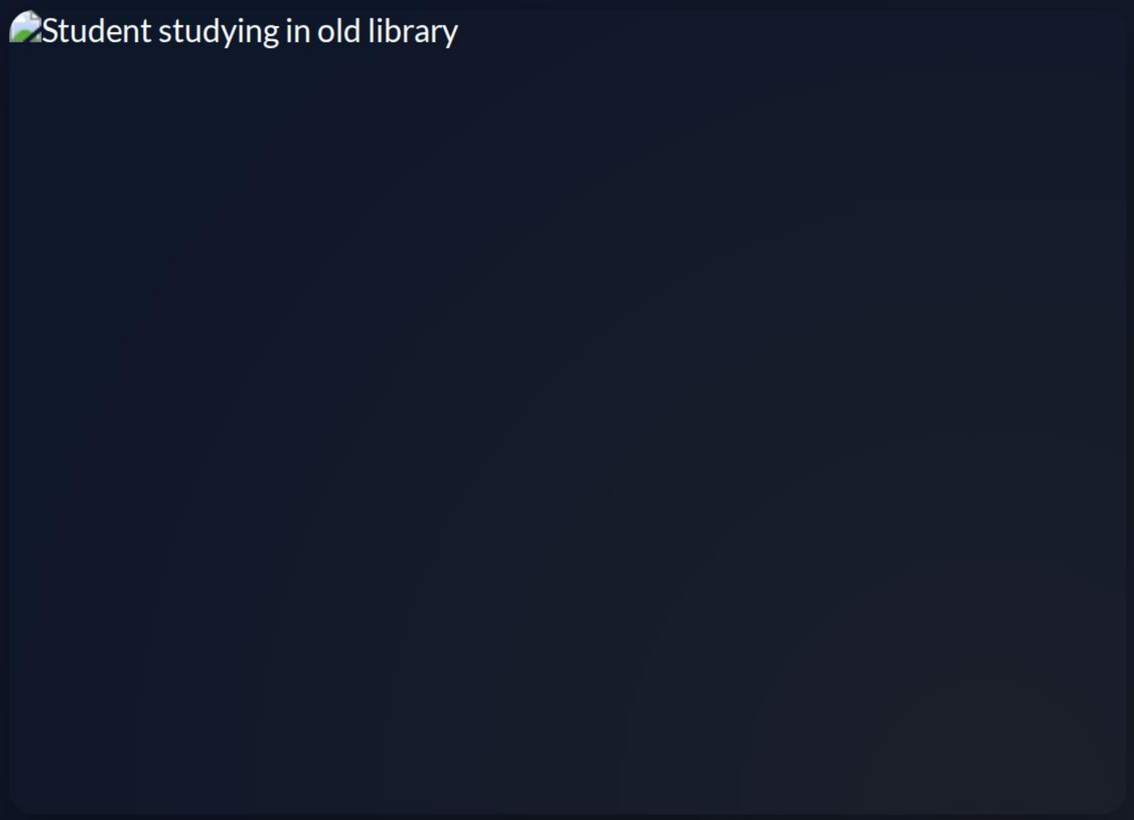


We are creating the first dictionary that tracks syntactic patterns, genealogy, and contact.

Application: Education

AthDGC resources aim at revolutionizing how Ancient Greek is taught.

- **Data-Driven Learning:** Students can see real usage stats, not just textbook rules.
- **Vocabulary:** Frequency lists based on actual historical strata.



Our data moves language pedagogy from intuition to empirical evidence.

Application: Digital Philology

Ground Truth

Our annotated corpus serves as the "Ground Truth" for optimizing computational algorithms.

Automation

It enables improved Automatic Glossing for digital libraries and refined computational analysis.

AthDGC provides the training data necessary to make digital tools useful for historical linguistics.

Phase II: Expanding Horizons

We plan to expand beyond Greek and Germanic to include:

- **Slavic:** Old Church Slavonic.
- **Celtic:** Old Irish.
- **Indo-Iranian:** Sanskrit.

Our future goal is to map these syntactic changes across the entire Indo-European family.

Commitment to Open Science

Open Access

All data, trees, and tools are released freely on GitHub/GitLab.

Community

We invite global collaboration to refine and expand the corpus.

AthDGC is a public good, designed to be accessible to researchers worldwide.

Final Conclusion

The Evidence

It proves that **translation drives change** and quantifies the massive historical drift to analyticity.

The Bridge

It bridges the gap between Philology, Diachronic Linguistics, and Computer Science.

AthDGC provides the digital infrastructure to understand our linguistic past.

Questions?

Athens Digital Glossa Chronos

Thank you for your attention.

We are now ready to address your questions.

Selected References (1)

- Nikiforidou, K. (1991). The meanings of the genitive: A case study in semantic structure. *Cognitive Linguistics*.
- Sitaridou, I. (2014). Word order. In G. K. Giannakis (Ed.), *Encyclopedia of Ancient Greek Language*.
- Eckhoff, H. M., et al. (2018). The PROIEL treebank family. *Language Resources and Evaluation*.

Foundational works on Treebanking, Semantics, and Diachronic Syntax.

Selected References (2)

- Lavidas, N. (2021). *The Diachrony of Written Language Contact*. Brill.
- Levin, B. (1993). *English Verb Classes and Alternations*.
- Tesnière, L. (1959). *Éléments de syntaxe structurale*.
- Trips, C. & Stein, A. (2019). Contact-induced changes in the argument structure of Middle English verbs.

Key literature on Valency, Language Contact, and Structural Change.

Image Sources

https://danielbwallace.com/wp-content/uploads/2015/09/cropped-img_bg_huge.jpg

Source: danielbwallace.com

https://external-preview.redd.it/national-and-kapodistrian-university-of-athens-building-a-v0-PNk16ZF-QRapAvMKGTUGcuS9jHrNj_xJeXRnffZqqOM.jpg?width=1080&crop=smart&auto=webp&s=85f3282e2b89bcf3c89a872193a64021165ae6bb

Source: www.reddit.com

https://img.freepik.com/premium-vector/abstract-hand-drawn-hourglass-time-clock-sand-doodle-concept-vector-design-outline-style_324137-6277.jpg

Source: www.freepik.com

https://media.istockphoto.com/id/517778587/photo/old-books-on-wooden-table.jpg?s=612x612&w=0&k=20&c=i0NIIzs1nWilgaao9wLa09AnYvqHDkwj4yIfbY7_lJ0=

Source: www.istockphoto.com

https://images.stockcake.com/public/d/a/2/da2b4f7f-be2d-458d-9a8a-2bf4f7b14b9f_large/ancient-library-interior-stockcake.jpg

Source: stockcake.com

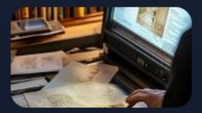
https://upload.wikimedia.org/wikipedia/commons/3/31/Codex_Argenteus_page.jpg

Source: en.wikipedia.org

Image Sources

https://img.freepik.com/premium-photo/abstract-network-interconnected-lines-with-glowing-nodes-against-blue-background_1174990-198028.jpg

Source: www.freepik.com



https://img.freepik.com/premium-photo/person-digitizing-old-documents-with-highresolution-scanner_995578-31512.jpg

Source: www.freepik.com

https://www.christs.cam.ac.uk/sites/default/files/inline-images/Christ%27s%20College%2C%20Cambridge%20Old%20Library%203.jpg

Source: www.christs.cam.ac.uk

https://st.depositphotos.com/2627021/2997/i/950/depositphotos_29974017-stock-photo-connected-world-europe-asia.jpg

Source: depositphotos.com