



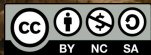
55 JAIIO

JORNADAS ARGENTINAS DE INFORMÁTICA



Desambiguación y enlazado de entidades en la poesía de Catulo: hacia una edición digital enriquecida y un grafo de conocimiento

Carlos Javier Nusch, Luciana Tanevitch, Diego Torres, Gimena del Rio Riande, Leticia Cecilia Cagnina, Leandro Antonelli, Marcelo Luis Errecalde



Esta obra está bajo una [Licencia Creative Commons](#)
Atribución-NoComercial-CompartirIgual 4.0 Internacional



GOBIERNO DE LA PROVINCIA DE
BUENOS AIRES



IIBICRIT



CONICET



UNIVERSIDAD
NACIONAL
DE LA PLATA

LIDIC



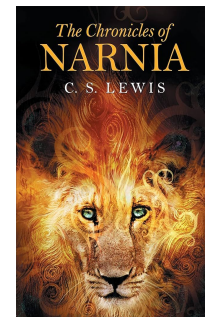
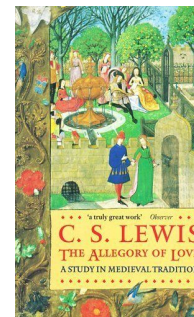
Universidad
Nacional de
San Luis

Motivación y Corpus: una lectura de C.S. Lewis



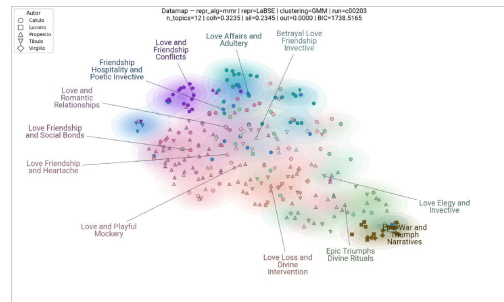
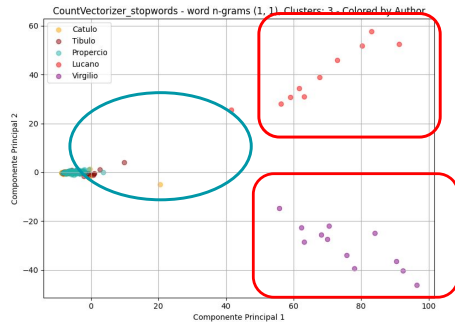
- Según Lewis, los motivos y el imaginario amoroso occidental se remontan a la poesía occitana del siglo XI.

¿Pero, qué ocurría antes?



Se trata de miles de textos en diferentes lenguas

- Evidentemente debíamos aprovechar las herramientas computacionales para la detección de patrones.
 - Reading Machines (Ramsay)
 - Distant Reading (Moretti, 2013)
 - Macroanálisis (Jockers, 2013)
- Aplicación de tareas de análisis automático de textos mediante técnicas de PLN sobre un corpus de textos latinos de los siglos I a. C. y I d. C. y españoles medievales y renacentistas.



Antecedente: De Catulo a Wikidata



De TXT a una edición TEI enriquecida

-  **Poemas .txt**
-  **Generación de **teiHeader****
-  **Segmentación y numeración de versos**
-  **NER automático con LatinCy**
-  **Lematización y clasificación** (*PERSON · LOC · NORP*)
-  **Consulta de autoridades** (*VIAF · Pleiades · Wikidata*)
-  **Detección de motivos amorios** (*Moreno Soldevila*)
-  **Anotación automática en XML-TEI**
-  **Validación TEI / Relax NG**
-  **Visualizaciones y análisis**

-  **TEI (Text Encoding Initiative - 1987)**
 - **estándar internacional basado en XML**
 - representar, describir, anotar y preservar textos digitales de forma interoperable y reutilizable.
 - **Comunidad académica internacional de la TEI Consortium**
 - Ofrece **Guidelines modulares**
 - Impulsado por **ACH, ACL y ALLC**
 - figuras centrales de su desarrollo, como **C. M. Sperberg-McQueen**, participaron posteriormente en la creación del **estándar XML del W3C**.

Antecedente: De Catulo a Wikidata

Objetivo: transformar textos planos en una edición digital estructurada, semánticamente enriquecida y reutilizable.

Catulo_Carmen_005.txt: Bloc de notas

Archivo Edición Formato Ver Ayuda

Vivamus, mea Lesbia, atque amemus,
rumoresque senum severiorum
omnes unius aestimemus assis.
soles occidere et redire possunt:
nobis, cum semel occidit brevis lux,
nox est perpetua una dormienda.
da mi basia mille, deinde centum,
dein mille altera, dein secunda centum,
deinde usque altera mille, deinde centum,
dein, cum milia multa fecerimus,
conturbabimus illa, ne sciamus,
aut ne quis malus invidere possit,
cum tantum sciat esse basiorum.



```
</fileDesc>
<encodingDesc>
  <appInfo>
    <application version="3.8.0" ident="latincy_la_core_web_lg">
      <label>la_core_web_lg v3.8.0</label>
      <ptr target="https://huggingface.co/latincy/la_core_web_lg"/>
    </application>
  </appInfo>
  <classDecl>
    <taxonomy xml:id="soldevila">
      <bibl>
        <author>Moreno Soldevila, R. (Ed.)</author>
        <date>2011</date>
        <title>Diccionario de motivos amorios en la Literatura Latina</title>
        <publisher>Universidad de Huelva</publisher>
      </bibl>
      <category xml:id="AOJAMIENTO_DE_LOS_AMANTES">
        <catDesc>AOJAMIENTO DE LOS AMANTES</catDesc>
      </category>
      <category xml:id="BESOS">
        <catDesc>BESOS</catDesc>
      </category>
      <category xml:id="ENVIDIA_HACIA_LOS_AMANTES">
        <catDesc>ENVIDIA HACIA LOS AMANTES</catDesc>
      </category>
      <category xml:id="INVITACION_AL_DISFRUTE_VITAL">
        <catDesc>INVITACIÓN AL DISFRUTE VITAL</catDesc>
      </category>
    </taxonomy>
  </classDecl>
</encodingDesc>
</teiHeader>
<standoff>
  <spanGrp type="topics">
    <span xml:id="span-INVITACION_AL_DISFRUTE_VITAL" from="1" to="6" ana="#INVITACION_AL_DISFRUTE_VITAL"/>
    <span xml:id="span-BESOS" from="7" to="13" ana="#BESOS"/>
    <span xml:id="span-AOJAMIENTO_DE_LOS_AMANTES" from="7" to="13" ana="#AOJAMIENTO_DE_LOS_AMANTES"/>
    <span xml:id="span-ENVIDIA_HACIA_LOS_AMANTES" from="12" to="13" ana="#ENVIDIA_HACIA_LOS_AMANTES"/>
  </spanGrp>
</standoff>
<text>
  <body>
    <l n="1" ana="#INVITACION_AL_DISFRUTE_VITAL">Vivamus, mea <placeName>Lesbia</placeName>, atque
    amemus,</l>
    <l n="2" ana="#INVITACION_AL_DISFRUTE_VITAL">rumoresque senum severiorum</l>
    <l n="3" ana="#INVITACION_AL_DISFRUTE_VITAL">omnes unius aestimemus assis.</l>
    <l n="4" ana="#INVITACION_AL_DISFRUTE_VITAL">soles occidere et redire possunt:</l>
    <l n="5" ana="#INVITACION_AL_DISFRUTE_VITAL">nobis, cum semel occidit brevis lux,</l>
    <l n="6" ana="#INVITACION_AL_DISFRUTE_VITAL">nox est perpetua una dormienda.</l>
    <l n="7" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES">da mi basia mille, deinde centum,</l>
    <l n="8" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES">dein mille altera, dein secunda centum,</l>
    <l n="9" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES">deinde usque altera mille, deinde centum,</l>
    <l n="10" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES">dein, cum milia multa fecerimus,</l>
    <l n="11" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES">conturbabimus illa, ne sciamus,</l>
    <l n="12" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES #ENVIDIA_HACIA_LOS_AMANTES">aut ne quis malus invidere
    possit,</l>
    <l n="13" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES #ENVIDIA_HACIA_LOS_AMANTES">cum tantum sciat esse
    basiorum.</l>
  </body>
</text>
</tei>
```

Antecedente: De Catulo a Wikidata

Objetivo: transformar textos planos en una edición digital estructurada, semánticamente enriquecida y reutilizable.

```
<text>
  <body>
    <lg type="poema">
      <l n="1" ana="#INVITACIÓN_AL_DISFRUTE_VITAL">Vivamus, mea <personName>Lesbia</personName>, atque amemus,</l>
      <l n="2" ana="#INVITACIÓN_AL_DISFRUTE_VITAL">rumoresque senum severiorum</l>
      <l n="3" ana="#INVITACIÓN_AL_DISFRUTE_VITAL">omnes unius aestimemus assis.</l>
      <l n="4" ana="#INVITACIÓN_AL_DISFRUTE_VITAL">soles occidere et redire possunt:</l>
      <l n="5" ana="#INVITACIÓN_AL_DISFRUTE_VITAL">nobis, cum semel occidit brevis lux,</l>
      <l n="6" ana="#INVITACIÓN_AL_DISFRUTE_VITAL">nox est perpetua una dormienda.</l>
      <l n="7" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES">da mi basia mille, deinde centum,</l>
      <l n="8" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES">dein mille altera, dein secunda centum,</l>
      <l n="9" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES">deinde usque altera mille, deinde centum,</l>
      <l n="10" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES">dein, cum milia multa fecerimus,</l>
      <l n="11" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES">conturbabimus illa, ne sciamus,</l>
      <l n="12" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES #ENVIDIA_HACIA_LOS_AMANTES">aut ne quis malus invidere possit,</l>
      <l n="13" ana="#BESOS #AOJAMIENTO_DE_LOS_AMANTES #ENVIDIA_HACIA_LOS_AMANTES">cum tantum sciat esse basiorum.</l>
    </lg>
  </body>
</text>
</TEI>
```

Ahora damos un paso más: enlazar las entidades

De TXT a una edición TEI enriquecida

 Poemas .txt

-  Generación de **teiHeader**
-  Segmentación y numeración de versos

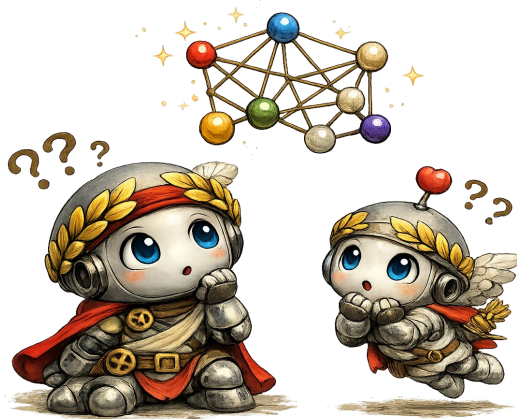
→  **NER**

-  Lematización y clasificación (*PERSON · LOC · NORP*)
-  Consulta de autoridades (*VIAF · Pleiades · Wikidata*)
-  Detección de motivos amorios (*Moreno Soldevila*)
-  Anotación automática en XML-TEI
-  Validación TEI / Relax NG
-  Visualizaciones y análisis.

→  **Entity Linking**

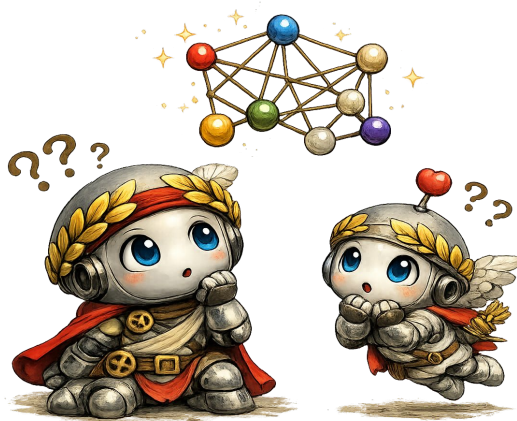
¿Por qué querríamos enlazar entidades?

-  **Identificar y distinguir referentes concretos** (personas, dioses, lugares y otras entidades).
-  **Conectar menciones del texto con identificadores persistentes**
 - QID de Wikidata.
-  **Relacionar entidades con motivos, tópicos y escenas del imaginario amoroso.**
-  **Construir una base de conocimiento orientada a grafos**
 - TEI + Wikidata.



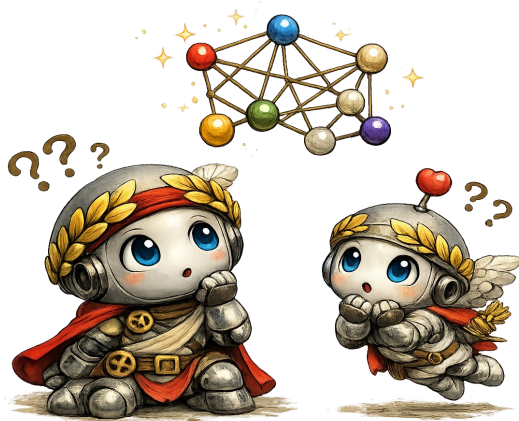
¿Por qué querríamos enlazar entidades?

- 🔍 **Realizar consultas complejas** sobre personajes, lugares, coocurrencias y relaciones entre poemas.
- 🧠 **Habilitar inferencias y descubrimiento de relaciones indirectas.**
- 📊 **Detectar patrones y redes recurrentes** que no son evidentes mediante lectura individual.
- ↻ **Integrar y reutilizar información externa** de Wikidata, VIAF, Pleiades u otros recursos.



¿Por qué querríamos enlazar entidades?

-  **Mantener trazabilidad editorial:** saber exactamente qué mención fue enlazada, con qué entidad y desde qué verso.
-  **Convertir la edición en una infraestructura de datos,** no solo en una representación digital del texto.



Del NER al Entity Linking: un dominio complejo



- El desafío es detectar una mención correcta y enlazarla correctamente.
- La idea general es que la tarea asista, no reemplace la curaduría filológica.

Fenómeno

Riesgo para EL

Flexión latina

La forma textual no coincide con Wikidata

Gentilicios

Puede remitir a pueblo, región o ciudad

Teónimos/personificaciones

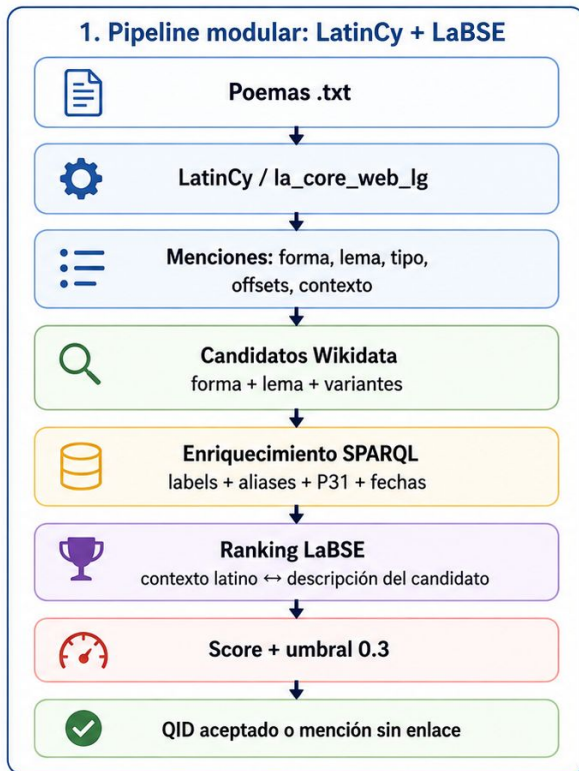
Oscilan entre persona, mito y lugar

Hipérbaton

Dificulta detectar spans completos

Wikidata moderno

Granularidad desigual para mundo clásico





Enfoque GPT-4.1: Prompt 1

You are a named entity recognition system for Latin texts.

Task:
Detect named entities.

Instructions:

- Detect ONLY: PERSON, LOC, NORP
- Mentions must be exact spans
- Offsets must be exact (0-based, end exclusive)
- Normalize classical Latin orthography (e.g., "uenus" → "Venus")
- If the entity is NORP (gentilicium), ALSO infer the associated place (country, region, or people)

Example:

"Romanus" → "Rome"

"Troianus" → "Troy"

Return ONLY JSON:

```
{  
  "mentions": [  
    {  
      "surface_form": "...",  
      "lemma": "...",  
      "ner_label": "PERSON",  
      "start": 0,  
      "end": 0,  
      "country": "..."  
    }  
  ]  
}
```

Text:

""

{TEXT}

""



Enfoque GPT-4.1: Prompt 2

You are an expert entity linking system for classical Latin texts.

Text:

""

{TEXT}

""

Mentions (with IDs):

{MENTIONS_JSON}

Candidates per mention (keyed by mention id):

{CANDIDATES_JSON}

Task:

For EACH mention, select the best matching Wikidata entity.

Rules:

- Only choose from the provided candidates
- Prefer historically relevant entities (Roman, mythological, ancient)
- Use context from the text
- If the mention is NORP (gentilicium):
 1. Prefer selecting the entity representing the people/group (e.g., "Romans", "Trojans")
 2. If such entity is NOT present among the candidates, select the related place (e.g., "Rome", "Troy")
- If none match, set qid = null
- Do NOT invent entities

Return ONLY JSON:...

Ground truth y evaluación

- Ground truth manual: 723 menciones en 108 poemas.
- Basado en ediciones críticas, comentadas y traducciones.
- Tipos: PERSON, LOC, NORP.
- NER: precision, recall, F1.
- EL: Accuracy@1, Recall@10, MRR; GPT solo Accuracy@1.

poem_id	mention	start_token	end_token	type	qid	wikidata_label_en	Gentilicio	Comments
Catulo_Carmen_001	Corneli	10	10	PERSON	Q109594	Cornelius Nepos		
Catulo_Carmen_001	Italorum	29	29	NORP	Q38	Italy	x	Gentilicio: le hacer metrica con y sin gentilicios
Catulo_Carmen_001	Iuppiter	38	38	PERSON	Q4649	Jupiter		
Catulo_Carmen_003	Veneres	3	3	PERSON	Q47652	Venus		
Catulo_Carmen_003	Cupidines	4	4	PERSON	Q5011	Cupid		
Catulo_Carmen_003	Orci	83	83	PERSON	Q860314,Q152262	Orcus, Pluto		Nos quedamos con Orcus (Q860314) el dios, aunque se use en el poema en el sentido de inframundo (no existe la categoría inframundo romano en Wikidata, sí existen, sin embargo: underworld (Q154365), Greek underworld (Q1521549) y Hades (Q112572697) que se llama igual que el dios que lo ocupa: Hades (Q41410). Pero que en la mitología romana se asocia con Plutón: Pluto (Q152262). -
Catulo_Carmen_004	Hadriatici	33	33	LOC	Q13924	Adriatic Sea		
Catulo_Carmen_004	Cycladas	37	37	LOC	Q171132	Cyclades		
Catulo_Carmen_004	Rhodum	38	38	LOC	Q43048	Rhodes		
Catulo_Carmen_004	Thraciam	43	43	LOC	Q41741	Thrace		
Catulo_Carmen_004	Propontida	44	44	LOC	Q35367	Sea of Marmara		
Catulo_Carmen_004	Ponticum	46	46	LOC	Q166,Q1913944	Black Sea, Euxineus Pontos		
Catulo_Carmen_004	Cytorio	59	59	LOC				Cytorus (Q5201382) es una ciudad, el poema se refiere al Monte Citorio de Plafagonia, Bitinia.
Catulo_Carmen_004	Amastri	68	68	LOC	Q455458	Amasra		
Catulo_Carmen_004	Pontica	69	69	NORP	Q621672	Pontus	x	Gentilicio: se refiere a Amastris, Amasris pontica

Resultados y errores: NER

- LatinCy: más cobertura y trazabilidad.
- GPT: mejor selección final, pero más conservador.

Modelo	Precision	Recall	F1
LatinCy	0.92	0.83	0.85
GPT-4.1	0.94	0.75	0.83

Resultados y errores

- EL: GPT mejora la selección final; LatinCy + LaBSE ofrece ranking auditable y mejor cobertura del espacio de candidatos.

Modelo	Accuracy@1	Recall@10	MRR
LatinCy + LaBSE	0.29	0.40	0.32
GPT-4.1	0.46	-	-

- **MRR — Mean Reciprocal Rank:** mide **qué tan arriba aparece la respuesta correcta en el ranking**. Si está primera vale 1; segunda, 1/2; tercera, 1/3; etc. Después se promedia. Como el Accuracy@1 es 0.29 y el MRR solo sube hasta 0.32, podemos interpretar que cuando la entidad correcta no está primera, **no aparece con mucha frecuencia en las posiciones inmediatamente siguientes**. Eso confirma que el ranking todavía tiene bastante margen de mejora.

Ejemplos de error

Etapas	Tipo de error	Ejemplo(s)	¿Qué ocurre?
NER – GPT-4.1	Omisiones / falsos negativos	Iuppiter, Diana, Cybele, Pontus, Brixia	GPT es más conservador: prioriza coherencia semántica y pierde cobertura.
NER – LatinCy	Sobredetección / falsos positivos	lugete, amabo	Detecta como entidades verbos o formas léxicas comunes.
NER	Span omitido	Tempe (C.64), Hamadryades (C.61)	La entidad existe en el ground truth, pero no se detecta.
NER	Tipo incorrecto	Sirmio: LOC → PERSON	El contexto o la forma superficial inducen una clasificación errónea. LatinCy detectó <i>Sirmio</i> como persona cuando antes detectó lugar. Quizá porque antes de esta instancia de Sirmio está "salve".
NER	Formas derivadas / gentilicios	Sestianus (C.44)	Una derivación de nombre propio puede parecer superficialmente un gentilicio.

Ejemplos de error

Etapas	Tipo de error	Ejemplo(s)	¿Qué ocurre?
Entity Linking	Generación de candidatos	Cytorio / Cytore (C.4)	La entidad correcta no se recupera o Wikidata representa mejor otra entidad relacionada.
Entity Linking	Ranking incorrecto	Castoris gemelle → Pólux, pero se prioriza Cástor	El candidato correcto está disponible, pero la similitud superficial domina al contexto.
Entity Linking	Ambigüedad ontológica	Tempe, Diae	Una mención puede admitir lecturas geográficas, históricas o míticas.
Entity Linking / Wikidata	Desajuste ontológico	Orci: deidad vs. inframundo; Sarapim: Serapis vs. templo	El QID es plausible, pero no refleja exactamente el uso poético.

Ejemplos de error

Orci: Nos quedamos con Orcus (Q860314) el dios, aunque se use en el poema en el sentido de inframundo (no existe la categoría inframundo romano en Wikidata, sí existen, sin embargo: underworld (Q154365), Greek underworld (Q1521549) y Hades (Q112572697) que se llama igual que el dios que lo ocupa: Hades (Q41410). Pero que en la mitología romana se asocia con Plutón: Pluto (Q152262). -

Ejemplos de error

Etapa	Tipo de error	Ejemplo(s)	¿Qué ocurre?
Latín poético	Morfología y polisemia	Pontica, Hibera, Troiugenum	Oscilan entre gentilicio, pueblo, lugar o modificador geográfico.
Latín poético	Personificación / superposición	Austri; Scylla, Charybdis, Syrtis	Una misma entidad puede funcionar simultáneamente como persona, mito o lugar.
Latín poético	Alusión compleja	Memnonis, Arsinoes (C.66)	Resolver el referente requiere conocimiento filológico y contextual.
Latín poético	Hipérbaton / entidad discontinua	Cinna est Gaius (C.10), Marrucine Asini (C.12)	Los componentes de una entidad pueden estar separados o contener formas de otras categorías.

CIENCIA DE DATOS: LA REALIDAD VS. LA FICCIÓN

Lo que mis amigos creen que hago



Analizando patrones complejos y prediciendo el futuro

Lo que mi jefe cree que hago



Tomando decisiones estratégicas y creando valor empresarial

Lo que la sociedad cree que hago



Hackeando el sistema con IA y algoritmos avanzados

Lo que REALMENTE hago



Lidiando con archivos CSV interminables, limpiando datos sucios y corrigiendo errores en Excel.

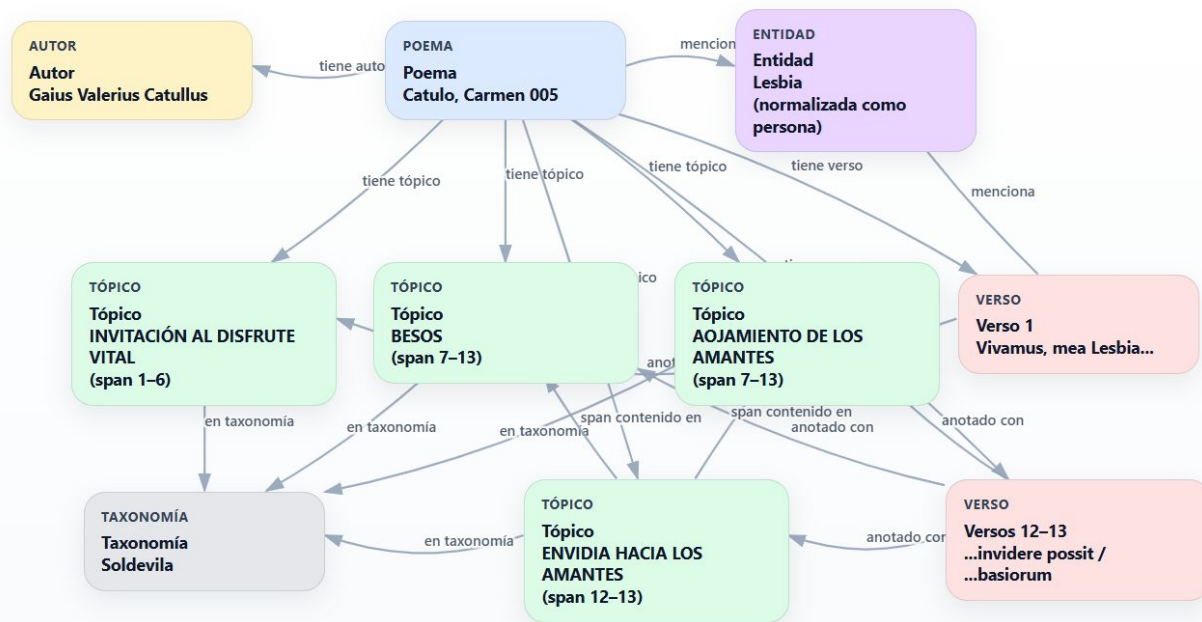
Conclusiones

- El **entity linking en poesía latina es viable como asistencia editorial**, pero no todavía como automatización completamente fiable.
- El **pipeline modular** ofrece mayor **trazabilidad, auditabilidad y control** sobre las distintas etapas del proceso.
- **GPT-4.1** integra mejor el contexto y logra una **mejor selección final**, aunque es más conservador y menos transparente.
- Ambos enfoques muestran **fortalezas complementarias**; una **arquitectura híbrida** aparece como la alternativa más prometedora.
- El **ground truth manual** constituye un recurso reutilizable para evaluar y mejorar futuros sistemas.
- El entity linking funciona como **punto entre XML-TEI y Wikidata**, permitiendo avanzar hacia una edición digital semánticamente enriquecida.

Trabajos Futuros

- Ampliar el corpus a **Tibulo y Propertio** para evaluar la generalización del enfoque.
- Mejorar **prompts, restricciones, abstención y reconocimiento de NIL**.
- Comparar con enfoques recientes de EL: **dense retrieval, re-ranking, enfoques generativos y LLM + retrieval**.
- Refinar la tipología para distinguir mejor **personas históricas, figuras míticas, deidades, lugares y colectivos**.
- Complementar Wikidata con **recursos especializados del mundo clásico**, como gazetteers, bases prosopográficas y LiLa.
- Profundizar la integración **TEI + Wikidata** y avanzar hacia un **grafo de conocimiento literario**.

Ventajas de una Edición Digital Enriquecida e Interoperable



TTL original

Incluyo un extracto para que puedas contrastar la vista visual con el grafo fuente.

```
@prefix rdf:
<http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix ex:
<https://example.org/aetates-amoris/> .
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/> .
@prefix dct:
<http://purl.org/dc/terms/> .
@prefix rdfs:
<http://www.w3.org/2000/01/rdf-
```

Esta vista prioriza la legibilidad: muestra los recursos principales como nodos y conserva en el panel los detalles textuales más importantes.



Ventajas de una Edición Digital Enriquecida e Interoperable



CATULLUS
Carmina

EdiciónAnotacionesEntidadesBúsquedaVisualización

CARMEN 5

1 Viuamus, mea **Lesbia**, atque amemus,
2 rumoresque senum seueriorum
3 omnes unius aestimemus assis.
4 soles occidere et redire possunt:
5 nobis, cum semel occidit breuis lux,
6 nox est perpetua una dormienda.
7 da mi basia mille, deinde centum,
8 deinde altera, dein secunda centum,
9 deinde usque altera mille, deinde centum,
10 deinde usque altera, dein secunda centum,
11 deinde altera mille, deinde centum,
12 deinde, cum milia multa fecerimus,
13 conturbabimus illa, ne sciamus,
13 aut ne quis malus inuidere possit,
14 cum tantum sciat esse basiorum.

AnotaciónTEI/XML

Forma: Lesbia
Lema: Lesbia
Tipo: PERSON
Inicio: 0,10
Fin: 0,16
Contexto:
... Viuamus, mea Lesbia, atque amemus,
rumoresque senum seueriorum ...

Legenda de entidades
PERSON
LOC
NORP
MISC

Notas editoriales

Aparato crítico

Edición digital basada en: R. A. B. Mynors (ed.), *Catullus*. Oxford, 1969.

TEI P5

Lesbia (Clodia Metelli)

Q132657 en Wikidata

Patricia romana del siglo I a. C., amada de Catulo.



ResumenPropiedadesEnlacesMenciones (12)

También conocida comoClodia Metelli

Sexo o génerofemenino

Instancia dehumano Q5

Fecha de nacimientoc. 95 a. C.

Fecha de muertec. 50 a. C.

PadresPublio Clodio Pulcro Q165325
Clodia Q189818

CónyugeQuinto Metelo Celer Q133060

Ocupaciónaristócrata

DescripciónPatricia romana, célebre por su belleza y por su relación con Catulo.

Enlaces externosVIAF 32172820
Bibliografía (5 recursos) ▾

Imágenes





Handwritten text in a cursive script, likely a Latin inscription related to the Vitruvian Man.



Carlos J. Nusch

carlosnusch@prebi.unlp.edu.ar



Esta obra está bajo una Licencia Creative Commons
Atribución-NoComercial-CompartirIgual 4.0 Internacional

OMNIA VINCIT AMOR
et nos credamus
amorem cordis



`print("Gratias vobis ago!")`

