Digital Processing of Textual Sources: Projects in Digital Humanities at CIHAM-UMR 5648

Olivier Brisville-Fertin[3, 4], Matthias Gille Levenson[1, 3, 4], Ariane Pinche[2, 3]

[1]ENC, [2]CNRS, [3]CIHAM, [4]ENSL

Digital Humanities and Romance Philology, April 4th, 2024, Lyon

Table of Contents

Presentation of CIHAM and the Corpor@tech Axis

- CIHAM - UMR 5648
- Research unit in medieval studies
- 52 researchers in History, Archaeology, Literature covering both Christian and Muslim civilizations.
- Particular attention to manuscript sources
- Long-standing expertise in Digital Humanities
  - XML TEI corpora modeling
  - Dissemination of open science principles
- A transversal research axis for digital humanities: Corpor@tech Open Science.

CIHAM and Digital Editions

- Pioneering work by Marjorie Burghart (CNRS Crystal Award 2013)
  - Digital editions: *Sermones.net*
  - Interactive paleography exercises: *Interactive Album of Mediaeval Palaeography*
  - Help to prepare digital editions: *TEI Critical Apparatus Toolbox*
  - Online Courses : *Digital Scholarly Editions: Manuscripts, Texts and TEI Encoding* with E. Pierazzo

Ongoing Digital Edition Projects

- DISTINGUO : a knowledge base on "distinctiones," master structures of medieval preaching – ANR 2019-2024
- LIBeR : The Decades of Bersuire, the first French translation of Tite-Live's *Roman History* – LiBer - ANR 2021-2026
- Fabliaux : digital corpus of French fabliaux from the Middle Ages, Biblissima+ project 2022-2023.

CIHAM and ATR

- The CATMuS Project
  - CATMUS for *Consistent Approaches to Transcribing ManuScripts*
  - International collaboration involving several CIHAM researchers
  - Funded by the BnF's dataLab and Biblissima+
  - Goal: The Consistent Approaches to Transcribing Manuscripts (CATMuS) guidelines, datasets, and models are the result of an international collaboration aimed at standardizing ATR practices for historic documents.
  - multilingual project: Latin, (Old) French, Occitan, Catalan, Castillan, Italian, venitian, Old English, Middle Dutch
  - To learn more about the project: Ariane Pinche, Thibault Clérice, Alix Chagué, Jean-Baptiste Camps, Malamatenia Vlachou-Efstathiou, Matthias Gille Levenson, et al. *CATMuS-Medieval: Consistent Approaches to Transcribing ManuScripts*. en. Dec. 2023. URL: https://inria.hal.science/hal-04346939 (visited on 01/08/2024)

CIHAM and ATR

- Le projet CATMuS : output
  - General transcribing guidelines (Work in progress)
  - Interoperable training datasets (Work in progress)
  - General model
    - A general model for medieval manuscripts : Ariane Pinche, Thibault Clérice, Alix Chagué, Jean-Baptiste Camps, Malamatenia Vlachou-Efstathiou, Matthias Gille Levenson, et al. "CATMuS Medieval". lat. In: (Nov. 2023). Publisher: Zenodo. URL: https://zenodo.org/records/10066219 (visited on 01/08/2024)
    - A general model for gothic prints : Sonia Solfrini and Simon Gabay. "CATMuS Gothic Print". frm. In: (Jan. 2024). Publisher: Zenodo. URL: https://zenodo.org/records/10599911 (visited on 03/27/2024)
    - A general model for prints : Simon Gabay and Thibault Clérice. "CATMuS-Print [Large]". fra. In: (Jan. 2024). Publisher: Zenodo. URL: https://zenodo.org/records/10592716 (visited on 03/27/2024)
    - A general model for model manuscripts (Work in progress)

Table of Contents

What is *Aljamiado*?

- Textual production of the medieval and early modern Muslim minorities, in the Kingdoms of Aragon and Castile, and in the Catalan Counties
- Translations into Romance vernacular (Aragonese, Castilian, Catalan) written in Arabic *abjad* (mostly in Aragon and Catalonia) (This is not Arabic =>)

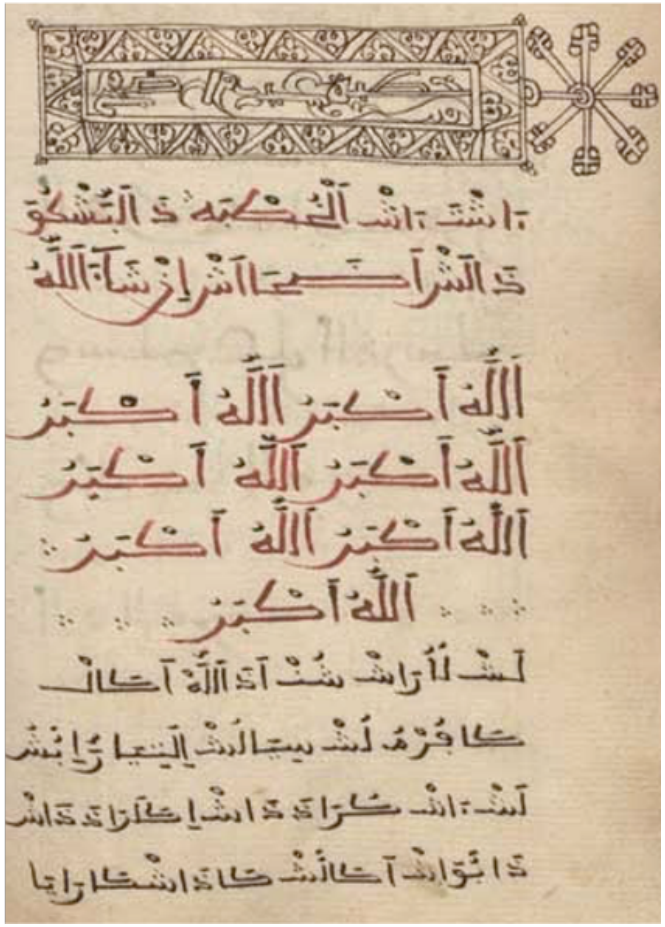


Figure: Madrid, CSIC-CCHS, BTNT, Resc/25, f.100v

How to edit such a document?

- Should we transcribe?

ءَاشْتَ ءَاشْ اَلْخُطْبَة ذَا لَبَّشْكُوَ ذَا لَشْ اَضَحَااَشْ[...]
لَشْ لُاْرَاشْ شُنْ اَذَ اَللَهْ اَكَالْ كَا فُرمَه لُشْ سِيَالُشْ اِلَتِيَارَّ

- Should we transliterate?
  - ʾāšta ʾāš alḫuṭbaï ḏe la paškuwa ḏe laš aḍaḥāas [...] laš luurāš šon aḏa Allah akāl kā forma loš siyāloš i la tiyārra
  - ʾešta ʾeš alḫoṭbaï ḏe la paškuwa ḏe laš aḍaḥeas [...] laš looreš šon aḏa Allah akel ke forma loš siyeloš i la tiyerra
- Which criteria should we follow?
  - Esta es alḫoṭbaï ḏe la pascūa ḏe las aḍaḥāas [...] Las loores son aḏa Allah Aquel que forma los ç̣īelos i la tīerra
  - Esta es aljotba de la pascua de las aḍaḥeas [...] Las loores son ada Allah Aquel que forma los cielos y la tierra
- How to conciliate textual fidelity and legibility?

Adopting DH: a sandbox project

- A bilingual version of "The City of Brass" (*Madīnat al-nuḥās*)
  - A popular Oriental tale on a Westward expedition to recover one of Salomon's vases, which belongs today to the modern *Arabian nights* (*Alf layla wa-layla*)
  - Story quite popular in the Medieval Islamic West: in the Hispanic Islamic production, several Arabic and Aljamiado versions
  - A bilingual version, copied in a late 15th-century manuscript, 89 pages.
  - Methodological, technical, linguistic, traductological, literary, cultural interests

Figure: Saragosse, FDHCA, L536, f. 73v

Using HTR to automatically acquire the transcription

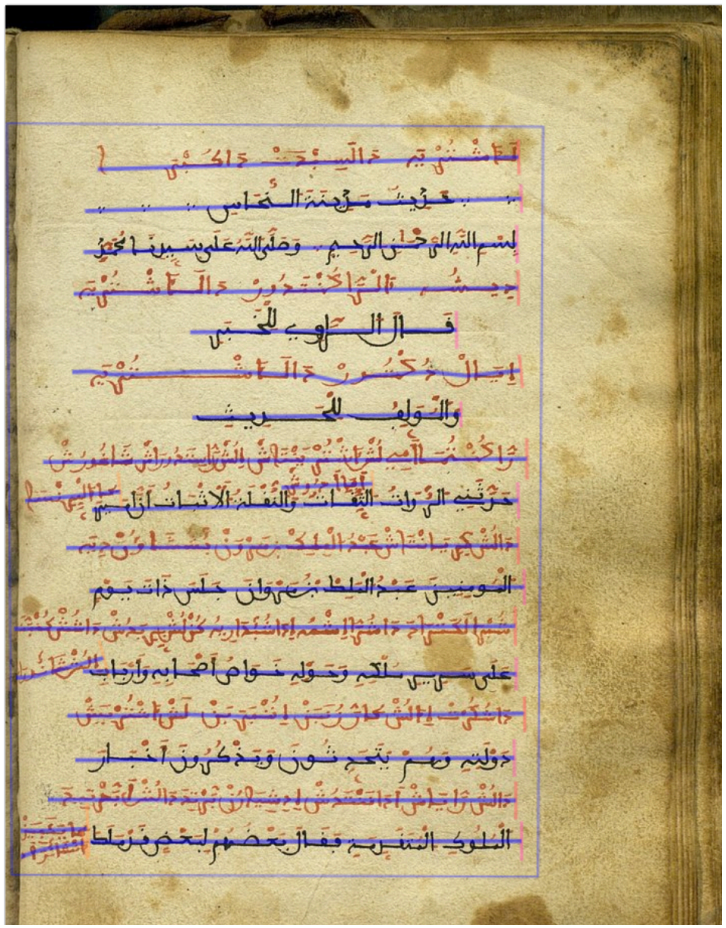


Figure: Hand-made segmentation on e-Scriptorium

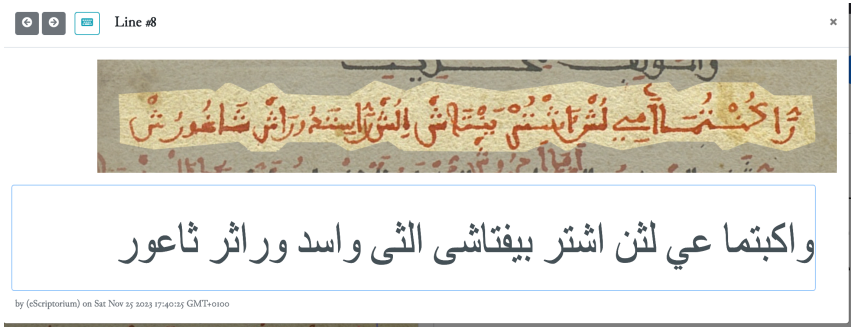


Figure: Prediction with a model trained for *al-ḫaṭṭ al-maġribī*

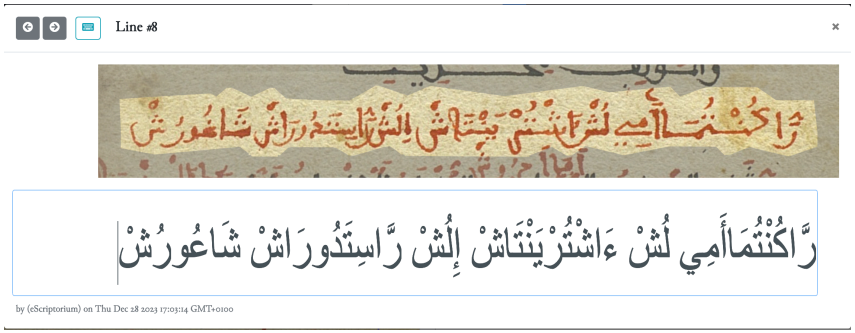


Figure: Prediction with a finetuned model after correction

Editing with TEI: how to define a viable encoding protocole?

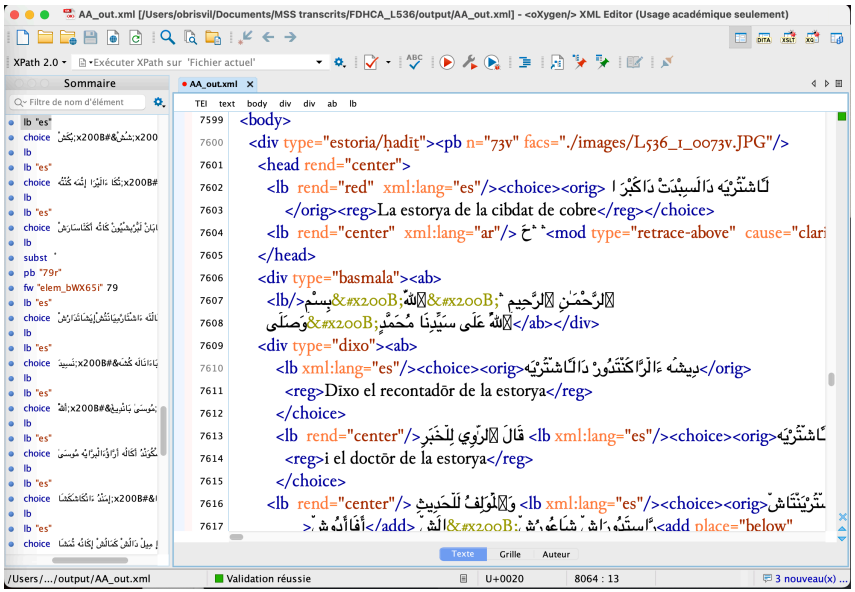


Figure: TEI encoding of the story

- How should we or could we encode?
  - a bilingual interlinear document
  - proposing both a diplomatic transcription and a regularized transliteration
  - in two different graphical systems
  - with a paleographical and philological apparatus
  - and no compatible fonts for Arabic and Latin paleography

Next steps: alignment and lexical-grammatical tagging?

- Alignment sequences to compare linguistic versions, i.e. the Arabic original and the Romance translation; more or less a synoptic edition.
- Lexical tagging of the tokens to establish patterns of translation; e.g. *al-amīr* <=> *el príncep*, *el prince*, *el príncipe*, *el capitán*, *el virrey*, *el visorrey*, i.e. 3 lemmas, with respectively 2, 1, and 3 variants.

Table of Contents

Tagging Medieval French Texts

- **Lemmatizing Old French Texts in Picard (2017) ?**
- Existing lemmatizers (Stanza, SpaCy, TreeTagger) not suitable for medieval French
- Experimentation with a new tool: Pie-Extended, *Manjavacas 2024*
- Creating training data
- **Results (2019)**
  - Deucalion Lemmatizer and POS-tagger for Old French (2019)
  - *See article, Camps et al. 2021*
- **Deucalion**, *Clérice et al. 2020*
- 1,132,849 tokens, Achieved approximately 97% accuracy in lemma and POS tagging
  - lemma: Tobler and Lommatzsch, Altfranzösisches Wörterbuch
  - POS and morph: CATTEX2009-max.

form	lemma	POS	morph
G'	je	PROper	PERS.=1|NOMB.=s|GENRE=m|CAS=n
irai	aler	VERcjg	MODE=ind|TEMPS=fut|PERS.=1|NOMB.=s
sor	sor2	PRE	MORPH=empty
eus	il	PROper	PERS.=3|NOMB.=p|GENRE=m|CAS=i
por	por2	PRE	MORPH=empty
lor	lor2	DETpos	PERS.=3|NOMB.=p|GENRE=f|CAS=r
terres	terre	NOMcom	NOMB.=p|GENRE=f|CAS=r
saisir	saisir	VERinf	MORPH=empty

Figure: Annotation Example

Tagging Medieval French Texts

Looking Ahead to 2024: New Challenges

- Enhancing interoperability of datasets
  - in synchrony
  - in diachrony
- Converting training data into Universal Dependencies format (special thanks to Lucence Ing at ENC)
  - **ils** s'enfuient
  - From Cattex : PERS.=6|NOMB.=p|GENRE=m |CAS=r
  - To UD : Case=Acc|Gender=Masc|Number=Plur|Person=6.
- Continuing work on comprehensive documentation (work in progress)

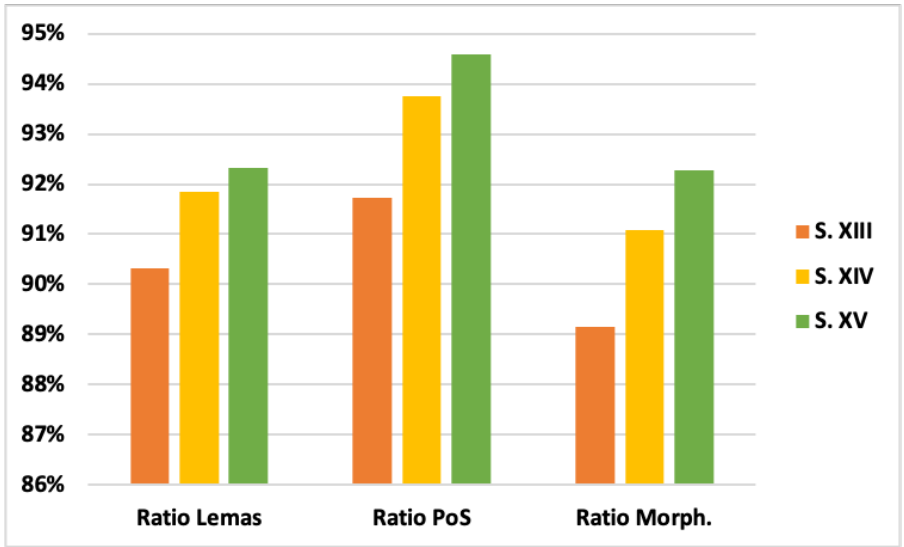
Tagging medieval Castilian text: state of art of annotated corpora and taggers

- Tagged corpora
  - *Corpus Del Diccionario Histórico de La Lengua Española* (CDH) `http://corpus.rae.es/cordenet.html`
  - *Old Spanish Textual Archive* (OSTA) `http://osta.oldspanishtextualarchive.org/`
  - *Corpus de documentos Españoles Anteriores a 1900* (CODEA+ 2022) `https://www.corpuscodea.es/corpus/corpus2022/consultas.php`
- Open source tagger(s):Freeling
  - Adapted to medieval castilian (Sánchez Marco 2012)
  - makes use of the EAGLES tagset.

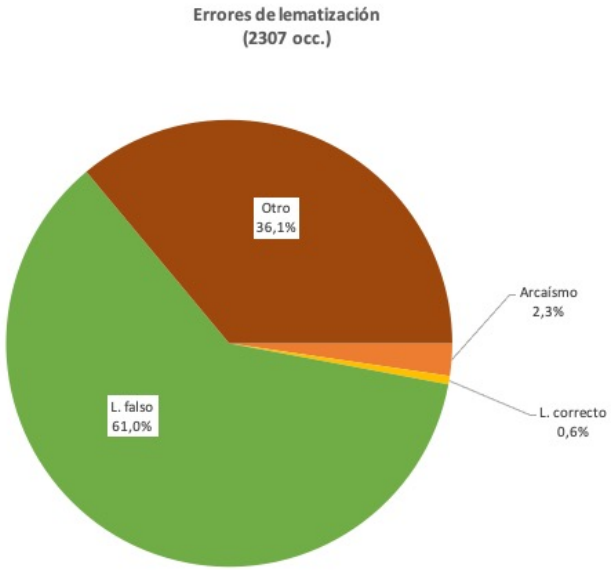
Quality evaluation of Freeling module

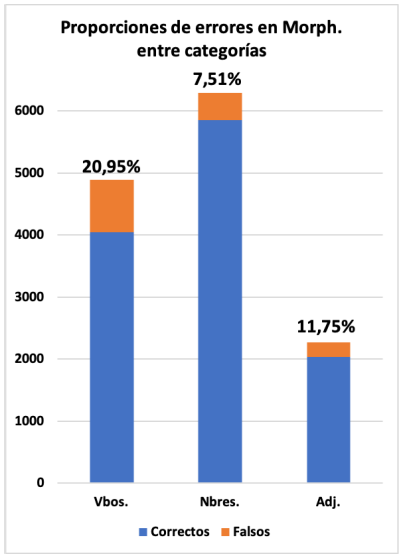
Matthias Gille Levenson, Olivier Brisville-Fertin, Maria Díez Yáñez, and Simon Gabay. "Construcción de Un Corpus de Evaluación de La Anotación Léxico-Gramatical Del Castellano Medieval (Siglos 13-15)". In: V Congreso de La Sociedad Internacional de Humanidades Digitales Hispánicas. Santiago de Compostela, 2021

- Sampling of modern editions with different norms, by extracting 200-250 words from 166 works (13th-15th c.) and manual correction:
  - 35 097 tokens;
  - 3 800 distinct lemmas;
  - 7 405 distinct forms.
- Results
  - lemmas: 92,28%;
  - PoS: 94,65%;
  - morph: 92,29%.



Qualitative assessment of Freeling module





- Allows to annotate text efficiently
- Performs poorly with homography
- Really dependant on the transcription norms of the corpus
- It should be possible to improve the quality of annotation with more moderns taggers
  $=>$ An exploratory project: "Étiquetage lexico-grammatical du castillan médiéval"

The e-CaM project: presentation

- Presenting the project:
  - A 19 member team in Medieval Philology, Linguistics, and DH
  - Two collective workshops
  - A first correcting test-campaign

- Our incentives:
  - Following the current advances in tagging for other Medieval languages
  - Reflecting on available tagsets, their potentials and pros and cons
  - Updating the free and open available tools for Medieval Spanish

E-CaM: objectives

- Planned achievements and foreseen challenges:
  - An established Medieval gold-corpus which implies to achieve a representative corpus (datation issues, examples size) of 1.5 M tokens
  - A documented and detailed manual for tagging: considering the updating of the EAGLES tagset and the conversion to Universal Dependencies
  - A first basis for correcting the whole corpus, in order to train lexical and grammatical tagging models

Thank you for your attention !

[1] Jean-Baptiste Camps, Thibault Clérice, Naomi Kanaoka, Ariane Pinche, Frédéric Duval, and Lucence Ing. "Corpus and Models for Lemmatisation and POS-tagging of Old French". en. In: (2021). URL: https://shs.hal.science/halshs-03353125 (visited on 10/11/2023).

[2] Thibault Clérice, Jean-Baptiste Camps, Ariane Pinche, Lucence Ing, Frédéric Duval, and Naomi Kanaoka. *Deucalion Model Ancien Français ENC*. original-date: 2018-10-29T11:06:15Z. Dec. 2020. URL: https://github.com/chartes/deucalion-model-af (visited on 10/29/2021).

[3] Simon Gabay and Thibault Clérice. "CATMuS-Print [Large]". fra. In: (Jan. 2024). Publisher: Zenodo. URL: https://zenodo.org/records/10592716 (visited on 03/27/2024).

[4] Matthias Gille Levenson, Olivier Brisville-Fertin, Maria Díez Yáñez, and Simon Gabay. "Construcción de Un Corpus de Evaluación de La Anotación Léxico-Gramatical Del Castellano Medieval (Siglos 13-15)". In: V Congreso de La Sociedad Internacional de Humanidades Digitales Hispánicas. Santiago de Compostela, 2021.

[5] Enrique Manjavacas. *emanjavacas/pie*. original-date: 2018-04-25T13:52:41Z. Mar. 2024. URL: https://github.com/emanjavacas/pie (visited on 04/02/2024).

[6] Ariane Pinche, Thibault Clérice, Alix Chagué, Jean-Baptiste Camps, Malamatenia Vlachou-Efstathiou, Matthias Gille Levenson, et al. "CATMuS Medieval". lat. In: (Nov. 2023). Publisher: Zenodo. URL: https://zenodo.org/records/10066219 (visited on 01/08/2024).

[7] Ariane Pinche, Thibault Clérice, Alix Chagué, Jean-Baptiste Camps, Malamatenia Vlachou-Efstathiou, Matthias Gille Levenson, et al. *CATMuS-Medieval: Consistent Approaches to Transcribing ManuScripts*. en. Dec. 2023. URL: https://inria.hal.science/hal-04346939 (visited on 01/08/2024).

[8] Cristina Sánchez Marco. "Tracing the Development of Spanish Participial Constructions: An Empirical Study of Semantic Change". Barcelona: Universitat Pompeu Fabra, 2012. URL: https://www.tdx.cat/bitstream/handle/10803/97044/tcsm.pdf?sequence=1 (visited on 09/16/2019).

[9] Sonia Solfrini and Simon Gabay. "CATMuS Gothic Print". frm. In: (Jan. 2024). Publisher: Zenodo. URL: https://zenodo.org/records/10599911 (visited on 03/27/2024).