Article

Minimal Computing and Weak AI for Historical Research: The Case of Early Modern Church Administration

Christoph Sander

German Historical Institute in Rome, 00165 Rome, Italy; c.sander@dhi-roma.it

Abstract

This paper introduces an AI-assisted human-centered and minimalist software stack and data model to structure and store early modern serial sources related to early-modern Catholic Church administration. The Vatican Archive preserves vast quantities of documents recording its administrative history. To date, the sheer volume and technical character of these Latin manuscripts have made systematic study appear nearly impossible. The multinational project GRACEFUL17 unfolds seventeenth-century Church governance on a large scale with the help of AI. It leverages simple but efficient NLP (NER, span categorizer, fuzzy searches) and classifier (gradient boost) techniques that run fast, reliably, and reproducibly to allow for multi-user offline work environments, as well as quick but controlled data modelling in a knowledge graph. By documenting this workflow, the paper enhances replicability and provides a rationale for specific design decisions beyond technical documentation. This paper advocates the use of "weak AI" on several grounds. Functionally, non-LLM pipelines offer stricter controllability and avoid many of the semantic biases introduced by large language models. They also require fewer training overheads and run locally with ease. Methodologically, the combination of simple AI models and symbolic reasoning underscores the indispensable role of human expertise: only experts can provide the ground truth necessary for models to reproduce and formalize complex semantic concepts and phenomena, rather than outsourcing this interpretive work to foundation models.

Keywords: serial sources; Latin; history; named entity recognition; classifier; digital history; minimal computing

1. Introduction

By the time of writing this article in summer 2025, large language models (LLMs) are increasingly discussed and used as tools in the digital humanities, dealing with textual, visual, and multimodal data. However, the historical humanities face a distinct set of challenges in this context (e.g., Simons et al. 2025; Valleriani 2025; Karjus 2025); relevant textual sources are often not available in machine-readable formats, the languages in which these sources are written are frequently underrepresented in LLMs, and semantic concepts in LLM pretrained data are not chrono-sensitive and may introduce hermeneutical biases (Oberbichler and Petz 2025). On an educational level, de-skilling scholars by rewarding specific usage of AI is a risk being discussed (Kosmyna et al. 2025; Selim et al. 2024; Marin and Steinert 2025). On a meta-scientific level, expectations within the scholarly community for reproducibility (Ries et al. 2024) and open algorithmic usage are rarely met by today's most powerful LLMs. On a societal level (Kieslich et al. 2024; Powers et al. 2025), LLMs

hardly live up to the principles associated with minimal computing, e.g., the avoidance of expensive or power-intense frameworks (Risam and Gil 2022; GO::DH Minimal Computing Working Group 2022). On a legal level, training LLMs on copyright protected scholarly publications is a grey area at best (Lehmann and Sichani 2025).

Many of these challenges are being addressed, among others, by curating large historical datasets for less biased training data (Langlais 2023), by refactoring cutting-edge AI engineering towards sustainability (Suchikova et al. n.d.) and by enforcing openness politically (TildeAI 2025; Langlais et al. 2025), etc. Instead of decisively working towards these ends, this article proposes an alternative approach—not a better one—that puts a premium on computational minimalism (software engineering), control and interpretability (knowledge engineering), and education-driven data curation (social engineering), without cutting off all of the potential AI offers to digital humanities scholarship.

It does so by taking the example of an ongoing multinational project, GRACEFUL17.[1] The project *required* the proposed design for specific reasons. It deals with massive amounts of unedited Latin serial sources that LLMs fail to properly understand linguistically and conceptually, because these sources have never been edited in large numbers before, let alone digitally. Source materials are protected by Vatican law, rendering automated text recognition legally difficult and expensive, not to mention their partly incoherent and complex document and layout structures, as well as paleographic and material heterogeneity.[2] Manual transcriptions thus had to be performed on-site, with no internet connection available, and bookings were to be made months in advance and with limited availability. The major research workload sits on the shoulders of three history PhD students who needed to familiarize themselves with the domain and methodology; no prior DH experience was taken for granted. Even more so, the requirements mentioned placed pressure on the timeline, pushing for efficient research and software design to be available on day one, while remaining flexible throughout.

This may sound a peculiar, if not dramatized, setup, and one that is not generalizable for DH projects.[3] It truly is not generalizable in many regards, but it is in others. It forces technical implementation to adjust to some realities of humanities research projects instead of the other way around. It has to take into account how emerging scholars not only reach project milestones but are also being educated according to their own preferences and those of the scientific communities with which they identify. It has to implement software to pace up workflows and support research but also work under conditions that may feel anachronistic but are not: namely, on slow and offline personal computers used by non-experts.

This article, however, is not a project report, but a problem-oriented discussion of one specific question: given the mentioned requirements, how can AI support the research, especially for related data engineering and modelling, but also for data analysis? I argue that so-called "weak" AI, including "good old-fashioned" symbolic AI (GOFAI), can efficiently generate structured data from texts. The software stack and resulting data may not be cutting-edge in themselves, but the swift, reliable, and low-cost implementation marks a recent advance. It enables a level of scalability that would have been unrealistic only ten years ago.

This article is mostly technical, with more detailed information provided in a data paper and future research papers (Sander et al. 2025b, Forthcoming; Sander and Hörnschemeyer 2025). By being technical, the article lives up to the need for (more) replicable research in the humanities (Peels 2019). It often is not enough to publish research results based on computational techniques that cannot dig into how data were created and which motivations and requirements were driving the upstream tasks. Similarly, code repository or software documentation does often conceal practical considerations that led to certain

design decisions. This article hence attempts to flesh out the design decisions that led to the project's current software implementation.

2. Tasks and Requirements

Put very simply, the Vatican Archive held Latin sources under scrutiny in the GRACE-FUL17 project report on administrative processes within the Curia of the Roman Catholic Church (Landau 1980; Viana 2018; Balavoine 2011). Drawing on two samples (1622–23, 1677–78)[4] and entailing roughly two thousand pages of densely written serial entries in the twenty thousands, each record ("entry") testifies to the provision or dispensation of certain entitlements: most importantly, the allocation of vacant lower church offices such as rectorships. They encompass administration, governance, and bureaucracy for the entire Catholic world (including non-European territories), and were recorded by the so-called Apostolic Dataria in Rome, subject to the Pope's authority (Fink and Mercati 1951; Pásztor 1970; Storti 1969). A typical entry of ca. 250 characters on average, for example, reports the death of a person holding a church office in a diocese and the provision of this church office to someone else, while indicating relevant dates, motivations, stakeholders, obligations, and legal circumstances. The underlying semantic model conceptualizes provisions as events and the offices provided as an (immaterial) objects (Sander and Boute 2025).[5]

From a data engineering perspective, these entries contain standardized (or standardizable) categorical, geographical, social, numerical, and temporal data, and there is little to no narration or stylistic diversity. Some of this contextual information describes the event (e.g., dates, persons, places), while others describe the event's object (e.g., institutional embeddings, certain categorical data). Yet, none of these data are delivered in tabular shape but in morphosyntactic sentences. For any desired heuristic text-as-data use, the workflow is transforming these linguistic expressions into structured knowledge, i.e., first tabular and then graph data. For doing so, researchers are ideally supported by software in several tasks (T):

T1: Extracting relevant semantic information (dates, persons in different roles, categorical information of different types, geographical entities, temporal information, etc.) to information atoms or semantic elements.

T2: Harmonizing, standardizing, and normalizing these elements across all their occurrences, e.g., by transforming dates into machine-readable formats, identifying identical persons, places, territories, etc., and potentially linking them to authority data (Wikidata, etc.).

T3: Reassembling the elements within one or across multiple entries to larger semantic or ontological structures (knowledge graph), such as combining elements that describe distinct events (provisions involving stakeholders, places, dates, etc.) and distinct objects these events deal with (church offices having a specific legal framing or institutional embedding).

T4: Using the resulting knowledge graph as a basis to analyze the semantic data to provide various insights, such as quantifications and rankings, geographical patterns, social networks, complex patterns and clusters, temporal trends, etc.

The modal, methodological, and epistemic requirements (R) for all these tasks include the following:

R1: All automation needs to be reproducible, fast, and available offline on multiple personal computers that synchronize and harmonize their individual data later on.

R2: Model predictions are being validated, and these validated and improved results must feed back into the training data to retrain a model. These predictions must also be interpretable and deterministic to the degree that they truly reflect the shape and quality of the training data and no other semantic context.

R3: Data models and ontology must truly and only conform to assumptions by domain experts, while conforming to formalized and machine-readable standards (e.g., OWL).

R4: Analytical or heuristic functions must be deterministic and simple enough to be validated by doctoral students with no expertise in mathematics or computer science.

While all these tasks and requirements are interdependent and must all be kept in mind for any digital solution stack, in what follows, the focus will be on T1–3 and R2. Details on software and hardware solutions (R1), on data modelling (T/R3), and on the analysis stack (T/R4) go beyond the scope of this article and will be sketched only briefly where they depend on the engineering design in focus here (Sander 2024a; 2024b).

2.1. Task 1: Extracting

With input as Latin sentences' data extraction itself is a natural language processing (NLP) task: named entity recognition (NER, one-label) and span categorization (SpanCat, multi-label). In essence, both pipelines predict an annotation layer applied to a text string by identifying tokens or character-offset spans as specific, predefined ontological classes or labels. Except for LLM-based pipelines (Hiltmann et al. 2025; Tudor et al. 2025; Xie et al. 2024; Jaskulski et al. 2025; Zhu et al. 2025), these methods do not require pretrained transformer architectures. In fact, the GRACEFUL17 spaCy models for NER and SpanCat depart from the language-agnostic blank models (Honnibal et al. [2014] 2020).

To truly fulfil R2, the models (Sander 2025f) themselves remain agnostic with regard to underlying ontological assumptions, except for those in the training data.[6] They are not confined to identifying just, say, persons or places, as they are statistical models that learn to recognize any linguistic pattern from the provided training data and apply these patterns to predict and extract relevant structures in new documents. Hence, the labeling schema for the GRACEFUL17 data goes beyond mere class annotations by also incorporating the entity's semantic context as roles, aspects, or regards—enabling a more nuanced interpretation of the entities in context. This is best seen from an example (Figure 1 and Table 1).[7]
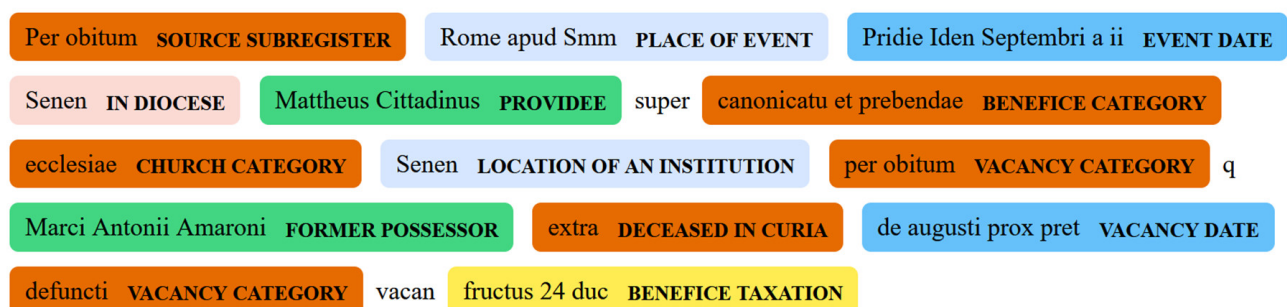

Figure 1. Rendering the named entity recognition spans with (Explosion [2016] 2025) as HTML. [created by Christoph Sander 2025]. Label tags are written in bold and underscript caps, colors represent ontological classes of recognized entities.

Table 1. Named entities and their meaning. Colors in the Class column match colors in Figure 1 and represent an entity's ontological class.

Class	Role/Label	Literal/Span	Encoded	Description
Person	Providee	*Mattheus Cittadinus*	Family: Cittadinus; Given: Mattheus	The name of the individual being appointed.
Person	Former Possessor	*Marci Antonii Amaroni*	Family: Amaronus; Given: Marcus Antonius	The name of the former possessor of the office.

Table 1. *Cont.*

Class	Role/Label	Literal/Span	Encoded	Description
Date	Event Date	*Pridie Id Septembri a ii*	1622-09-12 (i.e., 2nd year of pontificate Gregory XV)	The date of the decision (granting of the supplication).
Date	Vacancy Date	*de augusti prox pret*	1622-08	The date of the vacancy.
Place	Place of Event	*Rome apud SMM*	Rome, Santa Maria Maggiore	The administrative location of the dating/granting of the papal grace.
Place	Location of Institution	*Senen*	Siena	The location of the benefice's holding institution, e.g., a church.
Institution	In Diocese	*Senen*	Diocese of Siena	The diocese holding the benefice.
Type	Benefice Category	*Canonicatu et praebenda*	Canonship	The awarded benefice category.
Type	Church Category	*ecclesiae*	Church	The ecclesiastical institution to which the office is attached.
Type	Vacancy Category	*per obitum/defuncti*	Death of predecessor	The reason for the vacancy and office reassignment.
Type	Deceased in Curia	*extra*	Outside the Curia	Indicates whether the death occurred inside or outside the Curia.
Type	Source Subregister	*per obitum*	"Per obitum" sub-register	The sub-register from which the data originates.
Monetary Value	Benefice Taxation	*24 duc*	24 ducats	The tax valuation of the office in Apostolic Chamber's currency.

Instead of calling predicted annotations "entities", we refer to them as "elements", underscoring their wider semantic reference: atomic text segments (tokens or spans) tagged with semantic labels within a specific context and conceptual framework. Labels not only capture dates or persons (i.e., entities), but, amongst others, dates of different meaning and persons in different roles (i.e., elements). The mapping of these labels (e.g., "former possesor") to classes (e.g., "person", marked green in Figure 1 and Table 1) is subject to ontological definitions and, as such, not part of the machine learning pipeline but downstream ontology-based symbolic reasoning.

(spaCy's) NER architecture assigns at most one label per span, whereas (spaCy's) SpanCat can predict multiple labels per span, increasing the learning complexity significantly. Semantically, true multi-label predictions are rare but still necessary at times, which can make the standard NER pipeline unsuitable (e.g., when the category of a benefice is inferred from the category of the holding institution, or vice-versa, and hence one span for either is labelled twofold). In fact, chaining or stacking multiple NER models to approximate a multi-label setup has often yielded better results than using a single multi-label SpanCat pipeline. Potential labelled span overlaps only occur among certain labels, so that spectral clustering can find non-overlapping labels from a given training dataset and recommend label groups.[8] This allows for the training of partial NER models, focusing on specific labels only, and calling these models for processing sequentially, each adding further elements with disjunct labels. In this way, more than one label per span can feature in the result without having to rely on SpanCat and its less robust prediction results. A single NER model trained to predict all labels (often used for its faster runtime compared to a stacked pipeline) still achieves strong overall evaluation scores, with precision, recall, and F1 ranging between 0.91 and 0.93.[9]

The initial creation of training data is an inertia of the proposed pipeline (Dombrowski 2022), as it requires defining relevant labels and applying these as spans to textual entries.

However, neither of these steps can be outsourced to anyone but a domain expert and research team member: at least, not at the initial stage of establishing each label's semantics and scope. In fact, we created a few initial examples and then used an LLM for rapid prototyping. The ChatGPT o1 model few-shot generated training data were not stored in the project's database but were manually evaluated and only used for the initial training of the NER model. After a decent accuracy of our model, it was used and retrained with no further support from LLMs or external resources. Validating and correcting model predictions immediately and frequently retraining the model quickly led to a robust pipeline benefitting from all team members entering data.

2.2. Task 2: Normalizing and Linking

Following entity detection, the next step is entity linking (EL), which anchors semantic elements to named entities within a knowledge base (KB) (Cucerzan 2007; Hachey et al. 2013; Parravicini et al. 2019; Rao et al. 2013; Zwicklbauer et al. 2016). For well-known entities such as toponyms or historical figures in public KBs like Wikidata, machine learning-based linking might ensure high accuracy. In the GRACEFUL17 data, however, directly linking its semantic elements to resources in an existing KB is limited. Its abbreviated and Latin tokens are difficult to match with authority files and, more generally, many of the GRACEFUL17 entities are too domain-specific and even introduce most entities as authority data to the public sphere: persons, churches, and technical concepts, in particular (R3), are often unheard-of. Encoding Latin papal calendar dates and relative dates (from temporal adverbs such as "recently" or "ten or twenty years ago") into absolute XML schema dates is as challenging as mapping variously transcribed person names to identical persons or coping with multiple terms used for the same categorical concepts from Canon Law. Rule-based approaches can only help to prepare or validate data, and even the best LLMs lack the context-specific knowledge to achieve these tasks to the satisfaction of domain experts.

Instead, the GRACEFUL17 project uses a dictionary approach, in which domain experts initially map elements/tokens to an internal KB entity that might link it to a public resource or encode its information according to a semantic web schema. Building on this ever-expanding dictionary, a fuzzy Levenshtein-based (Levenshtein 1966) and customized thresholded matching suggests and links tokens/elements to these predefined KB entities, mapping various tokens to identical or different entities. Hence, over time, the system "learns" through dictionary expansion, as new resemblance patterns are added to the fuzzy matching framework. This matching is highly customized: for specific classes or labels, a lower threshold for fuzzy matching is more useful than for others. For example, place names might account for some fuzziness due to transcription errors or grammatical variance, while date strings must perfectly match, as one different character might have already denoted another day. Some matchings only apply for entries of specific archival sources, regular expressions can be plugged into token preprocessing, ontological rules may constrain possible target entities in the pool of target entities, and many other rules can govern this user-centered, semi-automatic, and bulk-optimized matching process. R1 and R2 are not only met but overfulfilled, as the matching mechanism requires no training whatsoever and is always available in real time. It is important to pool newly created entities immediately as targets for the next match, instead of depending on asynchronous recursive training. The costs on the performance side are mitigated by batch-processing elements, instead of performing this step at data-entry time. Currently, ca. 200,000 elements are linked to more than 10,000 entities (incl. dates).

This hybrid model allows for scalable and context-aware linking of structured textual elements to curated semantic data, which is particularly crucial in fields like archival

studies or prosopography, where custom taxonomies dominate, and data are often the initial ground truth for entities' existence in the semantic web.

2.3. Task 3: Grouping

Extracted elements linked to entities, however, remain flat data as a "bag of entities", as they simply map to the text they were extracted from. But there are deeper structures relating elements amongst each other. When an entry is reduced to its constituent elements, the morphosyntactic context in which they originally appeared is largely lost. While labels introduce contextual information, such as roles, and fuzzy matching/mapping enables cardinality and semantic identity, the identification of patterns and deeper relationships among these elements requires more complex understanding. As explained above, textual entries testify to one or many events that concern one object each. The events and objects are conceived as containers of elements (typically, n elements: n objects OR n events) and have a relationship with each other (typically, n events: 1 object).[10] The ground truth for these deep structures is the linguistic expression of the entry (Figure 2).
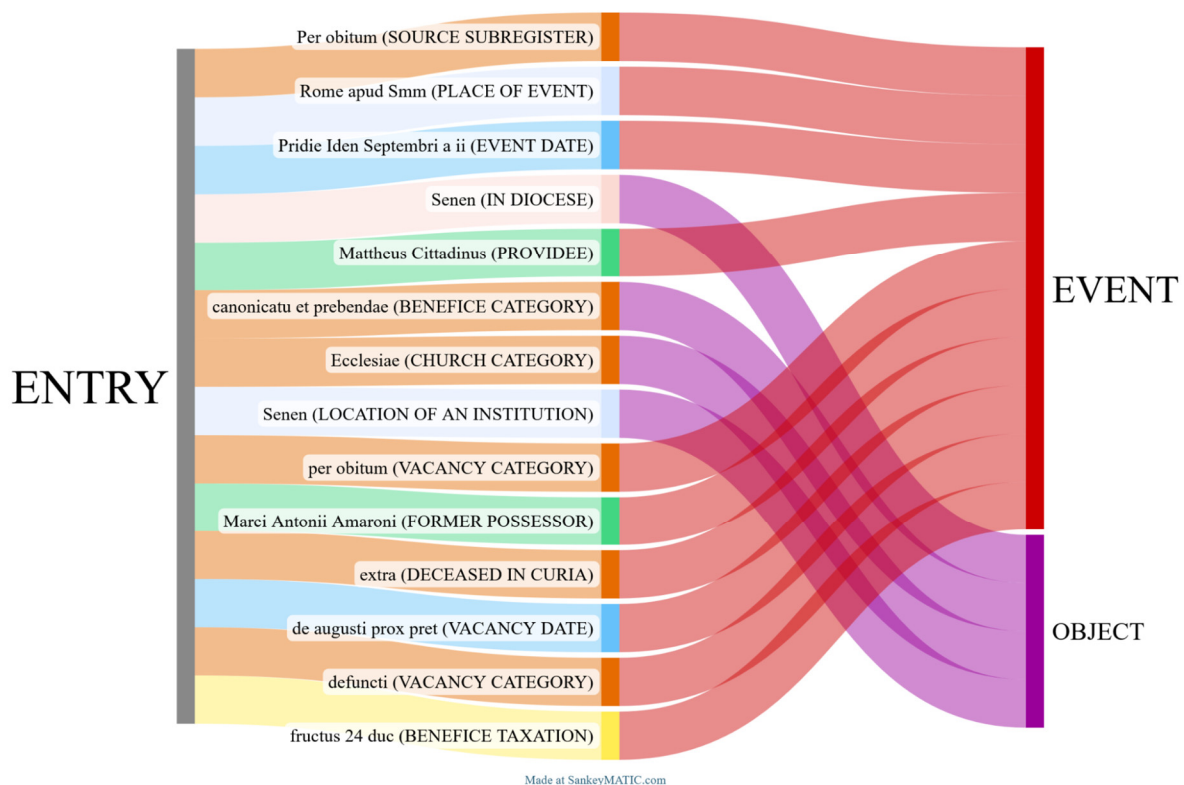

Figure 2. Figure of the assignment of elements to events/objects for a simple case of one object and one event. Sankey diagram (Bogart 2014) created by Christoph Sander.

An end-to-end modelling of these events and objects from the textual input directly risks violating R2. Using, for example, an LLM Graph Transformer (e.g., Chase 2022; Luo et al. 2025; You et al. 2025) that builds on the preexistent extracted elements or does this semantic extraction from text spontaneously would require some linguistic parsing of the Latin text—a requirement that is likely too ambitious to meet without compromising the pipeline's determinism and low-bias design. Training a blank generative language model for modeling deep structures (as for events/objects) is likewise unfeasible, given the insufficient volume of training data and the substantial computational resources that such a task would require.[11]

Instead, we reframe this task as a multilabel classification task. At first glance, this label assignment might miss the point: although complex linguistic semantics can in principle be encoded via embeddings, off-the-shelf classifiers struggle to leverage it directly. However, exactly this contextual information for ordinary classifiers appears to be hard to parse. The pipeline proposed here therefore combines an NLP-tuned feature preprocessing with a gradient boost classifier. It aims for a model that minimizes computational overhead while maximizing performance and interpretability.

The prediction target for each element is a particular deep structure (n elements to n events/objects), combined with a primary type that further specifies the event/object. This target variable is technically modeled as a single, composite label that unifies three pieces of information ("triplet"): the class (i.e., event or object), the specific subtype within that class (i.e., denoting the type of event/object, e.g., a provision or benefice), and an index per class for each entry (i.e., the ordinal number of an event/object per sample), e.g., "EVENT_APOSTOLIC-PROVISION_2nd" for an instance of the event event class, with the primary type "apostolic provision" with the index 2, meaning there is an entry testifying to more than one event and this element belonging to the second event. Each element can be assigned to multiple of these triplets.[12] This multi-label classification pipeline employs a multi-label binarizer (MLB) and a one-vs-rest classifier (OvR) (using Pedregosa et al. 2011).[13] It hence predicts multiple output labels per input data, i.e., the desired n elements to n events/objects mapping pattern.

What are the input data for this classifier to base its prediction on, or how does the model learn what is an element's correct triplet label? The model architecture is not specifically designed for linguistic input, but for vector and categorical data. To incorporate syntactic context into the classification task, a made-to-measure feature of preprocessing is required to transform heterogeneous information into a unified feature space (Listing 1).

Listing 1: Feature engineering for the CatBoost classifier, using Python packages pandas, catboost, and sklearn. See full code in src/Classifyer.py in (Sander 2024b).

```python
df['entry_length'] = df['entry'].apply(len)
df['start'] = df['start'] / df['entry_length']
df['end'] = df['end'] / df['entry_length']
df['entity_count'] = df.groupby('entry_ID')['text'].transform('count')
df['avg_start_position'] = df.groupby('entry_ID')['start'].transform('mean')
df['avg_end_position'] = df.groupby('entry_ID')['end'].transform('mean')
df['all_texts'] = df.groupby('entry_ID')['text'].transform(lambda  x: ' '.join(x))
df['all_labels'] = df.groupby('entry_ID')['label'].transform(lambda  x: ','.join(sorted(x)))
df['all_labels_count'] = df['entry_ID'].map(df.groupby('entry_ID')['label'].agg(list))
preprocessor =  ColumnTransformer(
    transformers=[
        ('text', TfidfVectorizer(token_pattern=r"(?u)\b\w+\b"), 'text'),
        ('label', OneHotEncoder(handle_unknown='ignore'), ['label']),
        ('all_texts', TfidfVectorizer(token_pattern=r"(?u)\b\w+\b"), 'all_texts'),
        ('all_labels', OneHotEncoder(handle_unknown='ignore'), ['all_labels']),
        ('all_labels_count', CountVectorizer(token_pattern=None,
tokenizer=lambda  labels: labels, lowercase=False), 'all_labels_count'),
        ('start_end', 'passthrough', ['start', 'end']),
        ('context_features', 'passthrough', [
            'entry_length', 'entity_count', 'avg_start_position', 'avg_end_position'
        ])
    ])
```

The element's own (short) text is vectorized by TF-IDF (Spärck Jones 1972) so that its most significant terms are captured ("text"), while a second TF-IDF representation of

the entire entry ("all_texts") injects broader contextual patterns and co-occurrences of all other element texts in the same entry. These natural language (pre)processing steps serve to capture semantics without a (larger) language model introducing semantics that derive from external linguistic data. The normalized significance of certain expressions either in the source element directly or its neighbor's inform decision-making based on linguistic patterns. By restricting TF-IDF to validated NER-extracted tokens, it ignores all the additional entry text that is not considered to convey relevant semantic information, effectively limiting the vector space and applying a domain-specific stop word list to the full entry text *ex positivo*.

In parallel, the categorical label of the respective element ("label") and the full collection of all entry labels per entry ("all_labels") are converted into one-hot vectors. The former ensures that local role information is available, which is important as most labels are only to occur in either events *or* objects. The latter presents a fingerprint comprising all labels per entry as a categorical vector feature, to allow inferences from exactly co-occurring label signals. Moreover, a count vectorizer ("all_labels_count") transforms all labels into a "bag of labels," exploding the one-hot fingerprint into a discrete and complex pattern of co-occurrences and frequencies. Finally, numerical context features, such as relative and average start and end positions within the entry, the overall entry length, and entity counts, are passed through directly, lending a notion of syntactic placement without actual linguistic parsing.

Once all features are assembled into a single matrix, the pipeline applies CatBoost as its core classifier (CatBoost 2025). Its gradient-boosting copes with class imbalances to mitigate unequal stratification in the training data by iteratively re-weighting misclassified instances. Wrapping it in the afore-mentioned OvR strategy, each binary CatBoost model (Sander 2025e) handles one label, class-imbalance corrections are applied per label, and weighted average evaluation scores range above 0.95.[14]

The pipeline also yields per-label confidence scores for interpretability as feature importance (on "explainable AI", see Aviyente and Karaaslanli 2022; Eberle et al. 2022; Lundberg et al. 2020), revealing that an element's own label ("label") dominates prediction, which is reasonable as the label quite likely at least determines whether to predict either an event *or* an object target label. In this way, the pipeline relaxes the unpredictability of deep learning, let alone of generative language models, or the combinatorial explosion of label-powerset approaches. Instead, it offers a deterministic, extensible framework that harmonizes local and global text context to infer deep event and object structures.

2.4. Task 4: Understanding

The structured and semantic data retrieved from semi-automatic pipelines covering NER, entity linking, and deep structure classification are, after all, serialized as RDF triples (RDF Core Working Group 2014) conforming to an OWL ontology (Sander and Boute 2025; Sander et al. 2025b). A rule-based parser transforms relational tables into a JSON-LD representation (Sporny et al. 2020), ingested by an Oxigraph RDF 1.2 and SPARQL 1.1 compliant triple store (Pellissier Tanon 2025). This transformation includes further ontology-driven processing and reasoning steps, but is largely based on the data created from the three core pipelines, in a reversed order: events and objects are described by semantic entities (T3) from a knowledge base which they themselves represent, or proxy the elements (T2) that derive from the textual sources (T1). Significantly simplified provisions are composed of actors in a spatiotemporal context, relating to offices of some provenance and type–semantic 'simple event model'-like compounds (van Hage et al. 2011), made of textual evidence via normalized elements.

The analysis of the resulting knowledge graph (ca. 30 million triples) starts with queries that count the occurrence of certain entities based on defined conditions and filters (Sander 2025b, Sander [2024] 2025c; Sander et al. 2025a). For example, researchers may wish to determine how many provisions occurred within a specific diocese for a particular office during a given period. Such research question-driven querying is not only intended to produce basic quantitative overviews of the data but is essential for substantiating the historical hypotheses and arguments developed by the project's PhD researchers. Their individual studies draw on a shared "global" data sample as a framework for examining the scope of Roman global governance, in relation to competing local dynamics. Accordingly, they require well-structured data to identify geographical and temporal trends and to compare administrative practices across regions and typological dimensions.

Such simple research questions are readily expressed as SPARQL queries, supported by visual query-building and result viewers (Sander 2025b; Harris and Seaborne 2013; Sparna 2025; TriplyDB [2014] 2025). In more complex (so far, only prototyped) scenarios, the focus may shift toward identifying trends through time series analyses, exploring statistical correlations or regressions, and performing clustering across specific aggregated dimensions. For instance, researchers ask whether there is a particular periodicity or seasonal pattern in the frequency of provisions over the course of a year, either in general or with regard to specific offices or dioceses. Do provisions for certain dioceses correlate with particular types of offices? Is the time lag between the occurrence of a vacancy and its eventual provision correlated with the geographical distance between Rome (the central ecclesiastical authority) and the remote diocese (in which the vacancy arose) (Sander 2025a)?

When aggregating provisions for dioceses (i.e., grouping individual events by the institution they relate to), a distinct profile emerges for each diocese, based on the semantic information about all aggregated provisions. While features such as dates or individual persons are difficult to compare, due to their low cardinality and present nature (i.e., they are largely unique to single provisions), categorical attributes lend themselves well to constructing what may be described as a "fingerprint" of each diocese within a given time frame. This fingerprint represents all recurring categorical information across the provisions of a diocese as a vector, with as many dimensions as there are relevant features. These fingerprints result from the combination and interaction of the categorical features associated with provisions (events) and ecclesiastical offices (objects). Using unsupervised machine learning techniques, such as HDBSCAN (McInnes et al. 2017), clusters can be efficiently identified within these high-dimensional feature spaces as vectorized fingerprints. Such clusters reveal, for example, dioceses that exhibit similar patterns of Roman governance—patterns whose interpretation ultimately depends on expert historians, who must assess whether these computational findings align with, contradict, or refine existing historiographical narratives.

While such research scenarios partly exceed the immediate scope of the project, they open avenues for future research or even trigger curiosity by means of technical feasibility. Computational analysis of the data is not to replace or predetermine historiographical judgement and contextualization, no matter which technology is employed. Rather, a research-driven design of tailored or adapted algorithms by leveraging the data as actual proxies of semantics and historical affairs has already proven valuable as an exploratory "case study scout." Outliers, averages, correlations, and patterns have made project researchers aware of inherent peculiarities of the data (including unintended and covert biases and flaws that were corrected afterwards), as well as insightful examples that would have remained hidden in the immense ocean of RDF data otherwise.

Analytical approaches thus range from relatively simple frequency counts, via structured queries, to more advanced statistical procedures and standard methods of unsuper-

vised machine learning. As with the data extraction process, the key principle applies here as well: the degree of control and determinacy of the results relative to the underlying data and the explicitly defined parameters or queries is of immense pedagogical and epistemic value. Researchers obtain the results they *formally* request, not a semantic approximation derived from a model interpreting the request's semantics. Any graph retrieval augmented generation scenario (Hu et al. 2025; Peng et al. 2024), for example, would need to enable deterministic counting and produce statistical measures. Findability, as in large text databases, is not a primary research goal, and if it is so, it is not as a narrative semantic search, but as a query of a specific concept, place, date, person, etc.

3. Discussion and Conclusions

In the GRACEFUL17 project, artificial intelligence supports transforming linguistic entries from historical sources into rule-based ontological models that are subsequently queried and analyzed computationally. Named entity recognition and classification are supervised and based on sub-symbolic AI, whereas entity linking is implemented through rule-based, and thus symbolic, approaches. Data modelling and analysis leverage ontological reasoning and SPARQL queries as symbolic AI, but also employ sub-symbolic unsupervised machine learning, such as cluster analysis. None of these methods require pretrained models, LLMs, or foundation models; they rely solely on domain-specific training data and/or axioms defined by domain experts. While this implementation partly emerged from the project's specific requirements, it is also driven by a larger argument for a human-in-control instead of a human-in-the-loop research and software design.

The initial need to create training data presents a trade-off, as it is undeniably time-consuming. However, this investment of time is justifiable, because it urges researchers to engage deeply with the phenomena and objects they investigate. In this critical phase, PhD researchers in particular acquire the epistemic competencies necessary to understand the material profoundly and to uncover its *relevant* contexts. While single-shot or few-shot learning, or, more generally, the use of pretrained models, may seem like attractive time-saving short-cuts, they risk bypassing a meaningful, if not essential, phase of human learning and academic qualification. This holds true in particular for projects that deal with highly specific resources that are not simply understood by 'common sense' but rather genuinely expand the semantic pool of a given research domain.

In addition to this pedagogical justification, there is also a methodological consideration. The decision to forgo pretrained language models that translate prompts into tasks ensures a level of control that is particularly important in scientific research. The definition of tasks and the parameters governing their execution are established through explicit programming code, in close consultation with domain experts. This affords a degree of control over the modality and framework of parsing, training, or prediction that cannot be equally guaranteed through natural language prompts. End-to-end remains just this, even with the best XAI, model alignment, and prompting techniques. Looking inside the alleged black box of transformers and foundation models is not the same as controlling the machinery inside. The AI models described here are also limited in this control, but training data engineering, feature engineering, and hyperparameter tuning provide considerably more control with less computational effort. Moreover, its models are deterministic, in the sense that, once trained, they consistently produce the same predictions. Especially in collaborative research settings, this determinism constitutes a key condition for ensuring the homogeneity of the generated research data and for avoiding dependency on variables beyond the researchers' control. The weak AI models employed are, of course, not free from bias, as they ultimately rely on statistical mathematics (Kleymann 2025) and inevitably reproduce the biases present in their training data. These training-induced biases are

not a flaw, but an intended feature: they are not to be "balanced" against alignments or project-external data. This deliberate dependency ensures that the AI does not "collaborate" with researchers, but, rather, mirrors the researcher's own epistemic horizon. To put the point provocatively, research that seeks to advance knowledge on the basis of previously unexplored sources should not be diluted by the prior knowledge of pretrained AI models trained on Wikipedia or Reddit.

Finally, LLM-based pipelines also deeply transform the analytical competence to formalize and operationalize a research question (also known as "a scientific problem"). The act of translating a research goal into a machine-readable form is not a cumbersome detour that only introduces friction between the original idea expressed in a natural language and its computational implementation. Rather, it constitutes a valuable intellectual discipline to articulate a question with *formal* precision. Commensuration without a naïve pursuit of quantification remains a critical intellectual and deeply creative part of research (Espeland and Stevens 1998; Halevi Hochwald et al. 2023; Healy 2017; Merry et al. 2015; Mau 2018). This act of "measuring phenomena" through formal expression is not simply bypassed when a natural-language prompt is used. The requirement for machine parsing does persist, but it is internalized, interpreted, and eventually executed by the LLM "under the hood." In fact, direct access to the original source data via natural-language prompts would effectively outsource this intellectual work to the latent embeddings of a large language model. LLMs' unintended biases or misinterpretations, leading to an inappropriate operationalization of the research question, may be detected and coped with. Yet, in this LLM-centered workflow, researchers would probably lose the cognitive momentum that comes from wrestling with a question until it can be precisely formalized, or until its resistance to formalization becomes epistemically significant in its own right. The LLM or reasoning model, in contrast, will do "its best" and reply, no matter what, always.

The agentic AI model architecture that integrates a researcher's linguistic input, formal computational coding, querying, narrativization, and visualization into a single workflow is certainly a promising prospect: one might imagine uploading thousands of scanned archival images and simply prompting in English, "which diocese performed best?". Imagine it working perfectly and delivering exactly what experts expect or desire. Imagine it as "de-fetishised AI" (Guest 2025), representing nothing but an "AI-based assistant capable of facilitating an accelerated science lab for in-depth historical research, interpretation, and reconstruction" (Eberle et al. 2024, p. 9).[15] Even in this close-to-wishful-thinking scenario, it remains a societal and meta-scientific question as to whether the best scholarly output takes precedence over cultivating the best human scholars. The use of AI in humanities scholarship, that is, the "if", the "how much", the "how", and the "what for", is a corollary to that normative question. It does not prescribe any single answer—let alone "AI veganism" (Joyner 2025)—but offers a meta-scientific framework for developing and assessing an AI-assisted research design.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: RDF research data and the GRACE OWL ontology are published on Zenodo (Sander and Boute 2025; Sander et al. 2025b). The ontology's namespace is https://w3id.org/grace/ontology. See its documentation at https://w3id.org/grace/widoco and https:

//w3id.org/grace/ontospy (accessed on 1 September 2025). Core modules of the codebase for the ML pipelines are published in a publicly released GitHub repository (Sander 2024b, 2025d). The spaCy and CatBoost models are published on Hugging Face (Sander 2025e, 2025f).

Notes

[1] GRACEFUL17: Global Governance, Local Dynamics. Transnational Regimes of Grace in the Roman Dataria Apostolica (17th Century) is a transnational, Franco-German research project, funded by the Deutsche Forschungsgemeinschaft and the Agence Nationale de la Recherche, and directed by Birgit Emich (Goethe Universität Frankfurt a. M.) and Olivier Poncet (École Nationale des Chartes in Paris). Other partner institutions include the Deutsches Historisches Institut in Rome, the École Française de Rome, and the Université de Reims-Champagne-Ardenne. The project's digital humanities component is based at the German Historical Institute in Rome.

[2] As a matter of fact, HTR technology is currently being tried, with some success. Yet, parsing the transcribed text into semantic units, also accounting for incoherent bindings of the paper sheets that were written on, poses challenges on yet another level.

[3] The projects "Repertorium (Academicum) Germanicum" are similar in data and analog workflows but are not born digital. Yet, they are an obvious application for the workflows presented here (Höing 1991; Esch 1991; Schwinges 2015; Gubler and Schwinges 2017; Beckstein et al. 2022; Hörnschemeyer and Voigt 2023; Schmugge 2023; Reimann 1991).

[4] Archivio Apostolico Vaticano (AAV) Dataria Ap. (Dataria Apostolica) Expeditiones 2 and 9.

[5] The GRACE ontology resembles the Simple Even Model (van Hage et al. 2009a, 2009b, 2011) in its core assumptions.

[6] Obviously, the tokenizer has to be taken into account, too, but no pretrained language models are used. For even more tacit and opaque algorithmic influences, see Kleymann (2025).

[7] Figure 1 and the synopsis are taken verbatim from a forthcoming data paper (Sander et al. Forthcoming).

[8] For the clustering, I used sklearn.cluster.SpectralClustering(n_clusters=n_clusters, affinity='precomputed', assign_labels='discretize').

[9] P: 0.9186346171867679, R: 0.9307807344458423, F1: 0.9246677907346924.

[10] Certain ontological axioms define the framework for these deep structures. By definition, every event requires an associated object. In such one-to-one relationships, the assignment of specific labels to either an object or an event is deterministic. Machine learning proves especially useful in more complex cases, where rule-based disambiguation reaches its limits.

[11] Yet, promising tests conducted with Jochen Büttner in fact suggest a viable pipeline for using fine-tuned foundation models to efficiently conduct the same task. A joint publication is in progress.

[12] Although one could imagine predicting class, subtype, and index as three separate outputs, doing so would effectively multiply the number of outputs by three and force the model to learn dependencies across them. As classes are currently binary (event vs. object), cardinalities for types are quite high (few types recurring frequently), and indexes do not exceed ten (max. events/objects per entry), this target triplet keeps the number of target dimensions manageable and far lower than the full Cartesian product of separate class, subtype, and index predictions. By encoding each valid combination of class, subtype, and index as a single multilabel triplet, the prediction task is reduced to m independent binary decisions (one per triplet). This approach eliminates the need for the model to output and reconcile three interdependent values (class, subtype, index) for each element. In practice, a MultiOutputClassifier would require three separate heads and learn the intricate dependencies between them, increasing complexity and the risk of inconsistent predictions. The OvR-triplet approach circumvents these issues by treating each composite role as its own binary label while preserving the model's inherent capacity to assign relevant combinations of deep structures.

[13] The MLB takes each sample's set of true triplets and transforms it into a fixed-length binary vector. If there are m triplet labels in the training data for one element, this element's vector has m positions, and it thereby converts a variable-sized label set into the uniform, numeric format required. Once the targets have been binarized, the OvR wrapper constructs m separate binary classifiers—one for each triplet label. Each binary model is trained to distinguish "element belongs to label i" versus "it does not." During training, OvR simply reads off the corresponding column of the binarized target matrix produced by the MLB. At inference time, each of the m classifiers casts an independent vote on whether its label applies. The collection of positive votes is then recombined into the final multilabel prediction for each element.

[14] F1: 0.9738834762666144, P: 0.9897332440073945, R: 0.9636759179906388, support: 45507.

15   Both references are not to suggest disagreement. In fact, Eberle et al. (2024) present an AI-assisted research case relying on largely unsupervised ML.

References

Aviyente, Selin, and Abdullah Karaaslanli. 2022. Explainability in Graph Data Science: Interpretability, Replicability, and Reproducibility of Community Detection. *IEEE Signal Processing Magazine* 39: 25–39. [CrossRef]

Balavoine, Ludovic. 2011. *Des Hommes et des Bénéfices: Le Système Bénéficial du Diocèse de Bayeux au Temps de Louis XIV*. Bibliothèque d'histoire moderne et contemporaine. Paris: H. Champion. Genève: Diff. Slatkine.

Beckstein, Clemens, Robert Gramsch-Stehfest, Clemens Beck, Jan Engelhardt, Christian Knüpfer, and Georg Zwilling. 2022. Digitale Prosopographie. Die automatisierte Auswertung des Repertorium Germanicum, eines Quellenkorpus zur Geschichte geistlicher Eliten des 15. Jahrhunderts. In *Digital History. Konzepte, Methoden und Kritiken Digitaler Geschichtswissenschaft*. Edited by Karoline Dominika Döring, Stefan Haas, Mareike König and Jörg Wettlaufer. Berlin: De Gruyter, pp. 151–69. [CrossRef]

Bogart, Steve. 2014. *SankeyMATIC: Build a Sankey Diagram*. SankeyMATIC, Released. Available online: https://sankeymatic.com/build/ (accessed on 1 May 2025).

CatBoost. 2025. *CatBoost* (V. 1.2.8). Available online: https://github.com/catboost/catboost (accessed on 1 May 2025).

Chase, Harrison. 2022. *LangChain*. Jupyter Notebook. Available online: https://github.com/langchain-ai/langchain (accessed on 1 May 2025).

Cucerzan, Silviu. 2007. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. Paper presented at 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic, June 28–30; Edited by Jason Eisner. Vienna: Association for Computational Linguistics, pp. 708–16. Available online: https://aclanthology.org/D07-1074/ (accessed on 1 May 2025).

Dombrowski, Quinn. 2022. Minimizing Computing Maximizes Labor. *Digital Humanities Quarterly* 16. Available online: https://dhq.digitalhumanities.org/vol/16/2/000594/000594.html (accessed on 1 May 2025).

Eberle, Oliver, Jochen Buttner, Florian Krautli, Klaus-Robert Muller, Matteo Valleriani, and Gregoire Montavon. 2022. Building and Interpreting Deep Similarity Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44: 1149–61.

Eberle, Oliver, Jochen Büttner, Hassan el-Hajj, Grégoire Montavon, Klaus-Robert Müller, and Matteo Valleriani. 2024. Historical Insights at Scale: A Corpus-Wide Machine Learning Analysis of Early Modern Astronomic Tables. *Science Advances* 10: eadj1719. [CrossRef]

Esch, Arnold. 1991. EDV-gestützte Auswertung vatikanischer Quellen des Mittelalters: Die neuen Indices des Repertorium Germanicum. Vorbemerkungen zum Thema. *Quellen und Forschungen aus Italienischen Archiven und Bibliotheken* 71: 241–42.

Espeland, Wendy Nelson, and Mitchell L. Stevens. 1998. Commensuration as a Social Process. *Annual Review of Sociology* 24: 313–43. [CrossRef]

Explosion. 2025. *Explosion/Displacy. JavaScript*. Released April 8. Available online: https://github.com/explosion/displacy (accessed on 1 May 2025). First published 2016.

Fink, Karl August, and Angelo Mercati. 1951. *Das Vatikanische Archiv: Einführung in die Bestände und ihre Erforschung*, 2nd ed. Rome: Regenberg.

GO::DH Minimal Computing Working Group. 2022. Minimal Computing. *DHCC*, May 17. Available online: https://go-dh.github.io/mincomp/ (accessed on 1 May 2025).

Gubler, Kaspar, and Rainer Christoph Schwinges. 2017. Repertorium Academicum Germanicum (RAG): Un nuovo Database per un'analisi basata sul Web e per la Visualizzazione dei Dati. *Annali di Storia dell'Università Italiane* 21: 13–24.

Guest, Olivia. 2025. What Does 'Human-Centred AI' Mean? *arXiv* arXiv:2507.19960. [CrossRef]

Hachey, Ben, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran. 2013. Evaluating Entity Linking with Wikipedia. *Artificial Intelligence* 194: 130–50. [CrossRef]

Halevi Hochwald, Inbal, Gizell Green, Yael Sela, Zorian Radomyslsky, Rachel Nissanholtz-Gannot, and Ori Hochwald. 2023. Converting Qualitative Data into Quantitative Values Using a Matched Mixed-Methods Design: A New Methodological Approach. *Journal of Advanced Nursing* 79: 4398–410. [CrossRef]

Harris, Steve, and Andy Seaborne. 2013. *SPARQL 1.1 Query Language*. With Eric Prud'hommeaux. Available online: https://www.w3.org/TR/sparql11-query/ (accessed on 1 May 2025).

Healy, Kieran. 2017. Fuck Nuance. *Sociological Theory* 35: 118–27. [CrossRef]

Hiltmann, Torsten, Martin Dröge, Nicole Dresselhaus, Till Grallert, Melanie Althage, Paul Bayer, Sophie Eckenstaler, Koray Mendi, Jascha Marijn Schmitz, Philipp Schneider, and et al. 2025. NER4all or Context Is All You Need: Using LLMs for Low-Effort, High-Performance NER on Historical Texts. A Humanities Informed Approach. *arXiv* arXiv:2502.04351. [CrossRef]

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. *spaCy: Industrial-Strength Natural Language Processing in Python*. Python. Released July 3. First published 2014. [CrossRef]

Höing, Hubert. 1991. Die Erschließung des Repertorium Germanicum durch EDV-gestützte Indices. Technische Voraussetzungen und Möglichkeiten. *Quellen und Forschungen aus Italienischen Archiven und Bibliotheken* 71: 310–24.

Hörnschemeyer, Jörg, and Jörg Voigt. 2023. Das 'Repertorium Germanicum'. Perspektiven einer Digitalen Prosopographie. In *Die Römischen Repertorien. Neue Perspektiven für die Erforschung von Kirche und Kurie des Spätmittelalters (1378–1484)*. Edited by Claudia Märtl, Irmgard Fees, Andreas Rehberg and Jörg Voigt. Bibliothek des Deutschen Historischen Instituts in Rom 145. Berlin: De Gruyter, pp. 135–58. [CrossRef]

Hu, Yuntong, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, and Liang Zhao. 2025. GRAG: Graph Retrieval-Augmented Generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*. Edited by Luis Chiruzzo, Alan Ritter and Lu Wang. Vienna: Association for Computational Linguistics. [CrossRef]

Jaskulski, Piotr, Tomasz Latos, Mariusz Ryńca, and Adam Zapała. 2025. Reliability of Large Language Models as a Tool for Knowledge Extraction from Biographical Dictionaries: The Case of the Polish Biographical Dictionary. *Digital Scholarship in the Humanities* 40: 538–48. [CrossRef]

Joyner, David. 2025. 'AI Veganism': Some People's Issues with AI Parallel Vegans' Concerns about Diet. *The Conversation*, July 29. Available online: http://theconversation.com/ai-veganism-some-peoples-issues-with-ai-parallel-vegans-concerns-about-diet-260277 (accessed on 1 May 2025).

Karjus, Andres. 2025. Machine-Assisted Quantitizing Designs: Augmenting Humanities and Social Sciences with Artificial Intelligence. *Humanities and Social Sciences Communications* 12: 277. [CrossRef]

Kieslich, Kimon, Marco Lünich, and Pero Došenović. 2024. Ever Heard of Ethical AI? Investigating the Salience of Ethical AI Issues among the German Population. *International Journal of Human–Computer Interaction* 40: 2986–99. [CrossRef]

Kleymann, Rabea. 2025. Taken for Granted? Investigating Constructivist Principles with Bayes' Theorem in Digital Humanities Scholarship. *Digital Scholarship in the Humanities*. ahead of print. [CrossRef]

Kosmyna, Nataliya, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task. *arXiv* arXiv:2506.08872. [CrossRef]

Landau, Peter. 1980. Beneficium, Benefizium. III. Kanonisches Recht und Kirchenverfassung. In *Lexikon des Mittelalters*. Munich: Artemis, vol. 1.

Langlais, Pierre-Carl. 2023. MonadGPT. *Hugging Face*, November 10. Available online: https://huggingface.co/Pclanglais/MonadGPT (accessed on 1 May 2025).

Langlais, Pierre-Carl, Pavel Chizhov, Mattia Nee, Carlos Rosas Hinostroza, Matthieu Delsart, Irène Girard, Anastasia Stasenko, and Ivan P. Yamshchikov. 2025. Pleias 1.0: The First Ever Family of Language Models Trained on Fully Open Data. *Procedia Computer Science* 267: 146–56. [CrossRef]

Lehmann, Jörg, and Anna-Maria Sichani. 2025. A Position Paper on AI and Copyrights in Cultural Heritage and Research (EU and UK). *Journal of Open Humanities Data* 11. [CrossRef]

Levenshtein, Vladimir I. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady* 10: 707.

Lundberg, Scott M., Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence* 2: 2522–5839. [CrossRef]

Luo, Kangyang, Yuzhuo Bai, Cheng Gao, Shuzheng Si, Yingli Shen, Zhu Liu, Zhitong Wang, Cunliang Kong, Wenhao Li, Yufei Huang, and et al. 2025. GLTW: Joint Improved Graph Transformer and LLM via Three-Word Language for Knowledge Graph Completion. In *Findings of the Association for Computational Linguistics: ACL 2025*. Edited by Wanxiang Che, Joyce Nabende, Ekaterina Shutova and Mohammad Taher Pilehvar. Vienna: Association for Computational Linguistics, pp. 11328–44. Available online: https://aclanthology.org/2025.findings-acl.591/ (accessed on 1 May 2025).

Marin, Lavinia, and Steffen Steinert. 2025. CTRL+ Ethics: Large Language Models and Moral Deskilling in Professional Ethics Education. In *Oxford Intersections: AI in Society*. Edited by Philipp Hacker. Oxford: Oxford University Press. [CrossRef]

Mau, Steffen. 2018. Die Quantifizierung des Sozialen. *Zeitschrift für Theoretische Soziologie* 7: 274–92. [CrossRef]

McInnes, Leland, John Healy, and Steve Astels. 2017. Hdbscan: Hierarchical Density Based Clustering. *Journal of Open Source Software* 2: 205. [CrossRef]

Merry, Sally Engle, Kevin E. Davis, and Benedict Kingsbury, eds. 2015. *The Quiet Power of Indicators: Measuring Governance, Corruption, and Rule of Law*, 1st ed. Cambridge: Cambridge University Press. [CrossRef]

Oberbichler, Sarah, and Cindarella Petz. 2025. Evaluating bias within an epistemological framework for AI-based research in the humanities. In *Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale, Proceedings del XIV Convegno Annuale AIUCD2025*. Edited by Simone Rebora, Marco Rospocher and Stefano Bazzaco. Verona: AIUCD, pp. 52–59. Available online: https://amsacta.unibo.it/id/eprint/8380/1/AIUCD2025_Proceedings.pdf (accessed on 1 May 2025).

Parravicini, Alberto, Rhicheek Patra, Davide B. Bartolini, and Marco D. Santambrogio. 2019. Fast and Accurate Entity Linking via Graph Embedding. Paper presented at 2nd Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA), Amsterdam, The Netherlands, June 30–July 5; New York: Association for Computing Machinery, pp. 1–9. [CrossRef]

Pásztor, Lajos. 1970. *Guida delle Fonti per la Storia dell'America Latina: Negli Archivi della Santa Sede e Negli Archivi Ecclesiastici d'Italia*. Vatican City: Archivio Vaticano.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and et al. 2011. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* 12: 2825–30.

Peels, Rik. 2019. Replicability and Replication in the Humanities. *Research Integrity and Peer Review* 4: 2. [CrossRef]

Pellissier Tanon, Thomas. 2025. *Oxigraph. Rust*. Zenodo. Released June 15. [CrossRef]

Peng, Boci, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2024. Graph Retrieval-Augmented Generation: A Survey. *arXiv* arXiv:2408.08921. [CrossRef]

Powers, Simon T., Neil Urquhart, Chloe M. Barnes, Theodor Cimpeanu, Anikó Ekárt, The Anh Han, Jeremy Pitt, and Michael Guckert. 2025. What's It Like to Trust an LLM: The Devolution of Trust Psychology? *IEEE Technology and Society Magazine* 44: 30–37. [CrossRef]

Rao, Delip, Paul McNamee, and Mark Dredze. 2013. Entity Linking: Finding Extracted Entities in a Knowledge Base. In *Multi-Source, Multilingual Information Extraction and Summarization*. Edited by Thierry Poibeau, Horacio Saggion, Jakub Piskorski and Roman Yangarber. Berlin/Heidelberg: Springer, pp. 93–115. [CrossRef]

RDF Core Working Group. 2014. RDF-Semantic Web Standards. HTML. *RDF-Semantic Web Standards*, February 25. Available online: https://www.w3.org/TR/rdf-schema/ (accessed on 1 May 2025).

Reimann, Michael. 1991. Neue Erschließungsformen kurialer Quellen: Das Repertorium Germanicum Nikolaus' V. und Calixts III. (1447–1458) mit computergestützten Indices. *Römische Quartalschrift für Christliche Altertumskunde und Kirchengeschichte* 86: 98–112.

Ries, Thorsten, Karina van Dalen-Oskam, and Fabian Offert. 2024. Reproducibility and Explainability in Digital Humanities. *International Journal of Digital Humanities* 6: 1–7. [CrossRef]

Risam, Roopika, and Alex Gil. 2022. Introduction: The Questions of Minimal Computing. *Digital Humanities Quarterly* 16. Available online: http://digitalhumanities.org/dhq/vol/16/2/000646/000646.html (accessed on 1 May 2025).

Sander, Christoph. 2024a. DATAria: Graceful17 Utilities Platform (App). Released. Available online: https://dataria.dhi-roma.it/ (accessed on 1 May 2025).

Sander, Christoph. 2024b. DATAria: Graceful17 Utilities Platform (Core Codebase). Released. Available online: https://github.com/ch-sander/dataria-core (accessed on 1 May 2025).

Sander, Christoph. 2025a. *GRACEFUL17: Article 1 Code (Nepotism/Lead Speed)*. Python. March 31, Released April 2. [CrossRef]

Sander, Christoph. 2025b. *GRACEFUL17 Explorer. Jinja*. April 16. DHI-Roma. Released April 19. Available online: https://github.com/DHI-Roma/g17-explorer (accessed on 1 May 2025).

Sander, Christoph. 2025c. *DATAria Python Utils*. Python. December 19, Released April 23. Available online: https://github.com/ch-sander/dataria-py-utils (accessed on 1 May 2025). First published 2024.

Sander, Christoph. 2025d. *DATAria: Graceful17 Utilities Platform (Core Codebase, Release)*. Zenodo, Released August 4. [CrossRef]

Sander, Christoph. 2025e. *G17_cat_boost. Version 5576ed0*. Hugging Face, April 28. [CrossRef]

Sander, Christoph. 2025f. *La_g17_all_tags. Version 25933a9*. Hugging Face, April 24. [CrossRef]

Sander, Christoph, and Bruno Boute. 2025. *GRACE Ontology* (Version 1.0.2). Zenodo, April 29. [CrossRef]

Sander, Christoph, and Jörg Hörnschemeyer. 2025. GRACEFUL17: A Scalable Digital Fast-Track Strategy: Mining, Modelling, and Mastering Early Modern Church Administration Data. Paper presented at Alliance of Digital Humanities Organizations Annual Conference, Lisbon, Portugal, July 14–18.

Sander, Christoph, Bruno Boute, Jörg Hörnschemeyer, Naomi Beutler, Filippo Sarra, Valentino Verdone, and Andrea Cicerchia. Forthcoming. GRACEFUL17 Data Paper. *Die Zeitschrift Für Digitale Geisteswissenschaften*. accepted for publication.

Sander, Christoph, Naomi Beutler, Filippo Sarra, Valentino Verdone, and Bruno Boute. 2025a. GRACEFUL17 Notebooks. Jupyter Notebook. DHI-Roma, released November 5. Available online: https://github.com/DHI-Roma/g17-notebooks (accessed on 1 May 2025).

Sander, Christoph, Naomi Beutler, Filippo Sarra, Valentino Verdone, Andrea Cicerchia, Bruno Boute, and Jörg Hörnschemeyer. 2025b. *Graceful17: Main Data Repository*. Zenodo, February 28. [CrossRef]

Schmugge, Ludwig. 2023. 'Repertorium Poenitentiariae Germanicum' und Digital Humanities. Eine fruchtbare Beziehung. In *Die Römischen Repertorien. Neue Perspektiven für die Erforschung von Kirche und Kurie des Spätmittelalters (1378–1484)*. Edited by Claudia Märtl, Irmgard Fees, Andreas Rehberg and Jörg Voigt. Bibliothek des Deutschen Historischen Instituts in Rom 145. Berlin: De Gruyter. [CrossRef]

Schwinges, Rainer Christoph. 2015. Das Repertorium Academicum Germanicum (RAG): Ein digitales Forschungsvorhaben zur Geschichte der Gelehrten des Alten Reiches (1250–1550). *Jahrbuch für Universitätsgeschichte* 16: 215–32.

Selim, Rania, Arunima Basu, Ailin Anto, Thomas Foscht, and Andreas Benedikt Eisingerich. 2024. Effects of Large Language Model-Based Offerings on the Well-Being of Students: Qualitative Study. *JMIR Formative Research* 8: e64081. [CrossRef]

Simons, Arno, Michael Zichert, and Adrian Wüthrich. 2025. Large Language Models for History, Philosophy, and Sociology of Science: Interpretive Uses, Methodological Challenges, and Critical Perspectives. *arXiv* arXiv:2506.12242. [CrossRef]

Sparna. 2025. *Sparnatural SPARQL Query Builder*. Released. Available online: https://github.com/sparna-git/Sparnatural (accessed on 1 May 2025).

Spärck Jones, Karen. 1972. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation* 28: 11–21. [CrossRef]

Sporny, Manu, Dave Longley, Gregg Kellog, Markus Lanthaler, Pierre-Antoine Champin, and Niklas Lindström. 2020. *JSON-LD 1.1*. V. 1.1. Released July 16. Available online: https://www.w3.org/TR/json-ld11/ (accessed on 1 May 2025).

Storti, Nicola. 1969. *La Storia e il Diritto della Dataria Apostolica dalle Origini ai Nostri Giorni*. Contributi alla storia del diritto canonico. Naples: Athena Mediterranea.

Suchikova, Yana, Natalia Tsybuliak, Jaime A. Teixeira da Silva, and Serhii Nazarovets. n.d.GAIDeT (Generative AI Delegation Taxonomy): A Taxonomy for Humans to Delegate Tasks to Generative Artificial Intelligence in Scientific Research and Publishing. *Accountability in Research*, 1–27. [CrossRef]

TildeAI. 2025. *TildeOpen-30b*. Hugging Face, June 6, Available online: https://huggingface.co/TildeAI/TildeOpen-30b (accessed on 1 May 2025).

TriplyDB. 2025. TriplyDB/Yasgui. TypeScript. May 31, TriplyDB, Released July 11. Available online: https://github.com/TriplyDB/Yasgui (accessed on 1 May 2025). First published 2014.

Tudor, Crina, Beata Megyesi, and Robert Östling. 2025. Prompting the Past: Exploring Zero-Shot Learning for Named Entity Recognition in Historical Texts Using Prompt-Answering LLMs. Paper presented at 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025), Albuquerque, NM, USA, May 4; Edited by Anna Kazantseva, Stan Szpakowicz, Stefania Degaetano-Ortlieb, Yuri Bizzoni and Janis Pagel. Albuquerque: Association for Computational Linguistics. [CrossRef]

Valleriani, Matteo. 2025. Large Language Models That Power AI Should Be Publicly Owned. Technology. *The Guardian*. May 26. Available online: https://www.theguardian.com/technology/2025/may/26/large-language-models-that-power-ai-should-be-publicly-owned (accessed on 1 May 2025).

van Hage, Willem Robert, Véronique Malaisé, Gerben de Vries, Guus Schreiber, and Maarten van Someren. 2009a. Combining Ship Trajectories and Semantics with the Simple Event Model (SEM). Paper presented at 1st ACM International Workshop on Events in Multimedia (EiMM '09), Beijing, China, October 23; pp. 73–80. [CrossRef]

van Hage, Willem Robert, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. 2009b. The Simple Event Model. Available online: https://semanticweb.cs.vu.nl/2009/11/sem/ (accessed on 1 May 2025).

van Hage, Willem Robert, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. 2011. Design and Use of the Simple Event Model (SEM). *Journal of Web Semantics* 9: 128–36. [CrossRef]

Viana, Antonio. 2018. Introducción histórica y canónica al oficio eclesiástico. *Ius Canonicum* 58: 709–40. [CrossRef]

Xie, Tingyu, Qi Li, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2024. Self-Improving for Zero-Shot Named Entity Recognition with Large Language Models. *arXiv* arXiv:2311.08921. [CrossRef]

You, Doohee, Andy Parisi, Zach Vander Velden, and Lara Dantas Inojosa. 2025. LLM-as-Classifier: Semi-Supervised, Iterative Framework for Hierarchical Text Classification Using Large Language Models. *arXiv* arXiv:2508.16478. [CrossRef]

Zhu, Yutao, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2025. Large Language Models for Information Retrieval: A Survey. *ACM Transactions on Information Systems*. ahead of print. [CrossRef]

Zwicklbauer, Stefan, Christin Seifert, and Michael Granitzer. 2016. Robust and Collective Entity Disambiguation through Semantic Embeddings. Paper presented at 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16), New York, NY, USA, July 7; pp. 425–34. [CrossRef]