MCKG: A Community-Driven Knowledge Graph on Medieval Charters

Semantic Web Journal
XX(X):1–9

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Jorge Álvarez-Fidalgo[1], Enrique Rodriguez-Martin[1] and Jose Emilio Labra-Gayo [1]

Abstract
Medieval charters serve as essential primary sources for understanding medieval societies, yet their analysis remains a labor-intensive process reliant on domain experts, leaving large digitized collections largely unexplored. To address this challenge, we present a Wikibase-based Medieval Charters Community-Driven Knowledge Graph (MCKG) that combines expert annotations with community contributions through a provenance-aware framework, ensuring data quality while enabling scalable data integration. Our solution features a hybrid data model that combines elements from CIDOC-CRM and the Wikidata data model to capture the complex legal, social, and biographical relationships in medieval charters. A standardized pipeline enables efficient corpus integration into the MCKG. We demonstrate this approach's effectiveness by populating the MCKG with a corpus of Spanish medieval charters using our integration pipeline, and resolving interdisciplinary, SPARQL-based competency questions, showcasing its research value for historical studies.

Keywords
Semantic Web, Linked Open Data, Medieval Charters, Cultural Heritage

1 Introduction

The preservation of historical texts is essential in order to understand how past societies were structured. Historians focus on tasks such as archival work, preservation and contextualization to achieve this goal.

Computer science contributes to this field by facilitating the work of domain experts: from the digitization of documents for non-physical preservation to developing tools that enable precise and systematic work about the original sources or derived ones.

The analysis and extraction of information from historical documents remains a manual and time-consuming task performed by domain experts, a challenge that grows with the size and complexity of the corpus.

In this paper, we focus on **medieval charters**: legal documents from the European Middle Ages that record grants of rights, property, or privileges. These texts serve as critical primary sources for understanding medieval societies, governance, and economic exchange. While these texts have been widely studied as diplomatic artifacts, we specifically analyze their **event-centric content**: the socio-legal transactions, participating entities, and relational networks they document, which remain under-explored in computational approaches.

Numerous collections of medieval charter texts exist, such as CBMA[1], Diplomata Belgica[*], HOME-Alcar[2], or CODEA[3] which contain tens of thousands of charters from Western Europe. While some, such as CBMA or HOME-Alcar, include partial or complete manual annotations of named entities, the vast majority remain unanalyzed due to limited expert contributions.

To address this challenge, we propose a **Community-driven Knowledge Graph (CKG)** that integrates both the information extracted manually by domain experts and the contributions that can be made by an open community, including automated extraction with NLP tools and inference mechanisms which could enable to infer new relationships using, for example, genealogical information.

A key feature of this proposal is **provenance tracking**, ensuring data accuracy by clearly distinguishing the source of each claim between expert judgement and community input. While the initial scale of this Linked Open Dataset is currently constrained by the limited availability of expert-annotated medieval charters, this work aims to establish a foundational collaborative framework –with the CKG as its central hub– to stimulate and facilitate future research in this domain.

Thus, the main **contribution** of this work lies in the construction of a **Medieval Charters Community-Driven Knowledge Graph** (MCKG)[†] by means of a methodological approach –detailed in Section 3– consisting of three phases:

1. Selection of technical infrastructure.
2. Data model specification.
3. Definition of a reusable pipeline to integrate digitized medieval charter corpora into the MCKG.

The remainder of this paper is structured as follows. Section 2 contextualizes our approach within the related work. In Section 4, we demonstrate the pipeline's application

[1]Dept. of Computer Science , University of Oviedo, Spain.

Corresponding author:
Jorge Álvarez-Fidalgo
Email: alvarezfjorge@uniovi.es

*https://www.diplomata-belgica.be/colophon_fr.html
†https://medievalcharterskg.wikibase.cloud

to populate the CKG. In Section 5, we discuss practical applications and key quality attributes of the MCKG. Finally, Section 6 contains the conclusions and future research directions.

2 Related work

The application of Knowledge Graphs (KGs) to historical research has seen significant development in recent years. Meroño-Peñuela et al.[4] provide a foundational survey identifying persistent challenges in historical document analysis, including multi-source integration, relationship modeling, and historical interpretation - many of which align well with Semantic Web solutions.

Several implementations demonstrate these approaches in practice. The WarSampo project[5] exemplifies comprehensive KG construction through its integration of diverse Finnish WWII materials (official records, personal correspondence, and cartographic sources) while proposing novel entity linking pipelines. Similarly, WW1LOD[6] tackles the complex temporal-spatial dynamics of integrating heterogeneous source datasets of changing geospatial boundaries and personal careers. At a larger scale, Europeana[7] represents a landmark achievement in cultural heritage Linked Open Data (LOD), aggregating millions of cross-domain artifacts through multinational collaboration.

Furthermore, the Enslaved project[‡] exemplifies two distinct contributions to historical knowledge representation. First, it developed an OWL-based ontology[8] to model the complex relationships in transatlantic slavery records. Secondly, the project implemented a CKG with the Enslaved ontology as the underlying schema and Wikibase as its infrastructure[9].

Wikibase is the foundational software behind Wikidata and has been adopted by several historical KG projects. The Biblissima Project[§] exemplifies this approach, using a Wikibase instance to aggregate and interlink authoritative sources for ancient through Renaissance works while aligning with the CIDOC-CRM ontology[10]. Another example would be FactGrid[¶], which employs Wikibase as a general-purpose historical platform, aggregating data across diverse periods and topics to facilitate cross-disciplinary research. These implementations demonstrate Wikibase's adaptability to historical applications.

Focusing on related work on medieval charters, NotaryPedia –a KG of Maltese notarial manuscripts[11]– represents the most directly comparable effort. It defines a lightweight ontology for representing deeds, reusing established vocabularies for all other entities (*foaf:Person, owl:Thing, schema:Place*) while depending on domain experts for information extraction. Opitz et al. implement a large-scale KG of digitized abstracts of medieval charters[12]. García-González et al. convert TEI-based XML transcriptions of medieval notarial manuscripts to RDF using ShExML[13]. Finally, HistRED[14] provides a manually annotated Relation Extraction dataset of historical Korean and Hanja documents, serving as a reference for non-Latin script annotation of entities and relationships.

3 Methodology

3.1 Technical Infrastructure

We have chosen **Wikibase** as the underlying infrastructure for our CKG. This selection is justified by the following factors:

- **Mature infrastructure**. Wikibase's stability is proven across major projects, most notably Wikidata which successfully manages over 100 million entities with high edit volumes.
- **Collaborative tools**. The platform provides essential features for academic collaboration including change management, dedicated discussion pages per entity, and configurable protection mechanisms.
- **Query engine**. Every Wikibase instance includes a built-in SPARQL endpoint supporting complex queries and federated queries across linked open data resources.
- **Validation**. Wikibase includes an extension for the definition of ShEx-based **Entity Schemas**[15,16] to describe and –externally– validate the entities of the model in RDF, thus facilitating compliance with the data model.

3.2 Data model

Medieval charters document legal acts –primarily grants of property– between historical actors. To represent the information contained in these charters, we adopt **event-based modeling** as their contents are fundamentally structured around the legal events they record, such as the participants -grantors, beneficiaries or witnesses- or the properties transferred.

Our MCKG data model (Figure 1) hybridizes the **CIDOC Conceptual Reference Model** (CRM) and the **Wikidata data model**. CIDOC-CRM, an ISO standard (ISO 21127:2023), provides a robust semantic framework for cultural heritage documentation. We use its classes extensively, as they align naturally with the entities and relationships described in medieval charters (see Subsection 3.2.1 for further details).

However, medieval charters also contain rich **genealogical** and **sociological** (occupations, public offices, geospatial relationships) information. CIDOC-CRM's event-centric framework fits these elements less naturally, as it requires convoluted workarounds. For example:

- Parent-child relationships must be modeled through a `crm:E67_Birth`[‖] event, with properties like `crm:P97_from_father` and `crm:P98_brought_into_life`.
- Sibling relationships are even more complex: they require two Birth events that share at least one parent property (`crm:P96_by_mother` or `crm:P97_from_father`).

These constraints pose practical challenges, as many charters mention kinship ties *without specifying parental*

[‡] https://enslaved.org/
[§] https://portail.biblissima.fr/en
[¶] https://database.factgrid.de/wiki/Main_Page
[‖] *crm:* stands for http://www.cidoc-crm.org/cidoc-crm/

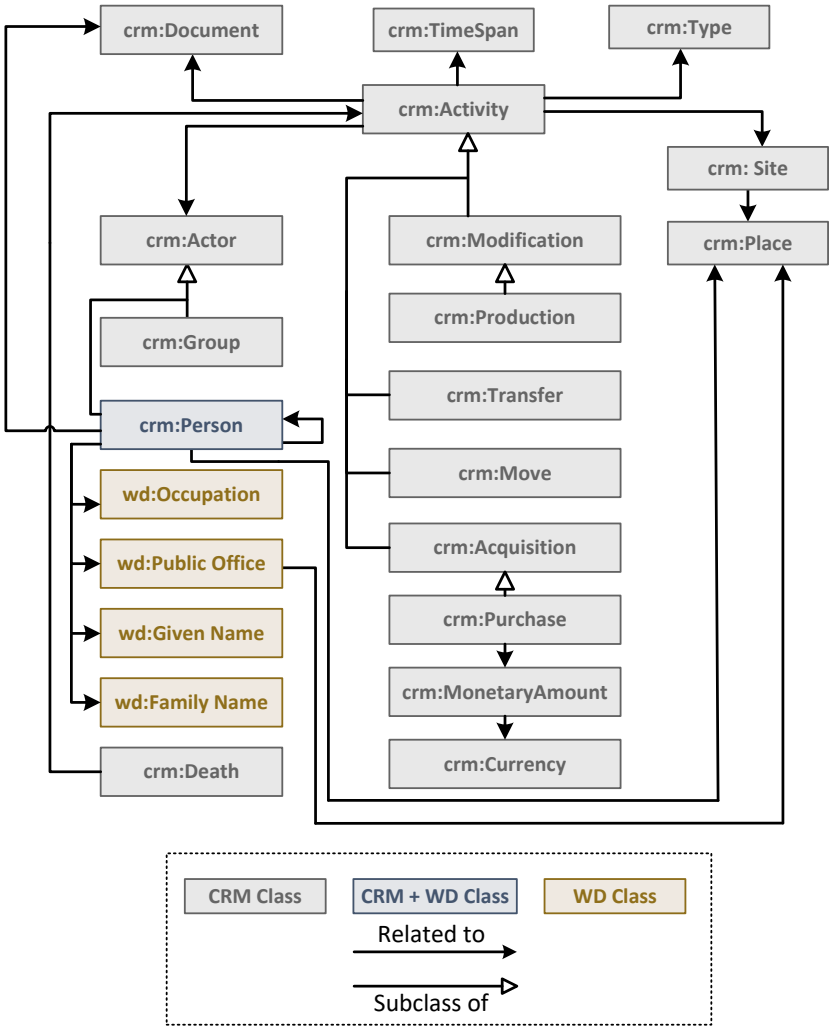

Figure 1. MCKG data model.

details; forcing the creation of blank nodes to represent the claim.

To streamline data representation of these complex relationships, we adopt Wikidata's person-related properties (e.g., `wd:P735-given name`[**], `wd:P22-father`, `wd:P106-occupation`), curated by WikiProject Genealogy[††]. This allows direct representation of names, kinship ties, and social roles –frequent in medieval charters– without CIDOC-CRM's intermediary events.

3.2.1 Core classes. The core classes of the data model are the following:

- **Activity and subclasses.** `crm:E7_Activity` represents the legal actions recorded in medieval charters. We include the following CRM classes to cover all anticipated cases:
  - `crm:E8_Acquisition`. Subclass of *Activity*.
  - `crm:E9_Move`. Subclass of *Activity*.
  - `crm:E10_Transfer_of_Custody`. Subclass of *Activity*.
  - `crm:E11_Modification`. Subclass of *Activity*.
  - `crm:E12_Production`. Subclass of *Modification*.
  - `crm:E96_Purchase`. Subclass of *Acquisition*.

  These entities form the core organizational structure of our model, with most classes relating to them.
- **Actor.** Activities are carried out by one `crm:E39_Actor` or more who may serve in various roles including *grantor, beneficiary, witness,* or *judge*. We distinguish between two types of Actors: `crm:E74_Group` (e.g. *'Children of Alfonso Fernández'*) or `crm:E21_Person`. As described previously, **Persons** represent a special case in our model. While they maintain the standard CRM hierarchy (as a subclass of Actor) and inherit certain CRM properties like `crm:P100_died_in`, we primarily utilize properties from the Wikidata data model –specifically those defined by WikiProject Genealogy– to streamline the representation of non-event based data.
- **Classes related to Wikidata properties.** The integration of Wikidata properties for the class Person requires the introduction of complementary classes

[**] *wd:* stands for `http://www.wikidata.org/prop/direct/`
[††] `https://www.wikidata.org/wiki/Wikidata:WikiProject_Genealogy`

Table 1. Wikidata Property-Class Mappings

Property	P-ID	Class	Q-ID
Occupation	P106	Occupation	Q12737077
Position held	P39	Public Office	Q294414
Given name	P735	Given Name	Q202444
Family name	P734	Family Name	Q101352

absent in CIDOC-CRM, as exposed in Table 1. Given the sparse documentation of these aspects in medieval charters, these classes are minimally defined with only a `native label` (`wd:P1705`) property to capture orthographic variations. The sole exception is **Public Office**, which additionally includes `applies to jurisdiction` (`wd:P1001`) to encode territorial scopes (e.g., *'King of Castile and León'*).

- **Place.** Medieval charters document numerous instances of (`crm:E53_Place`), primarily associated with persons (*place of residence*) and their trades (*place of work, jurisdiction of the position*) as well as to Activities and their subclasses through the class `crm:E27_Site`, which in CRM represents a set of physical features with a given configuration and location, and which we use to represent the transferred properties. Thus, a Site is located (`crm:P53_has_former_or current_location`) at one or several Places.
- **Document.** The content of medieval charters is physically embodied in a `crm:E31_Document` linked through the `crm:P70i_is_documented_in` property, and uniquely identified by domain experts.
  While the Document associated with a given Person can be retrieved through his participation in an Activity, this approach misses persons mentioned incidentally (e.g., third parties referenced for property geolocation) who are not directly related to the Activity.
  To avoid loss of information, we establish direct Document-Person relationships through the usage of `crm:P70i` claims. Additionally, this class also represents scholarly sources cited when creating statements.
- **Time-Span.** In CRM, all **Temporal Entities** (including Activities) are associated with a `crm:E52_Time-Span` through the `crm:P4_has_time-span` property. Typically, medieval charters are dated on a specific day, which we model by defining the Activity's Time-Span as occurring at some time within (`crm:P82_at_some_time_within`) the dated day.
- **Type.** Assigning a `crm:E55_Type` to an Activity through the property `crm:P2_has_type` enables precise modeling of legal nuances. For instance, a **Purchase** represents a legal transfer of ownership between actors that involves monetary compensation to the transferring party, but there are specific cases where the seller retains the right to repurchase within a defined period (*retrosale*). Thus, this typing mechanism allows differentiation between variants while maintaining the base classification of CIDOC-CRM.
- **Death.** Medieval charters frequently reference `crm:E69_Death` through formulaic expressions (e.g. *'May God forgive him'*) without providing specific temporal details. To address this limitation, we use the following structure:
  - The deceased person is associated through `crm:P100_died_in` to a `crm:E69` instance.
  - This Death event is connected through `crm:P183_ends_before_the_start_of` to a specific Activity.

  This approach establishes that the death occurred at some point before the related event.
- **Purchase-related classes.** As mentioned when discussing the Type class, Purchases involve the payment of a monetary amount, which is represented by the property `crm:P179_had_sales_price`. This property requires as a predicate an instance of the `crm:E97_Monetary_Amount` class, which in turn requires an instance of the `crm:E98_Currency` class. CIDOC-CRM's default interpretation is extended by including non-monetary compensation –such as livestock– since those goods are commonly used as payment in this context.

3.2.2 Qualifiers. To avoid loss of information in our CKG, we reify certain properties by means of qualifiers, which in the Wikibase ecosystem provide additional information to statements. They are the following:

- **Carried out by**. Activities have a property `crm:P14_carried_out_by` that expresses the active participation of an Actor in the activity. The CRM model allows for specifying the nature of such participation with the sub-property `crm:P14.1_in_the_role_of`. Since there exist dedicated properties to represent the roles of the main Actors (e.g. in an Acquisition we have the property `crm:P22_transferred_title_to` to designate the transferee of the ownership transfer) this qualification accomodates minor roles without dedicated properties, such as *witnesses*.
- **Is documented in**. Key properties (such as `native label` and those expressing family relationships like `father`, `mother`, `sibling`...) are qualified with the document in which those claims are made. This allows for a) precise contextualization of statements and b) enabling corpus-wide statistical analysis. For instance, documenting all charter occurrences of surname variants (e.g., *García* and *Garçía*) through qualified `native label` properties allows researchers to track orthographic distributions across time.
- **Kinship to subject**. The Wikidata `relative` property requires qualification through the property `kinship to subject` (`wd:P1039`) to specify exact relationships. Therefore, the same approach is adopted in our model.

3.3 Data Integration Pipeline

The proposed data integration pipeline -shown in Figure 2- consists of the following phases:

1. **Data collection.** We adopt a dual approach to ensure both truthfulness and scalable growth in the dataset. On one hand, domain experts (DE) apply rigorous manual annotation following their standard of choice;

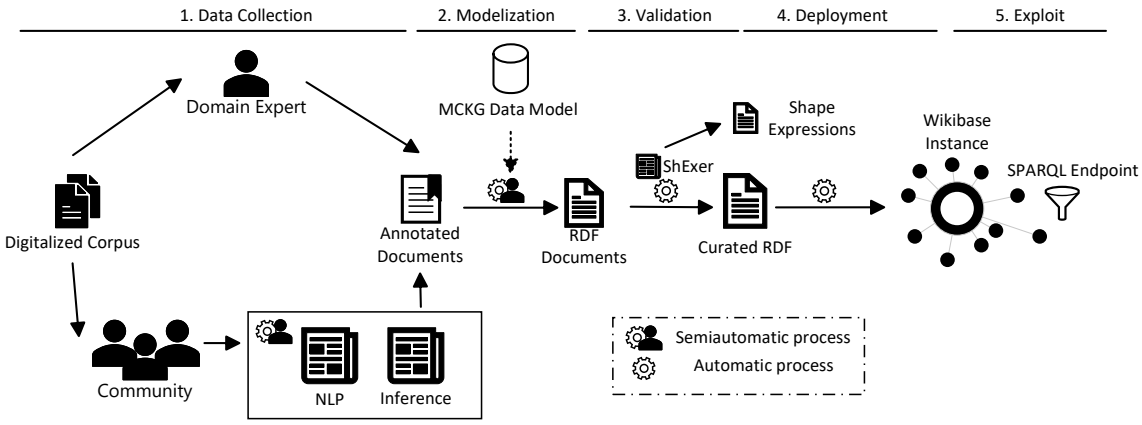

Figure 2. Data Integration Pipeline.

on the other hand, community contributors make use of NLP tools and inference mechanisms to expand existing annotations or create new ones. Thus, we aim to maintain academic integrity while enabling the integration of larger corpora into the MCKG.

2. **Modelization.** All annotated entities and relationships are formally structured as RDF triples conforming to the defined MCKG data model. This phase could be combined with the previous one if there is sufficient familiarity with the data model, by annotating directly using RDF triples.
3. **Validation.** *sheXer* is an automatic shape extractor based on graph mining capable of producing both ShEx and SHACL content[17].
   By applying *sheXer* on the RDF dataset we can infer its underlying structure: classes, properties, datatypes and cardinalities. This serves two critical functions:
   - Validating data integrity and compliance against our model. Rather than applying ShEx validation after KG population, we use *sheXer* to extract Shape Expressions from source data beforehand, enabling early inconsistency detection.
   - As described in Section 3.1, Wikibase's Entity Schemas are ShEx-based. Thus, generating ShEx shapes with *sheXer* provides a solid basis for the subsequent definition of such Entity Schemas.
4. **Deployment.** The validated RDF dataset is automatically integrated into the KG. Provenance tracking is implemented by means of two Wikidata data model properties:
   - `stated by authoritative source` for statements performed by DE.
   - `inferred from [source]` for community contributions.
5. **Exploit.** Once the graph is populated with the information structured according to the proposed data model, both DE and community members can execute SPARQL queries through the Wikibase Query Service to investigate diverse research questions. The formal structure of the data enables a comprehensive analysis of the corpus data (see Section 5 for applications).

4 KG Population

We evaluate our data model and methodology using the **AMSPO medieval charter corpus**[18], with authorization from its author.

This corpus comprises notarial records from medieval Spain, documenting the life and transactions of a northern goldsmith throughout his life. This collection has been extensively annotated by DE, detailing the characteristics of each event and elaborating toponymic and anthroponymic indexes that include most of the places and persons mentioned, as well as descriptions of their relationships and socioeconomic data.

Table 2. Entities in the MCKG by class.

Instance of	Frequency	Ratio (%)
Person	867	39.21
Family name	290	13.12
Site	130	5.88
Document	128	5.79
Monetary Amount	126	5.70
Place	123	5.56
TimeSpan	122	5.52
Purchase	121	5.47
Death	120	5.43
Given name	64	2.89
Occupation	62	2.80
Type	14	0.63
Kinship	12	0.54
Public Office	12	0.54
Acquisition	9	0.41
Currency	6	0.27
Group	2	0.09
Activity	2	0.09
Production	1	0.04
Total	**2211**	**100.00**

While domain experts have completed the primary Data collection, we still lack a contributing community. Therefore, the authors deliberately adopt the community's role to

participate in the annotation process. To this end, we have used a domain-specific Named Entity Recognition (NER) model for Western European medieval charters[19] that detects all entities in the corpus. Moreover, this enables us to retrieve all ortographic variants from the standardized name that the DE employs in the anthroponymic index.

As community contributors, we also make use of DEMel[20], a lexical resource mapping 700,000 medieval Spanish documented forms to modern lemmas (e.g., *'fillo de'* is a recurrent variant of *'hijo de'*). This enables consistent relation extraction (RE) between NER-detected entities as native Spanish speakers. Furthermore, we perform named entity linking (NEL) to Wikidata on a few entities, primarily for locations using Roberto Antuña's gazetteer[21], which documents modern equivalents for historical toponyms common in our corpus.

Table 3. Statements in the MCKG per each property.

Property	Frequency	Ratio (%)
Instance of	2211	17.79
Is documented in	1921	15.46
ID	1882	15.14
Native label	1029	8.28
Given name	862	6.94
Family name	842	6.77
Carried out by	594	4.78
Occupation	286	2.30
Transferred title to	218	1.75
Has former or current location	211	1.70
Transferred title from	201	1.62
Child	192	1.54
Residence	134	1.08
Has time span	133	1.07
Has type	133	1.07
Father	129	1.04
Transferred title of	129	1.04
Had sales price	126	1.01
Has currency	125	1.01
Has value	125	1.01
At some time within	122	0.98
Died in	120	0.97
Ends before the start of	120	0.97
Spouse	120	0.97
Relative	90	0.72
Mother	62	0.50
Sibling	62	0.50
Wikidata ID	59	0.47
Applies to jurisdiction	35	0.28
Position held	34	0.27
Subclass of	33	0.27
Employer	28	0.23
Nickname	28	0.23
Work location	22	0.18
Title	8	0.06
Has produced	1	0.01
Regnal ordinal	1	0.01
Was motivated by	1	0.01
Total	**12429**	**100.00**

During the **Modelization** phase, we transform all annotated data into two distinct RDF datasets; *'Domain Expert'* and *'Community'*. Using *sheXer*, we validated both datasets against our MCKG data model, resolving semantic inconsistencies and ultimately generating validated Shape Expressions for each final dataset (see *EntitySchemas* in the Wikibase instance).

Finally, in the **Deployment** phase, we process each RDF dataset for its integration into our Wikibase instance. To streamline this process, we first convert the RDF data to CSV format, enabling efficient bulk processing using the *WikidataIntegrator*[‡‡] tool. This intermediate step facilitates reliable data ingestion while maintaining the integrity of our structured knowledge.

Table 4. Properties that are qualifying other statements in the MCKG.

Property	Frequency	Ratio (%)
Is documented in	6430	90.35
In the role of	594	8.35
Kinship to subject	93	1.31
Total	**7117**	**100.00**

Table 2 lists the number of entities –with class distributions– in the populated MCKG, for a total of 2211 entities. Table 3 lists the number of existing statements existing for each property, for a total of 12429.

Table 5. Statements per type of reference in the MCKG.

Reference	**Statements**	**Ratio (%)**
Stated in auth. source + Inferred	4808	59.19
Stated in auth. source	1983	24.41
Inferred from	1332	16.40
Total	**8123**	**100.00**

Table 4 shows the number of properties that serve as qualifiers for other properties, for a total of 7117. As we can see, the vast majority (90.35%) specify the document in which the statement is asserted. Finally, Table 5 shows the reference distribution for statements. Note that the majority of the statements (59.19%) are endorsed by consensus; 24.41% are statements exclusive to DE, and the remaining 16.4% are contributions exclusive to the community (mainly NER-discovered entities that play a minor role in the events and the different orthographic variants of preexisting names).

5 Applications and quality attributes

In the **Exploit** phase, the MCKG data model enables the resolution of use cases –mainly competence questions– across various research domain that would otherwise require a laborious and time-consuming manual effort to obtain in the original corpus, such as the following. All these queries can be carried out in the MCKG Query Service, available as examples.

- **Charter-related queries**. MCKG enables precise analysis of matters related to the events represented in the charters. For instance, we can query that the total expenses of the goldsmith Alfonso Fernández in all his purchases were 3728.5 *maravedíes*, 40 *dineros*, 1 colt and 3/4 of a mare; or that the person who appears most often in AMSPO charters, apart from the main actor Alfonso, is his wife María Álvarez (73 activities).
- **Genealogical queries**. For instance, we can query that the earliest documented death date of María Álvarez

[‡‡] https://github.com/SuLab/WikidataIntegrator

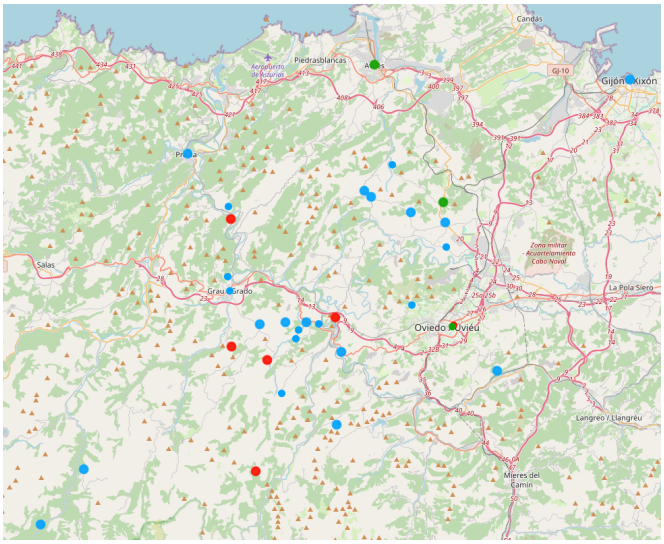

Figure 3. Fragment of the map result of the geospatial query that looks for places related to people.

(*AMSPO-P103*) is September 8th, 1346, or obtain all known relatives of a given person.

- **Sociographic queries**. For instance, querying for occupation distributions, the most common occupation in AMSPO is *'scribe'* (34 persons).
- **Etymological queries**. For example, an onomastic query: the orthographical variations of the given name *'Rodrigo'* (modern spelling) are *'Ruy'* (18 instances), *'Rodrigo'* (12 instances), *'Roy'* (7 instances) and *'Rodrig'* (3 instances).
- **Geospatial queries**. Although our graph lacks geospatial information, we can perform federated queries that geolocate our places by means of Wikidata external IDs. For example, Figure 3 shows the people in AMSPO corpus according to places of residence (blue), work locations (red) and jurisdictions of held positions (green) in a given area. If multiple people share a relationship for a place, MCKG Query Service allows for on-demand clustering at shared locations.
- **Other federated queries**. NEL to Wikidata can be used to retrieve other information lacking in our corpus. For instance, querying the incumbency of the bishops mentioned in AMSPO: *Fernán Álvarez* was Bishop of Oviedo during 1302-1322, *Odo* during 1324-1327 and *Don Juan* during 1328-1332.
- **Querying by provenance**. MCKG supports queries filtered by assertion certainty, depending on the property used for provenance tracking. For instance, we can query the witnesses who participated in a given purchase (*AMSPO-D120*), but only those with the domain expert (`stated in authoritative source`) as a reference: *Pedro Martínez, Pedro Fernández, Alfonso Yáñez* and *Fernán Nicolás*.

We consider that these applications are supported by key quality attributes that ensure the MCKG's research utility:

- **Data Stability.** The MCKG data model builds on mature ontologies and stable schemas with minimal anticipated changes. While those are not fully integrated into our data model, the **extensible** Wikibase framework allows seamless incorporation of new classes and properties without disrupting existing entities or statements. Given the formulaic nature of medieval charters, significant extensions to the data model are unlikely.
- **Scalability.** The proposed semi-automatic pipeline combines expert judgment and community input, while Wikibase's infrastructure enables efficient large-scale data ingestion and sustainable growth without compromising performance.
- **Data Integrity.** Wikibase features such as qualifiers, references, and version control ensure transparent provenance tracking, balancing an open, community-driven approach with truthfullness.
- **Interoperability.** Wikibase enables interaction with external knowledge bases through federated queries, supporting integration with other KGs in the Wikibase ecosystem –such as Wikidata or historical KGs like the aforementioned Biblissima and FactGrid– Moreover, given the MCKG's data model, its data is inherently compatible with CIDOC-CRM, ensuring interoperability with other CRM-based KGs. However, as detailed in Section 3.2, adapting properties inherited from Wikidata's data model into CIDOC-CRM requires verbose and structurally complex representations.

6 Conclusions and future work

Results lead to the following conclusions:

Knowledge Graph. We convert a corpus of digitized medieval charters into a structured KG, adding a new dimensionality to their analysis. This representation enables dynamic queries into the legal, socio-economic, and biographical aspects embedded in the texts, facilitating advanced research and exploration.

Community-Driven approach. Given the sheer volume of medieval charters requiring manual processing, we designed our KG as a **collaborative platform**. This approach mitigates the scarcity of domain experts while ensuring provenance transparency and the **rigorousness** of assertions. By involving diverse contributors, we balance scalability with scholarly accuracy.

Pipeline. The KG is deliberately **extensible**, with a defined pipeline for integrating additional medieval charter corpora (e.g., CODEA, HOME-Alcar). We successfully applied this standardized workflow to our initial corpus, demonstrating its viability for future expansions of our Linked Open Dataset.

6.1 Future work

The following key challenges emerged from our analysis, pointing to critical areas for improvement:

Information Extraction. While NER models are already applicable to medieval texts, **Relation Extraction** remains the primary bottleneck. The complexity and richness of interpersonal and socio-economic relationships in charters demand labor-intensive manual annotation, slowing corpus integration into the KG. Future work must prioritize the

development and training of RE solutions able to work in this context.

Entity Linking. Currently, linking entities to external knowledge bases (e.g., Wikidata) or across corpora within the MCKG is entirely manual. To address this, we propose a **semi-automatic matching pipeline** (e.g., using similarity metrics or graph embeddings) to suggest plausible links with human-in-the-loop validation. [22]

Acknowledgements

The authors thank Jorge Felpeto Cueva and Miguel Calleja Puerta from the DocuLab research group for their invaluable contribution and help with the AMSPO corpus.

This work has been partially funded by the project ANGLIRU: Applying kNowledge Graphs to research data interoperabiLIty and ReUsability, code: PID2020-117912RB from the Spanish Research Agency and by the regional project SEK-25-GRU-GIC-24-089.

References

1. Corpus de la Bourgogne du Moyen Âge. Cbma project - corpus burgundiae medii aevi. URL http://www.cbma-project.eu. Accessed: 05/05/2025.
2. Stutzmann D, Torres Aguilar S and Chaffenet P. HOME-Alcar: Aligned and Annotated Cartularies, 2021. DOI:10.5281/ZENODO.5600884. URL https://zenodo.org/record/5600884.
3. Corpus de Documentos Españoles Anteriores a 1800. *Corpus de Documentos Españoles Anteriores a 1800* 2015; DOI:10.37536/CODEA.2015. URL http://corpuscodea.es/.
4. Meroño-Peñuela A, Ashkpour A, Van Erp M et al. Semantic technologies for historical research: A survey. *Semantic Web* 2014; 6(6): 539–564. DOI:10.3233/SW-140158. URL https://journals.sagepub.com/doi/full/10.3233/SW-140158.
5. Koho M, Ikkala E, Leskinen P et al. WarSampo knowledge graph: Finland in the Second World War as Linked Open Data. *Semantic Web* 2021; 12(2): 265–278. DOI:10.3233/SW-200392. URL https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-200392.
6. Mäkelä E, Törnroos J, Lindquist T et al. WW1LOD: an application of CIDOC-CRM to World War 1 linked data. *International Journal on Digital Libraries* 2017; 18. DOI:10.1007/s00799-016-0186-2.
7. Isaac A and Haslhofer B. Europeana Linked Open Data – data.europeana.eu. *Semantic Web* 2013; 4(3): 291–297. URL https://eprints.cs.univie.ac.at/3732/. Number: 3 Publisher: IOS Press.
8. Shimizu C, Hitzler P, Hirt Q et al. The enslaved ontology: Peoples of the historic slave trade. *Journal of Web Semantics* 2020; 63: 100567. DOI:10.1016/j.websem.2020.100567. URL https://linkinghub.elsevier.com/retrieve/pii/S1570826820300135.
9. Shimizu C, Hitzler P, Gonzalez-Estrecha S et al. The Wikibase Approach to the Enslaved.Org Hub Knowledge Graph. In Payne TR, Presutti V, Qi G et al. (eds.) *The Semantic Web – ISWC 2023*, volume 14266. Cham: Springer Nature Switzerland. ISBN 978-3-031-47242-8 978-3-031-47243-5, 2023. pp. 419–434. DOI:10.1007/978-3-031-47243-5_23. URL https://link.springer.com/10.1007/978-3-031-47243-5_23. Series Title: Lecture Notes in Computer Science.
10. Frunzeanu E, Robineau R and MacDonald E. Biblissima's Choices of Tools and Methodology for Interoperability Purposes. *CIAN-Revista de Historia de las Universidades* 2016; 19: 115. DOI:10.20318/cian.2016.3146. URL https://e-revistas.uc3m.es/index.php/CIAN/article/view/3146.
11. Ellul C, Azzopardi J and Abela C. NotaryPedia: A Knowledge Graph of Historical Notarial Manuscripts. In Panetto H, Debruyne C, Hepp M et al. (eds.) *On the Move to Meaningful Internet Systems: OTM 2019 Conferences*, volume 11877. Cham: Springer International Publishing. ISBN 978-3-030-33245-7 978-3-030-33246-4, 2019. pp. 626–645. DOI:10.1007/978-3-030-33246-4_39. URL http://link.springer.com/10.1007/978-3-030-33246-4_39. Series Title: Lecture Notes in Computer Science.
12. Opitz J, Born L and Nastase V. Induction of a Large-Scale Knowledge Graph from the Regesta Imperii. In Alex B, Degaetano-Ortlieb S, Feldman A et al. (eds.) *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Santa Fe, New Mexico: Association for Computational Linguistics, pp. 159–168. URL https://aclanthology.org/W18-4518/.
13. Garcıa-Gonzalez H, Albarran-Fernandez E, Emilio J et al. Converting Asturian Notaries Public deeds to Linked Data using TEI and ShExML. In *Workshop: Workshop on Humanities in the Semantic web - WHISE III as ESWC 2020*.
14. Yang S, Choi M, Cho Y et al. HistRED: A Historical Document-Level Relation Extraction Dataset, 2023. DOI:10.48550/arXiv.2307.04285. URL http://arxiv.org/abs/2307.04285. ArXiv:2307.04285 [cs].
15. Prud'hommeaux E, Labra Gayo JE and Solbrig H. Shape expressions: an RDF validation and transformation language. In *Proceedings of the 10th International Conference on Semantic Systems, SEMANTICS 2014*. ACM, pp. 32–40.
16. Thornton K, Solbrig H, Stupp GS et al. Using shape expressions (ShEx) to share RDF data models and to guide curation with rigorous validation. In *The Semantic Web*. Springer International Publishing, 2019. pp. C1–C1. DOI:10.1007/978-3-030-21348-0_40.
17. Fernandez-Álvarez D, Labra-Gayo JE and Gayo-Avello D. Automatic extraction of shapes using sheXer. *Knowledge-Based Systems* 2022; 238: 107975. DOI:10.1016/j.knosys.2021.107975. URL https://linkinghub.elsevier.com/retrieve/pii/S0950705121010972.
18. Felpeto Cueva J. *El archivo de un artesano del siglo XIV: el orfebre Alfonso Fernández de Oviedo*. doctoral thesis, University of Oviedo, 2023. URL https://digibuo.uniovi.es/dspace/handle/10651/71402. Accepted: 2024-02-16T11:28:22Z.
19. Aguilar ST. Multilingual named entity recognition for medieval charters using stacked embeddings and bert-based models. In *Proceedings of the second workshop on language technologies for historical and ancient languages*. pp. 119–128.
20. DEMel - Diccionario del Español Medieval electrónico. URL https://demel.uni-rostock.de/?lang=en. Accessed: 05/05/2025.

21. Antuña Castro R. *Notariado y documentación notarial en el área central del señorío de los obispos de Oviedo (1291-1389)*. PhD Thesis, 2014.
22. Version 7.1.3 — cidoc crm. URL https://cidoc-crm.org/version/version-7.1.3.