Curated Datasets For Literary Tourism: A Case Study In Knowledge Graph Creation

M. Begliuomini, M. Crisan, E. Daga, Rossana Damiano, Laurence Roussillon-Constanty, M.A. Stranisci, C. Trinchero

Curated Datasets For Literary Tourism: A Case Study In Knowledge Graph Creation

Miriam Begliuomini[1], Marius Crisan[2], Enrico Daga[3,*], Rossana Damiano[4,7,*], Florin Nechita[5], Laurence Roussillon-Constanty[6], Marco Antonio Stranisci[4,*] and Cristina Trinchero[1,7]

[1]Dipartimento di Lingue e Letterature Straniere e Culture Moderne, Università di Torino, Turin, via S. Ottavio 18, Italy
[2]Teacher Training Department, West University of Timișoara, Bvd. Vasile Pârvan 4, Romania
[3]Knowledge Media Institute, The Open University, Walton Hall, Milton Keynes, MK7 6AA, United Kingdom
[4]Dipartimento di Informatica, Università di Torino, Turin, corso Svizzera 185, Italy
[7]Digital Scholarship for the Humanities (DISH) Centre, Università di Torino
[5]Universitatea Transilvania, Brașov, Romania
[6]Université de Pau et des Pays de l'Adour, Avenue de l'Université, Pau, France

Abstract

European mountains have inspired generations of writers whose works can play a significant role in determining the touristic potential of the area. However, fragmentation of research and cultural initiatives about European mountains hinder this potential, even in the digital age. In this paper, we describe the use of the World Literature Knowledge Graph (WL-KG) to integrate the curated data sets of writers and works created by a set of research projects about European mountains as part of the CON.NE.C.T W.O.N.D.E.R.S. project using the SPARQL Anything library for triplification. The goal of the project is two-fold: on the one side, it aims at bridging local repositories of literary data, remodeling them according to a common model when needed, to overcome the fragmentation of the otherwise underrepresented research about the mountain areas across Europe. On the other side, it aims at creating applications that leverage the networked representation of literary, geographical and temporal data for the discovery and exploitation of new paths and connections in the field of literary tourism.

Keywords

Knowledge graphs, Triplification, SPARQL Anything

1. Introduction

The intertwining between European mountains and their re-telling across centuries represents one of the most significant heritages of our continent. These places have inspired generations of writers whose works about the European mountain regions have a significant role in determining the touristic potential of the area. Such a potential is not fully expressed, though, since fragmentation of the many research and cultural initiatives about European mountains hinders its discovery by a wider target. In addition, writers who have written about European mountains are often underrepresented in the official narrative, and their works are little known to the general public and outside of academia. The main aim of CON.NE.C.T W.O.N.D.E.R.S (CONnecting NEw Cultural Tourism WAys Open to Networking Digital Experience in Representing Sites) project[1] is to enhance the cultural heritage of European mountains through a research framework that relies on the networked representation of writers, works and places.

In CON.NE.C.T W.O.N.D.E.R.S the networked representation of relevant resources at the literary, geographical and historical level is provided by the World Literature Knowledge Graph (WL-KG: https://literaturegraph.di.unito.it) [1]. The World Literature Knowledge Graph (WL-KG) currently

SemDH 2025: Second International Workshop of Semantic Digital Humanities. Co-located with ESWC 2025, June 02, 2025, Portoroz, Slovenia.

*Corresponding author.
miriam.begliuomini@unito.it (M. Begliuomini); marius.crisan@e-uvt.ro (M. Crisan); enrico.daga@open.ac.uk (E. Daga); rossana.damiano@unito.it (R. Damiano); florin.nechita@unitbv.ro (F. Nechita); laurence.roussillon-constanty@univ-pau.fr (L. Roussillon-Constanty); rossana.damiano@unito.it (M. A. Stranisci); cristina.trinchero@unito.it (C. Trinchero)
0000-0002-3184-5407 (E. Daga); 0000-0001-9866-2843 (R. Damiano); 0000-0001-9866-2843 (M. A. Stranisci)

[1]https://connectwonders.di.unito.it

includes 194,346 authors and 971,120 works linked to Wikidata, Goodreads, and Open Library. Open to the integration with literary archives of any size and scope, the WL-KG relies on the Underrepresented Network Ontology (UR-ON) and is available through a visualization platform specifically conceived for non-expert users that allows the discovery of writers and their work. In the context of the project it has been adopted as a background to align existing digital resources about the project topics and to integrate the new outputs emerging from the cooperation between the partners of the consortium.

The ultimate goal of integrating datasets from ongoing and completed research projects about European mountains in the WL-KG is aimed at two main objectives. First, it seeks to connect local literary data repositories, restructuring them into a unified model when necessary, to address the fragmentation of research on mountain regions across Europe. By doing so, common patterns and connections between datasets are expected to emerge, with benefits for scholars that preserve, study and disseminate Europe's mountain heritage. Secondly, it aims to develop applications that use the interconnected representation of literary, geographical and temporal data to discover and explore new paths and connections able to reach a wider audience for the benefit of literary tourism.

The integration of curated datasets collected by different research groups to research local literary traditions can a be a bottleneck for knowledge graph maintenance, since it cannot be directly managed by researchers from the humanities who are not familiar with RDF and graph-based representations in general. In turn, this limitation can hinder the exploration of similarities and points of contacts between literary data from different areas. In this paper, we describe the data gathering and integration process designed and implemented for CON.NE.C.T W.O.N.D.E.R.S., which relies on the use of SPARQL Anything, an open-source project that supports the triplification of diverse data sources without the need of a domain vocabulary.

The paper is structured as follows. After introducing the topic of digital archives in the literary field in Section 2, we describe the project in Section 3. Section 4 provides the background about the WL-KG; the data modeling and gathering, and the triplification process are described in Section 5. Section 6 provides an example of data integration. Conclusion and Future Work end the paper.

2. Background and Related Work

Several digital resources that provide information about literary works and writers are available online. Wikidata [2] is a general-purpose KG which includes knowledge about writers and their works. Other archives are domain-specific: Goodreads is a social cataloging website owned by Amazon, where readers share their impressions about books. Open Library is a project of the Internet Archive[2] where users can borrow books. Among these three archives, only Wikidata relies on the Linked Open Data paradigm. Open Library exposes its data through APIs, while Goodreads dismissed its APIs in 2020. This leads to issues in data gathering and mapping, since there is no unified model to align these resources.

Some digital archives are monographic and curated by teams of experts. It is the case of The European Literary Text Collection[3] [3], a multi-lingual dataset of novels written from 1848 to 1920; DraCor[4] [4], a collection of plays corpora in multiple languages; MiMoText[5], a parallel corpus of French and German novels published from 1750 to 1799.

Other resources are more oriented to explore the intersection between people and society. The Japanese Visual Media Graph[6] [5] gathers data about Japanese visual media (including manga and visual novels) from communities of fans. The Orlando Textbase[7] [6] is a KG developed to explore feminist literature. WeChangeEd[8] [7] is a KG of $1,800$ female editors born between 1710 and 1920 aligned with Wikidata.

[2] https://archive.org
[3] https://www.distant-reading.net/eltec
[4] https://dracor.org
[5] https://mimotext.github.io
[6] https://jvmg.iuk.hdm-stuttgart.de
[7] https://www.artsrn.ualberta.ca/orlando
[8] https://www.wechanged.ugent.be

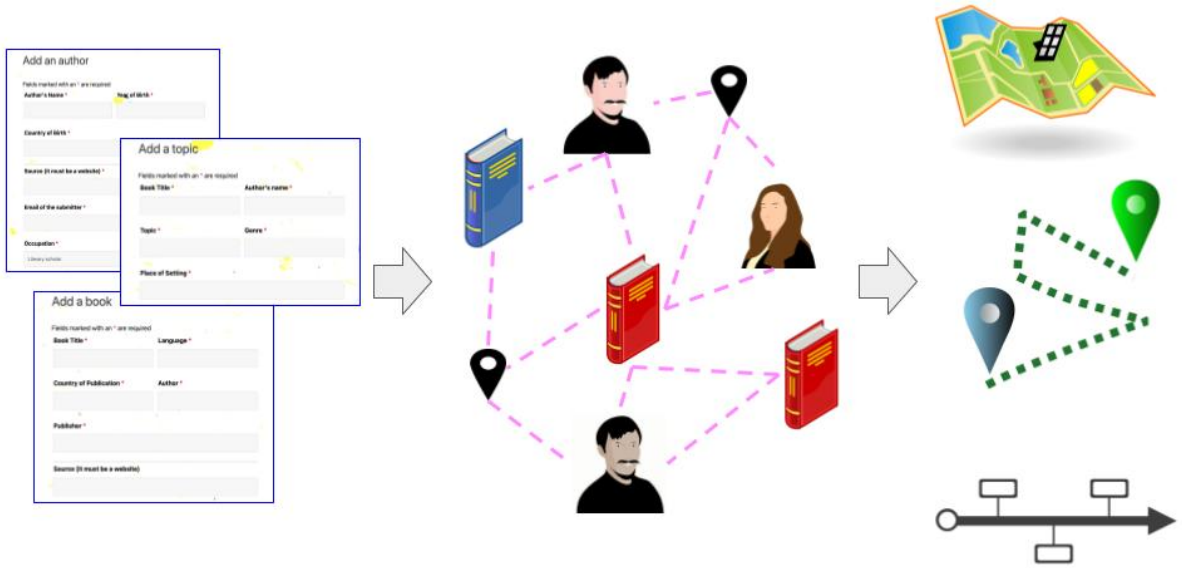

Figure 1: A sketch of the CON.NE.C.T W.O.N.D.E.R.S project structure: on the left, the apparatus for data collection; middle, the knowledge graph where connections between writers, places and locations are represented; right, the geographical and temporal representations extracted from the graph.

In CON.NE.C.T W.O.N.D.E.R.S, we tap from this research methodology by using the World Literature Knowledge Graph to alleviate two main issues. On the one side, we aim at bridging local repositories of literary data, remodeling them according to a common model when needed, to overcome the fragmentation of the research about the rural, mountain, cross-border areas across Europe. On the other side, we aim at creating applications that leverage the networked representation of literary, geographical and temporal data for the discovery and exploitation of new paths and connections in the field of literary tourism.

3. CON.NE.C.T W.O.N.D.E.R.S: Overview

CON.NE.C.T W.O.N.D.E.R.S[9] gathers four partners from the UNITA consortium[10], namely University of Turin (Italy), University of Pau (France), West University of Timisoara and Transilvania University (Romania), all located in rural, mountain, and cross-border regions across Southern, Western and Central-Eastern Europe. The overview of the project structure can be seen in Figure 1.

3.1. Project objectives and methods

To achieve the objective of experimenting a novel research framework that relies on the networked representation of writers, works and places, each partner has identified a case study connected with the research interests of its local team and the characteristics of its territory. The ultimate goal of the project is to leverage this representation not only to study the local patterns of writers, works and places and the connection over them at the cross-regional and cross-national level, but also to design novel proposals for the valorization of the territory through literary tourism initiatives. The latter may include, for example, itineraries consisting of locations mentioned in a given work, or point of interests connected with the biographical events of a writer, with a preference for local and less represented works and writers, with the ultimate goal of inspiring the creation of novel itineraries and

[9] https://connectwonders.di.unito.it/
[10] https://univ-unita.eu/

experiences that join territory and literature. To support the exploration of the case studies and the design of tools for the valorization of literary heritage on a geographical basis, the consortium will prototype a set of applications that leverage the networked representation of writers, works and places to create itineraries, timelines and other types of interactive visualizations from the knowledge graph in a semi-automatic fashion.

3.2. Case studies

In the initial phase of the project, the case study explored by each partners have been identified in cooperation with local stakeholders (museum, archives, and associations and professionals operating in the touristic field):

Case study 1 – Tradition Meets Technology: Innovative Solutions at Casa Mureșenilor Museum. This case study revolves around Casa Mureșenilor Museum in Brașov, home to one of Romania's most significant family archives—comprising over 25,000 documents (letters, official records, publications, photographs). Casa Mureșenilor Museum[11] is at the forefront of integrating technology to preserve and promote cultural heritage. The Mureșianu family, notable as the proprietors of Gazeta Transilvaniei, the first political newspaper of Transylvanian Romanians, takes center stage through groundbreaking projects that enhance the visitor experience through technology (use of virtual reality technologies to offer a captivating encounter with the museum's archives, an interactive AI-powered avatar delivering personalized information about the museum, exhibits, and history, a gamified Virtual Tour inviting audiences to unravel the secrets of the Mureșianu family through an interactive digital narrative).

Case study 2 – Revisiting the Pyrenees in sounds and pictures. This case study taps from the "RESPYR" project, initiated in 2021, which studies the mountain landscape by crossing several approaches and several views and by proposing a dialogue between specialists of different eras and disciplines [8]. The starting point of the project is anchored in the nineteenth century and focuses on the study of accounts, poems and drawings by British travelers in the Pyrenees and on the representation of the mountains by major writers of the time, such as John Ruskin. The scientific challenge of this project is twofold: to carry out research on a very local scale on the British presence in the Pyrenees and to participate in larger-scale projects on the representations of the mountain landscape on a European scale in order to participate in the development of landscape studies in the field of Anglophone studies.

Case study 3 – Transylvania and the Banat in British travel writing. Seen through British travellers' eyes in the nineteenth century, the Carpathians in the Banat region and in Transylvania are sources of historical, geographic and ethnographic richness [9]. English travel accounts have many common features, ranging from the wilderness of the landscape, the greatness of the mountains and their sublime, depicted in Major E. C. Johnson's "On the Track of the Crescent", to the melancholy feeling stirred in Charles Boner's Transylvania: its products and its people. Some narratives are enriched with personal sketches of animate or inanimate features, military men, local peasants or milk women dressed in simple traditional costumes coming to Herculesbad or Băile Herculane, or the "magnificent scenery" on the banks of the Danube, the bubbling waters and whirlpools through the Kasan Pass. On the other hand, Transylvanian castles, such as Hunyadi Castle and the fortress of Deva, are depicted as imposing places which fall into ruin and desolation.

Case study 4 – Travel and literature: practices and authors in the French-speaking world of yesterday and today. This case study proposes a reflection on tourism and literature encompassing two main areas. The first focuses on literary tourism, with particular attention to the practices of trekking and literary walks, which combine the physical experience of walking with the discovery of places evoked by literature. The second explores the contribution of the Swiss writer Rodolphe Töpffer (19th century), known for his "voyages en zigzag" in the Alps and the resulting writings, which interweave narration, geography and autobiography, giving rise to an interesting reflection on the experience of the traveller and the tourist [10, 11].

In the current phase of the project, partners are carrying out the integration of the data collected within each case study in the World Literature Knowledge Graph using a set of web forms specifically

[11] https://muzeulmuresenilor.ro/

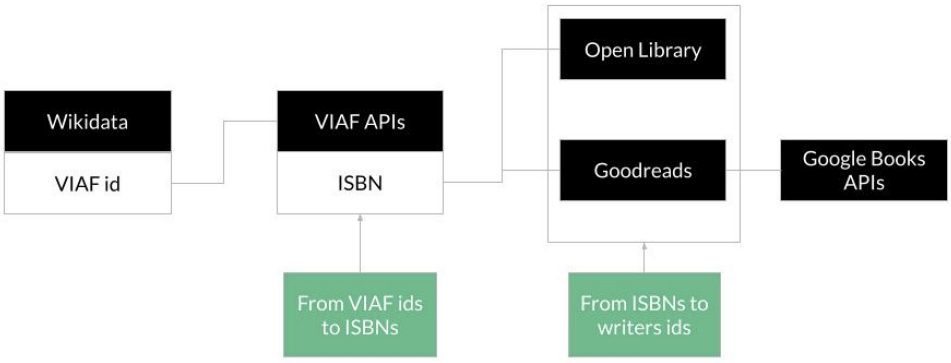

Figure 2: Semantic alignment of resources in the construction of the WL-KG

designed to alleviate the task of translating the knowledge about literary facts into an RDF format, leaving to conversion procedures the task of carrying out the translation from the ingestion format to the format required by the knowledge graph. In parallel, we are developing and testing APIs that allow applications to retrieve paths connecting writers, works and places from the knowledge graph.

4. The World Literary Knowledge Graph

The ecosystem of digital archives of literature is vast, but fragmented, and not all resources acknowledge the Linked Data paradigm. For instance, there is no systematic mapping of writers' pages on Wikidata onto other sources such as OpenLibrary[12] and Worldcat[13]. In addition, underrepresentation of minorities is a long-lasting problem that affects both digital and traditional media. Silencing practices [12] relegated ethnic minorities and non-Western people to a marginal role in textbooks [13], movies [14], and digital archives [15].

4.1. Knowledge Graph construction

Developed to provide a data-driven representation of World Literature while reducing at the same time the gap between mainstream literature and underrepresented writers and works, the World Literature Knowledge Graph[14] (WL-KG) [16, 1, 17] is a knowledge base of writers and their works gathered from Wikidata and aligned with three external archives: OpenLibrary, Goodreads, and Google Books.

The creation of the WL-KG relied on two main strategies. The first was based on the semantic alignment of resources: writers extracted from Wikidata were aligned with two other public archives, Open Library[15] and Goodreads[16], and mapped onto the identifiers from VIAF and Open Library; the resulting graph was further augmented with literary works from Open Library and Goodreads (see Figure 2). Aligning literary facts from different platforms in a single semantic resource allows for a richer representation of the World Literature, with a more balanced knowledge about writers from different areas, also thanks to the inclusion of the readers' communities. The second strategy was based on the automatic extraction of the writers' biographical triples from English Wikipedia pages. The pipeline combines the methodology described in [18], which relies on an annotated corpus of writers' biographies called WikiBio, with an approach based on Lexico-Semantic Patterns [19] to automatically extract relations belonging to four career-relevant properties for writers on Wikidata: 'educated at' (P69), 'employer' (P108), 'award received' (P166), and 'nominated for' (P1411).

[12] https://openlibrary.org
[13] https://www.worldcat.org
[14] Funded by the by Next Generation Internet (NGI) Search Programme - "Change the way we use and experience, search and discover data and resources on the internet and web", 2022-2023.
[15] https://openlibrary.org/
[16] https://www.goodreads.com/

4.2. The UR-ON ontology network

Reflecting the commitment of the WL-KG towards the mitigation of underrepresentation in the literary domain, writers and books in the WL-KG are represented according to the Under-Represented Ontology Network (UR-ON)[17] [1], a network of two domain ontologies aimed at encoding life events of potentially under-represented writers and their works: The Under-Represented Writers Ontology (URW-O), which allows drawing a link between a writer's life and their cultural production, and the the Under-Represented Books Ontology (URB-O), which introduces the publication event, a concept that encodes a number of information about a work and its production process.

URW-O provides the implementation of the biographical patterns to in order to represent two main situations, namely the process of migrating, and the status of a person in a given country. Both are embodied in a specific time interval, and this relation of time-dependency need to be formally expressed for two reasons: on one side, it is essential to order life events in a chronological fashion; on the other side, it allows drawing a link between a writer's life, and their cultural production.

URB-O is mapped onto the Functional Requirements for Bibliographic Records (FRBR) [20], a standard for modeling the relationship between a work (FRBR:Work), its expressions (FRBR:Expression), and manifestations (FRBR:Manifestation). Following the FRBR ontology, we defined a work as an instance of type FRBR:Expression, which is described as the 'intellectual or artistic realization of a work in the form of alpha-numeric, musical, or choreographic notation'. We then defined the concept of URB:Edition as a subclass of FRBR:Manifestation, namely 'the physical embodiment of an expression of a work'. These two concepts are linked through the property **frbr:embodiment**. Each semantic relation between an expression and its edition is wrapped in a URB:Publication pattern, which is a subclass of a DUL:Event, an event in DOLCE can be used as a reification to provide rich descriptions of something that happens or occurs. Finally, the model integrates the PIM:Reception pattern with a number of attributes that are specific to the reception of literary works. Depending on the source of knowledge from which a work is derived, it may have an average rating (**urb:rated**), a number of ratings (**urb:numberOfRatings**), or a number of readers (**urb:numberOfReaders**)

4.3. Graph visualization

The fruition of the WL-KG is supported by a dedicated graphical interface, designed with the goal of promoting the exploration of the connections between literary works. The navigation flow that starts with an initial search for a topic of interest. Once a relevant topic is found, the user can drag the resource onto the central board and explore its relationships with other objects and predicates, creating a visual representation of the connections.

By clicking on resources of type "Person" (as visible in Figure 3), the user can access information about an author, including both direct relationships such as published works and indirect relationships such as all the topics covered in their works, or a map of all the locations where their works were published. Clicking on resources of type "Expression" (as visible in Figure 3) displays information specific to a particular work, such as editions, languages, and readers ratings.

The platform also allows subject-based navigation: users can browse all works linked to a specific item from the URB:Folksonomy. The graph-based navigation encourages serendipitous discovery, allowing users to stumble upon unexpected connections and relationships: for example, the Italian writer Italo Calvino and the American writer Stephen King share the genre termed "speculative science fiction novel" to which their respective books "The Nonexistent Knight" and "The Dark Tower" belong.

The visualization platform has been evaluated with the help of domain experts who have used it to perform search task and have been subsequently requested to fill in a questionnaire about their experience (details can be found in [1]).

[17] https://purl.archive.org/urwriters, https://purl.archive.org/urbooks

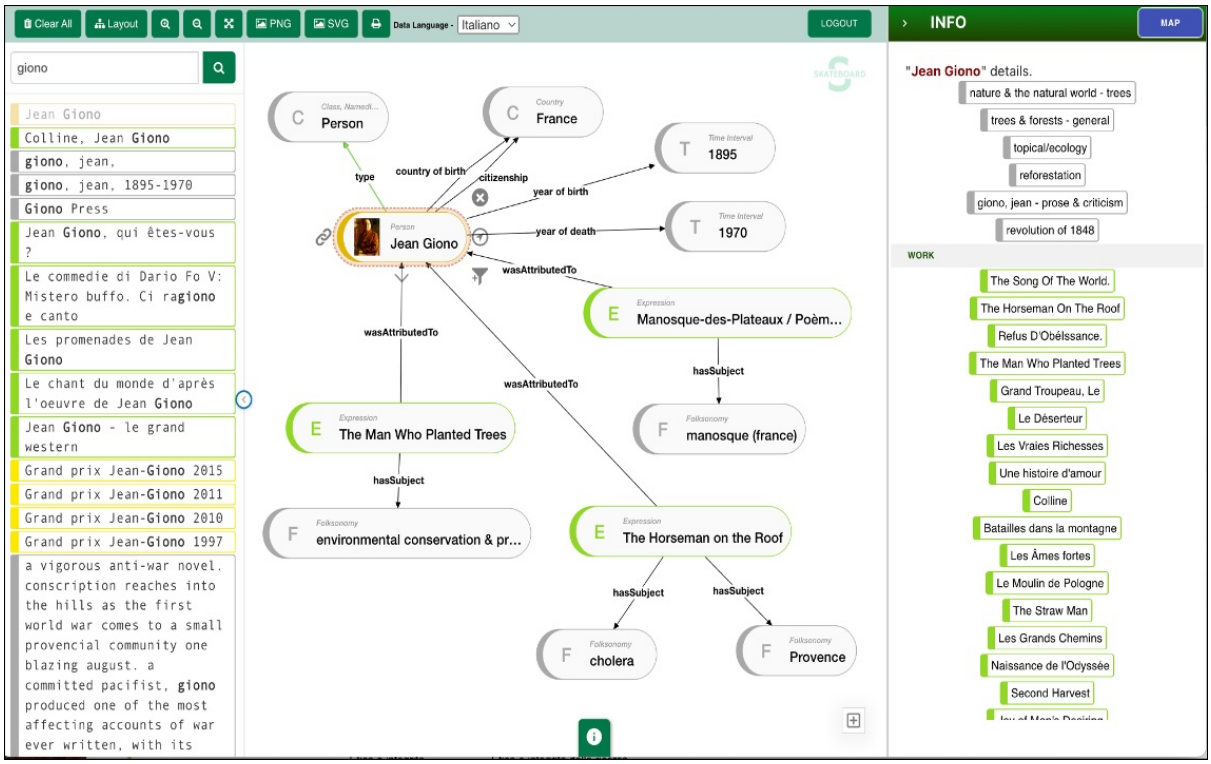

Figure 3: A snapshot of the visualization platform showing the French writer Jean Giono and a selection of his books about the Provence region. On the left, the search box; in the middle, the whiteboard where entities can be dragged; on the right, info pane about the selected entity

5. Integrating curated datasets in the WL-KG

5.1. Data modeling and gathering

Each case study in CON.NE.C.T W.O.N.D.E.R.S brings with itself variations in the description of the domain entities, which need to be reconciled with the data model provided by URW-O and URB-O on a case by case basis. For example, the case study about Transylvania and the Banat does not include websites as sources for data, differently from the other case studies. Or else, the publication date is not always known in the case study about the Alps.

To do so, we have designed a preliminary set of forms for each of the core domain entities using the

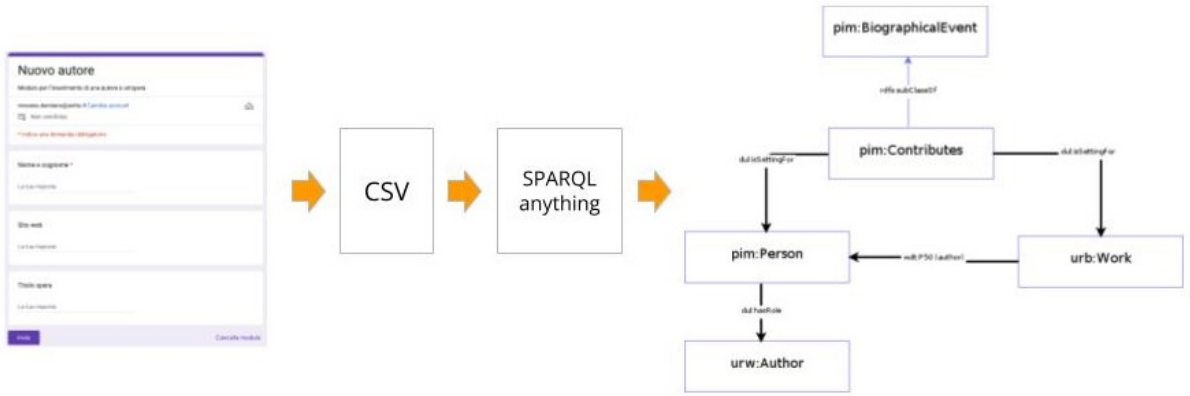

Figure 4: Data ingestion and integration pipeline. On the left, the form for a record of type Writer, exported in CSV format and then processed using SPARQL Anything (middle) to obtain a set of triples that match the ontology pattern on the right.

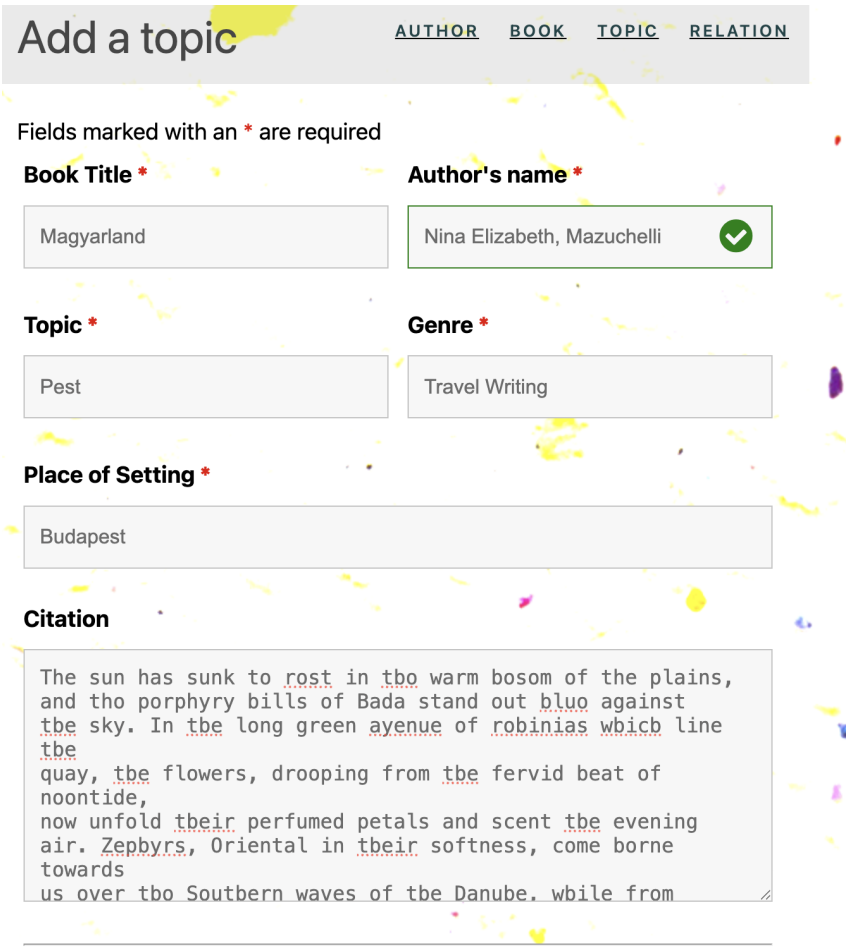

Figure 5: The template for the ingestion of literary facts in the KG. In the example the template is filled with information about the work *"Magyarland": Being the Narrative of Our Travels Through the Highlands and Lowlands of Hungary*, its genre, the place of setting, and topic

domain ontologies as a reference, asking each partner to provide a set of example entries and map them onto the basic forms to let discrepancies emerge. Based on this feedback, all requirements have been merged to create a set of forms that include all the required fields for each case study, asking partners to produce a set of guidelines for the ingestion of the various entries (books, writers, places) for each case study (see Figure 5 for an example). For each form, a mapping with the URW-O and URB-O ontologies has been defined. The rationale behind the decision to create a single set of forms is not only the need to make the integration of records into the knowledge graph simpler, but also to create a common ground to all case studies that facilitates the emergence of connections and similarities between them since the data modeling and gathering phase.

Once exported in CSV formats, records are imported into the knowledge graph using a set of scripts that rely on SPARQL Anything for translating the input records into a set of RDF triples that fit the patterns defined for each entity type, according to the pipeline described in Figure 4.

5.2. Data integration

The W3C standard SPARQL 1.1 [21] is the reference language for interacting with RDF knowledge graphs. SPARQL has methods for selecting, filtering, and aggregating data into tabular form. In addition, SPARQL can also project the result into an RDF template using the CONSTRUCT query type. For this reason, research has explored the application of SPARQL to a range of use cases broader than querying, specifically, the integration of heterogeneous data sources [22, 23, 24]. In practice, approaches rely on extending SPARQL to access data in non-RDF formats.

```
PREFIX urw: <https://purl.archive.org/urwriters#>
PREFIX urb: <https://purl.archive.org/urbooks#>
PREFIX prov: <http://www.w3.org/ns/prov#>

urb:magyarland a urb:Expression;
    prov:wasAttributedTo urw:nina-elizabeth-mazuchelli;
    urb:subject urb:pest;
    urb:genre urb:travel-writing;
    urb:setting urw:budapest;
    prov:wasDerivedFrom https://archive.org/details/magyarlandbeing01unkngoog
```

Figure 6: A snapshot of materialized triples about the work *"Magyarland": Being the Narrative of Our Travels Through the Highlands and Lowlands of Hungary* in the format of the WL-KG after the data ingestion done by the domain experts.

Recent research [25] proposes to rely on an intermediate RDF model, named Façade-X, whose components can be transparently mapped to various file formats. This method allows building software that provides *indirect access* to source data as-RDF, relieving knowledge engineers from the task of dealing with the variety of formats and related languages they rely upon – *re-engineering* (i.e. transforming resources by minimising domain considerations and focusing on the syntactical meta-model), and letting them focus on the semantic lifting – *remodelling* (i.e. re-framing the original domain model into a new one) [25]. Façade-X enables uniform access to a wide range of data formats *as-if* they were RDF, including the popular CSV, JSON, HTML, and XML and it was successfully applied to many complex scenarios, from scraping the content of Web sites, joining data from multiple sources, and even building knowledge graphs from Music scores [26]. Façade-X can, in principle, represent any format expressed in a BNF grammar, as well as the relational data model [27]. The Façade-X approach is at the basis of the SPARQL Anything open source project, which also constitutes the reference implementation[18]. The software has been applied in many real-world scenarios and is receiving increasing attention from the KG community in both academia and industry[19]. The method is accessible as a command-line interface, as a server, and from Java and Python code [28][20].

The data acquisition and integration process is described in Figure 4. Users input information in a Google form. The data is exported in tabular format. Next, a SPARQL Anything query extracts and maps the data to the URW ontology, generating the output knowledge graph. Figure 7 shows the query developed to extract information about books.

6. Example

In this section we present an example of KG population based on the *Case study 3 - Transylvania and the Banat in British travel writing*. Through the engagement with domain experts on this topic we identified a gap in Wikidata about the literary work *"Magyarland": Being the Narrative of Our Travels Through the Highlands and Lowlands of Hungary* [29].

Despite the existence of Nina Elizabeth Mazuchelli in Wikidata[21], there are no mentions of her work in this knowledge base. Therefore, partners from the Universitatea de Vest din Timisoara integrated the World Literature KG with additional information about this literary work. To make their contribution easier, the project team created a set of templates for data ingestion, are available on the official website of the project[22].

As it can be observed in Figure 5, the web form contains facts about this work: besides the mention of the work itself, experts added the topic, the genre, the place of setting and a relevant quote from

[18] http://sparql-anything.cc
[19] See the activity on the open-source project page on GitHub: http://github.com/sparql-anything/sparql.anything
[20] See also the extensive online documentation: https://sparql-anything.readthedocs.io/
[21] https://www.wikidata.org/wiki/Q56025933
[22] https://connectwonders.di.unito.it/contribute-kg/

```
PREFIX fx: <http://sparql.xyz/facade-x/ns/>
PREFIX xyz: <http://sparql.xyz/facade-x/data/>
PREFIX schema: <http://schema.org/>
PREFIX lwt: <https://litetaryworktopics.com/litetaryworktopics2025#>
CONSTRUCT {
    ?bookEntity a lwt:Edizione .
    ?bookEntity dc:Title ?cleanEditionTitle .
    ?bookEntity lwt:ScrittoDa ?authorEntity .
    ?bookEntity lwt:PubblicatoDallOrganizzazione ?publisherEntity .
    ?bookEntity lwt:HaDataPubblicazione ?publicationDate .
    ?authorEntity a lwt:Autore .
} WHERE {
    SERVICE <x-sparql-anything:location=book_details.csv,csv.headers=true> {
        ?row <http://sparql.xyz/facade-x/data/Titolo%20Edizione> ?editionTitle .
        BIND (REPLACE(?editionTitle, "\"", "") AS ?cleanEditionTitle)
        BIND (fx:entity(lwt:, REPLACE(?cleanEditionTitle, " ", "_", "i"))
        AS ?bookEntity)
        ?row <http://sparql.xyz/facade-x/data/Nome%20Cognome%20Autore> ?authorName
        BIND (REPLACE(?authorName, "\"", "") AS ?cleanAuthorName)
        BIND (fx:entity(lwt:, REPLACE(?cleanAuthorName, " ", "_", "i"))
        AS ?authorEntity)
        ?row <http://sparql.xyz/facade-x/data/Data%20di%20pubblicazione> ?publicat
            .
        ?row <http://sparql.xyz/facade-x/data/Casa%20Editrice> ?publisherName .
        BIND (REPLACE(?publisherName, "\"", "") AS ?cleanPublisherName)
        BIND (fx:entity(lwt:, REPLACE(?cleanPublisherName, " ", "_", "i"))
        AS ?publisherEntity)
} }
```

Figure 7: SPARQL Anything query to extract information from a book record and generate the related knowledge graph.

the book. After the insertion we used SPARQL Anything to convert the knowledge provided by the experts in triples (Figure 6) that were subsequently added to the KG, thus filling a gap in the existing knowledge base.

7. Conclusion and Future Work

In this paper, we illustrated the process by which information about writers, books and locations concerning mountain areas are integrated in an existing knowledge graph of world literature, the World Literature Knowledge Graph, as part of the CON.NE.C.T W.O.N.D.E.R.S. project involving four European universities from the UNITA alliance located in rural and mountain areas. In particular, we described the procedure for ingesting the information about the different type of domain entities through web based forms that alleviate the data gathering process for the domain experts, leaving to a set of SPARQL Anything scripts the task of triplifing the input data according to the model provided by the reference ontologies. The rationale behind this process in to overcome the fragmentation of the landscape of mountain-related literature and make it available for the study and development of rural, mountain and border areas across Europe. Most of the data sources for this endeavor, in fact, are currently not represented in digital form or maintained in local repositories.

As future work, we envisage two main activities: first, searching for common patterns and connections between the case studies by leveraging the integrated representation provided by the World Literature Knowledge Graph; second, using the extracted paths to create novel literary and touristic itineraries in the areas studied by the project in a semi-automatic fashion.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] M. A. Stranisci, E. Bernasconi, V. Patti, S. Ferilli, M. Ceriani, R. Damiano, The world literature knowledge graph, in: International Semantic Web Conference, Springer, 2023, pp. 435–452.

[2] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Communications of the ACM 57 (2014) 78–85.

[3] C. Schöch, M. Eder, C. Odebrecht, M. Kestemont, A. Primorac, J. Tonra, K. M. Poniž, C. Kanellopoulou, Distant reading for european literary history. a cost action, Proceedings of DH2018 (2018).

[4] F. Fischer, I. Börner, M. Göbel, A. Hechtl, C. Kittel, C. Milling, P. Trilcke, Programmable corpora: Introducing dracor, an infrastructure for the research on european drama, Digital Humanities 2019 (2019) 5.

[5] M. Pfeffer, M. Roth, Japanese visual media graph: Providing researchers with data from enthusiast communities, in: International Conference on Dublin Core and Metadata Applications, 2019, pp. 136–141.

[6] J. Simpson, S. Brown, From xml to rdf in the orlando project, in: 2013 International Conference on Culture and Computing, IEEE, 2013, pp. 194–195.

[7] M. Van Remoortel, J. M. Birkholz, M. Alesina, C. Bezari, C. D'Eer, E. Forestier, Women editors in europe, Journal of European Periodical Studies 6 (2021) 1–6.

[8] L. Roussillon-Constanty, Revoir les pyrénées: Le voyage aux eaux sous la plume des voyageuses britanniques, Oltre la crisi. Il patrimonio ambientale e culturale transfrontaliero: sfide, potenziale, prospettive (2024) 51–65.

[9] M. Crişan, 19th century oradea: The reflections of a multiethnic city in british travel literature., Romanian Review on Political Geography/Revista Română de Geografie Politică 13 (2011).

[10] C. Trinchero, et al., «je ne suis pas un touriste»: i patrimoni culturali del viaggio in italia di jean giono, in: Valorizzazione della macroarea italo-francese per un turismo sostenibile. Riflessi culturali, sociali ed economici, volume 8, Edizioni della Associazione Culturale Antonella Salvatico-Centro …, 2023, pp. 115–141.

[11] M. Begliuomini, et al., Zig-zag fra le alpi di rodolphe töpffer, in: Valorizzazione della macroarea italo-francese per un turismo sostenibile. Riflessi culturali, sociali ed economici, Edizioni della Associazione Culturale Antonella Salvatico, 2023, pp. 87–96.

[12] G. C. Spivak, Can the subaltern speak?, in: Colonial discourse and post-colonial theory, Routledge, 2015, pp. 66–111.

[13] A. Wolf, Minorities in us history textbooks, 1945–1985, The Clearing House 65 (1992) 291–297.

[14] M. Erigha, Race, gender, hollywood: Representation in cultural production and digital media's potential for change, Sociology compass 9 (2015) 78–89.

[15] J. Adams, H. Brückner, C. Naslund, Who counts as a notable sociologist on Wikipedia? Gender, race, and the "professor test", Socius 5 (2019) 2378023118823946.

[16] M. A. Stranisci, V. Patti, R. Damiano, et al., Representing the under-represented: A dataset of post-colonial, and migrant writers, in: D. Gromann, G. Sérasset, T. Declerck, J. P. McCrae, J. Gracia, J. Bosque-Gil, F. Bobillo, B. Heinisch (Eds.), 3rd Conference on Language, Data and Knowledge, LDK 2021, September 1-3, 2021, Zaragoza, Spain, volume 93 of *OASIcs*, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021, pp. 7:1–7:14. URL: https://doi.org/10.4230/OASIcs.LDK.2021.7.

[17] M. A. Stranisci, V. Patti, R. Damiano, User-generated world literatures: a comparison between two social networks of readers, in: A. Falcon, S. Ferilli, A. Bardi, S. Marchesin, D. Redavid (Eds.), Proceedings of the 19th The Conference on Information and Research science Connecting to Digital

and Library science, IRCDL 2023, Bari, Italy, February 23-24, 2023, volume 3365 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 38–46. URL: https://ceur-ws.org/Vol-3365/short3.pdf.

[18] M. A. Stranisci, R. Damiano, E. Mensa, V. Patti, D. Radicioni, T. Caselli, et al., Wikibio: a semantic resource for the intersectional analysis of biographical events, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, Association for Computational Linguistics, 2023, pp. 12370–12384.

[19] M. A. Stranisci, V. Basile, R. Damiano, V. Patti, et al., Mapping biographical events to odps through lexico-semantic patterns, in: Proceedings of the 12th Workshop on Ontology Design and Patterns (WOP 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021), volume 3011 of *CEUR WORKSHOP PROCEEDINGS*, CEUR-WS, 2021, pp. 1–12. URL: https://ceur-ws.org/Vol-3011/paper3.pdf.

[20] B. B. Tillett, Frbr and cataloging for the future, Cataloging & classification quarterly 39 (2005) 197–205.

[21] S. Harris, A. Seaborne, SPARQL 1.1 Query Language, W3C Recommendation, W3C, 2013. Https://www.w3.org/TR/2013/REC-sparql11-query-20130321/.

[22] R. Cyganiak, Tarql (sparql for tables): Turn csv into rdf using sparql syntax, Technical Report, Technical Report, 2015. Available at: http://tarql.github.io, 2015.

[23] M. Lefrançois, A. Zimmermann, N. Bakerally, A sparql extension for generating rdf from heterogeneous formats, in: Proc of ESWC, Springer, 2017, pp. 35–50.

[24] F. Michel, C. Faron-Zucker, F. Gandon, SPARQL micro-services: lightweight integration of web APIs and linked data, in: LDOW@ WWW, 2018.

[25] E. Daga, L. Asprino, P. Mulholland, A. Gangemi, Facade-X: an opinionated approach to SPARQL anything, Studies on the Semantic Web 53 (2021) 58–73.

[26] M. Ratta, E. Daga, Knowledge graph construction from musicxml: An empirical investigation with sparql anything, Proceedings of the first workshop on Musical Heritage Knowledge Graphs (MHKG), co-located with the 21st International Semantic Web Conference (ISWC) (2022).

[27] L. Asprino, E. Daga, A. Gangemi, P. Mulholland, Knowledge graph construction with a façade: A unified method to access heterogeneous data sources on the web, ACM Trans. Internet Technol. (2023). URL: https://doi.org/10.1145/3555312. doi:10.1145/3555312.

[28] PySPARQL Anything Showcase, Adjunt proceedings of the Extended Semantic Web Conference (ESWC), Posters and Demos (2024).

[29] N. E. Mazuchelli, " Magyarland;": Being the Narrative of Our Travels Through the Highlands and Lowlands of Hungary, volume 1, London: S. Low, Marston, Searle, & Rivington, 1881.