THE LEGEND OF THE SWORD IN THE STONE

PROJECT

This **ontology-based** project was developed for the Information Science and Cultural Heritage course at the University of Bologna, taught by Professor Francesca Tomasi and Professor Marilena Daquino.

It draws inspiration from the centuries-old Arthurian legend to collect, analyze, and semantically interlink a curated corpus of **cultural heritage items** —including manuscripts, artworks, historical documents, and audiovisual resources—related to the Arthurian cycle at various levels.

Through **Link Open Data** methodologies, the project constructs a dynamic knowledge graph, in which entities, drawn from diverse sources, are interconnected across disciplines and institutions. This not only enriches academic research but also ensures that the Arthurian legacy—from medieval chronicles to modern film adaptations—is preserved, accessible, and constantly reinterpreted.

LOD are the foundation of the **Semantic Web**, an evolution of the traditional web that aims to make data understandable to machines as well. The goal is to transform the web into a network of semantically interconnected data.

To achieve this, the metadata provided by the institutions holding the selected items were transformed into **semantic triples** composed of a subject, a predicate, and an object. This is a key step in our project, and it's where **RDF** becomes essential. RDF is one of the fundamental technologies of the Semantic Web, used to structure and link data on the web. In addition to RDF, other technologies like **TEI/XML** were used in the text analysis section.

STUDY OF THE DOMAIN

IDEA

"It is said that once upon a time,
England flourished with brave knights;
the good king died without an heir,
and so everyone coveted power.
Only a miracle could save
the kingdom from war and destruction:
it was the sword in the stone
that appeared one day".- The Sword in the Stone, 1963

The project's conceptualization began with a collective discussion driven by our shared interest in the **Arthurian cycle**. From this initial exploration, the myth of **The Sword in the Stone** emerged as a particularly compelling subject. This theme not only provided a strong foundation for our collaborative work but also offered a rich narrative framework for a detailed analysis of its characters, symbols, and concepts. Our unanimous enthusiasm for this topic became a driving force, motivating us throughout the project's development.

The legend is one of the foundational myths of Western tradition, rich in symbolic and narrative elements that have endured through the centuries. Its origins lie in Celtic traditions and the oral tales of ancient Britain, gradually evolving into a true saga thanks to medieval reinterpretations, particularly starting with Geoffrey of Monmouth's Historia Regum Britanniae (12th century). This gave rise to the tradition of the Breton cycle.

At the heart of the legend stands **King Arthur**, a just and noble ruler, chosen according to tradition by drawing the mythical sword Excalibur from the stone—a symbol of divine legitimacy and power. At his side appears the wise and mysterious **Merlin**, enchanter and advisor, embodying arcane knowledge and magical forces that guide and protect the destiny of the king and his realm: the mythical Camelot.

The popularity of the legend did not fade in the Middle Ages. The Arthurian cycle enjoyed vast literary and artistic success, eventually leading to modern adaptations in fantasy novels, graphic novels, films, and television series. In cinema and TV, works such as Excalibur (1981), King Arthur (2004), and the BBC series Merlin have breathed new life into these characters, demonstrating the enduring relevance of the myth.

ITEMS

To ground our project in tangible cultural heritage, we conducted targeted research to identify relevant key cultural artifacts. This process was crucial for bridging the gap between the abstract narrative of the myth and its material representations, providing a valuable dataset for subsequent analysis and interpretation.

For example artworks were selected from the digitized collection of the Metropolitan Museum of Art in New York, while the soundtrack comes from MusicBrainz, an open music encyclopedia that collects and makes musical metadata available.

KNOWLEDGE ORGANIZATION

Metadata Analysis

The first activity carried out by the team was the identification of the **metadata standard** adopted by the providers of the selected items, with the goal of ensuring **interoperability**. In cases where the institution had not explicitly stated the standard in use, the team decided to apply either the standard commonly adopted by similar institutions or the one deemed most suitable and relevant to the nature of the object, according to semantic accuracy and the context of usage of the formal grammar chosen. In such cases, the symbol * was used.

While an institution like, for example, the Metropolitan Museum of Art provides an API for its metadata but does not explicitly declare the use of a pre-existing metadata standard, in our project we made the conscious decision to adopt **CDWA (Categories for the Description of Works of Art)**, the typical content standard for the museum context. This choice was motivated by the fact that the metadata provided by The Met appeared to faithfully reflect the categories defined by the CDWA standard—with fields such as Title, Style, Measurements, Materials, Current Location... This led us to deduce that, in the absence of an explicit declaration, The Met was likely inspired by this standard to structure its information.

The team also highlighted the distinction between a general classification of the items and a more specific and detailed description, allowing for a more precise definition of the type of object actually held by the institution or of its concrete artistic manifestation. For example: is it a printed or a electronic edition (es.eBook)? Is the soundtrack in CD or vinyl format? Can the item be broadly classified as a work of art, and more specifically as a statue, a painting, or a tapestry…?

Theoretical Model

The team built the theoretical model starting from the descriptions of individual items provided by the institutions.

The model was developed and structured on the basis of **triple statements in natural language**, following the subject–predicate–object format. This approach makes it possible to highlight both the connection of the items to the project's theme – the Legend of the Sword in the Stone – and the relationships among the different items. As stated in the section Metadata Analysis, not every institution provided a metadata standard that the team could use to describe each item, but even when this was possible, some connections – predicates – of some items have been altered for the sake of readability and controlled authority forms' recognition.

Additional information about the entities correlated to the items was also included to enrich the data and illustrate ulterior possible connections branching out among and from them. The team worked to make the network among information surrounding our project's concept as **interconnected** and differentiated as possible, for this reason different colours and shapes have been used to categorise the kind of information described (whether an item, an entity or a literal).

Furthermore, entities have been split into two categories specifically: the sharp-edged blue boxes for entities of any kind and the round-edged brown boxes for entities deemed most relevant for the topic and that were witnesses of many interconnections both between entities and the items themselves (such as Arthurian Romance, King Arthur, Excalibur, Merlin, etc…).

Conceptual Model

The definition of our Conceptual Model began with the identification and formalization of the relationships between the metadata of our selected cultural heritage items and the core theme of The Sword in the Stone legend. The objective was to elevate the entities, relationships, and concepts from our theoretical model to a **formal and abstract level**, assigning to each an appropriate **class and property** from existing vocabularies and ontologies.

As a starting point, we created **custom URIs** (Uniform Resource Identifier) to uniquely identify the entities in our project, establishing our base URI: **https://github.com/The-Sword-in-the-Stone-LOD/The-Sword-in-the-Stone-LOD/**.

To formally represent the subject–predicate–object relationships, we began with our project's items as the subjects. For each predicate, we sought the best way to express the relationship identified in the theoretical model by using the most suitable properties from existing ontologies. The object of the relationship was then modeled as a concrete entity.

The choice of which specific vocabulary, schema, or ontology to use for a given class or property was made on a **case-by-case** basis, depending on the nature of the entity being described and the metadata standards that might be relevant. We relied on vocabularies pertinent to our domain and the entities we wanted to represent. Specifically, for our bibliographic resources (such as the manuscript we analyzed), we primarily used the **BIBFRAME** schema, which is the standard for describing bibliographic information. For the museum and artistic items, we relied on **CIDOC CRM**, while for people, places, and organizations we employed **FOAF and Schema.org**.

This combination ensured excellent flexibility while maintaining semantic coherence across various data source.

The full list of namespace prefixes used in our ontology includes:

- @prefix **sits:** https://github.com/The-Sword-in-the-Stone-LOD/The-Sword-in-the-Stone-LOD/.
- @prefix **bf:** http://id.loc.gov/ontologies/bibframe/.
- @prefix **crm:** http://www.cidoc-crm.org/cidoc-crm/.
- @prefix **dc:** http://purl.org/dc/elements/1.1/.
- @prefix **dct:** http://purl.org/dc/terms/.
- @prefix **dbo:** http://dbpedia.org/ontology/.
- @prefix **dbr:** http://dbpedia.org/resource/.
- @prefix **foaf:** http://xmlns.com/foaf/0.1/.
- @prefix **gn:** http://www.geonames.org/ontology#.
- @prefix **gndo:** http://d-nb.info/standards/gndo#.
- @prefix **mo:** http://purl.org/ontology/mo/.
- @prefix **owl:** http://www.w3.org/2002/07/owl#.
- @prefix **rda:** http://rdaregistry.info/Elements/.
- @prefix **rdf:** http://www.w3.org/1999/02/22-rdf-syntax-ns#.
- @prefix **schema:** https://schema.org/.
- @prefix **skos:** http://www.w3.org/2004/02/skos/core#.

KNOWLEDGE REPRESENTATION

CSV FILES

We created a collection of **.csv files** (one table for item) with the full description of our 10 chosen cultural heritage items in natural language.

For some items, like specified in the section Metadata Analysis of our project, the types of predicates were specified by the institutions providing some kind of formal description of the item, whereas for some other the mapping was done by the members of the team according to personal choices that took into consideration: the accuracy of the predicate based on the kind of being described and the generalisation of the idea behind the semantic connection of the subject to the object for a following most accurate mapping in formal language.

RDF PRODUCTION

CSV Files in Formal Language

Starting from the CSV files produced in natural language, which represent the raw metadata of our items, we created CSV files in **formal language** (which can be accessed in a folder named ‘formal_language’ (< csv_files) on our GitHub repository) that we used as the starting point to generate our RDF dataset. The description in the CSV files named ‘[item-name]_formal.csv’ follows the same structure of the description of the chosen items in the natural language CSV files (that we collected in another folder titled ‘natural_language’ on our GitHub repository) which is ‘subject-predicate-object’ in order to clearly map the statements we had in natural language to the ones we were going to produce in formal language: specifically, we mapped the predicates to properties derived from **schemas, vocabularies, and ontologies** — the same ones already employed in the conceptual model — while the objects were linked to instances mapped with **authority-controlled forms** in order to provide disambiguation.

We also produced two files titled ‘additional’ which contain the description of connections and statements that weren’t directly related to the items but more so to entities linked to them. We have decided to describe through **custom URIs** entities that we deemed particularly relevant to the concept of our project such as Arthurian Romance, Middle Ages, King Arthur and Merlin, and so on – these entities are the same that can be seen pictured in a brown box in the theoretical model.

We used ontologies such as Dublin Core (dcterms), BIBFRAME, and CIDOC-CRM (crm) for the predicates, while for entity disambiguation we linked our URIs to authority resources such as VIAF (for authors and people), Wikidata (for concepts), and Geonames (for places).

To see our CSV files **click here**.

Python Script for Transformation

The production of the **RDF dataset** of all the items was done automatically through **Python** with the usage of libraries like rdflib, csv and DictReader.

What the script does specifically is:

- First of all, it initialises a graph which will contain all the statement triples of the items and the ‘additional’, and defines a dictionary of prefixes and their connected namespaces’ URIs;
- It then accesses the CSV ‘formal_language’ folder and takes all files of extension ‘.csv’, reads them through the function DictReader and then iterates through each of them and appends them to a list;

- Since we set up the structure of our CSV files with a header **‘Subject,Predicate,Object’**, the information are processed as rows and for each column analysed accordingly: the subjects and predicates are firstly split in prefixes and their names and full URIs are created for each of them; the objects instead are categorised whether or not they have a ‘:’ inside them or not. If there is, a full URI is created; if not, a parsing depending on the kind of object (plain text, integers or date years) creates a Literal of the according data type;
- The information are then added as triples in the graph initialised at the beginning and the script produces an output file of the triples (**‘full_dataset.ttl’**) serialised through **Turtle format** – that we chose for human-readability purposes.

TEXT ANALYSIS

XML/TEI document

Our encoding work focused on the selection of specific excerpts from Carlo Haugwitz’s work, *Il mago Merlino: Memorie, traduzioni, leggende*. We encoded the text in its **original 1865 edition**, excluding later versions. The encoding does not cover the entire work but only carefully **selected passages**.

In particular, we chose the sections that explicitly mention Merlin and Arthur, with special attention to Merlin’s figure and his **prophecies** concerning Arthur. This makes the project more targeted and relevant to its main theme: the legend of the Sword in the Stone.

To validate the XML/TEI document, we used an online validator and the XML/TEI plugin for VS Code.

Main sections of the XML/TEI document:

<teiHeader>

Contains all metadata relating to the work and its digital edition. It is divided into several subsections:

- **<fileDesc>** — Describes the **electronic resource**, including the title of the digital edition, the author, the digital editor, the licenses, and so on. It also provides a detailed description of the **original source**, with precise references to the **encoded pages** and the list of **works cited**, all linked to Wikidata.
- **<encodingDesc>** — Explains the editorial principles and encoding choices. For example, we wanted to use *lb/* to accurately represent the layout of the first edition of the work.
- **<profileDesc>** — Details non-bibliographic features of the text, such as language (Italian), text type (essay) and so on.

  It also includes structured metadata such as a list of organizations, three separate lists of people (authors and editors, historical figures, and legendary characters cited), and a list of places (historical and mythical). In the text, the names of **people or places** are marked with @ref, which points to the xml:id in the lists of the teiHeader, and from there they can also be linked to external identifiers, just as it was done. Viaf, Geonames, Pleiades and Wikidata have been used.

Furthermore, keywords considered meaningful for the project—such as celtic traditions and sword—are also included (both in Italian and English) to highlight central ideas of the Sword in the Stone legend.

<text>

Contains the actual text of the work.

- **<front>** — Includes the preliminary parts, such as the title page and part of the preface.
- **<body>** — Contains the selected excerpts from the main body of the text. Proper names are enclosed in *persName* and place names in *placeName*. Peoples or ethnic groups appear in *name type="ethnonym"*; abstract concepts relevant to the project are encoded with *term type="concept"*; while bibliographic sources cited in the text are marked with *bibl*.

From XML to HTML

We have produced a **stylesheet** with **XSL extension**, a file that contains the rules for identifying **XML nodes** in the source document (using **XPath**), specifying how to manipulate them, and saving them in a new HTML file.

In particular, **XSLT** processes the XML document by following its hierarchical tree structure, starting from the root (**/tei:TEI**) and recursively applying templates to all child elements. The XSL file defines templates that specify the rules for transforming specific XML elements into corresponding HTML elements.

The **XSLT** transformation process was structured to extract and reorganize the metadata from the TEI header of the XML document, converting them into dedicated and well-formatted HTML sections. All descriptive information has been extracted and presented: metadata from both the original 1865 edition and the current digital edition, as well as a specific section dedicated to displaying the list of cited **works, people and places** mentioned, linked to external resources such as **VIAF**, **Wikidata**, etc., along with the **keywords** highlighted in the text, all organized for clear and immediate consultation.

Regarding the processing of the actual text, specific **templates** were created to handle all structural elements (such as divisions into chapters and sections...) and textual elements (paragraphs, citations,...), preserving the original semantic attributes.

The adopted approach maintains a clear separation between content and presentation: the **XSL** stylesheet deals exclusively with structure and semantic transformation, while all visual formatting rules are delegated to embedded CSS, thus ensuring a well-structured and aesthetically coherent HTML output.

Once the XSL file was produced, both it and the XML file were uploaded to an online tool called Free Formatter to generate the corresponding **HTML file**, which was then converted into a **browsable HTML page**.

XML/TEI to RDF

To transform a TEI/XML file into an **RDF file**, a Python script was used (developed with the support of ChatGPT). The goal was to convert the data from its original hierarchical structure to a graph-based structure, which is characteristic of Linked Data.

The process took place in a single, but complex, operation. The script used a parsing library, such as lxml, to analyze the XML file's tree structure. This allowed it to navigate the document and identify key elements, such as the metadata within the and the main text in the . Subsequently, the program created an RDF graph in memory using the rdflib library, which it then populated with the extracted information.

Every relevant piece of information, from people to events, was mapped, assigning it a unique URI for identification, and linking it with properties from standard vocabularies like dcterms:creator or foaf:name. Once the mapping was complete, the script serialized the graph into readable formats such as Turtle (.ttl) and RDF/XML (.rdf), making the data ready for publication and sharing. Following this, the syntax of the files was checked for correctness using the W3C RDF Validator, and the result of the check was positive, confirming that the conversion adhered to standards; furthermore, another check for correctness was performed with RDF Grapher.

Click here to see XML/TEI to RDF transformation!