



Building a Knowledge Graph of Chinese Kung Fu Masters From Heterogeneous Bilingual Data

RESEARCH PAPER

YUMENG HOU

LIN YUAN

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

Various endeavours into semantic web technologies and ontology engineering have been made within the organisation of cultural data, facilitating public access to digital assets. Although models for conceptualising objects have reached a certain level of maturity, only a few have delved into the tacit roles of individuals in the development of knowledge. Simultaneously, the field of cultural analytics demands practical methods for integrating diverse multilingual materials into a consistent data representation. In this context, our work addresses a human-centred perspective to construct a knowledge graph presenting historical Chinese kung fu masters, aptly named **MA²KG** (**M**artial **A**rts **MA**sters **K**nowledge **G**raph). The data workflow is built upon an established ontology model that describes traditional martial arts. It aggregates information from heterogeneous bilingual sources through direct connections and rule-based inference, incorporating data from English Wikidata and Chinese Baidu Baike and complemented with manual annotations. In addition, we describe our methodology and process in making the dataset available with scripts for reproducing similar agent-mediated contexts, data application and inspection cases are provided to discuss our findings and concerns regarding the use of linked open data strategies.

CORRESPONDING AUTHOR:

Yumeng Hou

Laboratory for Experimental
Museology, EPFL, Lausanne,
Switzerland

yumeng.hou@epfl.ch

KEYWORDS:

knowledge graphs; social
networks; Chinese martial
arts; linked open data; digital
humanities

TO CITE THIS ARTICLE:

Hou, Y., & Yuan, L. (2023).
Building a Knowledge Graph of
Chinese Kung Fu Masters From
Heterogeneous Bilingual Data.
*Journal of Open Humanities
Data*, 9: 27, pp. 1–12. DOI:
[https://doi.org/10.5334/
johd.136](https://doi.org/10.5334/johd.136)

1 INTRODUCTION

Traditional martial arts, known as kung fu in Chinese, represent significant systems of mind-body practices and treasures of human knowledge. Globally renowned, partly thanks to the blossoming of kung fu movies in the past decades, these long-lived traditional practices have multiple registrations on UNESCO's lists of **Intangible Cultural Heritage (ICH)**. Even so, kung fu practices have gradually lost their public appeal with an ever-decreasing number of practitioners. There is an urgent need to rekindle public interest, especially among the younger generations, who have grown up as digital natives, and engage them in knowing and exploring martial arts knowledge.

To this end, semantic web technologies and ontology engineering have enabled various means to preserve ICH, creating public access to digitised and born-digital cultural records through united data pathways. While models that conceptualise material and object-based cultural aspects have reached a certain degree of maturity, few efforts have probed into the roles that individual people have played throughout knowledge development. This humanistic dimension is particularly relevant to martial arts, a distinct practice incorporating embodied, performative and ideological understandings, wherein knowledge transmission heavily relies on in-person contact. Consequently, numerous practitioners have contributed to the development and recreation of martial arts knowledge. Between the practitioners, there exist certain explicit or implicit relations likely to have influenced stylistic development.

In an attempt to tackle the human-centred perspective above, we constructed a **knowledge graph (KG)** representing the inter-relations between distinguished practitioners, namely the masters, and investigate its application to facilitate knowledge access as well as to study (culture) contacts through the various relations between individuals in history.¹ Specifically, this effort focuses on Chinese martial arts and introduces the **MA²KG (Martial Arts MAsters Knowledge Graph)**, an ontology-based data resource representing the networks of historical Chinese kung fu masters.² The computational pipeline (Figure 1) is built upon the data model of **Martial Art Ontology (MAon)** (Hou, 2023). It combines ontological linkage with rule-based inference to gather data from diverse sources into a consistent graph structure. The data sources include online structured databases of Wikipedia and its Chinese alternative Baidu Baike, in addition to manual transcription of the texts in the *Hong Kong Martial Arts Living Archive*, accordingly establishing a bilingual resource in English and Chinese.³

The MA²KG dataset has been published online with script examples for replication in similar agent-mediated contexts. In this article, we present our methods of building the KG, results, data application and inspection, which are followed by a further discussion of findings and concerns regarding the employment of **linked open data (LOD)** strategies.

2 RELATED WORK

As an increasing number of **cultural heritage (CH)** materials transition to the digital realm, challenges have emerged, especially for the GLAM (galleries, libraries, archives, and museums) sectors, in their efforts to assemble heterogeneous data into an operable resource that can be searched, studied, and presented (Bikakis, Hyvönen, Jean, Markhoff, & Mosca, 2021). Conventional approaches to this end have largely centred on cataloguing. However, their effectiveness is contingent on users understanding the logic behind data tagging, which seldom holds for the general public.

¹ Culture contact refers to contact between peoples with different cultural backgrounds, often leading to changes in both systems, such as in artefacts, customs, and beliefs.

² This article will use “kung fu” to denote traditional martial arts originating in the regions known as Greater China. The intention is to differentiate this specific system from others within the broad range of martial arts, e.g., karate, (classical) fencing, mixed martial arts, etc.

³ Wikidata is a free and open knowledge base hosted by the Wikimedia Foundation. <https://wikidata.org>. Baidu Baike (a collaborative Chinese encyclopedia) is considered the equivalent of Wikipedia in mainland China. <https://baike.baidu.com/>. The *Hong Kong Martial Arts Living Archive* has been an international research collaboration since 2012. The project encompasses a comprehensive digital strategy, including employing state-of-the-art motion capture systems and audio-visual archiving tools, amongst many others, to record and annotate the living kung fu traditions in Hong Kong (Chao, Delbridge, Kenderdine, Nicholson, & Shaw, 2018).

Seeking ways to operate and interoperate such various cultural objects has led to the rising use of semantic web technologies. Notably, LOD and KG have emerged as promising solutions to tackle the shortcomings of manual cataloguing. A common approach involves creating a network-like knowledge description model, known as domain ontology, and its application to connect data sources through explicit relationships or rule-based inferences. This new path has shown promise in facilitating public access to cultural collections, enabling casual users to initiate queries and explore data through linked connections without prior knowledge of the content. Achieving this in practice requires effective KG engineering, which involves structuring data elements into a formal conceptual framework and establishing connections between the data through techniques such as named entity recognition and data classification for querying (Rejeb et al., 2022).

KG is not a new approach. Formed as a network composed of nodes, edges, and labels (or properties), KG has been widely applied to illustrate real-world concepts via linking, relating, and analysing entities in massive datasets via semantic relationships (Dong et al., 2014). Interdisciplinary researchers have embraced this concept to enrich CH and ICH studies by conceptualising cultural entities using semantic standards. Concurrently, ontological engineering has been increasingly used in structuring cultural materials into programmatic structures, enabling data representation relating tangible and intangible identities detected from textual (Dou, Qin, Jin, & Li, 2018), visual (Caraffa, Pugh, Stuber, & Ruby, 2020), iconographical (Carboni & De Luca, 2019) and audiovisual (Meghini, Bartalesi, & Metilli, 2021) features in diverse contexts.⁴

Despite these existing efforts, humans have not been adequately represented as informative nodes, even though they consistently play a traceable role in interconnecting knowledge exchanges and communications. An exception is ArCO, which establishes a comprehensive ontology model describing Italian CH, incorporating a context description module that delineates humanistic information such as authors, collectors, copyright holders and inventories in relation to a CH entity (Carriero et al., 2019). Likewise, WarSampo KG provides a semantic infrastructure spanning dimensions such as *Persons* (soldiers), *Army Units*, *Places*, and *Events*, facilitating the coherent presentation of distributed data sources related to the Second World War (Koho et al., 2021).

At the same time, LOD has fostered new prosopography and social network analysis methods. Such formal network approaches are important for examining the relationships between intellectuals and the evolution of communities, a phenomenon researchers interpret as a small-world effect resulting from entangled webs of influences, interdependencies, and inspirations (Cline, 2020; Petz, Ghawi, & Pfeffer, 2022). In comparison to conventional network construction from a single literature or scholarly documentation, knowledge engineering allows researchers to extract data from a broader range of sources to build more complex networks. For example, the China Biographical Database (CBDB) integrates approximately 491,000 individuals (as of May 2021) whose lifespans range from the seventh through nineteenth centuries, along with over 228,000 biographical articles (S. Chen & Wang, 2022; Fuller & Wang, 2021). This forms a valuable resource for further studies, such as analysing kinship and themes based on trivial items of news or facts (Blouin, Magistry, & Van den Bosch, 2021). When these relational patterns of pivotal concepts, including eponyms, individuals, places, things, and times, are extrapolated appropriately, they serve as helpful tools for historical examination and supporting or disputing historical arguments (Bingenheimer, 2021; Breure & Heiberger, 2019). These technical and knowledge advances have laid the foundation for our investigation.

3 DATASET DESCRIPTION

- Object name: **MA²KG** (**M**artial **A**rts **MA**sters **K**nowledge **G**raph).
- Format names and versions: RDF data (TTL syntax).
- Creation dates: March to August 2022.
- Dataset creators: Yumeng Hou, Lin Yuan.

⁴ For more examples, see (Golub & Liu, 2022; Hou, Kenderdine, Picca, Egloff, & Adamou, 2022; Ziku, 2020).

- Language: English, Chinese.
- Repository name: Zenodo, <https://zenodo.org/record/8211203>.
- Publication date: 2023-08-03 (first publish date on GitHub: 2022-10-09).

4 METHODS AND RESULTS

As depicted in [Figure 1](#), our workflow involves data acquisition and knowledge generation to build a KG for both machine operation and human-led analysis. The computational process begins by extracting domain-specific entities and properties from LOD and importing the structured RDF triples into a graph database. Subsequently, we integrate knowledge elements from manual annotations into a property graph, forging new links through explicit relationships and rule-based inference. In the final stage, visual computation is applied to examine the use of data to aid in analytics. This also involves assigning visual attributes to facilitate the study of each entity or broader patterns.

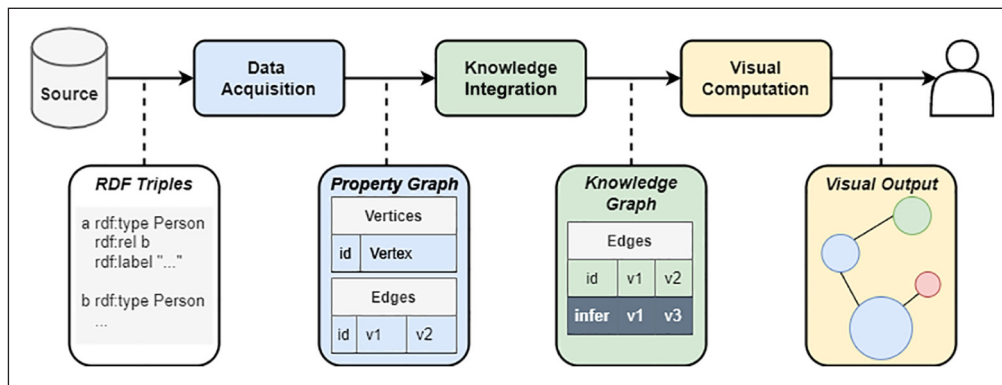


Figure 1 Illustration of the engineering pipeline in building the MA²KG.

4.1 THE DATA MODEL

The data rationale underpinning these steps relies on the Martial Art Ontology (MAon), an ontology model that conceptualises the epistemic aspects of traditional martial arts. As illustrated in [Figure 2](#), the schema adopts a human-centred perspective and pivots on the class **MA_Master** to associate different classes and properties in describing martial arts masters. The network structure also makes it possible to illustrate interpersonal contact and single out implications arising from exchanges between individuals, techniques, or styles. More specifically, explicit relations, such as **is_master_of**, **is a student of** and **is_clan_member_of**, can be extracted from LOD materials and transformed into ontological data to complement manual annotations. Implicit relations, i.e. **share_place** and **share_time**, can be constructed through rule-based inference based on relevant attributes. The rationale behind establishing such “share” connections is that two masters may have interacted if they lived in the same area during the same time period, potentially resulting in martial arts exchanges that influenced a specific style.

4.1.1 Principal entities in the MA²KG

The following paragraphs outline the principal classes and properties in the MA²KG schema and elucidate the design motivation through typical instances.

class MA_master This class denotes the distinguished martial art practitioners who have attained impressive martial arts skills and hold significance in transmitting knowledge to a group of students, often referred to as a clan. The honorary titles of masters primarily confer public recognition, rather than rank, upon these practitioners for their accomplishments within their respective regions, styles (or systems), and communities. Instances of this class include `master:Ip Man`, `master:Bruce Lee`, and `master:Lam Sai Wing`, to name a few.

The class inherits the assertions of the CIDOC-CRM `E21.Person` to link with other datasets and incorporate additional descriptions about a person. Moreover, a range of properties has been devised to link up the entities of class `MA_master`, representing both explicit and inferred clues about interpersonal exchanges between the masters. These comprise explicit lineage and

master-disciple relations, such as [master:Ip Man] - (is_master_of) - [master:Bruce Lee], and conversely, [master:Bruce Lee] - (is_student_of) - [master:Ip Man]. Implicit relations can be inferred from the persons' overlap in place and lifetime, such as [master:Ip Man] - (share_place) - [master:Wong Fei-hung] and [master:Ip Man] - (share_time) - [master:Huo Yuanjia].

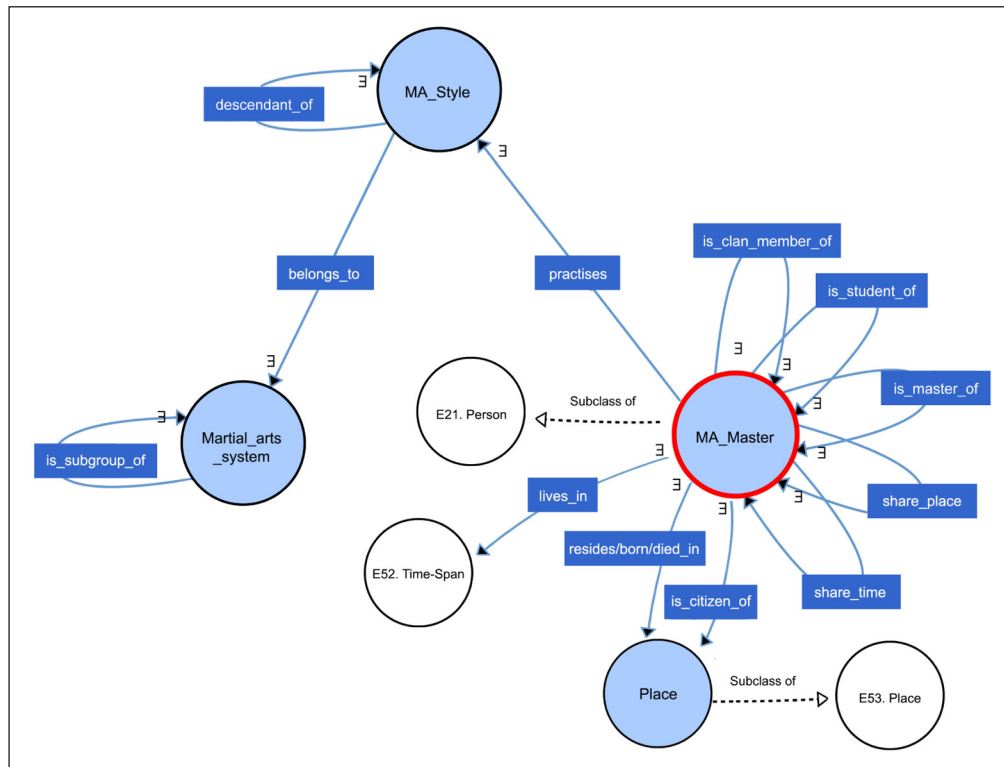


Figure 2 Schematic illustration of the Master module instantiated by the MA²KG.

class MA_style A martial art style signifies a compilation of principles, techniques and training methods that differentiate one style from another. Instances within this class are overarching denoting a specific martial art style, usually characterized by shared kinesthetic qualities, norms, and principles. Yet in some cases, there may be a symbolic element that acts as an inspiration or descriptor of its stylistic characteristics. Examples include `style:Fujian White Crane` (or known as `style:Yong Chun White Crane`), `style:Hung Gar Kuen`, `style:Wing Chun`, and many more. When a practitioner is identified to have practised a particular martial style or styles according to the data source, a `practises` property gets established, for example, [master:Ip Man] - (practises) - [style:Wing Chun].

Certain relationships also exist between the martial art styles and systems, for which we have devised a series of properties. Respectively, `belongs_to` indicates a martial art style is part of a martial art system, whilst `is_descendant_of` implies that a certain style can be considered as descending from another. For instance, [style:Lam Family Hung Gar] - (is_descendant_of) - [style:Hung Gar] and [style:Hung Gar] - (belongs_to) - [system:Southern Chinese styles]. The relational property `is_subgroup_of` applies to the hierarchical interpretation between the systems, such as [system:Southern Shaolin Styles] - (is_subgroup_of) - [system:Southern Chinese Styles].

class Martial_art_system Martial art styles can be categorised under a common theme according to factors like geographic origins, ethnic groups, documented lineage, or other criteria. This is conveyed through the multilayered conceptualisation of `Martial_art_system`, for instance, `system:Southern Chinese martial arts`, `system:Shaolin kung fu`, and `system:Aikido`. Styles within the same system typically adhere to a uniform technical and philosophical framework.

class Place is an instantiation of the CIDOC-CRM class `E53.Place`, which denotes a specific geographical location. For example, it could be the Blue House - situated on Stone Nullah Lane, Hong Kong, where Lam Sai Wing once operated a school alongside a clinic. Other examples

include the city of Foshan in China and Hong Kong. This class provides information about the locations where a master's major activities have taken place and complements records of their birthplace (via relation `born_in`), death place (`died_in`), residence (`resides_in`) and citizenship (`is_citizen_of`). For example, `[master:Ip Man] - (born_in) - [place:Foshan]` and `[master:Ip Man] - (is_citizen_of) - [place:Hong Kong]`, `[master:Bruce Lee] - (resides_in) - [place:Seattle]` and `[master:Bruce Lee] - (died_in) - [place:Kowloon Tong]`.

4.2 OPEN DATA ACQUISITION AND KG CONSTRUCTION

The MA²KG has been created through a combination of data sources, including structured data obtained from Wikidata, Wikipedia pages, and manual annotations provided by the authors in both English and Chinese. These annotations were based on the content of books and exhibition panel texts produced from the *Hong Kong Martial Arts Living Archive* (Chao, 2018; Chao, Shaw, & Kenderdine, 2016; Kenderdine, Shaw, & Chao, 2018). In this section, we debrief the workflow considerations and implementation steps. Additionally, the source codes, core ontologies, RDF resources and scripts have been made publicly available on GitHub.⁵

4.2.1 Open data acquisition

Valuable data for CH representation can exist in a wide range of formats, including open knowledge bases, historical documents, audio-visual archives, and social media records. Given the diverse nature and scale of the data, different approaches to data acquisition may apply. In our attempt to construct a bilingual KG of kung fu masters, our objective was to acquire structured data from open knowledge bases and integrate manual additions into a unified structure.

The acquisition of open data runs primarily on the Wikidata Query Service, a collaborative knowledge graph platform that offers a SPARQL endpoint for developers to retrieve RDF triples using semantic queries.⁶

To automate the process, we created a SPARQL template for executing queries that extract a chain of `Person` entities (see Listing 1). A fundamental consideration guiding the code design is that a relationship could have probably existed between any pair of masters historically. In cases of indirect relationships, such as shared membership in a specific lineage of practitioners, a multi-step connection could be found through overlap in location or time. Accordingly, the query runs with a specified `Person` entity, denoted as Q_x , and then searches for other `Person` entities, denoted as P_x , which have at least one relationship to or from the `Person` Q_x . This process incorporates data fields that are crucial for identifying a master, such as art name, birth and death year, citizenship and occupation, as properties within the entity graph.

Listing 1 Excerpt of the SPARQL template to construct a chain of `Person` entities.

```

1 PREFIX neo: <neo4j://voc#>
2 CONSTRUCT {
3   # construct a Person entity with properties
4     ?person a neo:master .
5     ?person neo:name ?personName .
6   # construct a related person entity
7     ?relatedPerson a neo:master .
8   # specify personal relationship
9     ?person neo:relation ?relatedPerson .
10  } WHERE {
11   # scrape a chain of related entities
12     ?person (wdt:Px)* wd:Qx .
13     ?person wdt:Px ?relatedPerson .
14     OPTIONAL {?person rdfs:label ?personName .}
15     SERVICE wikibase:label {bd:serviceParam wikibase:language "en" .}
16  }

```

⁵ MA²KG GitHub repository: <https://purl.org/ma2kg/git>.

⁶ Wikidata Query Service: <https://query.wikidata.org/>.

4.2.2 Knowledge integration and KG construction

Knowledge inference operates on the property graph acquired through the steps described in Section 4.2.1. The objective is to deduce implicit relationships from explicit ones, leveraging the existing entity-relationship triples and inference rules. For instance, the inferred relationships named “Share Time” and “Share Place” signify overlaps in years and locations between two masters, which indicates a potential interaction (like competition and friendship). When the same place-related entity is detected in the entity network of two masters, e.g., **Ip Man** and **Wong Fei-hung** both have the relationship (Is Citizen Of)-[Hong Kong], the Share Place relationship will be established between them. Likewise, the Share Time relationship is created to denote overlapping lifetimes.

However, inference computations may generate false relationships that lead to inaccurate knowledge, which makes knowledge validation necessary. This process primarily involves consultation with domain experts to ensure the credibility of the knowledge graph by referencing the original data sources during the workflow process. Subsequently, we sent the pre-checked dataset to a scholar specialising in Southern Chinese traditional martial arts based in Hong Kong for review and potential amendments.

Table 1 lists key metrics for the final MA²KG. The graph comprises 594 nodes and 14,289 relationships, a representation that reasonably aligns with the statistics of Chinese kung fu masters identifiable from relevant documentation and online sources. It is well connected, as indicated by the metrics of WCC and SCC.⁷ And 98.9% of the edges (relations) within the MA²KG are connected to a master, as the primary goal of data integration was to enhance contextual knowledge about the masters, guided by a human-centric rationale.

Nodes	594
Master nodes	241 (0.406)
Edges	14,289
Master-related edges	14,132 (0.989)
Nodes in the largest WCC	448 (0.754)
Nodes in the largest SCC	279 (0.470)

Table 1 Essential structural metrics of the MA²KG dataset.

5 APPLICATION AND INSPECTION

During the visual analysis of the MA²KG, we implemented a graph database using Neo4j, a graph data management system equipped with a range of analytical capabilities. Specifically, we employed Cypher (a SQL-like language) for data query, inference and integration, Neo4j Graph Data Science Library for computing relevant graph metrics, and Neo4j Bloom for building an interactive visualisation interface, as demonstrated in **Figure 3**.⁸

5.1 LINEAGE ANALYSIS

A commonly explored theme in the study of historical exchanges, lineage analysis typically examines the relationships between people of different generations or martial art styles. In this context, **Figure 4** shows the lineage diagram of the prominent practitioners of Wing Chun, a style that gained global recognition chiefly through kung fu movies. The diagram visually portrays the intricately connected lineages and styles of the masters. Notably, the two most eye-catching nodes, Master **Ip Man** and his student **Bruce Lee**, hold the highest centrality

⁷ A weakly connected component (WCC) is a graph where there is a path between every two vertices in the underlying undirected graph. A strongly connected component (SCC) is a directed graph in which there is a path from each vertex to another vertex.

⁸ Neo4j is a graph database management system offering high-performance storage and a suite of analytical capabilities, such as Bloom - a data visualisation tool to interact with Neo4j’s graph data platform with no coding required - and Graph Data Science Library, which includes computations from a set of graph metrics. Neo4j.com: <https://neo4j.com/product/neo4j-graph-database/>. Neo4j Graph Data Science Library: <https://neo4j.com/product/graph-data-science/>. Neo4j Bloom: <https://neo4j.com/product/bloom/>.

scores in the graph. These visual outcomes are consistent with the widely held perception of these masters' significance. For example, Ip Man is well known as a legendary Wing Chun master whose numerous students went on to develop new styles or sub-styles of martial arts. Bruce Lee, arguably the most famous disciple of Ip Man, established Jeet Kune Do, a style known for inheriting the fundamental concept of Wing Chun, which emphasises efficiency in both time and movement via single fluid motions that attack while defending (Wikipedia contributors, 2023b).

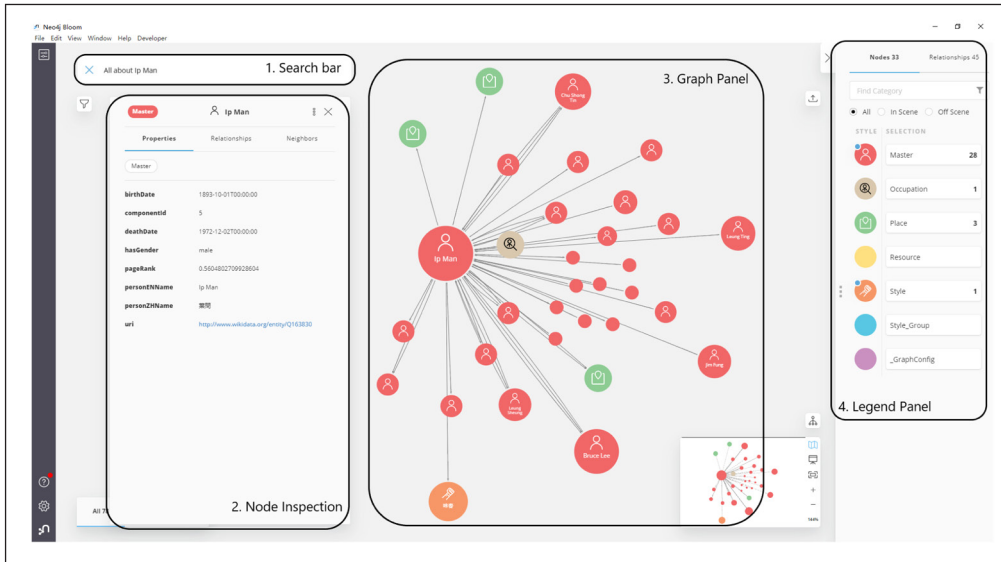


Figure 3 Visual implementation of MA²KG in Neo4j Bloom, where (1) the Search Bar enables natural-language keyword search using entity names, pre-prescribed tags or query blocks; (2) the Node Inspection panel presents descriptive information about a selected node and is extensible to accommodate complete ontological annotations; (3) the Graph Panel visualises the queried sub-graph; and (4) the Legend Panel provides flexibility to adjust the styling features according to visual attributes.

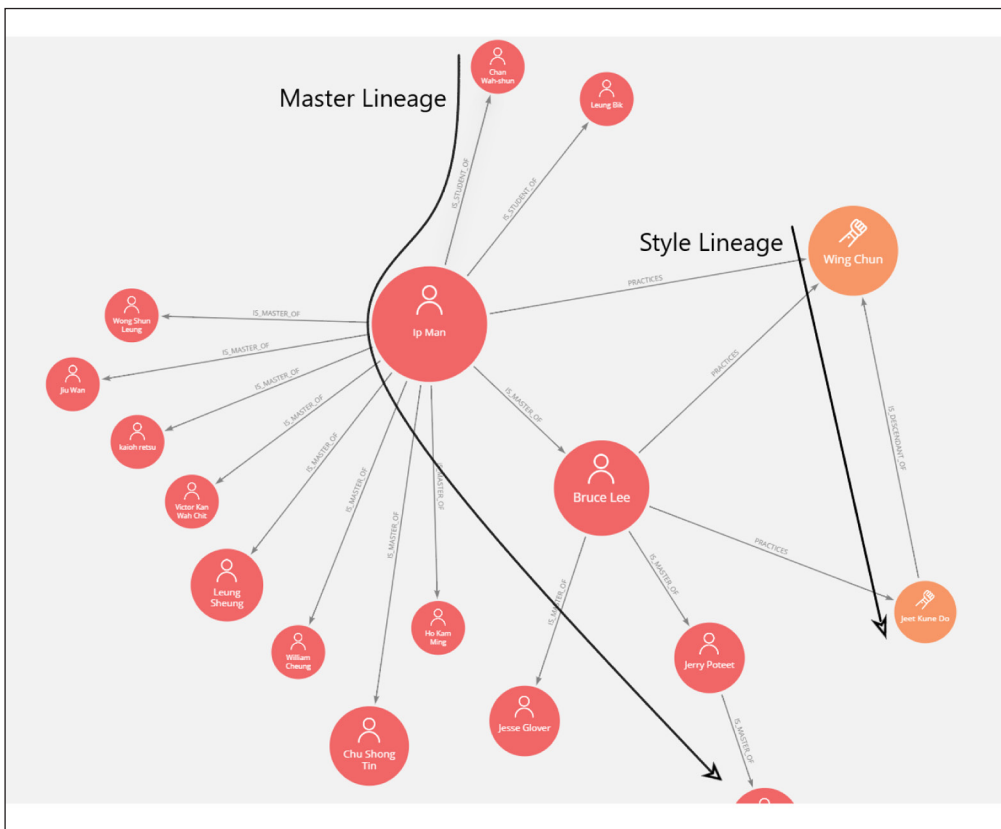


Figure 4 The master-style lineage diagram of the Wing Chun style, as computed from the MA²KG.

As the sample suggests, individual data sources often provide only partial information and a more comprehensive understanding can be achieved by supplementing this information with related data across different sources. However, it's worth mentioning that open data sources, like Wikidata, are not always scholarly or accurate due to various factors, such as a lack of historical documentation, the interference of language and political events, and the private systems of knowledge transmission within certain clans. Nevertheless, by integrating

and cross-referencing data from different sources - both in English and Chinese, online and offline - the MA2KG modality demonstrates its potential to discover implicit associations and spot errors, thus helping researchers uncover and validate certain facts. Such network features permit users with little prior knowledge to explore and keep exploring databases. Moreover, it is promising to enable a recommendation system serving the public dissemination of cultural information.

5.2 INFLUENCE AND CENTRALITY

In detecting influential masters, we utilised PageRank and Degree centrality algorithms to assess the centrality of masters' nodes and determine their significance. Specifically, PageRank evaluates the importance of nodes based on the number of incoming relationships, taking into account the importance of the corresponding source nodes. Degree centrality, on the other hand, measures the connectivity by counting the number of links connected to a node, whether they are incoming, outgoing, or both. In addition, Wikidata's Pageview stats were collected to supplement the measurement of masters' popularity. These were calculated from the visit statistics of each Wikipedia page bearing a master's name.⁹

The results, shown in Table 2, reveal variations in the ranking of "central" masters based on different metrics. For instance, Wong Fei-hung, a master of Hung Gar (or Hung Kuen), attains the highest Degree centrality score, likely due to his extensive teaching and interaction with many masters throughout his life. Meanwhile, Tung Ying-Chieh, the instructor of Yang Cheng-Fu and a renowned teacher of Yang-style Tai Chi, tops the PageRank metric, reflecting the cumulative importance of a node and its neighbouring connections.¹⁰ According to Wikidata's Pageview stats, Bruce Lee appears as the most significant martial artist, probably due to his impact on the media and movie sectors.

MASTER	DEGREE	PAGERANK	PAGEVIEW
Wong Fei-Hung	173	0.16	155,375
Leung Sheung	165	0.24	5,062
Lam Sai-Wing	158	0.18	15,214
Ip Man	144	0.35	1,306,847
Bruce Lee	127	0.44	6,238,349
Tung Ying-Chieh	22	2.67	2,830

Table 2 The most influential masters in the MA²KG based on three distinct metrics.

These findings suggest a possible bias in the data, possibly arising from different ways of measuring "influence" in digital social records and inherent biases within the data sources. For instance, Lam Cho, a notable master in the history of Hung Gar who refined the Lam Family Hung Gar lineage, failed to stand out in all three measures, possibly because his online profile is not as popular. This underscores the importance of involving domain expertise in graph construction. Scholars from diverse backgrounds should be involved to foster KG's inclusiveness and integrity.

5.3 THE NO-NAMES

Public attention tends to gravitate towards well-documented narratives and known figures. In contrast, individuals who have historically played a role in the transmission or evolution of martial arts may remain relatively obscure. To enhance the visibility of lesser-known masters,

⁹ For more information, see Neo4j's PageRank algorithm: <https://neo4j.com/docs/graph-data-science/current/algorithms/page-rank/>, and Degree Centrality algorithm: <https://neo4j.com/docs/graph-data-science/current/algorithms/degree-centrality/>. Wikipedia Pageview stats is a tool to analyse how many visits have occurred to a page during a given time. <https://pageviews.toolforge.org/>.

¹⁰ Wong Fei-hung, a Chinese martial artist, physician, and folk hero, became famous as the protagonist in numerous martial arts films and television series. Yang Cheng-Fu was one of the best-known masters of Yang-style Tai Chi, which belongs to the five primary families of Tai Chi and is the most widely practised style in contemporary times.

we harnessed Dijkstra's shortest path algorithm (J.-C. Chen, 2003) to reveal the potential linkages between a given master and all other nodes in the graph of MA²KG.

Figure 5 illustrates a network featuring three masters: Wong Fei-hung, Wong Shun Leung, and Barry Pang. The latter two, although not widely recognised, came to our attention when we explored the various paths within a length of three from the significant figure, Master Wong Fei-hung. In this example, implicit relationships provide valuable insights that can unveil intriguing connections between entities that might otherwise hardly be observable. For instance, Master Wong Shun Leung, who was active during the era of Master Wong Fei-hung, became visible. Similarly, through the master-disciple relationship and stylistic influence, Master Barry Pang emerges as a notable figure, credited with making substantial contributions to the development of martial arts in Australia.¹¹

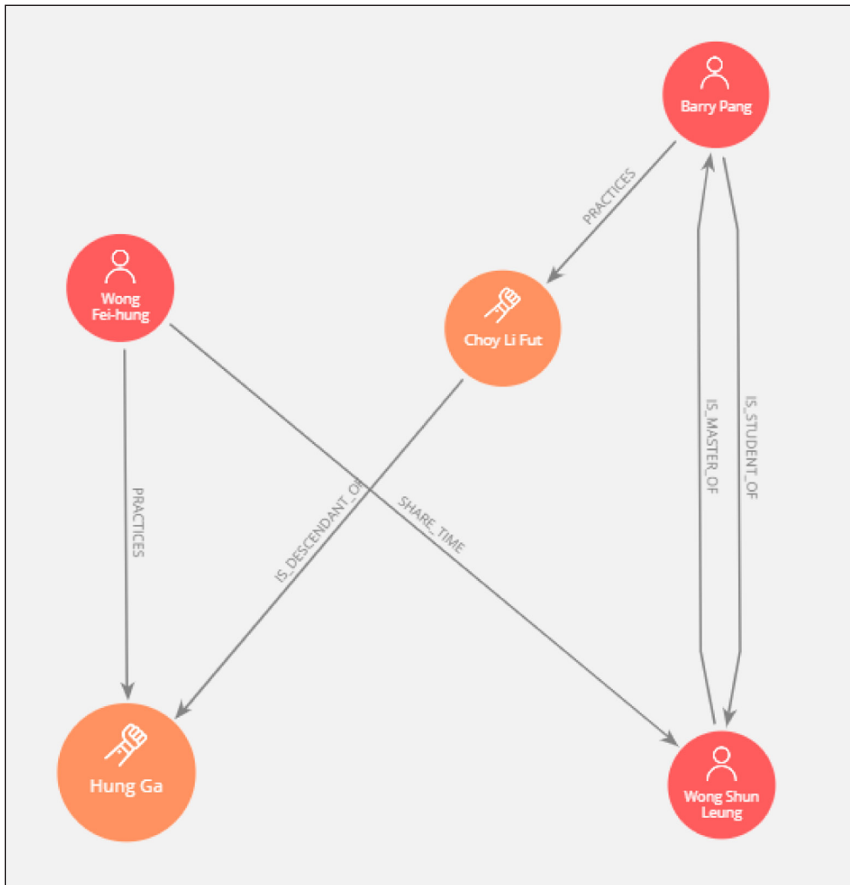


Figure 5 Illustration of explicit and implicit relationships connected to Master Wong Shun Leung.

6 CONCLUSION

To achieve the goal of organising heterogeneous materials into an informative and interoperable model, this article presents an effort that automatically acquires martial arts knowledge from dispersed data sources in different languages. This is accomplished by creating the MA²KG, a knowledge graph encompassing 241 Chinese kung fu masters based on an extended ontology of martial arts. Additionally, a visual exploration interface is implemented, allowing casual users to interact with the conceptual entities of kung fu masters.

In addition to introducing the MA²KG graph dataset and engineering methods, this work expresses a fresh perspective on curating cultural data that places a strong emphasis on the individuals and their interconnections. The approach responds to the frustration concerning “ways of access” that cultural collections face today. With this effort, we aim to highlight the pivotal roles that people play in the (re)creation and transmission of ICH knowledge and make their contributions resonate with the people today.

¹¹ Choy Lee Fut is famous for incorporating techniques from various styles, primarily the Shaolin system based on Choy Gar, Li Gar, and Hung Gar. The style is also renowned for its effectiveness in fighting multiple persons, as recognised by Bruce Lee (Wikipedia contributors, 2023a). Barry Pang: https://en.wikipedia.org/wiki/Barry_Pang.

Nonetheless, the reported work retains a potential for future improvement. During the data application and inspection stage, as we dig into new interpretations enabled by the creation of LOD datasets, concerns about the reliability of these interpretations grow. Depending solely on publicly crowd-sourced data instead of scholarly materials inevitably raises credibility issues. Yet, a potential solution may involve integrating human-led academic validation and machine-operated cross-referencing checks into the data integration workflow. This underscores another limitation in the current data acquisition process, which is that taking data only from structured data pools like Wikidata is insufficient. Therefore, extending the approach to extract knowledge from unstructured data is imperative, where a combination of named entity recognition and text mining holds promise.

ACKNOWLEDGEMENT

We extend our profound gratitude to Professor Sarah Kenderdine, director of the Laboratory for Experimental Museology (eM+) at EPFL, for her guidance and support throughout the course of this research. We also thank the reviewers for their insightful comments towards improving our manuscript.


COMPETING INTERESTS


The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Yumeng Hou (Methodology, Investigation, Data Curation, Formal analysis, Writing, Review/Editing), Lin Yuan (Software, Investigation, Formal analysis, Writing).

AUTHOR AFFILIATIONS

Yumeng Hou  orcid.org/0000-0002-7908-0693
Laboratory for Experimental Museology, EPFL, Lausanne, Switzerland

Lin Yuan  orcid.org/0000-0003-1151-4261
Section of Computer Science, EPFL, Lausanne, Switzerland

REFERENCES

- Bikakis, A., Hyvönen, E., Jean, S., Markhoff, B., & Mosca, A.** (2021). Special issue on semantic web for cultural heritage. *Semantic Web*, 12(2), 163–167. DOI: <https://doi.org/10.3233/SW-210425>
- Bingenheimer, M.** (2021). The historical social network of chinese buddhism. *Journal of Historical Network Research*, 5(1). DOI: https://doi.org/10.17928/jjadh.5.2_84
- Blouin, B., Magistry, P., & Van den Bosch, N.** (2021). Creating biographical networks from chinese and english wikipedia. *Journal of Historical Network Research*, 5(1). DOI: <https://doi.org/10.25517/jhnr.v5i1.120>
- Breure, A. S., & Heiberger, R. H.** (2019). Reconstructing science networks from the past: eponyms between malacological authors in the mid-19th century. *Journal of Historical Network Research*, 3, 92–117. DOI: <https://doi.org/10.25517/jhnr.v3i1.52>
- Caraffa, C., Pugh, E., Stuber, T., & Ruby, L. W.** (2020). Pharos: A digital research space for photo archives. *Art Libraries Journal*, 45(1), 2–11. DOI: <https://doi.org/10.1017/alj.2019.34>
- Carboni, N., & De Luca, L.** (2019). An ontological approach to the description of visual and iconographical representations. *Heritage*, 2(2), 1191–1210. DOI: <https://doi.org/10.3390/heritage2020078>
- Carriero, V. A., Gangemi, A., Mancinelli, M. L., Marinucci, L., Nuzzolese, A. G., Presutti, V., & Veninata, C.** (2019). Arco: The italian cultural heritage knowledge graph. In *International semantic web conference* (pp. 36–52). DOI: https://doi.org/10.1007/978-3-030-30796-7_3
- Chao, H.** (2018). *Lingnan hung kuen: Kung fu in cinema and community*. Hong Kong: City University of HK Press.
- Chao, H., Delbridge, M., Kenderdine, S., Nicholson, L., & Shaw, J.** (2018). Kapturing kung fu: Future proofing the hong kong martial arts living archive. In *Digital echoes* (pp. 249–264). New York: Springer. DOI: https://doi.org/10.1007/978-3-319-73817-8_13
- Chao, H., Shaw, J., & Kenderdine, S.** (Eds.). (2016). *300 years of hakka kung fu: Digital vision of its legacy and future*. International Guoshu Association. (Research Unit(s) information for this publication is provided by the author(s) concerned.)

- Chen, J.-C.** (2003). Dijkstra's shortest path algorithm. *Journal of formalized mathematics*, 15(9), 237–247.
- Chen, S., & Wang, H.** (2022). China biographical database (cbdb): a relational database for prosopographical research of pre-modern china. *Journal of Open Humanities Data*, 8, Article 4. DOI: <https://doi.org/10.5334/johd.68>
- Cline, D. H.** (2020). Athens as a small world. *Journal of Historical Network Research*, 4, 36–56. DOI: <https://doi.org/10.25517/jhnr.v4i0.84>
- Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., . . . Zhang, W.** (2014). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th acm sigkdd international conference on knowledge discovery and data mining* (pp. 601–610). DOI: <https://doi.org/10.1145/2623330.2623623>
- Dou, J., Qin, J., Jin, Z., & Li, Z.** (2018). Knowledge graph based on domain ontology and natural language processing technology for chinese intangible cultural heritage. *Journal of Visual Languages & Computing*, 48, 19–28. DOI: <https://doi.org/10.1016/j.jvlc.2018.06.005>
- Fuller, M., & Wang, H.** (2021). Structuring, recording, and analyzing historical networks in the china biographical database. *Journal of Historical Network Research*, 5(1). DOI: <https://doi.org/10.25517/jhnr.v5i1.123>
- Golub, K., & Liu, Y.-H.** (2022). *Information and knowledge organisation in digital humanities: Global perspectives*. New York: Taylor Francis. DOI: <https://doi.org/10.4324/9781003131816>
- Hou, Y.** (2023). *The martial art ontology (maon) version 1.1*. Retrieved from <https://purl.org/maont/techCorpus> (last accessed: 2023-08-01).
- Hou, Y., Kenderdine, S., Picca, D., Egloff, M., & Adamou, A.** (2022). Digitizing intangible cultural heritage embodied: state of the art. *Journal on Computing and Cultural Heritage (JOCCH)*, 15(3). DOI: <https://doi.org/10.1145/3494837>
- Kenderdine, S., Shaw, J., & Chao, H.** (2018). *Kung fu motion*. EPFL ArtLab Lausanne. Retrieved from https://kungfumotion.live/wp-content/uploads/2018/04/kungFuMotion_ENFR.pdf (last accessed: 2023-07-25).
- Koho, M., Ikkala, E., Leskinen, P., Tamper, M., Tuominen, J., & Hyvönen, E.** (2021). Warsampo knowledge graph: Finland in the second world war as linked open data. *Semantic Web*, 12(2), 265–278. DOI: <https://doi.org/10.3233/SW-200392>
- Meghini, C., Bartalesi, V., & Metilli, D.** (2021). Representing narratives in digital libraries: The narrative ontology. *Semantic Web*(Preprint), 1–24. DOI: <https://doi.org/10.3233/SW-200421>
- Petz, C., Ghawi, R., & Pfeffer, J.** (2022). Tracking the evolution of communities in a social network of intellectual influences. *Journal of Historical Network Research*, 7(1), 114–154. DOI: <https://doi.org/10.25517/jhnr.v7i1.146>
- Rejeb, A., Keogh, J. G., Martindale, W., Dooley, D., Smart, E., Simske, S., . . . others.** (2022). Charting past, present, and future research in the semantic web and interoperability. *Future Internet*, 14(6), 161. DOI: <https://doi.org/10.3390/fi14060161>
- Wikipedia contributors.** (2023a). *Choy li fut* — *Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Choy_Li_Fut&oldid=1104309123 (last accessed: 2023-07-25).
- Wikipedia contributors.** (2023b). *Jeet kune do* — *Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Jeet_Kune_Do&oldid=1105839521 (last accessed: 2023-07-25).
- Ziku, M.** (2020). Digital cultural heritage and linked data: Semantically-informed conceptualisations and practices with a focus on intangible cultural heritage. *Liber Quarterly*, 30(1). DOI: <https://doi.org/10.18352/lq.10315>

TO CITE THIS ARTICLE:

Hou, Y., & Yuan, L. (2023). Building a Knowledge Graph of Chinese Kung Fu Masters From Heterogeneous Bilingual Data. *Journal of Open Humanities Data*, 9: 27, pp. 1–12. DOI: <https://doi.org/10.5334/johd.136>

Submitted: 22 August 2023

Accepted: 12 October 2023

Published: 24 November 2023

COPYRIGHT:

© 2023 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.